Avens Publishing Group
J Clin Investigat Dermatol
September 2020 Volume 8, Issue 2

Inter- and Intra-Physician Variation in Quantifying Actinic Keratosis Skin Photodamage

Keywords: Rating scale; Dermatology; Rater reliability; Rctinic damage; Kappa; Physician variation

Abstract

We investigated the variations in physician evaluation of skin photodamage based on a published photodamage scale. Of interest is the utility of a 10-level scale ranging from none and mild photodamage to actinic keratosis (AK). The dorsal forearms of 55 adult subjects with various amounts of photodamage were considered. Each forearm was independently evaluated by 15 board-certified dermatologists according to the Global Assessment Severity Scale ranging from 0 (less severe) to 9 (the most progressed stage of skin damage). Dermatologists rated the levels of photodamage based upon the photographs in blinded fashion. Results show substantial disagreement amongst the dermatologists on the severity of photodamage. Our results indicate that ratings could be more consistent if using a scale of less levels (3-levels). Ultimately, clinicians can use this knowledge to provide better interpretation of inter-rater evaluations and provide more reliable assessment and frequent monitoring of high-risk populations.

Open Access **Research Article**

Journal of Clinical & Investigative Dermatology

Schmeusser B[1], Borchers C[1], Travers JB[1,3,4], Borchers S[3], Trevino J[3], Rubin M[3], Donnelly H[3], Kellawan K[3], Carpenter L[3], Bahl S[3], Rohan C[3], Muennich E[3], Guenthner S[5], Hahn H[3], Ali Rkein[3], Darst M[6], Mousdicas N[7], Cates E[1], Sunar U[2] and Bihl T[1,2*]

[1]*Department of Pharmacology & Toxicology, Boonshoft School of Medicine, USA*

[2]*Department of Biomedical, Industrial & Human Factors Engineering, USA*

[3]*Department of Dermatology, Boonshoft School of Medicine, USA*

[4]*Dayton Veterans Administration Medical Center, USA*

[5]*The Indiana Clinical Trials Center, PC, USA*

[6]*Charlotte Dermatology, Charlotte, USA*

[7]*Richard L. Roudebush VA Medical Center, Indianapolis, USA*

***Address for Correspondence**

Bihl T, Department of Pharmacology & Toxicology, Boonshoft School of Medicine, Wright State University, Dayton, OH, 45435, USA, Fax: 937-775-7221, Tel: 937-775-2463, Email: trevor.bihl@wright.edu

Submission: 19 August, 2020
Accepted: 1 September, 2020
Published: 6 September, 2020

Abbreviations

AK: Actinic Keratosis; BCC: Basal Cell Carcinoma; CV: Coefficient of Variation; CI: Confidence Interval; FFPAS: Dermatologic Assessment Form Forearm Photographic Assessment Scale; NMSC: Non-Melanoma Skin Cancer; SCC: Squamous Cell Carcinoma; UV: Ultraviolet

Introduction

Non-melanoma skin cancers are the most common form of malignancy within North America, whose prevalence is only rising with nearly 3.5 million cases diagnosed within the United States alone each year. This is associated with a substantial financial impact, currently estimated at $5 billion to treat non-melanoma skin cancers (NMSCs), including basal cell carcinomas (BCCs) and squamous cell carcinomas (SCCs) [1]. Major causes of skin pathologies are exposure to ultraviolet (UV) radiation, commonly from sunlight and artificial sources such as tanning beds. The term actinic neoplasia is used for AKs and NMSC to denote the role of UV and advanced age. Skin pathologies related to photodamage include precancerous actinic keratosis, which can progress into basal cell carcinoma and squamous cell carcinoma.

Although classified as pre-cancerous, AK progression to NMSC is variable. A majority of NMSC arises from AKs, but a majority of AKs do not become cancer or will even be clinically present in 1-5 years [2]. Most AKs are diagnosed clinically, yet the current gold standard for diagnosis of an AK is a classified as an invasive procedure, involving a biopsy of the lesion with subsequent histopathology [3]. NMSC are diagnosed by histology.

Given the considerable morbidity and occasional mortality associated with actinic neoplasia, strategies including use of cyclooxygenase-2 inhibitor celecoxib [4] and nicotinamide [5] have been tested. To assess risk factors for actinic neoplasia, as well response to rejuvenation techniques, there is a need for a reliable method to evaluate photodamage. In particular, presence of photodamage is strongly associated with actinic neoplasia. One of the primary methods for quantifying photodamage is the 10-point Dermatologic Assessment Form Forearm Photographic Assessment Scale (FFPAS) of [6].

Recent studies have suggested that noninvasive imaging of skin could detect and monitor precancerous lesions using hemoglobin contrast [7]. However, complicating both manual and computer detection of AKs, is the complexity of the FFPAS evaluation mechanism. In performing a clinical trial (N=55 subjects; 110 dorsal forearms; N=15 physicians), we noticed significant disagreement across diagnoses. Such disagreements in the community, if realized through conflicting second opinions, could result in some receiving overly aggressive treatments while others receive insufficient treatment. In this paper, we examine these results and pose various suggestions to improve the FFPAS process.

Materials and Methods

A clinical trial was performed on subjects who are patients in the Wright State University Department of Dermatology clinics. This was performed under an institutional review board-approved protocol, and informed consent was obtained from all the patients

ISSN: 2373-1044

Figure 1: Examples of mild, moderate, severe photodamage analyzed by clinical dermatologists.

before the measurements.

Selection of patients and photos

The patients were 35 years old or older with "fair" skin (Fitzpatrick scale I or II) [8] and did not have recent (< 6 month) history of use of a tanning bed/significant sun exposure. A total of 55 subjects were recruited and these subjects expressed various levels of photodamage, including clinically-apparent AKs. Each subject had each of their forearms photographed, resulting in a total of 110 photos of arms to be evaluated. Examples of a forearm exhibiting mild, moderate, and severe photodamage is shown in Figure 1.

Selection of raters and rating process

Each forearm was evaluated for photodamage by board-certified dermatologists (N=15 physicians) trained in evaluating actinic damage. This group consisted of dermatologists from both academic (4) as well as private practice (11) backgrounds who had a minimum of 5 years of post-residency experience. For evaluation in this study, the dermatologists used the 10-point FFPAS of McKenzie et al. (2011) [6]. As shown in Table 1, FFPAS is a subjective measure to examines clinical signs of UV-induced skin damage along four dimensions: fine wrinkling, coarse wrinkling, abnormal pigmentation, and a global assessment [6]. In using the FFPAS approach, each individual clinical sign is scored, and a global assessment is provided to rank the overall actinic damage [6].

Dermatologists in this study were trained in FFPAS by reviewing the examples in the published McKenzie scale [9]. Once trained, the dermatologists individually, independently, and separately evaluated each arm of participants from a PowerPoint presentation of photographs which consisted of not only the 110 forearms from the 55 subjects, but an additional 20 forearm pictures duplicated to assess intra-rater reliability. The raters provided the global assessment for each arm. No identification of the patient, arm, or initial assessment was provided and the arms were randomly organized in their presentation. Each dermatologist was also given unlimited time for assessment. The raters were blinded to clinical information and the source of the photos, and were not allowed to discuss their observations with other raters.

Down Sampling Ratings

Travers et al.[7] down sampled, or pooled, the 10 FFPAS categories into three groups (Mild/none, Moderate, and Severe). Notionally, these 3 down sampled groups followed the general groupings of Table 1, where scores of 0, 1, 2, and 3 are mild, scores of 4, 5, or 6 are moderate, and scores of 7, 8, and 9 are severe. Consistent with [7], scores of 0 are grouped into mild due there being few observations of 0 in the study. These groups were in the implicit groupings of Table 1 and were used in [7] to develop a three-class machine learning classifier for actinic damage classification. This down sampled FFPAS scores are used herein to understand how different multitudes of ratings might affect rater reliability.

Statistics

The data was analyzed using JMP (SAS), Matlab 2019a a (Mathworks, Boston, MA) and the Fleisses Kappa software package in Matlab [10]. Inter- and intra-rater reliability was assessed using coefficient of variation (CV) [11], Fleiss's κ [12], Cohen's κ [13], and graphical means. For any confidence interval (CI) or hypothesis test, $\alpha = 5\%$ was used.

The CV between each dermatologist for each of the arms rated with the equation:

$$CV = s/\bar{x}$$

which scales the sample variance for a given arm by the sample mean [11]

Intra-rater reliability was analyzed using Cohen's κ [13]. The κ values were calculated for each dermatologist for their assessment of the 20 duplicated samples. Inter-rater reliability was analyzed using Fleiss's κ [12], an extension of Cohen's κ. In addition to the confidence intervals and hypothesis tests of both methods, κ has a further interpretation with general hierarchy of [14], seen in Table 2.

Results

The results from this study are presented in Figure 2 which presents a heat map where subjects are the rows and the rating dermatologists (consistent throughout this study) are the columns. Two rows are presented for each subject, for left (L) and right (R) arms. The colors in Figure 2 range from blue, for 0, to red, for 9. The scores are further provided in each cell for the rating each dermatologist gave a specific arm. While, overall, scores tended towards the middle values (Figure 3a), it is visually apparent in Figure 2 that some subjects generally have more severe actinic damage than others. Considering the range of scores by each patient-arm pairing, Figure 3b, it is further apparent

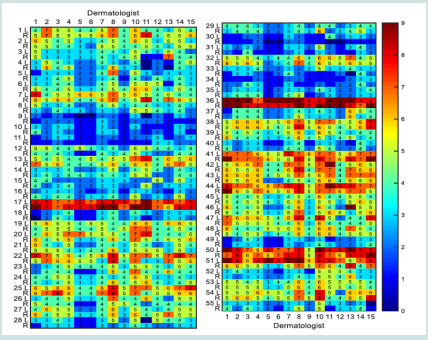

Figure 2: Physician scores (0-9) associated with each arm, numbers along rows indicate patient number along with left ("L") or right ("R") arm. Columns indicate which physician evaluated the patient. Colors and numbers in each cell are the score given.

ISSN: 2373-1044

Table 1: The Dermatologic Assessment Form Forearm Photographic Assessment Scale (FFPAS) of McKenzie et al. (2011) [6]

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: General interpretation of κ [14].

Fleiss' κ	Interpretation
<0.01	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair Agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

that there is general disagreement by raters. An example illustrates this disagreement using patient 1's left arm; in row 1 of Figure 2 we see scores ranging from 3 (mild), by dermatologist 13, to 7 (severe), by dermatologists 2 and 8, giving a range of 4, as seen in Figure 3b. Collectively, Figure 3b illustrates the general differences in diagnosis with a mean range of 3.57 across all arms in this study. Since each of the groups of FFPAS (Mild, Moderate, Severe) encompass 3 scores, the average range of scores in this study indicates very different diagnoses were given for each arm (e.g. Mild to some, Moderate to others). Similarly, this could possibly result in the prognosis would greatly changing for a subject, for example, depending on which dermatologist a subject would visit.

Inter-rater reliability

The inter-rater reliability was overall slight for the data in Figure 2 (κ = 0.114, CI 0.111-0.116). For the hypothesis that the raters provide equal ratings, the null was rejected at a 5% level of significance with $P < 0.001$. Using the hierarchy of [14] in Table 2, we would find that there is only slight agreement between raters. Considering the CV, we further see a high degree of variability between the dermatologists. Utilizing a maximum acceptability CV of 10% (good/low variability) only 4/110 ratings of the patient arms studied by the 15 dermatologists met criteria. If utilizing a maximum acceptability CV of 20% (okay/medium variability), still only 18/110 of the rating of the patient arms by the dermatologists met criteria. A CV of less than 30% (bad/high variability) represented 59/110 ratings. The remaining 51/110 ratings had CV greater than 30%, indicating unacceptable variability. Overall, 80.1% of the CV met standards of high-unacceptable variability. These results can be summarized below in Table 2.

Intra-rater reliability

The intra-rater reliability differed heavily by dermatologist, as visualized by the heatmap in Figure 4. Overall, 20 arm pictures were repeated and provided in the study with one arm given three times. The heatmap in Figure 4a show ratings for the repeated images and the heatmap in Figure 4b provides the original values from Figure 2 for direct comparison. Arm 51 right is listed twice in Figure 4b to aid comparison against Figure 4a as arm 51 right as dermatologists rated this arm 3 times. Figure 4 is evaluated by comparing columns in 4a to columns in 4b to look for consistency; for example, dermatologist 1 frequently did not rate consistently whereas dermatologist 6 almost always provided the same rating. Evaluating the intra-rater reliability with Cohen's κ found the mean intra-rater reliability to be moderate (κ=0.473, 95% CI 0.377-0.570).

Down Sampling Rating

With the three down sampled groups, the dermatologists had moderate agreement (κ = 0.41, 95% CI 0.3968-0.4060 Examining the CV for this grouping, we find the results in Table 3 which illustrate that the overall variability is much lower (45.5% of patients getting highly variable results versus 83.7% before) when using the down sampled rating scale. Intra-rater reliability for the down sampled ratings was found to have substantial agreement (κ=0.753, 95% CI 0.684 - 0.823).

Discussion

Assessment of dermatological conditions is often highly subjective in nature and the result of the complex interaction between standards, experience, training, personalities, as well as patient medical histories and overall health. Thus, understanding the severity of actinic damage is a challenge in daily clinical practice. Such challenges are exacerbated as telemedicine increases in use for triage [15] with photography-based prerounds recommended for dermatological assessment [16]. Although the Form Forearm

Table 3: Results from coefficient of variation analysis of dermatologist ratings. Results indicate a high variability/CV, with 19.9% of the data having Okay to Good (low) variability.

CV	Variability	Arm Ratings	% of Data
<10%	Good (low)	4	3.6%
<20%	Okay (med)	18	16.3%
<30%	Bad (high)	59	53.6%
>30%	Unacceptable	51	46.4%

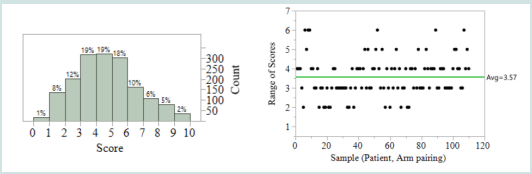

Figure 3: (left) Physician clinical scores of skin damage. (right) Range of scores from dermatologists. From (Travers, et al., 2019) [7].

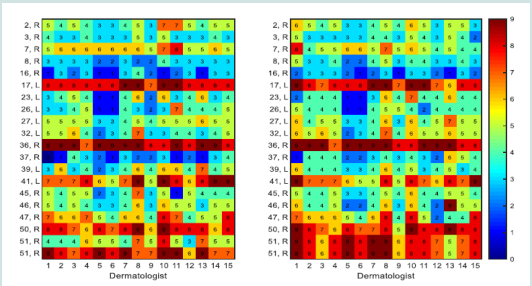

Figure 4: Physician scores (0-9) associated with each arm for (left) repeated samples with second/third look and (right) for the original ratings from Figure 2. Numbers along rows indicate patient number along with left ("L") or right ("R") arm. Columns indicate which dermatologist evaluated the patient. Colors and numbers in each cell are the score given. Notably, subject 51's right arm was examined 3 times by all dermatologists (2 repeated evaluations). The original scores for 51, R is repeated twice in b) to facilitate readability.

ISSN: 2373-1044

Table 4: Results from coefficient of variation analysis of dermatologist ratings with a revised FFPAS for 3 levels only. Results indicate lower variability/CV than with 10 levels, with 31.8% of the data having Okay or Good (low) variability.

CV	Variability	Arm Ratings	% of Data
<10%	Good (low)	10	9.09%
<20%	Okay (med)	25	31.82%
<30%	Bad (high)	31	60%
>30%	Unacceptable	44	40%

Photographic Assessment Scale (FFPAS) [6] is used clinically in actinic damage assessment, to the best of our knowledge, its inter- and intrarater reliability has never been determined.

This study considered FFPAS in a clinical study of N=55 patients and ND = 15 board certified dermatologists. To evaluate these results, the authors used both graphical and statistical methods. Graphical heat maps were used to visualize the ratings of dermatologists by sample and kappa statistics were used to evaluate inter- and intra-rater reliability. As noted in [17], heatmaps are seldom used to visualize patient data across repeated visits despite their value in visualizing patient progress; the work presented herein illustrates the value of heatmaps for similar purposes, including high-level assessment of agreement by an external observer.

When considering the 10 level FFPAS scoring, the authors found slight inter-rater agreement and moderate intra-rater agreement by dermatologists in the FFPAS ratings, which were both on the lower side of rater reliability assessment. The result of such a problem is that severe actinic damage could go untreated if rated low by a dermatologist. This is a recurring phenomenon in dermatology due to the subjective nature of some aspects and the fact that every dermatologist has varying amounts of experience and clinical expertise.

In addition to the standard 10 levels of FFPAS, the authors further down sampled the ratings into 3 groups, the mild, moderate, and severe high level groups of [6] as used in the prior work of [7]. When considering the down sampled groups, we found moderate interrater agreement and substantial intrarater agreement, both an improvement over the 10 levels of FFPAS.

While studies suggest that there is no optimal number of levels in a Likert-like questionnaire [17,18] and FFPAS is consistent with such recommendations, the results indicate the possibility that FFPAS has too many levels. Thus, it appears that more consistent results would be possible with a simpler, i.e. less levels, assessment scale.

The authors do acknowledge some limitations in this study. The selection of pictures may have not represented all possible actinic damage conditions seen in clinical practice; additionally, these pictures do not precisely represent the normal anatomical distribution [19]. The authors were also unable to collect the mean time the raters spent on completing the questionnaire since this was emailed to participants. Additionally, while the authors illustrated some benefit in both inter- and intra-rater reliability using a down sampled FFPAS scoring system, this study did not query the participating dermatologists on using a revised scale, the authors cannot definitively say that FFPAS scoring with 3 scales is better, but such a simplification warrants further study.

References

1. Rohrbach DJ, Zeitouni NC, Muffoletto D, Saager R, Sunar U (2015) Characterization of nonmelanoma skin cancer for light therapy using spatial frequency domain imaging. Biomed Opt Express 6: 1761-1766.
2. Criscione VD, Weinstock MA, Naylor MF, Luque C, Eide MJ, et al. (2009). Actinic keratoses: Natural history and risk of malignant transformation in the Veterans Affairs Topical Tretinoin Chemoprevention Trial Actinic keratoses. Cancer 115: 2523-2530.
3. Hames SC, Sinnya S, Tan J, Morze C, Sahebian A, et al. (2015). Automated Detection of Actinic Keratoses in Clinical Photographs. PLoS ONE 10: p.e0112447.
4. Elmets CA, Viner JL, Pentland AP, Cantrell W, Hui-Yi, et al. (2010). Chemoprevention of nonmelanoma skin cancer with celecoxib: a randomized, double-blind, placebo-controlled trial. Journal of the National Cancer Institute 102: 1835-1844.
5. Chen AC, Martin AJ, Damian DL (2016). Nicotinamide for Skin-Cancer Chemoprevention. N Engl J Med 374: 790.
6. NE McKenzie, K Saboda, LD Duckett, R Goldman, C Hu, et al. (2011). Development of a photographic scale for consistency and guidance in dermatologic assessment of forearm sun damage. Arch Dermatol 147: 31-36.
7. Travers J, Poon C, Bihl T, Rinehart B, Borchers C (2019). Quantifying skin photodamage with spatial frequency domain imaging: statistical results. Biomedical optics express 10: 4676-4683
8. Rosendahl C, Tschandl P, Cameron A, Kittler H (2011). Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J Am Acad Dermatol 64: 1068-1073.
9. Wallace V, Crawford D, Mortimer P, Ott R, Bamber J (2000). Spectrophotometric Assessment of Pigmented Skin Lesions: Methods and Feature Selection for Evaluation of Diagnostic Performance. Phys Med Biol 45: 735-751.
10. Cardillo G (2007). Fleisses kappa: compute the Fleiss'es kappa for multiple raters. Matlab Central, File Exchange 15426.
11. Phelps CE, Mooney C(1993). Variations in medical practice use: causes and consequences. University of Rochester.
12. Fleiss JL (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin 76: 378-382.
13. Cohen J(1960). A coefficient of agreement for nominal scales. Educational and psychological measurement 20: 37-46.
14. Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics 33: 159-174.
15. Hollander JE, Carr BG (2020). Virtually perfect? Telemedicine for COVID-19. New England Journal of Medicine 382: 1679-1681.
16. Trinidad J, Kroshinsky D, Kaffenberger BH, Rojek NW (2020) Telemedicine for inpatient dermatology consultations in response to the COVID-19 pandemic. J Am Acad Dermatol 83: 69-71.
17. Roosan D, Karim M, Chok J, Roosan M (2020). Operationalizing Healthcare Big Data in the Electronic Health Records using a Heatmap Visualization Technique. HEALTHINF 361-368.
18. Matell MS, Jacoby J(1971). Is there an optimal number ofalternatives for Likert scale items? Study I:Reliability and validity. Educational and psychological measurement 3: 657.
19. Lee J, Paek I (2014). In search of the optimal number of response categories in a rating scale. Journal of Psycho educational Assessment 32: 663-673.

Acknowledgement

Ohio Third Frontier to the Ohio Imaging Research and Innovation Network (OIRAIN) 667750 (US); National Institutes of Health ES020866 (JBT), AG048946 (JBT); and Veteran's Administration Merit Awards 5I01BX000853 (JBT) and 1101CX000809 (JBT).