Reliability of Clinician Erythema Assessment grading scale

Jerry Tan, MD,[a,b] Hong Liu, PhD,[c] James J. Leyden, MD,[d] and Matthew J. Leoni, MD, MBA[c]
Windsor, Ontario, Canada; Princeton, New Jersey; and Philadelphia, Pennsylvania

Background: Facial erythema is a clinical hallmark of rosacea and often causes social and psychological distress. Although facial erythema assessments are a common endpoint in rosacea clinical trials, their reliability has not been evaluated.

Objective: The objective of this study was to evaluate the inter- and intrarater reliability of the Clinician's Erythema Assessment (CEA), a 5-point grading scale of facial erythema severity.

Methods: Twelve board-certified dermatologists, previously trained on use of the scale, rated erythema of 28 rosacea subjects twice on the same day. Interrater and intrarater agreement was assessed with the intraclass correlation and κ statistic.

Results: The CEA had high interrater reliability and good intrarater reliability with an overall intraclass correlation coefficient (ICC) for session 1 and session 2 of 0.601 and 0.576, respectively; the overall weighted κ statistic for session 1 and session 2 was 0.692.

Limitations: Raters were experienced dermatologists and there may be a risk of recall bias.

Conclusion: When used by trained raters, CEA is a reliable scale for measuring the facial erythema of rosacea. (J Am Acad Dermatol 2014;71:760-3.)

Key words: Clinician Erythema Assessment; erythema; rating scale; rosacea; scale validation.

Rosacea, a common chronic skin disorder, affects almost 16 million people in the United States as reported by the National Rosacea Society. Rosacea is often characterized by flushing, persistent erythema, visible blood vessels, papules, and pustules.[1,2] Rosacea primarily affects the central facial area (ie, cheeks, nose, chin, forehead, and eyes).[3] A standard classification system, based on primary clinical features, divides rosacea into 4 subgroups: erythematotelangiectatic, papulopustular, phymatous, and ocular.[2] Erythematotelangiectatic rosacea is primarily characterized by flushing and persistent central facial erythema. The latter is a primary feature of rosacea that occurs almost universally among patients with rosacea.[2,4] Currently, there is only 1 Food and Drug Administration–approved pharmacologic treatment that directly targets erythema of rosacea, leaving a potential unmet need among patients with rosacea.

The development of clinical therapies necessitates reliable, responsive, discriminatory, and validated outcome measures to standardize diagnosis, evaluation, and treatment. In reality, however,

From the University of Western Ontario, Schulich School of Medicine and Dentistry (Windsor campus)[a]; private practice, Windsor[b]; Galderma Research and Development, Princeton[c]; and University of Pennsylvania.[d]

Supported by Galderma Research and Development, Princeton, NJ, and Galderma Laboratories LP, Fort Worth, TX.

Disclosure: Dr Tan has served as a consultant, investigator, speaker, and on advisory board, and has received grants and honoraria from Galderma. Drs Leoni and Liu are employees of Galderma Research and Development. Dr Leyden has served as a consultant and on advisory boards, and has received honoraria from Galderma, Allergan, Anacor, Unilever, Sol-Gel, and Acleris and has served as a consultant and received honoraria from Cutanea.

Accepted for publication May 8, 2014.

Reprint requests: Matthew J. Leoni, MD, MBA, 5 Cedar Brook Dr, Suite 1, Cranbury NJ 08512. E-mail: matthew.leoni@galderma.com.

Published online July 7, 2014.
0190-9622/$36.00

http://dx.doi.org/10.1016/j.jaad.2014.05.044

the clinimetric properties of outcome measures used in clinical studies are often unreliable or imprecise.[5]

A variety of methods to evaluate the erythema of rosacea exist. The National Rosacea Society proposed a single scale for erythema grading in rosacea, but it was not formally evaluated. Investigator Global Assessment scales for rosacea, which often contain an erythema section, have been used in multiple clinical studies without prior clinimetric evaluation or validation.[4,6] Other examples include reflectance spectrophotometry and a visual analog scales.[7,8] Reflectance spectrophotometry provides objective measurements of redness only on selected anatomical sites, and is therefore not suitable for evaluating the entire face. A visual analog scale is a sensitive scale usually used for pain assessment, but has not been validated for the assessment of erythema. Therefore, the evaluation of facial erythema in rosacea still lacks a validated and clinimetrically rigorous scale.

The objective of this study was to evaluate the reliability of a Clinician Erythema Assessment (CEA) scale by determining interrater and intrarater agreement.

CAPSULE SUMMARY

- Rosacea is a common, chronic skin disease marked by central facial erythema.
- Although facial erythema is often evaluated in clinical trials, the reliability of the assessment scales have not been evaluated.
- This study evaluated the interrater and intrarater reliability of the Clinician Erythema Assessment scale and found that the Clinician Erythema Assessment is a reliable scale for measuring facial erythema of rosacea.

METHODS

This study was evaluated in accordance with federal and local regulatory requirements. The study was reviewed and approved by institutional review boards. All subjects provided written informed consent before entering the study.

A common method to demonstrate scale reliability is to determine interrater and intrarater agreement. One way to assess interrater agreement is to calculate the κ statistic, often denoted as κ. Kappa measures pairwise agreement among observers; weighted κ, in this case Fleiss κ, a generalization of the simple κ coefficient, uses weights to quantify the relative difference between grades. Weighted κ is especially useful for ordinal data with more than 2 grades.[9] Interrater and intrarater agreement can also be assessed using a version of intraclass correlation coefficient (ICC) that measures the consistency of grades among observers.[10]

The CEA scale was evaluated in a 1-day, single-center study involving men and women aged 18 years or older. Subjects had a clinical diagnosis of rosacea and no more than 2 facial inflammatory lesions (papules/pustules) at both the screening and day-1 visits. During the screening visit, the principal investigator ensured an appropriate distribution of each CEA category of erythema (grades 1-4) in the study population.

Twelve US board-certified dermatologists participated in a consensus training and harmonization session with reference photographs (Table I) on the use of the CEA before live subject evaluations. Physicians discussed results of the evaluations. Training was repeated, if necessary, until the physicians achieved consensus. During rating sessions using CEA, each dermatologist evaluated each subject twice with at least a 2-hour interval between evaluation sessions. Subjects were renumbered and reordered in the second session to reduce recall bias.

Weighted κ statistic for interrater reliability for each evaluation session and intrarater reliability for each rater was calculated using the method of Fleiss and Cohen.[11] ICCs of interrater reliability for each evaluation session and intrarater reliability for each rater were calculated using the method of Shrout and Fleiss.[10] Interrater agreement was assessed by comparing each rater's score with the mean of all raters' scores for the same subject. Intrarater agreement (test-retest) was assessed by comparing each rater's first and second session scores for the same subject. To determine sample size, assuming the obtained interrater and intrarater correlation coefficients were approximately 0.8, approximately 10 raters and 25 subjects were required to ensure that the width of the confidence interval of both coefficients was at least 0.2.[12]

There are no universally accepted standards for reliability, but the following criteria have been proposed by Landis and Koch[9] for interpreting agreement with the weighted κ statistic: less than 0, poor; 0 to 0.2, slight; 0.2 to 0.4, fair; 0.4 to 0.6, moderate; 0.6 to 0.8, substantial; and 0.8 to 1, almost perfect. Fleiss suggests that an ICC of 0.4 or greater but less than 0.75 indicates a fair to good reproducibility, and 0.75 or greater indicates an excellent reproducibility.[5]