

A Comprehensive Critique and Review of Published Measures of Acne Severity

^aTAMARA AGNEW, BA (HONS), PhD CANDIDATE; ^bGARETH FURBER, BPSYCH (HONS), PhD (CLINICAL PSYCHOLOGY); ^aMATTHEW LEACH, RN, BN(HONS), ND, DipClinNutr, PhD;
^bLEONIE SEGAL, B. Econ, M. Econ PhD

^aSchool of Nursing & Midwifery, University of South Australia, SAHMRI, Adelaide, South Australia; ^bHealth Economics & Social Policy Group, Centre for Population Health Research, University of South Australia, SAHMRI, Adelaide, South Australia

ABSTRACT

Objective: Acne vulgaris is a dynamic, complex condition that is notoriously difficult to evaluate. The authors set out to critically evaluate currently available measures of acne severity, particularly in terms of suitability for use in clinical trials. **Design:** A systematic review was conducted to identify methods used to measure acne severity, using MEDLINE, CINAHL, Scopus, and Wiley Online. Each method was critically reviewed and given a score out of 13 based on eight quality criteria under two broad groupings of psychometric testing and suitability for research and evaluation. **Results:** Twenty-four methods for assessing acne severity were identified. Four scales received a quality score of zero, and 11 scored ≤ 3 . The highest rated scales achieved a total score of 6. Six scales reported strong inter-rater reliability ($ICC > 0.75$), and four reported strong intra-rater reliability ($ICC > 0.75$). The poor overall performance of most scales, largely characterized by the absence of reliability testing or evidence for independent assessment and validation indicates that generally, their application in clinical trials is not supported. **Conclusion:** This review and appraisal of instruments for measuring acne severity supports previously identified concerns regarding the quality of published measures. It highlights the need for a valid and reliable acne severity scale, especially for use in research and evaluation. The ideal scale would demonstrate adequate validation and reliability and be easily implemented for third-party analysis. The development of such a scale is critical to interpreting results of trials and facilitating the pooling of results for systematic reviews and meta-analyses. (*J Clin Aesthet Dermatol.* 2016;9(7):40–52.)

Acne vulgaris (acne) is a polymorphic skin condition, characterized by physical symptoms (i.e., lesions, nodules, cysts,¹ scarring² and psychological sequelae.^{3–10} The condition is dynamic and complex, with constantly fluctuating acute and chronic symptoms¹¹ and inconsistent distribution.¹² The complexity of these attributes makes it inherently difficult to evaluate.^{13,14} In fact, the measurement and grading of acne severity is a recognized challenge impeding high quality research.^{15–20} This issue was most recently highlighted in a Cochrane review of the efficacy and safety of minocycline, where authors concluded that “Its efficacy.... could not be reliably determined due to the poor methodological quality of the trials and lack of consistent choice of outcome measures.”²¹

Approaches to the assessment of acne severity. There are four broad approaches to the assessment of acne severity: 1) lesion counting, 2) global acne severity grading, 3) subjective self-assessment, and 4) multimodal digital

imaging. These techniques are described below.

Acne lesion counting involves tallying the number of different lesion types and is typically done *in-situ*. It is described as precise, objective, and highly discriminative (describing severity down to the individual lesion level),¹⁹ and can provide continuous data for statistical analysis.^{16,18} However, it is also time-consuming, intrusive for the patient/subject and does not capture various clinical aspects of symptoms including concentration, distribution and size of lesions, or skin redness (erythema).^{17,18} In addition, counting requires specialist knowledge and training to administer, and assessment is dependent on variables, such as lighting, the assessor’s visual capacity and underlying skin quality.²⁰

Global severity grading (grading) is a universal assessment of acne in which a client’s presentation is compared against text descriptions or photographs. Grading is promoted for use within a clinic setting as it is practical and easy to use¹⁸; graders are also able to evaluate

DISCLOSURE: The authors report no relevant conflicts of interest.

ADDRESS CORRESPONDENCE TO: Tamara Agnew, Centre for Population Health, University South Australia, 8th Floor, South Australian Health & Medical Research Institute (SAHMRI) Building, North Terrace, Adelaide SA 5000 (Postal: WEA-48, GPO Box 2471, Adelaide SA 5001); E-mail: tamara.agnew@unisa.edu.au

a range of aspects pertinent to severity, including number, type, and size of lesions, but also the presence and coverage of inflammation, erythema, and seborrhea.^{19,20} This approach is often criticized for being subjective, less sensitive to change, and too simplistic to provide useful insight^{16,22}; but these criticisms seem to reflect aspects of individual instruments (i.e., the interpretation of text description or limited categories in a scale), rather than a problem that is inherent in the approach. Some have noted that the introduction of photographic scales has provided an easier to use and more precise system relative to older text-based scales.²³

Subjective self-assessment is variously portrayed as simply the identification of the condition²⁴ or as a global evaluation of severity category provided by the patient.^{25,26} It has been identified as an unreliable approach to the assessment of severity,^{24,26} although studies have reported agreement between perceived acne severity and poor quality of life.²⁷

Multi-modal imaging is the use of specialist photographic equipment and computer algorithms to capture and analyze lesion types, extent of erythema, and pigmentation changes. Described as “objective,”^{28,29} this new technology relies on specific equipment, which may include purpose-built technology or require multiple items of specialist equipment including ultraviolet (UV) A lamps, fluorescent lights, polarizers, filter wheels, and a digital camera.²⁸

Quality of existing scales. A number of expert reviews have explored the range and quality of existing acne severity rating scales. Witkowski and Parish¹⁶ published an expert evaluation of scales published prior to 1997, including a description of two main approaches, and their own opinion regarding the superiority of lesion counting over grading. Tan et al¹⁸ published an expert review in 2008, which identified scales published between 1982 and 2007; but, the review did not critically appraise the quality of the scales. Notwithstanding, the authors of this article did describe scales for acne scarring and quality of life, and did recognize the complex nature of the condition and the necessity for a holistic scale that combines all of the physical and psychosocial aspects of acne vulgaris. In a more recent, but somewhat brief, overview of the most frequently used scoring systems for acne assessment, Adityan et al³⁰ concluded comparably to Tan et al,¹⁸ and Witkowski and Parish,¹⁶ that the nature of acne makes the condition difficult to assess, and that the measures currently available are unable to detect multiple aspects of the condition. A scale that is able to capture these aspects, which is quick, easy to use, and accurate, would be beneficial in clinical practice and research. Witkowski and Parish¹⁶ went further to suggest that the use of a counting method would provide the best results for epidemiological and clinical research, while a grading method would be best suited for the clinical setting.

Previously, Lehmann et al¹⁹ had conducted a methodologic review to establish the effectiveness of acne treatments by reviewing trials conducted before 1999. This

study did not specifically set out to identify outcome measures, and it does not formally evaluate them. Nevertheless, during the course of this review, the authors identified 25 different assessment methods, including 19 different approaches to counting lesions.

More recently, Barratt et al²⁰ conducted a systematic review of acne scales published prior to 2001. The 2009 review²⁰ investigated the various measurement properties of nine different methods, including three novel technologies, such as polarized light or fluorescence photography. Using a dichotomous (yes/no) scoring scheme (capturing only whether evidence was reported in the paper), each method was evaluated in two major categories (i.e., construction and evaluation) and eight subcategories, including scale development, piloting, item description and reproducibility, inter and intra-rater reliability, validity, and responsiveness to change. Barratt et al²⁰ concluded that the plethora of mostly nonvalidated scales made secondary analysis and the interpretation of results difficult. The information provided by the authors of this systematic review was limited to whether or not the acne scales provided evidence of reliability or validity, rather than making an assessment about the quality of this information. The review was also limited to nine scales published prior to 2001.

Recognizing the concerns with existing measures of acne severity, and the inadequate appraisals of acne severity scales to date, the authors conducted a critical review of original published acne scales to formally evaluate their quality against a set of predetermined criteria. The aim was to assess the applicability of current acne scales for use within the clinical or community trial setting as distinct from clinical practice. This review represents the first published critical appraisal of acne severity scales. Building on information provided by previous authors in the field,^{16,18-20,30} it aims to examine not only the usefulness of the instrument in a clinical or community trial setting, but also evaluate the methodology used to describe various psychometric outcomes.

METHODS

Literature search. A systematic search of the literature was conducted to identify publications that described methods for grading acne. The search was initially performed in November 2013, and updated in August 2015. MEDLINE, CINAHL, Scopus, and Wiley Online were searched using various combinations of the search terms acne vulgaris, assessment, diagnosis, classification, measurement outcome, evaluation, and scale. To focus these searches, terms such as validation, reliability, and psychometric were also included. Papers were limited to those published in the English language and in peer-reviewed, scholarly journals and dermatology textbooks. There were no limits placed on date of publication. Also searched were the reference lists of expert opinion pieces, research articles, and literature/systematic reviews.^{16,18-20,30} The search was conducted by TA. The articles were first screened by title and abstract, and then a full-text copy of the article was retrieved to confirm eligibility.

TABLE 1. Quality criteria—acne severity instruments

	CRITERION	SCORE
I. Psychometrics		
CORE CRITERIA	Validity—has the scale been subject to formal validity testing?	2—Excellent correlation (ICC ≥ 0.75) ³² 1—Fair to good (ICC 0.41–0.74) ³² 0—no evidence provided or evidence of fair or poor correlation (ICC < 0.40) ^{32,33} ; fails to meet basic criteria
	Inter-rater or between rater reliability	2—Excellent correlation (ICC ≥ 0.75) ³² 1—Fair to good (ICC 0.41–0.74) ³² 0—no evidence provided or evidence of fair or poor correlation (ICC < 0.40) ^{32,33} ; fails to meet basic criteria
	Intra-rater or within rater reliability	2—Excellent correlation (ICC ≥ 0.75) ³² 1—Fair to good (ICC 0.41–0.74) ³² 0—no evidence provided or evidence of fair or poor correlation (ICC < 0.40) ^{32,33} ; fails to meet basic criteria
	Reports data suggesting that the instrument is sensitive to change	2—yes, evidence provided and is sensitive to change 1—authors suggest instrument is sensitive to change, but not supported 0—not provided or evidence provided and weak or not sensitive to change
II. Suitability for use in research and evaluation		
DESIRABLE CRITERIA	Population representativeness	2—test sample drawn from the general population 1—test sample drawn from limited population (e.g., those attending a clinic or school children) 0—limited description or test population not described
	Ease of use/feasibility: Based on feedback from raters about clarity and ease of use of the instrument	1—yes, feedback provided and described as good 0—negative feedback, no feedback provided
	Independent third-party assessment and verification; may include a clearly defined protocol for centralized assessment, independent ratings and verification purposes.	1—yes 0—No
	Non-expert evaluation—evidence of testing by non-expert evaluators	1—yes 0—no

Inclusion criteria. Articles were included if they described an approach for measuring acne severity either in a clinical or research setting. Articles were excluded if they assessed acne scarring, measured only one pathophysiological aspect of acne (such as sebum production), or described photographic techniques for capturing acne symptoms, which did not provide a method for visual assessment (e.g., fluorescence digital photography or polarized light photography).

Data extraction. In keeping with the aim of the review, TA extracted data related to the criteria set out below including evidence of validity, inter and intra-rater reliability, sensitivity to change, population, ease of use,

independent remote assessment, and non-expert evaluation.

Quality criteria for assessment. Quality criteria were developed based on previous research that identified necessary components for a novel acne severity scale.³¹ The aim of the current review was to establish whether published acne severity instruments are useful for research and evaluation in a community-based research setting. Collectively, the quality assessment criteria fell under two broad groupings (Table 1):

1) The psychometric properties of the scale—the purpose of these criteria was to answer specific questions about the symptoms of acne, with evidence of validation,

inter-rater (across) and intra-rater (within) reliability, and sensitivity to change forming the core criteria; and

2) Suitability for use in research and evaluation—these are the “desirable” qualities, those that are advantageous or beneficial, including ease of use, suitability for independent third-party assessment and verification, non-expert evaluation, and appropriateness of test populations (Table 1).

The application of scores based on individual criteria, and an accumulated, overall quality score enabled the authors to recognize quality published or supporting evidence for each scale, and to rank the papers accordingly. A scoring system was devised to rate each quality attribute as described below.

Core criteria—psychometrics. Three levels of performance were set for each core criteria; validity, inter-rater reliability, intra-rater reliability, and sensitivity to change. If the evidence in a category was strong, the scale received the maximum score of 2 points. If there was weak evidence of validity, reliability, or sensitivity, then the assigned score was 1. If there was no evidence of any psychometric testing in a category, or the instrument performed poorly, the score assigned was 0.

Desirable criteria—suitability for use in research and evaluation. *Population Representativeness.* A three-level scale was applied to assess the likely representativeness of the population used to create the scale and thus, likely generalizability. If there was no description of the population used to develop the scale, a score of 0 was allocated. If the population was narrowly defined (e.g., dermatology clinic patients, or school children), the score assigned was 1. If the scale was developed and tested on a general, random sample, the allocated score was 2.

Ease of use, independent third-party assessment, and non-expert evaluation. In these criteria, a score (based on a simple yes [1] or no [0]) was awarded if the authors provided descriptive evidence within the articles.

Scoring of acne scales. Each criteria was allocated a score of 0, 1, or 2 (where relevant), which were summed to yield a total quality score ranging from 0 to 13, with higher scores indicative of a higher quality measure of acne severity. Scoring was completed by the lead author (TA), and reviewed and confirmed by all authors.

RESULTS

Figure 1 summarizes the outcomes of the literature search. In total, 211 documents were retrieved, of which 54 duplicates were removed, and 119 articles excluded following title review. The abstracts of 38 articles were read and 16

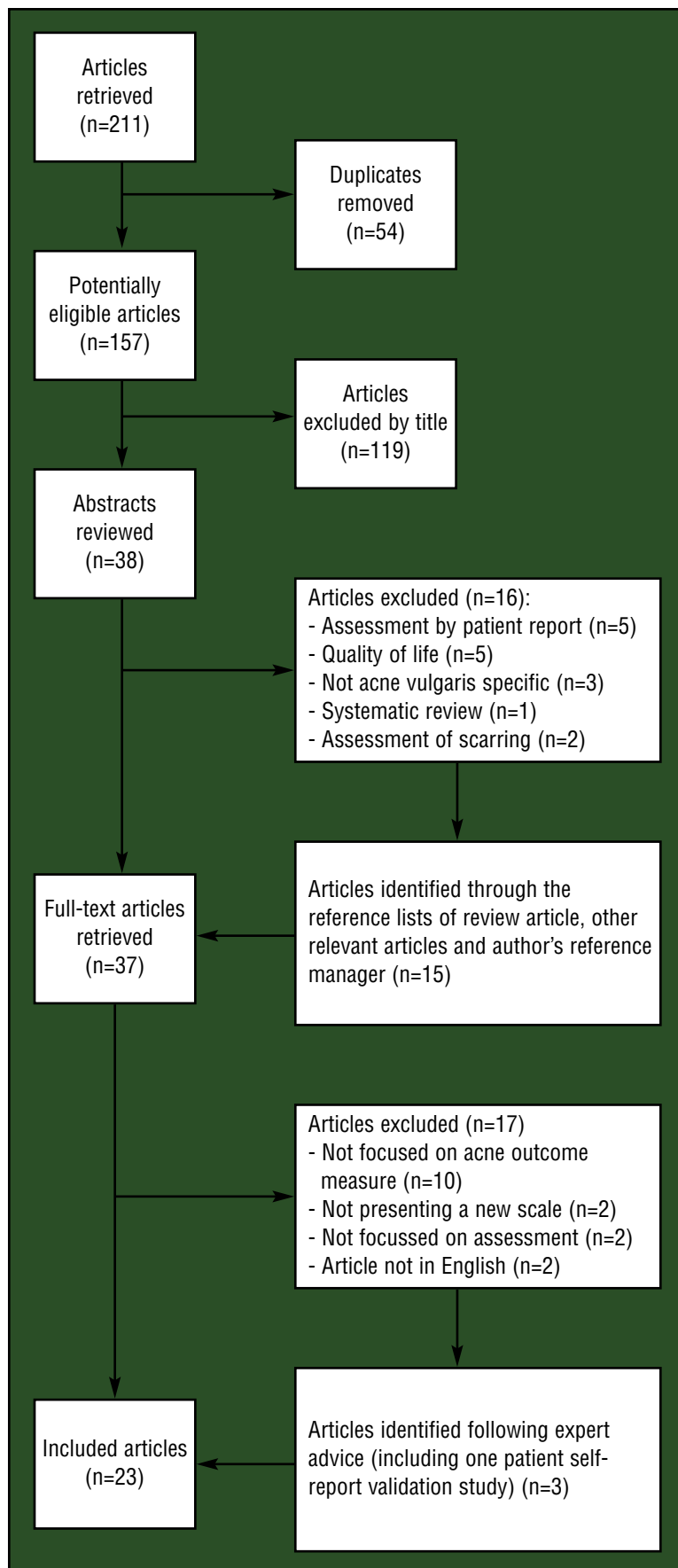


Figure 1. Summary of search results

subsequently excluded (not reporting on a new acne scale). A hand search of the reference lists of reviews and articles identified 15 additional articles. After reviewing the content of 37 documents, 17 were excluded as they did not focus on acne outcome instruments, did not present a novel approach or idea, did not focus on assessment of physical symptoms, or were not reported in the English language. Following comments from experts, three more documents were added. Twenty-four methods were included in the final review.

Of the 24 published methods for assessing acne severity, seven (29%) were lesion counting methods, including one multimodal imaging article, which described a method for capturing and counting acne lesions, and 17 (71%) were global severity grading scales, including one patient self-report method (Table 2). For the purpose of this critique, Burke and Cunliffe's Leeds technique³⁴ was evaluated as two different methods (counting and grading) as each provided unique psychometric testing results. Where independent validity or feasibility studies have been conducted, results have been treated as a single data source for that scale; further, the data were not additive, and assignment of scores was based on the highest performing results as they were reported.

Of the counting methods, two provided an overall grade based on the number of a particular lesion type, multiplied by a severity factor^{35,36}; two attempted to reduce counting fatigue by providing novel approaches to counting sections of face^{37,38}; one relied on the assessment of acne by a computer program²⁹; one counted the lesions based on type across the whole face,³⁴ while another counted inflammatory and noninflammatory lesions and applied a grade based on this.³⁹ Two grading methods were based solely on photographic images,^{40,41} four were based on text only,^{34,42-44} and seven provided both text and images,^{22,45-49} although one⁵⁰ did not publish the images. Two of the global scales are modified versions of the original Investigators Global Assessment (IGA),⁴⁶ first published in 2005, revised in 2007,⁵¹ and variously validated in 2007⁵² and 2011.⁴⁵ A six-point scale published in 2005⁴⁷ is an adapted version of the original published in 1982.¹³ Finally, the revised version of the Leeds global grading scale appears to be a version of the original black and white images from the original article,³⁴ re-published in color. The authors included one multimodal imaging measure, which provided some evidence of validity and reliability testing for the assessment of acne signs and symptoms.

A summary of the basic attributes and the quality assessment of each of the 24 included scales are reported in Table 2. The maximum quality score was 13; the highest score, 6, was achieved by three measures—the Leeds Revised Acne Grading (LRAG),^{40,53} The Global Acne Severity Scale (GEA),⁴⁵ and The Escala de Gravedada del Acné Española (EGAE). Sixty-five percent (n=15) of the measures generated a total quality score ≤ 3 , with four of these (27%) scoring 0. Performance in relation to the two broad quality areas is discussed below.

Core score—psychometric properties. Inter-rater

reliability was reported for 14 (58%) methods; of these, seven (33%) reported strong correlation scores of ≥ 0.75 .^{23,29,32,34,41,46,48,54} Intra-rater reliability was reported for six (30%) methods; four (71%) reported strong ICCs ranging from 0.80 to 0.97^{23,34,38,54} and one (14%) reported evidence of moderate positive correlation.³⁵ One instrument reporting intrate positive correlation⁵³ was not assigned a score for this category as there were concerns about methodology. Specifically, the authors did not conduct the intra-rater assessment in a traditional test/re-test design; instead they rated participants whose condition appeared to remain “stable”, at two different time points, weeks apart, and compared the results. Test/re-test reliability relies on the same test being administered on two occasions⁵⁷; *in situ* test/re-test can be difficult to achieve given the polymorphic nature of acne. In order to overcome this, the design should ensure that either an image is captured and the same image is re-assessed; or *in situ* assessment occurs on the same day (see Lucky et al³⁸) with hours separating rather than days or weeks in order that the same lesions are counted twice and not different presentations of the same condition.

Three scales (12.5%)^{41,48,52} reported sensitivity to change; Cook et al⁴⁸ reported clinically meaningful change, but the authors did not clearly describe how this was evaluated. Puig et al (2013)⁴¹ reported the EGAE as sensitive enough to detect change among a group of participants who were compliant with treatment, but not for participants who were not adherent with the protocol of the study.

The psychometric property of validation is one of the most important criteria for determining the quality of an instrument (i.e., does the scale measure what it is supposed to measure?). It is especially difficult to determine validity for acne severity assessment tools as there is no gold standard instrument against which validity can be assessed. Regardless, four (16.6%) instruments were assessed against existing, unvalidated scales.

Authors of the Comprehensive Acne Severity Scale (CASS) assessed their grading method against the Leeds grading technique.⁵² They found that there were significant positive correlations between the CASS and the Leeds technique (Spearman's correlation coefficient, $r_s 0.823$); however, the Leeds technique is itself not validated and has been described as skewed to the severe end by the same authors.¹⁸ Validity was also reported in a study that assessed the feasibility of the LRAG in Spain.^{40,53} The validity of the LRAG in the Spanish population was based on a comparison with a counting method which was not described by the authors and so outcomes were difficult to corroborate, and results should be interpreted with caution. Participating dermatologists performed lesion counts prior to categorizing acne symptoms on the scale which may have influenced the grading outcome.

Desirable score—suitability. Scores were also generated to acknowledge factors that make the scale appropriate for use within a research setting and this category could potentially generate five additional points. Four (16.6%) assessment methods did not achieve any score in this major category,^{22,39,55,56} while 11 (46%) scored

TABLE 2. Quality assessment of acne measurement methods

Citation	Assessment Method	CORE QUALITIES				DESIRABLE QUALITIES				SCORES			Notes to Assessment
		Validity	Inter-rater	Intra-rater	Sensitivity	Population	Feasibility	Assessment	Evaluation	Core Score	Desirable Score	Score Total	
DRÉNO ET AL 2011⁴⁵	Global Assessment: The Global Acne Severity Scale (GEA Scale) is a 6-point photo-numeric scale with descriptive text. Assessment <i>in situ</i> and remote assessment	0	2	2	0	1	0	1	0	4	2	6	(d)
O'BRIEN ET AL 1998⁴⁰ (ORIGINAL) GUERRA-TAPIA ET AL 2010⁵³ (VALIDATION)	Global Assessment: Leeds Revised Acne Grading Scale (LRAG) is a revision of the original Leeds Grading technique. ³⁴ A photo-numeric scale with 13 categories	1	1	0*	1	1	1	1	0	3	3	6	(a) (b)
PUIG ET AL 2013⁴¹	Global Assessment: The Escala de Gravedada del Acné Española (EGAE) is a photo-numeric acne for face, chest, and back. Assessment <i>in situ</i>	1	2	0	0	1	1	1	0	3	3	6	(i)
BURKE & CUNLIFFE 1984³⁴	Counting: Counting of inflammatory and noninflammatory lesions. Assessment <i>in situ</i>	0	2	2	0	1	0	0	0	4	1	5	(e)
BURKE & CUNLIFFE, 1984³⁴ BERGMAN ET AL 2009⁵⁴	Global Assessment: The Leeds technique is an 11-point scale, supported by black and white images. Assessment <i>in situ</i> . Validation of Tool ⁵⁴ : Leeds Global Assessment by remote assessment	0	2	2	0	1	0	0	0	4	1	5	(f) (g)
TAN ET AL 2007⁵²	Global Assessment: The Comprehensive Acne Severity Scale (CASS) is a 6-point scale with text description of each category. Assessment <i>in situ</i> .	2	0	0	2	1	0	0	0	4	1	5	(h)
PATWARDHAN ET AL 2013²⁹	Counting: Multimodal facial imaging system captures very high resolution images; classification algorithms pre-programmed for computer-based analyses	0	2	0	0	1	0	1	1	2	3	5	(c)

Continued on the next page

TABLE 2 cont. Quality assessment of acne measurement methods

Citation	Assessment Method	CORE QUALITIES				DESIRABLE QUALITIES				SCORES			Notes to Assessment
		Validity	Inter-rater	Intra-rater	Sensitivity	Population	Feasibility	Assessment	Evaluation	Core Score	Desirable Score	Score Total	
LUCKY ET AL 1996 ³⁸	Counting: Five zone counting according type and number of lesions. Assessment <i>in situ</i>	0	1	2	0	1	0	0	0	3	1	4	(j)
CDER, 2005 ⁴⁶ – GLOBAL ASSESSMENT BERGMAN ET AL 2009 ⁵⁴	Global Assessment: The Investigators Global Assessment Scale (IGA) is a 5-point severity scale with descriptive text. Validation of Tool ⁵⁴ : IGA by remote assessment	0	2	0	0	1	0	1	0	2	2	4	(l) (m)
COOK ET AL 1979 ⁴⁸	Global Assessment: Photographic Standard 8-point scale supported by text description and 5 black and white photographs. Remote assessment	0	2	0	1	0	0	0	0	3	0	3	(k);
LEYDEN ET AL 2005 ⁴⁷	Global Assessment: A revised version of scale presented by Allen and Smith ¹³	0	1	0	0	1	0	1	0	1	2	3	(y)
SAMUELSON 1985 ⁴²	Global Assessment: 9-point grading supported by text description. Assessment <i>in situ</i>	0	1	0	0	1	0	1	0	1	2	3	(r)
DE ALMEIDA ET AL 2013 ²⁶	Self-report: Level of severity light (≤ 10), moderate (11–30), and severe (31+ lesions) compared to counting by four dermatologists	0	0	0	0	2	0	0	1	0	3	3	(t)
HAYASHI ET AL 2008 ⁴⁹	Global Assessment: 4-point photographic scale; remote assessment	0	1	0	0	0	0	1	0	1	1	2	(n)
PASCOE ET AL 2015 ⁴³	Global Assessment: Physician Global Assessment (PGA) 5-point scale (0–4) where 0 is clear and 4 is severe	0	1	0	0	1	0	0	0	1	1	2	(o)

Continued on the next page

TABLE 2 cont. Quality assessment of acne measurement methods

Citation	Assessment Method	CORE QUALITIES				DESIRABLE QUALITIES				SCORES			Notes to Assessment
		Validity	Inter-rater	Intra-rater	Sensitivity	Population	Feasibility	Assessment	Evaluation	Core Score	Desirable Score	Score Total	
LIDEN ET AL 1980³⁵	Lesion Counting/Grading: Assessment made by defining the type of lesion and multiplying the number of those lesions by the severity index factor	0	0	1	0	0	0	1	0	1	1	2	(q)
BURTON ET AL 1971⁴⁴	Global Assessment: The Burton Scale is a 6-point scale based on overall impression of acne	0	1	0	0	1	0	0	0	1	1	2	(s)
MICHAËLSSON ET AL 1977³⁶	Counting: Each type of lesion has a factor of between 0.54.0; lesions counted and then multiplied by the factor. Improvement is graded on a 5-point scale	0	0	0	0	1	0	0	1	0	2	2	(v) (w) (x)
BEYLOT ET AL 2010⁵⁵	Global Assessment: Simple 3-grade acne classification (mild, moderate and severe). Remote assessment	0	1	0	0	0	0	0	0	1	0	1	(p)
CHRISTIANSEN ET AL 1976³⁷	Counting: 5cm ring placed on the most severely affected area; papules, pustules and comedones within the ring are counted	0	0	0	0	1	0	0	0	0	1	1	(v) (w) (x)
BLANEY AND COOK 1976⁵⁰	Global Assessment: 9-point scale supported by text description and photograph (images not published)	0	0	0	0	0	0	0	0	0	0	0	(v) (w) (x)
DOSHI ET AL 1997⁵⁶	Global Assessment: Global acne grading system (GAGS) 5-point scale based on most prominent lesion in any of 6 regions, multiplied by a factor of 1, 2, or 3 to achieve a global score	0	0	0	0	0	0	0	0	0	0	0	(v) (w)
PLEWIG ET AL 1975³⁹	Counting: Recording type and number of lesions and categorized according to whether they are comedonal or papulopustular. Grade based on type and number of lesion	0	0	0	0	0	0	0	0	0	0	0	(u) (v) (w)

Continued on the next page

TABLE 2 cont. Quality assessment of acne measurement methods

Citation	Assessment Method	CORE QUALITIES				DESIRABLE QUALITIES				SCORES			Notes to Assessment
		Validity	Inter-rater	Intra-rater	Sensitivity	Population	Feasibility	Assessment	Evaluation	Core Score	Desirable Score	Score Total	
POCHI ET AL 1991²²	Global Assessment: Simple mild, moderate, or severe scale for inflammatory lesions	0	0	0	0	0	0	0	0	0	0	0	(u) (v) (w)

- a) The original publication did not provide any description of statistical testing for validation purposes; scores for this instrument are derived from subsequent assessment
- b) Inter-rater reliability ICC=0.72 (95% CI, 0.62-0.80); *Authors provide statistical evidence of intrarater reliability as Cronbach α = 0.82; however, we have concerns about the methodology associated with achieving the scores and so we have awarded this criteria 0; Validity (compared to counting) at baseline for facial acne vulgaris: noninflammatory lesions ($p<0.050$) and inflammatory lesions ($p<0.012$); and at follow-up: inflammatory lesions ($p<0.000$) and noninflammatory ($p = 0.070$)
- c) Inter-rater correlation coefficient for inflammatory lesion counts compared to 3 physicians reported 0.711, 0.828, and 0.877, respectively. Noninflammatory lesion count inter-rater correlation coefficient reported as 0.92
- d) Inter-rater reliability ICC (photographic evaluation) = 0.8057 (95% CI: 0.75100.8494; $p<0.0001$); inter-rater agreement (*in situ* evaluation) R = 0.8385.⁴⁵; Intrarater reliability ICC = 0.7982 (95%CI: 0.7559 to 0.8339; $p<0.0001$)⁴⁵
- e) Inter-rater reliability (face) non-inflamed r = 0.83 (<0.002) inflamed r = 0.86 (<0.001); Intrarater reliability (face) non-inflamed r = 0.87 (<0.002) inflamed r = 0.92 (<0.001)
- f) Inter-rater reliability (face) r = ICC 0.89; Intrarater reliability (face) r= ICC 0.92
- g) Inter-rater reliability K = 0.879 - 1.0; Intrarater reliability K = 0.381 (or K=0.74 for “close agreement”)
- h) Concurrent validity: Current scale and Leeds Revised Grading including photography: Spearman’s correlation = 0.823; $p<0.001$; Responsiveness to change; face 0.762 ($p<0.001$)
- i) Inter-rater reliability assessed as Kendall’s W = 0.773, $p<0.001$; Concurrent validity: Current scale and LRAg including photography: Spearman’s correlation = 0.889 (CI 95%; 0.8740.902; $p<0.001$); Sensitivity to change assessed, but not significant: effect size 0.12 (95% CI 0.025 – 0.26) Cohen’s d = 0.19 $p = 0.102$; 95.6% of dermatologists found the instrument easy to use in clinical practice
- j) Inter-rater reliability (total lesion count) ICC = 0.61 (95% CI); Intrarater reliability range ICC = 0.82 to 0.97 (95% CI); Feasibility is described as ‘a function of the experience of the rater’; comfort with the instrument not a predictor of individual reliability
- k) Inter-rater correlation coefficients (Weeks 3 and 12) .785 and .891, respectively ($p<0.01$); clinically significant change for *in situ* assessment = 1.5
- l) Draft guidance for industry
- m) Inter-rater reliability K = 0.606 – 0.926; Intrarater reliability K = 0.3119 (or K=0.96 for “close agreement”)
- n) Total agreement in 45.1% (n=176) cases; 2/3 agreement in 23.6% (n=92) cases; 1/3 agreement in 14.9% (n=58) and no agreement in 16.4% (n=64) (Hayashi); Validity compared to counting lesions ICC: comedones 0.011 (CI 95% 0.007-0.015 $p<0.0001$); papules + pustules 0.034 (CI 95% 0.030-0.038, $p<0.0001$); nodules + cysts 0.110 (CI 95% 0.082 – 0.138, $p<0.0001$)
- o) Inter-rater agreement K=0.68 (0.61-0.90)
- p) Inter-rater reliability kappa = 0.2 (“minimal” acne) to 0.6 (“severe” acne)
- q) Intrarater reliability equal assessment reviewer A: 20/37 (0.54 95% CI 0.38-0.70) and reviewer B: 25/37 (0.68 95% CI 0.53-0.83)
- r) Frequency of agreement (within one grade) between investigator and participant = 88%; agreement between independent reviewers = 74%
- s) Total agreement described as a percentage: 77.2% (n=1204); one grade difference for 22.5% (n=351) and 0.3% (n=5) rejected as more than 1 degree of difference
- t) Dermatologist reported acne present in 85.76% of the population; self-report was 76.12%; sensitivity self-report = 80.95%; specificity self-report = 52.88%; tests of significance not reported
- u) Published in text book or conference paper
- v) No testing described
- w) No description provided of how the instrument was developed
- x) Developed *de novo* in clinical or community based trial (with slight or no description of testing of instrument)
- y) Inter-rater reliability r= 0.648-0.836 with the majority reported as $r \geq 0.61$; tests of significance not reported. Intra-rater reliability tested, but not reported

only one out of a possible five.^{34,35,37,38,42,47-50,52} Six methods (25%) generated the majority of the overall score in this category, which may exaggerate the benefit of the method within the research or evaluation setting; results should be read in conjunction with scores from the core qualities section.^{26,29,36,37,42,47}

Population representativeness. Seven assessment methods (30%) did not describe their test population,^{22,35,37,39,46,49,55,56} and two gave very limited descriptions.^{36,48} No scale scored a maximum of two points in this category as test populations were mainly convenience samples drawn from dermatology settings. Where scales provided visual representation of a category, the images were homogenous in terms of cultural representation and age, making it difficult to determine whether the images would be suitable in a different study population or environment.

Ease of use/feasibility. Puig et al⁴¹ provided evidence that raters found the scale easy to use; Lucky et al³⁸ asked raters to describe their comfort using the technique and found that 25 percent (n=3) were not comfortable, eight percent (n=1) were fairly comfortable, 42 percent (n=5) were comfortable, and 25 percent (n=3) were very comfortable; however, levels of comfort were not a predictor of reliability among this group; and reliability scores varied widely among those who had similar years of experience in rating acne. Guerra-Tapia et al⁵³ reported that the LRAg was easy to use for 89.5 percent of dermatologists, but the authors did not report the number of dermatologists in the study.

Independent third-party assessment and verification. There were nine reports of third-party assessment. Dréno et al²³ described a process where seven investigators assessed 22 participants *in situ*, and then 34 images were assessed by the same observers (Inter-rater ICC 0.8057, 95% CI: 0.7510-0.8494; $p < 0.0001$; Intra-rater ICC 0.7982, 95% CI: 0.7559-0.8339; $p < 0.0001$); *in situ* assessment occurred following assessment of photographic images, and it is not clear whether the images were of the same people assessed by the observers. Cook et al⁴⁸ clinically assessed participants at baseline and subsequent visits. Slides were taken and projected for assessment by panelists who were asked to determine whether there had been any clinical change (improvement or worsening) and whether they thought this was clinically significant, but no outcomes or grading were reported. In Samuelson's⁴² study, participants reported the use of *in situ* clinical assessment and photography, of which images were projected onto a wall for independent assessment. Overall, the reviewers scored the images 1 to 2 grades lower than *in situ* assessment. The feasibility study by Guerra-Tapia et al⁵³ reported "statistically significant" findings for independent assessment of images despite not reporting any p values, and presenting an ICC value (0.72), which was lower than the authors own suggested acceptable limit (ICC 0.80), as well as being in the moderate range according to Streiner et al.³² Blaney and Cook⁵⁰ used images to provide a comparison between the first and last treatment within a clinical trial

where authors described the results of the trial, but did not report any analysis of psychometric or suitability testing. Three dermatologists assessed 244 images in the study by Hayashi et al,⁴⁹ which describes how a grading scale was established, but does not show results of psychometric testing of the grading system, while the study by Beylot et al⁵⁵ was based on remote assessment, where eight experts assessed 10 images and categorized them as either mild, moderate, or severe with agreement ranging from κ 0.232 to 0.615.

Finally, a study by Bergman et al⁵⁴ assessed the feasibility of four published methods for independent, remote assessment of acne including the Leeds technique,³⁴ the IGA scale, and two methods of assessing inflammatory lesion counts (ILC, total ILC [TILC], and frontal ILC [FILC]), for use in the remote assessment of acne. K scores were reported for intra-rater and inter-rater reliability. Intra-rater reliability results were strong for all methods (Rater A TILC κ 0.9891; FILC 0.9897; Leeds 1.0; IGA A 0.926; Rater B TILC 0.9077; FILC 0.9325; Leeds 0.879; IGA 0.606); however, inter-rater reliability was much weaker for the grading methods, compared to the counting methods (TILC 0.8706; FILC 0.8449; Leeds 0.381; IGA 0.3119).

Non-expert evaluation. No scales provided evidence of non-expert assessment; while Michaëlsson et al did describe the assessment of clinical images by "laymen", they did not report any findings.³⁶

DISCUSSION

This review identified 24 measures of acne severity. Critical appraisal of these tools highlighted the generally poor quality of most published measures of acne severity in terms of validity and reliability and suitability for use in a clinical or community trial setting. This is consistent with the findings of earlier reviews.¹⁹⁻²¹

In the current review, the global grading scales scored better than counting scales in the subcategory ease of use; however, given the overall poor performance of most of the scales, there was little evidence of any one approach being superior to another in either of the psychometric or suitability categories.

The LRAg⁵³ achieved a total quality rating of six out of a possible 13, but there are a number of concerns regarding the methodology and the reporting of the Validación Escalas de Gravedad de Acné (VEGA) study,⁵³ which was the source of all scores for this update to the Leeds global assessment method. These results should be interpreted with caution. To elaborate, there is no description of how the scale was administered in the Spanish dermatology setting (i.e., any changes to the original scale or instructions for the participant dermatologists). The authors did not state how many dermatologists were involved in this study, or whether they had completed training in the method. In this study, the lesions were counted and graded at baseline and follow-up, in both cases the counting occurred prior to global assessment at the same appointment; it is not clear who conducted the counting of the lesions or which method they adopted. The method for assessing intra-observer

reliability is questionable, and the reported results for inter-rater reliability are described as significant (ICC=0.72) despite the authors themselves defining an acceptable outcome level of ICC 0.80. Finally, the presentation of the results is confusing and difficult to interpret.

The Leeds technique counting method³⁴ is widely used in clinical trials and most often used for validation purposes, despite not being formally validated against a gold standard measure itself. Both inter-rater (non-inflamed lesions 0.87 [$p<0.002$]; inflamed lesions 0.92 [$p<0.001$]), and intra-rater (non-inflamed lesions 0.83 [$p<0.002$]; inflamed lesions 0.86 [$p<0.001$]) reliability estimates are good. This tool, like other counting methods, requires expert dermatological training or extensive training in the counting method. To achieve the reliability outcomes reported above, the authors, both experienced clinical and academic dermatologists, completed the validation process following 18 months of training to achieve a correlation of ≥ 0.80 on two consecutive days.³⁴ By contrast, Lucky et al³⁸ found this method of counting demonstrated only moderate between rater reliability (ICC 0.61) in the raters, who were dermatology physicians and nurses.

Psychometric testing by Bergman et al⁵⁴ of the grading method by Burke and Cunliffe demonstrated poor inter-rater reliability outcomes ($\kappa=0.381$), but stronger intra-rater results ($\kappa=0.879-1.0$) for remote digital assessment of inflamed lesions among women with mild-to-moderate acne.

Poor image quality means that some earlier scales are now obsolete. For example, the image quality of the published Leeds technique,³⁴ the LLAG scale,⁴⁰ and Samuelson's photographic method⁴² makes the scales difficult to read; this is in contrast to more recent scales, including those by Hayashi et al⁴⁹ and Dréno et al,²³ which utilize modern photographic technology to produce digital images with a higher resolution.

The use of digital photography for acne assessment offers two important benefits. First, it provides a platform for independent third-party assessment, and second, it delivers a record for verification purposes. While there is some evidence of third-party testing, only one study reported an analysis directly comparing two approaches (i.e., *in situ* and remote assessment) in one population.⁴² Although the findings revealed a difference in reported outcomes between the two approaches of 1 to 2 grades, it is possible that these differences may have been influenced by the quality and/or administration of the color images.

The authors did not include efficiency in terms of time and cost as part of this critique. Studies have identified time spent in consultation with a medical professional as being a positive aspect of care, but it is important to distinguish between doctor/patient communication and the nonverbal act of acne assessment. One study found that seven percent of clients who were satisfied with aspects of medical care (i.e., communication, shared decision making, and instruction) were not satisfied with the time spent in the dermatology examination room.⁵⁸ Time and cost associated with the administration of a particular test is relative to the budget and setting of evaluation research and clinical trials.

Typically, a method that requires categorization of acne based on a global assessment scale should take less time than counting, and therefore, cost less to administer. In selecting a suitable acne severity instrument, the researcher might consider efficiency in the context of their budget and proposed delivery method.

In the absence of any gold standard acne assessment scale, clinicians, researchers, and others may resort to highly simplified methods. For example, it is not uncommon to find acne classified as mild, moderate, or severe in acne treatment guidelines,⁵⁹⁻⁶¹ or in the reporting of treatment outcomes by manufacturers of acne treatments.⁵⁵ Results of inter-rater reliability for this simplified method are reported as weak to moderate.⁵⁵

The multimodal imaging system shows great potential for providing robust clinical outcomes though the fixed nature of the equipment means that it may not be suitable for all research conducted in the community. Further evaluation by independent and impartial researchers may provide necessary support to the current findings.

The authors noted other areas of concern with the research describing the acne severity scales, including potential problems surrounding sampling, the quality of the statistical analysis, and reporting of findings.

Their review of the literature revealed issues with the selection of analytical tests. For example, there was some incorrect use of tests for measuring types of data, such as Pearson's coefficient for discrete data; analysis of internal consistency for scales categorizing attributes⁶²; or incorrectly reporting sensitivity to change with tests of effect size. Second, reporting bias was evident in some articles. Authors make inferences about data without reporting findings, or report confidence intervals without p values. In many articles, ICC are used without specifying the type (i.e., ICC 1, 2, or 3), each having distinct implications on findings.

The authors also note that many studies have used convenience samples (e.g., clinic samples), which can limit the generalizability of the findings or the transferability of the scale to another research or clinical setting.⁶³ It is necessary therefore, that a rich description of the study population and detail of the necessary skill set of the raters is required.

There are some limitations to this review. The authors were unable to access original sources of some very early scales described in the literature, and as such, were not able to include these measures in this review.^{64,65} The data extraction was primarily conducted by TA, but with input from ML, GF, and LS wherever there was some lack of clarity in possible interpretation.

While they are aware of the increasing use of fluorescence or polarized photography, these were not included in the review as they are methods for capturing data rather than an original scale for assessing outcomes.

There are several strengths of this review. To the authors' knowledge, this is the first formal quality assessment of acne severity scales that draws on data from validation studies as well as that provided by the original authors. It is also the

first to consider specifically, applicability to the clinical trial setting, by including categories such as independent assessment and suitability for non-expert raters. It is also the first review to provide an overall quality score based on the application of objective criteria.

CONCLUSION

There is still much to be learned about acne. For instance, there is no clear definition of acne severity that is likely to contribute to the vast array of poorly validated acne severity scales and the inconsistencies among reported outcomes. Further, there is a weak understanding of the natural history of acne as well as patterns of natural symptom resolution.^{15,66} These knowledge gaps add to the complexity of acne assessment.

The future development of acne scales for research purposes should be mindful of the range of academic disciplines with interests in studying this field. A tool that can be utilized by experts and non-experts alike will be more useful for assessment, interpretation of results, and pooling of data for systematic reviews and meta-analyses. Given the extremely high prevalence and burden of acne, an assessment tool suitable for use in community-based and clinical trial settings should be a priority.

The absence of an internationally accepted measure of acne severity impedes quality clinical research and the adoption of best practice,^{17,19,20,31} with potential implications for the person experiencing acne. This paper concludes as others have, that a robust scoring system to assess acne severity is required. What this study has contributed is the ranking of existing scales according to objective quality criteria, highlighting in the process shortcomings in methodology and reporting underpinning most published scales. Future development of acne scales, or further investigations of currently available scales, is needed to rectify these issues and in order to move closer to development of a valid and reliable “gold standard” instrument.

REFERENCES

1. Fabbrocini G, Padova M, Cacciapuoti S, Tosti A. Acne. In: Tosti A, Grimes PE, De Padova MP, eds. *Color Atlas of Chemical Peels*. Springer Berlin Heidelberg; 2012:95–105.
2. Dreno B, Khammari A, Orain N, et al. ECCA grading scale: an original validated acne scar grading scale for clinical practice in dermatology. *Dermatology*. 2006;214:46–51.
3. Loney T, Standage M, Lewis S. Not just “skin deep”: psychosocial effects of dermatological-related social anxiety in a sample of acne patients. *J Health Psychol*. 2008;13:47–54.
4. Niemeier V, Kupfer J, Gieler U. Acne vulgaris—psychosomatic aspects. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*. 2006;4:1027–1036.
5. Purvis D, Robinson E, Merry S, Watson P. Acne, anxiety, depression and suicide in teenagers: a cross-sectional survey of New Zealand secondary school students. *J Paediatr Child Health*. 2006;42:793–796.
6. Kameran I, Khalis M-A. Quality of life in patients with acne in Erbil city. *Health and Quality of Life Outcomes*. 2012;10:60.
7. Magin P, Adams J, Heading G, et al. Experiences of appearance-related teasing and bullying in skin diseases and their psychological sequelae: results of a qualitative study. *Scand J Caring Sci*. 2008;22:430–436.
8. Gupta MA, Gupta. Depression and suicidal ideation in dermatology patients with acne, alopecia areata, atopic dermatitis and psoriasis. *Br J Dermatol*. 1998;139:846–850.
9. Picardi A, Mazzotti E, Pasquini P. Prevalence and correlates of suicidal ideation among patients with skin disease. *J Am Acad Dermatol*. 2006;54:420–426.
10. Mallon, Newton, Klassen, et al. The quality of life in acne: a comparison with general medical conditions using generic questionnaires. *Br J Dermatol*. 1999;140:672–676.
11. Robinson JK, Bahatia AC, Callen JP. Protection of patients’ right to privacy in clinical photographs, video and detailed case descriptions. *JAMA Dermatology*. 2014;150(1):14–16.
12. Gollnick HPM, Finlay AY, Shear N, on behalf of the Global Alliance to Improve Outcomes in A. Can we define acne as a chronic disease?: If so, how and when? *Am J Clin Dermatol*. 2008;9:279–284.
13. Allen BS, Smith G. Various Parameters for Grading Acne Vulgaris. *Arch Dermatol*. 1982;118:23–25.
14. Chiang A, Hafeez F, Maibach HI. Skin lesion metrics: role of photography in acne. *J Dermatol Treat*. 2014;25:100–105.
15. Williams HC, Dellavalle RP, Garner S. Acne vulgaris. *Lancet*. 2012;379:361–372.
16. Witkowski JA, Parish LC. The assessment of acne: An evaluation of grading and lesion counting in the measurement of acne. *Clinics in Dermatology*. 2004;22:394–397.
17. Tan JKL, Jones E, Allen E, et al. Evaluation of essential clinical components and features of current acne global grading scales. *J Am Acad Dermatol*. 2013;69:754–761.
18. Tan JK. Current measures for the evaluation of acne severity. *Expert Rev Dermatol*. 2008;3:595–603.
19. Lehmann HP, Robinson KA, Andrews JS, et al. Acne therapy: a methodologic review. *J Am Acad Dermatol*. 2002;47:231–240.
20. Barratt H, Hamilton F, Car J, et al. Outcome measures in acne vulgaris: systematic review. *Br J Dermatol*. 2009;160:132–136.
21. Garner SE, Eady A, Bennett C, et al. Minocycline for acne vulgaris: efficacy and safety. *Cochrane Database Syst Rev*. 2012(8):CD002086.
22. Pochi PE, Shalita AR, Strauss JS, et al. Report of the Consensus Conference on Acne Classification. Washington, D.C., March 24 and 25, 1990. *J Am Acad Dermatol*. 1991;24:495–500.
23. Dréno B, Poli F, Pawin H, et al. Development and evaluation of a Global Acne Severity Scale (GEA Scale) suitable for France and Europe. *J Eur Acad Dermatol Venereol*. 2011;25:43–48.
24. Menon C, Gipson K, Bowe WP, et al. Validity of subject self-report for acne. *Dermatology*. 2008;217:164–168.
25. Magin PJ, Pond CD, Smith WT, et al. Correlation and agreement of self-assessed and objective skin disease severity in a cross-sectional study of patients with acne, psoriasis, and atopic eczema. *Int J Dermatol*. 2011;50:1486–1490.
26. de Almeida H, Cecconi J, Duquia RP, et al. Sensitivity and specificity of self-reported acne in 18-year-old adolescent males. *Int J Dermatol*. 2013;52:946–948.
27. Kilkenny M, Merlin K, Plunkett A, Marks R. The prevalence of common skin conditions in Australian school students: 3. Acne

- vulgaris. *Br J Dermatol*. 1998;139:840–845.
28. Bae Y, Nelson JS, Jung B. Multimodal facial color imaging modality for objective analysis of skin lesions. *Journal of Biomedical Optics*. 2008;13(6).
 29. Patwardhan S, Kaczvinsky JR, Joa JF, Canfield D. Auto-Classification of Acne Lesions Using Multimodal Imaging. *J Drugs Dermatol*. 2013;12:746–756.
 30. Adityan B, Kumari R, Thappa DM. Scoring systems in acne vulgaris. *Indian J Dermatol Venereol Leprol*. 2009;75:323–326.
 31. Tan J, Wolfe B, Weiss J, et al. Acne severity grading: Determining essential clinical components and features using a Delphi consensus. *J Am Acad Dermatol*. 2012;67:187–193.
 32. Streiner D, Norman G, Cairney J. Reliability. *Health Measurement Scales: A Practical Guide to their Development and Use*. 5th ed. Oxford, UK: Oxford University Press; 2014.
 33. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall; 1991.
 34. Burke BM, Cunliffe WJ. The assessment of acne vulgaris—the Leeds technique. *Br J Dermatol*. 1984;111:83–92.
 35. Lidén S, Göransson K, Odsell L. Clinical evaluation in acne. *Acta Dermatovener (Stockholm)*. 1980;Suppl. 89:47–49.
 36. Michaëlsson G, Juhlin L, Vahlquist A. Effects of oral zinc and vitamin A in acne. *Arch Dermatol*. 1977;113:31–36.
 37. Christiansen J, Holm P, Reymann F. Treatment of acne vulgaris with retinoic acid derivative Ro 11-1430. A controlled clinical trial against retinoic acid. *Dermatologica*. 1976;153:172–176.
 38. Lucky AW, Barber BL, Girman CJ, et al. A multirater validation study to assess the reliability of acne lesion counting. *J Am Acad Dermatol*. 1996;35:559–565.
 39. Plewig G, Kligman AM, SpringerLink. *Acne Morphogenesis and Treatment*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1975.
 40. O'Brien SC, Lewis JB, Cunliffe WJ. The Leeds revised acne grading system. *J Dermatolog Treat*. 1998;9:215–220.
 41. Puig L, Guerra-Tapia A, Conejo-Mir J, et al. Validation of the Spanish Acne Severity Scale (Escala de Gravedad del Acne Espanola-EGAE). *Eur J Dermatol*. 2013;23:233–240.
 42. Samuelson J. An accurate photographic method for grading acne: initial use in a double-blind clinical comparison of minocycline and tetracycline. *J Am Acad Dermatol*. 1985;12:461–467.
 43. Pascoe VL, Enamandram M, Corey KC, et al. Using the Physician Global Assessment in a clinical setting to measure and track patient outcomes. *JAMA Dermatology*. 2015;151:375–381.
 44. Burton JL, Cunliffe WJ, Stafford I, Shuster S. The prevalence of acne vulgaris in adolescence. *Br J Dermatol*. 1971;85:119–126.
 45. Dreno B, Poli F, Pawin H, et al. Development and evaluation of a Global Acne Severity Scale (GEA Scale) suitable for France and Europe. *J Eur Acad Dermatol Venereol*. 2011;25:43–48.
 46. CDER. *Draft Guidance - Acne Vulgaris: Developing Drugs for Treatment*. Rockville, MD: United States Food and Drug Administration; September 2005.
 47. Leyden JJ, Shalita A, Thiboutot D, et al. Topical Retinoids in Inflammatory Acne: A retrospective, investigator-blinded, vehicle controlled, photographic assessment. *Clin Ther*. 2005;27:216–224.
 48. Cook CH, Centner RL, Michaels SE. An acne grading method using photographic standards. *Arch Dermatol*. 1979;115:571–574.
 49. Hayashi N, Akamatsu H, Kawashima M, Acne Study G. Establishment of grading criteria for acne severity. *J Dermatol*. 2008;35:255–260.
 50. Blaney DJ, Cook CH. Topical use of tetracycline in the treatment of acne: a double-blind study comparing topical and oral tetracycline therapy and placebo. *Arch Dermatol*. 1976;112:971–973.
 51. Thiboutot DM, Weiss J, Bucko A, et al. Adapalene-benzoyl peroxide, a fixed-dose combination for the treatment of acne vulgaris: Results of a multicenter, randomized double-blind, controlled study. *J Am Acad Dermatol*. 2007;57:791–799.
 52. Tan JK, Tang J, Fung K, et al. Development and validation of a comprehensive acne severity scale. *J Cutan Med Surg*. 2007;11:211–216.
 53. Guerra-Tapia A, Puig-Sanz L, Conejo Mir J, et al. Feasibility and reliability of the Spanish version of the Leeds Revised Acne Grading Scale. *Actas Dermo-Sifiliograficas*. 2010;101:778–784.
 54. Bergman H, Tsai KY, Seo SJ, et al. Remote assessment of acne: the use of acne grading tools to evaluate digital skin images. *Telemed J E Health*. 2009;15:426–430.
 55. Beylot C, Chivot M, Faure M, et al. Inter-observer agreement on acne severity based on facial photographs. *J Eur Acad Dermatol Venereol*. 2010;24:196–198.
 56. Doshi A, Zaheer A, Stiller M. A comparison of current acne grading systems and a proposal of a novel system. *Int J Dermatol*. 1997;36:416–418.
 57. Polgar S, Thomas SA. *Introduction to Research in the Health Sciences*. 5th ed. Edinburgh [Scotland]: Churchill Livingstone Elsevier; 2008.
 58. Camacho F, Balkrishnan R, Khanna V, et al. How Happy Are Dermatologists' Patients? *The Dermatologist*. 2013;21(4).
 59. Goulden V. Guidelines for the management of acne vulgaris in adolescents. *Pediatr Drugs*. 2003;5:301–313.
 60. Katsambas AD, Stefanaki C, Cunliffe WJ. Guidelines for treating acne. *Clin Dermatol*. 2004;22:439–444.
 61. Strauss JS, Krowchuk DP, Leyden JJ, et al. Guidelines of care for acne vulgaris management. *J Am Acad Dermatol*. 2007;56:651–663.
 62. Streiner D, Norman G, Cairney J. Selecting the items. *Health Measurement Scales: A Practical Guide to their Development and Use*. 5th ed. Oxford, UK: Oxford University Press; 2014.
 63. Polit DF, Beck CT. Generalization in quantitative and qualitative research: Myths and strategies. *Int J Nurs Stud*. 2010;47:1451–1458.
 64. Pillsbury D, Shelley W, Kligman A. *Dermatology*. Philadelphia: Saunders; 1956.
 65. Witkowski J, Simons H. Objective evaluation of demethylchlortetracycline hydrochloride in the treatment of acne. *JAMA*. 1966;196:397–400.
 66. Farrar MD, Ingham E. Acne: inflammation. *Clin Dermatol*. 2004;22:380–384. ●