Development and Initial Validation of a Multidimensional Acne Global Grading System Integrating Primary Lesions and Secondary Changes

Elena Bernardis, PhD; Haochang Shou, PhD; John S. Barbieri, MD, MBA; Patrick J. McMahon, MD; Marissa J. Pirmen, MD; Leigh Ann Rola, PA-C; Jenna L. Streicher, MD; James R. Treat, MD; Leslie Castelo-Soccio, MD, PhD; Albert C. Yan, MD

IMPORTANCE The qualitative grading of acne is important for routine clinical care and clinical trials, and although many useful systems exist, no single acne global grading system has had universal acceptance. In addition, many current instruments focus primarily on evaluating primary lesions (eg, comedones, papules, and nodules) or exclusively on signs of secondary change (eg, postinflammatory hyperpigmentation, scarring).

OBJECTIVES To develop and validate an acne global grading system that provides a comprehensive evaluation of primary lesions and secondary changes due to acne.

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study created a multidimensional acne severity feature space by analyzing decision patterns of pediatric dermatologists evaluating acne. Modeling acne severity patterns based on visual image features was then performed to reduce dimensionality of the feature space to a novel 2-dimensional grading system, in which severity levels are functions of multidimensional acne cues. The system was validated by 6 clinicians on a new set of images. All images used in this study were taken from a retrospective, longitudinal data set of 150 patients diagnosed with acne, ranging across the entire pediatric population (aged 0-21 years), excluding images with any disagreement on their diagnosis, and selected to adequately span the range of acne types encountered in the clinic. Data were collected from July 1, 2001, through June 30, 2013, and analyzed from March 1, 2015, through December 31, 2016.

MAIN OUTCOMES AND MEASURES Prediction performance was evaluated as the mean square error (MSE) with the clinicians’ scores.

RESULTS The scale was constructed using acne visual features and treatment decisions of 6 pediatric dermatologists evaluating 145 images of patients with acne ranging in age from 0 to 21 years. Using the proposed scale to predict the severity scores on a new set of 40 images achieved an overall MSE of 0.821, which is smaller than the mean within-clinician differences (MSE of 0.998).

CONCLUSIONS AND RELEVANCE By integrating primary lesions and secondary changes, this novel acne global grading scale provides a more clinically relevant evaluation of acne that may be used for routine clinical care and clinical trials. Because the severity scores are based on actual clinical practice, this scoring system is also highly correlated with appropriate treatment choices.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Elena Bernardis, PhD, Department of Dermatology, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce St, 2 E Gates Bldg, Room GAO2060, Philadelphia, PA 19104 (elena.bernardis@pennmedicine.upenn.edu).
Acne vulgaris is one of the most common skin disorders managed by pediatricians and dermatologists, affecting 85% of adolescents, and is responsible for a greater global burden of disease than psoriasis, cellulitis, and melanoma. Although qualitative grading of acne is important in the clinic to select appropriate treatment and as an outcome for clinical trials, acne grading remains a topic of debate with no universally accepted global acne severity scale. Most acne scoring systems can be divided into quantitative lesion counting and qualitative acne global grading systems. Lesion counting involves quantifying noninflammatory (ie, comedones) and inflammatory (ie, papules, pustules, and nodules) lesions. In contrast, global grading systems seek to provide a qualitative estimate of disease severity that can complement quantitative lesion counting. Several dozen useful acne grading systems have been developed, most of which categorize severity through a text description and/or photographic templates. These qualitative scales have been shown to be more practical and clinically relevant than lesion counting alone. However, the varying and sometimes inconsistent definitions used in these tools reflect the inherent challenges of classifying the visual complexity of acne lesions on the skin. Notably, most of these systems do not account for the presence of scarring, which can have important implications for treatment decisions and thus can significantly alter the clinical severity of the acne. Although several scales are available to assess acne scarring, these typically do not have the ability to simultaneously consider other acne lesion types (eg, comedones, papules, pustules, and nodules). In addition, these acne global grading systems often do not incorporate measurements for postinflammatory changes or erythema, which can often have significant relevance to patients with respect to their perceived severity of disease. Given the limitations of current grading systems, we sought to develop and validate a novel acne global grading system that could account for primary lesions (eg, comedones, papules, and nodules) as well as secondary changes (eg, postinflammatory hyperpigmentation, scarring).

Figure 1. Overview of Developing and Validating the New Grading System

The scale development was divided into the following main steps: (1) high-level image analysis; (2) data collection; (3) pattern analysis to condense the information down to a clinically intuitive, low-dimensional severity table; and (4) scale validations.

Key Points

**Question** Is it possible to develop a global grading system for acne assessing primary lesions (eg, comedones, papules) and secondary changes (eg, postinflammatory changes, scarring)?

**Findings** This diagnostic study presents a multidimensional acne grading system linking the most relevant visual features of acne to a severity score and validates the scale with correlation of treatment decisions by 6 clinicians evaluating images of patients with acne. The proposed 2-dimensional scale achieved an overall mean square error smaller than the mean within-clinician differences.

**Meaning** This grading system pilot provides a possible groundwork to reformulate acne severity as a dimensionality reduction problem.

Methods

This diagnostic study was deemed exempt from institutional review board approval and informed consent by the Children’s Hospital of Philadelphia institutional review board because of the use of deidentified patient images. The study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.

An overview of the methods used to create and validate the new grading system is summarized in Figure 1. The process was divided into 4 main steps: (1) high-level image analysis to understand how clinicians interpret acne on a visual inspection; (2) data collection, in which clinicians quantified acne visual cues and chose a corresponding treatment intensity; (3) feature analysis and thematic content analysis to condense the information into a clinically intuitive, low-dimensional severity table; and (4) validation of the scale. We start with a brief overview of development of the scale and refer the reader to the eMethods in the Supplement for complete details.

To capture the clinicians’ thought process when analyzing acne on a visual inspection, we started by creating a new
representation for acne severity (step 1); a multidimensional space linking acne’s visual features to classes of increasing treatment intensities (eFigure in the Supplement and Table). This was achieved by designing experiments that asked clinicians to evaluate images of patients with acne to define new acne severity groups in terms of increasing treatment intensities as well as to extract and collect features indicative of acne that influenced decisions of how to treat the patients (step 2). In step 3, we validated which features were most significantly associated with the severity scores via statistical models (eg, eFigure2 in the Supplement). Features included scars (acne-related and overall), inflammation, number of papules, and number of nodules, followed by the combined presence of postinflammatory color changes due to the resolving acne or acne treatment (eTable 2 in the Supplement). By using pattern analysis—in particular choosing feature combinations via content thematic analysis and validating the choices via visualizations to ensure a continuous and clinically intuitive distribution of the final severity levels—the multidimensional space of acne features that was discovered was then reduced to 2 dimensions (step 4). The first dimension captures primary lesions (comedones, papules/pustules, and nodules), whereas the second dimension encodes all other visual cues indicative of acne: (1) inflammation, defined as areas of redness associated with primary acne activity (excluding areas within the acne primary lesions); (2) scars; and (3) postinflammatory color variations associated with resolving activity or ongoing treatment, manifested as presence of focal color changes and/or erythema not associated with primary lesion inflammation (eg, macular erythema, postinflammatory erythema), hyperpigmentation, and dryness, redness, and/or color changes due to treatment.

To validate the scale, 6 clinicians (5 board-certified dermatologists, 4 of whom had an additional board certification in pediatric dermatology, and 1 physician’s assistant [L.A.R.]) who did not participate in the data collection phase to create the scale rated 40 new test images of acne patients’ skin (including face, back, and chest) that were not used in the creation of the scale itself. Patient consent for all images used in this study was waived by the institutional review board. Each clinician was provided with a hard copy of the scale (Figure 2) and instructed to choose 1 of 9 treatment intensities (Table) and to use the proposed table to provide a severity score. Clinicians were approached individually and provided with a form to fill out in private to minimize any bias in their choices. Prediction performance was evaluated as the mean squared error (MSE) comparing the proposed scales with the treatment choices of the clinicians. The MSE is a measure of the expected deviation between the predicted and observed scales. It assesses the accuracy of the scale’s prediction by quantifying the mean squared difference between the clinicians’ scores from the 2 clinical interviews (the mean score of images evaluated within a group of images vs images evaluated individually) and those obtained by using the proposed scale. The smaller the MSE, the more closely the predicted value is to the desired one.

In addition, we evaluated the predictions of our proposed scale for the 145 images used during the development phase and compared them with the predictions obtained via the data-driven learning models as well as the 1-dimensional version of our scale based only on active lesion count (eFigure 3C in the Supplement). Prediction performance was again evaluated as the MSE with the mean of clinicians’ ratings. For the regression models, to avoid overfitting, we cross-validated the prediction models with 80% randomly selected images and calculated the MSEs on the remaining 20% of testing samples, repeating the procedures 1000 times. The intra-rater variability of each clinician was derived by computing the mean squared difference between the scores obtained from 2 sets of ratings given by the same clinicians (images evaluated within a group vs images evaluated individually). Data were collected from July 1, 2001, through June 30, 2013, and analyzed from March 1, 2015, through December 31, 2016.

Table. Severity Label Definitions and Corresponding Treatment Intensity Groups Obtained From Initial Clinical Assessments*

<table>
<thead>
<tr>
<th>Acne Severity Level</th>
<th>Treatment Intensity Group</th>
<th>Associated Severity Label by Clinicians</th>
<th>Final Clinical Severity Label Used in the Proposed Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No treatment necessary</td>
<td>Clear</td>
<td>Clear</td>
</tr>
<tr>
<td>2</td>
<td>Benzoyl peroxide wash and/or a mild topical retinoid</td>
<td>Almost clear</td>
<td>Almost clear</td>
</tr>
<tr>
<td>3</td>
<td>Benzoyl peroxide wash and a topical retinoid</td>
<td>Mild</td>
<td>Mild</td>
</tr>
<tr>
<td>4</td>
<td>Benzoyl peroxide wash and a stronger topical retinoid or a topical retinoid and consideration of an oral antibiotic</td>
<td>Mild to moderate</td>
<td>Mild to moderate</td>
</tr>
<tr>
<td>5</td>
<td>Benzoyl peroxide wash, topical treatment, and an oral antibiotic</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>6</td>
<td>Same as 5, but start considering isotretinoin</td>
<td>Moderate to severe</td>
<td>Moderate to less severe</td>
</tr>
<tr>
<td>7</td>
<td>Same as 5, but recommend isotretinoin</td>
<td>Severe</td>
<td>Less severe</td>
</tr>
<tr>
<td>8-9</td>
<td>Should be on isotretinoin (more severe = 8) (case 1) or, in very severe cases, start with a lower dosage of isotretinoin (to avoid flaring) and add corticosteroids (case 2)</td>
<td>More severe</td>
<td>Severe (case 1) or very severe (case 2)</td>
</tr>
</tbody>
</table>

* Acne images were grouped into sets of lowest to highest perceived severity along with corresponding treatment intensities; severity labels were assigned to each group by clinicians in the initial experiments, and the final severity labels were definitions used in the proposed scale.

Results
All images used in this study were taken from a retrospective longitudinal data set of 150 patients diagnosed with acne in the dermatology clinic, ranging across the entire pediatric
Primary lesions and secondary changes can be encoded into a score by first counting lesions (nodules, papules/pustules, and comedones), choosing the highest corresponding row, and then selecting the degree of secondary changes (and choosing the correct combination). The dashed lines indicate any possible entry (i.e., the final severity is independent of that specific feature/variable). For example, if papules or nodules are present, the number of comedones is no longer relevant to the score. Similarly, if scars and inflammation are both present, the degree of postinflammatory hyperpigmentation is also no longer relevant to the score. For nodule quantifiers, few indicates 2 to 3; some, 4 to 6; and many, more than 6. For papule/pustule quantifiers, few indicates 1 to 3; some, 4 to 8; and many, more than 8. For comedone quantifiers, few indicates 1 to 3; some, 4 to 12; and many, more than 12. Scars and inflammation should be categorized into none, mild/moderate, or severe. Postinflammatory encodes any postinflammatory color changes and should be marked yes if any of the following cues are present: focal color changes and/or diffuse erythema not associated with primary acne lesion activity; hyperpigmentation; and redness, dryness, or color due to treatment. Mild + allows for differentiation of severe comedonal acne. If comedones are covered, the classification becomes severe comedonal; otherwise, the classification is the same as mild. An additional (lower-left) entry includes older acne cases that have completely cleared (only valid with >0 scars).
The proposed 9-point, multidimensional acne severity scale is presented in Figure 2. By having the raters count active lesions and detect the presence of secondary changes (binary in the case of postinflammatory changes and divided into 3 levels of severity in the case of inflammation and scars), it provides a way to map acne's visual appearance into the following 9 categories: clear, almost clear, mild, mild to moderate, moderate, moderate to less severe, less severe, severe, or very severe (Table). Each severity label corresponds to an increased treatment intensity group (Table). In addition, we added a mild plus category to allow for differentiation of severe comedonal acne, which is usually treated without the use of any oral antibiotics.

Details on how to use the scale are included in the caption of Figure 2, whereas Figure 3 shows some examples of acne grading using the proposed new scale. By providing specific guidelines, the scale eliminates potential overlaps between the groups and, by design, counteracts counting uncertainties, for example, those due to lesions that may appear similar (eg, larger papules vs smaller nodules). Because the development of the scale was based on mapping what clinicians saw to how they would treat what they saw, it provides an assessment of the area that is being inspected assuming nothing else is present.

Comparing the proposed scale with the clinicians' scores on the 40 new test images, we achieved an overall MSE of 0.821. In other words, the mean mistake of acne classification was less than 1 point when comparing it with the criterion standard (clinical scores). The MSE of the proposed scale and the clinicians' scores falls below the MSE of within-clinician differences measured in the development phase. In addition, for comparison with experiments presented in the eMethods in the Supplement, the interrater reproducibility is also higher when using the scale, yielding an intraclass correlation coefficient of 0.91.

A comparison of severities predicted by the model (ie, severity levels in direct correspondence with treatment intensity groups) vs treatment groups reported by clinicians are illustrated in Figure 4, where treatment intensities (y-axis), or clinical scores, are plotted against the scale scores (x-axis). We highlight this comparison in the plot by using 2-colored markers when the treatments differ: the marker's outer color represents the treatment associated with the clinical score, whereas the inner color represents the one associated with the score from the new scale. Marker sizes represent the number of images in each group. Circles below the diagonal line indicate overestimated scores, whereas the ones above the line indicate the underestimated ones. There are only 5 colors for the markers. As can be noted in the Table, although 9 categories were needed to properly account for all acne scenarios, the actual treatment intensities that the patients would be receiving (Table) are limited to 5 groups (eg, moderate and moderate to less severe would both have the same treatments).

The proposed 2-dimensional scale also showed the best prediction performance on the 145 images used during the development phase when compared with the predictions using the data-driven learning models as well as the 1-dimensional version of the scale based only on active lesion count. The MSE obtained using the proposed new acne severity scale (Figure 2)
Global acne grading scales are generally limited by focusing on lesional activity or secondary changes, such as scarring, which prevents a complete assessment of the burden of acne. The proposed scale presented herein represents a novel global acne grading system that incorporates primary lesions and secondary changes into a single comprehensive instrument. In addition, by mapping visual cues to disease severity as assessed by the treatment decisions of clinicians, we were able to create a scale that reduces subjectivity when using this instrument.

This scale has several potential applications. In clinical practice, the scale has the potential to facilitate more accurate assessment of the complete burden of acne to improve medical decision-making (because it provides an entry point for treatment, which can then be adjusted by the treating clinician depending on patient history or preferences) or to monitor for improvement after treatment. In a research setting, by providing a more comprehensive evaluation of the patient, it will be possible to better stratify patients for enrollment in clinical trials and to more accurately capture the improvement of their acne with therapy, enabling more robust studies. In addition, because the instrument includes scales for primary lesions and secondary changes, it is possible to use each of these subscales to monitor for improvement of these separate categories simply by tracking the component scores from each row and column.

Limitations
The results of this study should be interpreted in the context of the study design. Because this validation study only included clinical images evaluated by pediatric dermatologists at 1 health center, future validation studies are needed to confirm the validity and generalizability of this instrument, including among live patients. For example, although the conclusions were validated in a pediatric population—and we would anticipate that this scale could be extrapolated to an adult population—further work should also explore the applicability of this system in adults. Furthermore, the treatment correspondence was based on the clinical practice in a major children's hospital of board-certified dermatologists and dermatology clinicians with US-based treatments in mind. Although the ordering of the severity levels should be universally valid, in future work, it would be important to investigate country-specific adaptions for the treatment intensity correspondences.

Finally, the table itself is relatively complex and may be cumbersome to administer in a busy clinical setting. We have created a software version of the table that can be downloaded from the project website (https://dermatology.upenn.edu/labs/codelab/acne). In the future, the table may be integrated with automated lesion detection image analysis algorithms to further reduce the burden of administration and improve standardization.17

Conclusions
Acne vulgaris continues to be an area of active research and is one of the most common diseases worldwide. Currently available grading systems provide limited utility in terms of how to translate existing disease states into appropriate therapeutic plans because they do not account for overall disease activity. Following additional development and validation, this novel acne severity scale, which also captures information on secondary postinflammatory changes and scarring, has the potential to improve the care of patients with acne in the clinical and research settings.
ARTICLE INFORMATION

Accepted for Publication: December 4, 2019.
Published Online: January 29, 2020.

Author Affiliations: Department of Dermatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Bernards, Barbieri, McMahon, Perman, Rola, Streicher, Treat, Castelo-Soccio, Yan); Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Shou); Division of Pediatrics, Section of Dermatology, The Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania (McMahon, Perman, Rola, Streicher, Treat, Castelo-Soccio, Yan); Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (McMahon, Perman, Rola, Streicher, Treat, Castelo-Soccio, Yan).

Author Contributions: Drs Bernardis and Yan had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: Bernardis, Yan. Acquisition, analysis, or interpretation of data: All authors. Drafting of the manuscript: Bernardis, Shou, Barbieri, Yan. Critical revision of the manuscript for important intellectual content: Bernardis, Barbieri, McMahon, Perman, Rola, Streicher, Treat, Castelo-Soccio, Yan. Statistical analysis: Bernardis, Shou. Obtained funding: Bernardis, Yan. Administrative, technical, or material support: Bernardis, Rola, Yan. Supervision: Castelo-Soccio, Yan.

Conflict of Interest Disclosures: Dr Yan reported receiving personal fees from Ortho Pharmaceutical, Johnson & Johnson, and Valeant Pharmaceuticals outside the submitted work. No other disclosures were reported.

Funding/Support: This study was supported in part by the Children’s Hospital of Philadelphia’s 2013-2015 Chair’s Initiative Award (Drs Bernardis and Yan), award T32-AR-007465 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (Dr Barbieri), and a Pfizer Fellowship grant to the Trustees of the University of Pennsylvania (Dr Barbieri).

Role of the Funder/Sponsor: The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We thank the clinicians in the Section of Dermatology who participated in the clinical validation of the scale and the Penn Libraries’ Biomedical Library and the Library’s Innovation Intern, James Bigbee, for the implementation of the severity scale as a software interface.

REFERENCES