Supplementary Online Content

Bernardis E, Shou H, Barbieri JS, et al. Development and initial validation of a multidimensional acne global grading system integrating primary lesions and secondary changes. *JAMA Dermatol.* Published online January 29, 2020. doi:10.1001/jamadermatol.2019.4668

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Creating the New Multidimensional Acne Severity Scale

In order to create the multidimensional acne severity scale, we conducted a series of experiments with clinicians within the Section of Dermatology at Children's Hospital of Philadelphia (CHOP) to understand how acne is evaluated in the clinical setting in order to define severity levels that were in correspondence with treatment intensities as well as gather information on which features are used to determine the severity (eMethods 1.1). This allowed us to create a multidimensional acne severity space, which we then populated by collecting data from pediatric dermatologists (eMethods 1.2).

To reduce the acne severity to a manageable scale, we then reduced the dimensionality of the space by validating the choice of most relevant acne visual features via statistical methods (eMethods 2.1), and by studying the space of acne features and how severity levels varied as a functions of different feature combinations (eMethods 2.2). The final 2 dimensional table compacts 6+ dimensional feature space and is presented in the main manuscript. An overview of the methodology is visualized in Figure 1.

Datasets. All images used in this study were taken from a retrospective longitudinal dataset of 150 patients diagnosed with acne in the dermatology clinic, 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 (eTable 1). Informed consent was waived by the institutional review board. All images were taken from off-the-shelf digital cameras and mobile devices and, for each body area, shown to clinicians at comparable resolutions.

1. Designing an acne severity feature space

1.1 Defining new acne severity levels and finding visual acne features

We presented 7 pediatric dermatologists and 1 physician assistant with a set of 220 images of facial lateral views, de-identified and standardized by cropping out the largest rectangular skin area bounded by ear, mouth, and eye. Each clinician was asked to divide the images (recursively) into groups of increasing *severity* levels. In the first scenario (eFigure 1a), clinicians provided a *severity* tag for each folder. In the second scenario (eFigure 1b), clinicians labeled the image folders with *treatment-intensity* tags instead (topical retinoid, oral antibiotic, *etc.*). By counting the most frequently used words, we ranked the treatment options and mapped them to a 1-8 scale as reported in Table 1b-c. The scores obtained by mapping the treatment groups (eFigure 1b) had an overall better inter-rater reliability [1,2] when compared to scores obtained by mapping the severity groups (eFigure 1a): Kendall's coefficient increased from 0.864 to 0.869 and the intra-class coefficient (ICC) from 0.792 to 0.826. Among all the cases, mild to early severe categorization had the greatest improvement. We thus divided severity into 9 possible levels (Table 1d), each one with an associated treatment intensity level (Table 1b).

Many factors count in deciding how to treat acne. From a very quick glimpse at an acne affected portion of the skin, dermatologists can quickly give an initial treatment suggestion and have a remarkable capacity of seeing the features that matter most to them in making that decision while filtering out all other visual information. To gather which cues counted most and how the clinicians quantified them (eTable 2), we observed how clinicians analyzed the acne details as they reasoned through their treatment choices via a supervised questions-answers reinforcement learning approach, until questions (for the first) and answers (for the second) reached a consensus for all clinicians.

1.2. Collecting clinical data: acne features and corresponding intensity groups

Six pediatric dermatologists were provided with 145 images (105 lateral views taken from the previous experiment; 40 new forehead views representing more cases of comedonal, or in general less severe acne, standardized by selecting areas between the hairline and eyebrows) and asked to identify acne activity. Clinicians had to assign also one of the 9 treatment intensities to the current acne (assuming no prior treatment and that it was the only acne present on the entire face) and identify cases where visual cues hinted the presence of treatment. The data from the clinicians was collected via a Matlab graphical user interface developed by us.

Image sets	Clear	Almost clear	Mild	Mild to Moderate	Moderate	Moderate to Severe	Severe	More Severe	Very Severe
220 (initial assessments)	9.9%	15.6%	15.1%	28.8%	16.0%	10.8%	2.4%	1.4%	
145 (acne data collection)	3.8%	7.5%	24.5%	13.2%	30.2%	6.6%	9.4%	2.8%	1.9%
40 (clinician validation)	5.0%	27.5%	17.5%	12.5%	15%	7.5%	5%	2.5%	7.5%
Acne statistics in the clinic (50 patients over 1 month time)	2.0%	8.0%	42.0%		32.0%		10%		6%

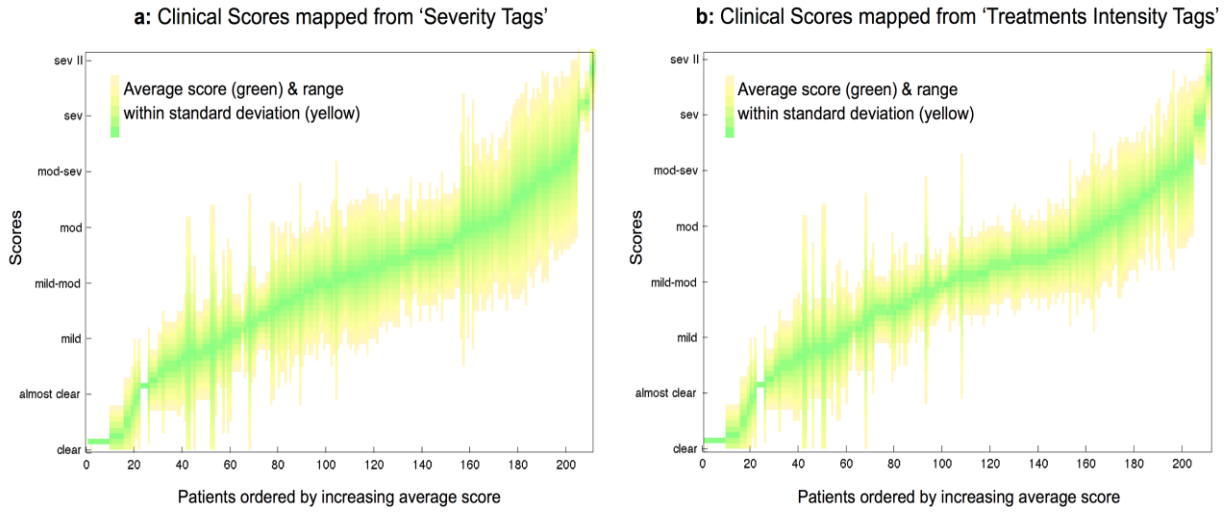
Image Sets	Phototypes I-III (self-described Caucasian)	Phototypes IV-VI (self-described Caucasian, African-American, Central Asian or Other Descent)	Phototypes III-IV (self-described East Asian Descent)
220 (initial assessments)	68.2%	13.8%	9.8%
145 (acne data collection)	78.1%	14.3%	7.6%
40 (clinician validation)	75%	20%	5%

eTable 1. Image Statistics Divided by Acne Severity and Skin Phototypes
Top: Acne severity statistics of images used. Because acne type was the most important aspect in terms of patient population, we selected representative images to adequately span the range of acne types encountered in the dermatology clinic. Bottom: Phototypes of acne patients.

2. Reducing dimensionality to create the acne severity scale

2.1. Validating choice of visual acne features via statistical methods

We validated the choice of the most relevant imaging features in predicting the final scores with three statistical learning methods including tree-based model, linear regression, and mixed effects model [21,22].



eFigure 1. Clinicians' Acne Assessments Interrater Variability

Subjects (x-axis, ordered by increasing average severity) vs. severity scores (y-axis, green represents the average score and yellow the range their standard deviation) are plotted in each panel. **a: Clinicians' scores derived from *severity* tags**. While there is low inter-rater variability in scoring for clear and very severe cases, there are still many images with high inter-rater variability, highlighting a lack of consensus for these ratings. **b: Clinicians' scores derived from *treatment-intensity* tags**. In contrast with a), mapping treatment intensities to obtain a severity score drastically reduces the overall inter-rater variability.

The tree-based model [3] estimated the hierarchical decision path and provided the hierarchical ranking for features that contributed most to distinguish discrete severity levels. eFigure 2 illustrates the estimated tree model using the whole sample. From top to bottom, at each node of the tree, a criterion was set based on one of the variables included in the model. If the statement is true (Y), the decision path goes to the left side; if false (N), then goes to the right side. At the bottom of each path (leaves), a severity score will be predicted to the image that reaches the corresponding leaf. The top layers highlight most relevant features including nodules, papules and scars. The length of the branches is proportional to the decrease of impurity from its parent node to the next, where impurity measures the ability of separating different categories. The shorter the branches are, the most discriminating the top layer features, while later features capture subtle differences among neighboring severity categories. The tree model was fit using the 'tree' package in R with expert clinical severity ratings as outcome and the following input predictors: scars; inflammation; #C, #N; #P; post-inflammatory (local) color changes, (focal or diffuse) erythema (not associated with inflammation), hyper-pigmentation (presence of pigment due to melanin); as well as dryness, redness, and/or color changes due to treatment.

To avoid over-fitting, we used cross-validation by splitting samples into training and testing sets. Our training set included clinical severity ratings provided by 6 pediatric dermatologists along with the features recorded during the decision making process. We constructed the prediction models using the tree methods with 80% randomly selected images and evaluated the prediction performance (calculated the MSE) on the remaining 20% testing samples. The procedures were repeated 1000 times and the results were compared with the proposed scale. For tree model, even though the estimated tree structures might vary depending on the sample split, the major features, #N, #P and scars, were selected in each of the 1000 iterations, while inflammation and #C were selected 81.8% and 88.2% times out of the 1000 iterations. This further confirmed our ordering of the feature importance. Note that the tree structure predicts the non-linearity of the severity levels: after the first four leaves, they no longer appear in increasing order, until we reach the final most severe level.

The linear mixed effects model [4,5] evaluated the associations between imaging features and severity scores from each image while accounting for potential correlations of ratings provided for the same

image. We fit the linear mixed effects model using 'lme4Test' package in R with an ID variable indicating the subjects that the images belonged to. eTable 1 shows the estimated p-values for each image feature. Similarly to the tree model, we also conducted cross-validation for the linear mixed effects model. Consistent with the tree model, features that most significantly associated with the scores (p value<0.05) among all 1000 iterations included scars, inflammation, #P, and #N. Our analyses also show agreement with the relevant clinical components consensus reached in Tan *et al*. [6,7].

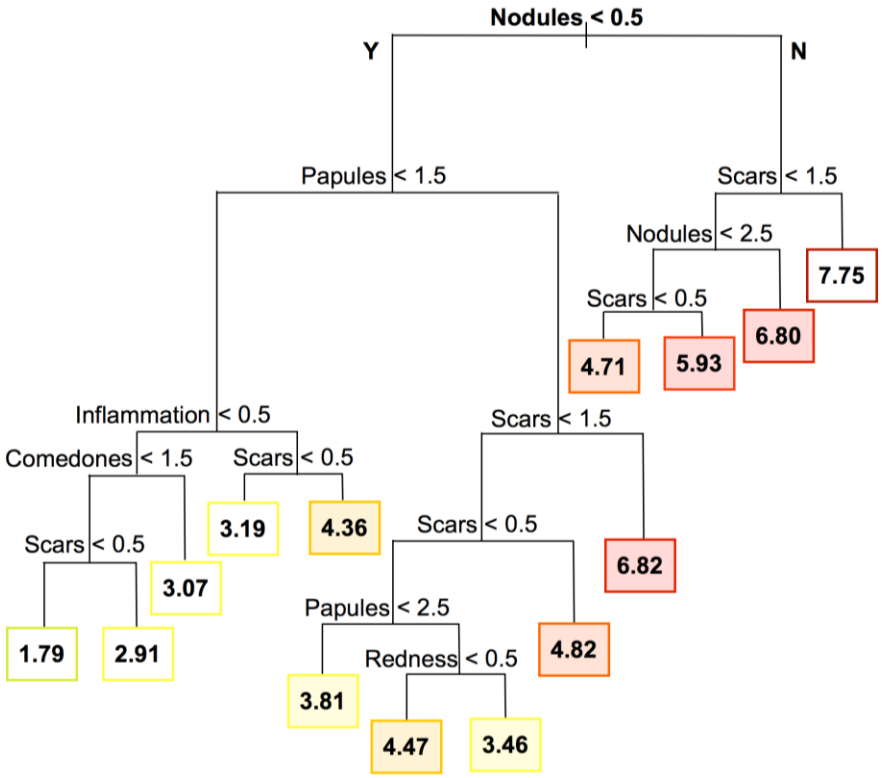
a: Acne feature	b: Quantifier options	c: Coeffs.	d: Std. Error	e: p-value
#C	None (0); few (1-3); some (4-12); many (12+); covered.	0.023	0.044	0.603
#N	None (0); one small; few (2-3); some (4-6); many (6+).	0.684	0.065	**<0.001**
#P	None (0); few (1-3); some (4-8); many (8+); covered.	0.479	0.053	**<0.001**
(Acne) Inflammation	None; mild/moderate; severe.	0.487	0.099	**<0.001**
Scars	None; mild/moderate; severe.	0.805	0.076	**<0.001**
Post-inflammatory (combined)		0.124	0.089	0.164
Cluster/location concern	Yes; no.	0.95	0.275	**0.001**
Pigment/melanin	Yes; no.	0.075	0.179	0.674
Post-inflammatory (focal) color changes	Yes; no.	0.293	0.118	**0.014**
Color change due to treatment	Yes; no.	0.386	0.266	0.147
Dryness due to treatment	Yes; no.	0.18	0.19	0.344
Redness due to treatment	Yes; no.	-0.305	0.135	**0.025**

eTable 2. Relevant Image Features Validated by Data-Driven Learning Models

a: Feature to measure, divided in active lesions, background active inflammation, scars, post-inflammatory activity. b: Quantifiers options give for each feature. c-e: Corresponding p-values from the mixed effects model, where the smaller the p-value, the more relevant the feature. We highlight in bold features with significant associations with the severity scores (p value<0.05), which include: scars, inflammation, #P, and #N. The result is consistent with the findings from the tree model shown in eFigure 2. As can be noticed, weights correlate between the two models reinforcing the relevance of the top features: #N, scars and inflammation, followed by #P and evidence of past activity.

2.2. Creating the scale by linking acne visual features with treatment choices

To reduce the dimensionality of the acne severity space to the format of the proposed scale (Figure 2), we started by creating a one-dimensional space to capture lesion count, keeping in mind previously developed acne global grading scales, dividing cases into two scenarios: presence or absence of papules/pustules (#P) and/or nodules (#N). When these were present, the number of comedones (#C) became almost irrelevant in terms of both severity and treatment. As #P and #N were not independent variables, we wanted a combination of #P and #N that increased proportionally with treatment intensity. Furthermore, we observed that one nodule alone was insufficient to catalogue the acne as 'severe'. We thus treated the presence of several papules similarly to the FDA's 'one small nodule'. Since comedonal acne is typically treated predominantly with topical retinoids even in the setting of significant burden, we grouped the cases of 'many' and 'covered' comedones together, but highlighted their difference in the final scale, by adding a 'Mild +' category to allow for cases of 'severe comedonal acne'.



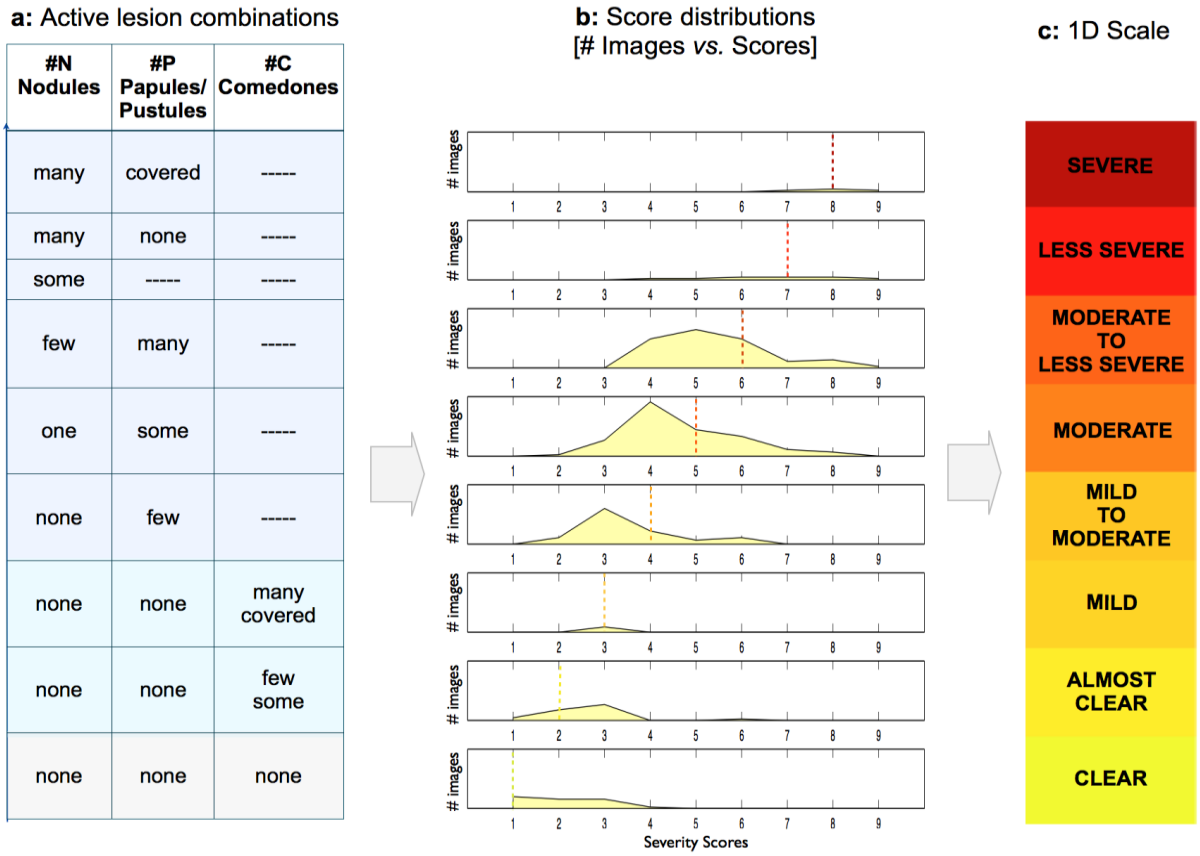
eFigure 2. Estimated Decision Tree Model from Acne Data

From top to bottom, at each node of the tree, a criterion was set based on one of the variables included in the model. If the statement was true (Y), the decision path goes to the left side; if false (N), it goes to the right. For active lesions, criteria quantifiers correspond to the ones used in the acne table, *e.g.* for nodules (0: none, 1: one nodule, 2: few or 1-2 nodules; 3: some or 3-4 nodules, 4: many or 6+ nodules). For background activity, mapping was as follows: 0-none, 1- mild/moderate, 2-severe, for scars and inflammation; and 0-none, 1-yes (present) for redness. Note how most final severities associated with each branch are not in increased ordering (boxes with colored background), further highlighting the non-linearity of acne severity.

A second dimension encodes the level of overall inflammation due to active acne primary lesions together with the historical presence of damage manifested as scarring or post-inflammatory color, such as erythema (not associated with inflammation), hyper-pigmentation, as well as visual cues that hint past or ongoing treatment. From the clinical interviews, we observed that presence of scarring alone (‘scars’) or significant inflammation (‘inflam’) caused clinicians to skew toward stronger treatment options. Scarring alone however, without inflammation, indicated presence of past activity (when no active lesions are present) or undergoing treatment and possibly resolving acne (when active lesions are present), so treatment intensity varied depending on the amount of inflammation present. Thus, we created three possible scarring/inflammation levels and added an additional level to capture resolving acne, *i.e.,* when only post-inflammatory color (‘post-inflam’) changes are present. Note, ‘post-inflam’ indicates past and/or resolving activity, usually trumped by presence of scars, and is most relevant in the absence of active inflammation.

To assign a severity score to each lesion count combination, we identified all images (from all clinicians) with the corresponding combination of features (#N, #P, #C, Inflam, Scars, Post-inflam), and plotted, for each feature combination group, their severity distribution. We then computed the average for each group to infer the final severity labels. For cases in which we did not have images for the possible combination pair, final intensities were interpolated from the adjacent scores as well as adapted by clinical judgment.

Collecting more data in the future might aid in refining the table further. eFigure 3 includes a visualization of this process applied to the first dimension only (using only lesion count).



eFigure 3. Constructing the 1-Dimensional Scale (Based Only on Lesion Count)
a: Lesion counting dimension (see Figure 2 for details on quantifiers). **b:** Severity distributions for each lesion count combination in a). Note that the range of severities for each distribution shows a large number of the images are underestimated. Each distribution should only have one clear peak and quickly tail off after a small standard deviation. **c:** One-dimensional version of our scale, with colors corresponding to the average score of each distribution, and possible severity labels that could be associated with each one.

References

1. Bates D, Maechler M, Bolker B, Walker S, Fitting Linear Mixed-Effects Models Using lme4. J of Statistical Software, 2015. 67(1), 1-48.
2. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. URL https://www.R-project.org/.
3. Fox J and Weisberg S. An {R} Companion to Applied Regression, Second Ed. Thousand Oaks CA: Sage, 2011.
4. Hastie T, Tibshirani R, and Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.
5. Shrout, PE, Fleiss, JL, Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 1979. 86, 2:420-8.
6. Tan JKL, Wolfe B, Weiss J, Stein-Gold L, Bikowski J, Del Rosso J, et al. Acne severity grading: determining essential clinical components and features using a Delphi consensus. J Am Acad Dermatol. 2012;67(2):187–93
7. Tan JKL, Jones E, Allen E, Pripotnev S, Raza A, Wolfe B. Evaluation of essential clinical components and features of current acne global grading scales. J Am Acad Dermatol. 2013;69(5):754–61