OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Cosmetic

Crowd-Sourced Reliability of an Assessment of Lower Facial Aging Using a Validated Visual Scale

Jason D. Kelly, PhD*
Bryan Comstock, MS†
Timothy M. Kowalewski, PhD*
James M. Smartt, MD‡

Background: Reliable and valid assessments of the visual endpoints of aesthetic surgery procedures are needed. Currently, most assessments are based on the opinion of patients and their plastic surgeons. The objective of this research was to analyze the reliability of crowdworkers assessing de-identified photographs using a validated scale that depicts lower facial aging.

Methods: Twenty photographs of the facial nasolabial region of various non-identifiable faces were obtained for which various degrees of facial aging were present. Independent crowds of 100 crowd workers were tasked with assessing the degree of aging using a photograph numeric scale. Independent groups of crowdworkers were surveyed at 4 different times (weekday daytime, weekday nighttime, weekend daytime, weekend nighttime), once a week for 2 weeks.

Results: Crowds assessing midface region photographs had an overall correlation of R = 0.979 (weekday daytime R = 0.991; weekday nighttime R = 0.985; weekend daytime R = 0.997; weekend nighttime R = 0.985). Bland–Altman test for test-retest agreement showed a normal distribution of assessments over the various times tested, with the differences in the majority of photographs being within 1 SD of the average difference in ratings.

Conclusions: Crowd assessments of facial aging in de-identified photographs displayed very strong concordance with each other, regardless of time of day or week. This shows promise toward obtaining reliable assessments of pre and postoperative results for aesthetic surgery procedures. More work must be done to quantify the reliability of assessments for other pretreatment states or the corresponding results following treatment. *(Plast Reconstr Surg Glob Open 2020;8:e3315; doi: 10.1097/GOX.0000000000003315; Published online 25 January 2021.)*

INTRODUCTION

Accurately assessing the objective degree of facial aging in human anatomy has proved to be difficult, usually leaving patients to rely on either their surgeon or their own subjective opinion, which could be highly unreliable.[1] The majority of postoperative outcomes are patient-reported, which can possibly lead to emotional and psychological issues, especially in patients with a history of anxiety or depression.[2,3] With 17.7 million cosmetic surgery procedures in 2018, and the number of cosmetic procedures continually increasing, a more objective assessment technique is needed.[4]

Techniques attempting to measure facial aging through a variety of methods have been researched. One such method used by Glogau utilizes photographs illustrating progressive degrees of photoaging with 4 classification levels.[5] Other methods (such as the Global Aesthetic Improvement Scale) have used subjective characteristics based on the perceived level of improvement, as determined by the physician.[6] An additional approach has been using scales that are derived from the patient's level of satisfaction with the procedure.[7] Some of these techniques can be susceptible to certain levels of bias, as detailed in the study by Pannucci and Wilkins.[8]

*From the *Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minn.; †Department of General Internal Medicine, University of Washington, Seattle, Wa.; ‡Bucky Plastic Surgery, Philadelphia, Pa.*

Received for publication September 15, 2020; accepted October 24, 2020.

DOI: 10.1097/GOX.0000000000003315

Disclosure: *At the time the research was conducted the corresponding author was a full-time employee of the University of Minnesota. All work was completed with funding from the National Science Foundation. Each of the co-authors now co-founding members of a company researching an area related to the research included in this article. Jason D. Kelly is a full-time employee and holds equity in this company. Timothy M. Kowalewski, Bryan Comstock, and James M. Smartt each hold equity ownership in the company.*

Of the anatomical facial regions, the lower face is typically one of the most common areas of concern for patients seeking facial rejuvenation.[9] Day et al previously published a validated 5-grade photograph numeric scale, which assessed the severity of lower facial aging.[10] This is known as a "visual guide" scale, in which photograph examples are shown to the reviewer that depict each possible state in the continuum. This is extremely useful, but any individual's opinion could lead to biased results. It would be critically important to obtain evaluations that could be repeatedly measured without losing accuracy (Fig. 1).

Large groups of independent peoples have been shown in the past to provide an excellent method of obtaining a group consensus that is more accurate than individual opinions, coined "crowdsourcing."[11] Moreover, studies have shown diverse crowds made of independent raters can repeatedly provide near identical results.[12] Previous research has shown that using crowdsourcing in healthcare domains can prove beneficial in settings where multiple domain experts may not be available.[13] Coordinating with physicians to evaluate before and after photographs of surgical procedures can be difficult, time-consuming, and expensive. However, using the knowledge of crowds can provide a cheap and efficient way of obtaining a plethora of these evaluations. Using crowds for obtaining evaluations of photographs of human anatomy may be a valuable way to create more objective assessments for cosmetic surgery patients. One can imagine a plastic surgeon's practice potentially using various crowdsourcing platforms to provide patients with a detailed analysis of the level of improvement they have received.

METHODS

Photographs of the lower face region were collected using publicly available images from cosmetic surgery online libraries and open source image sets. Twenty photographs were used in which the photographs were cropped and positioned to remove all personally identifiable information and display only the nasolabial area. The demographics of the pictured peoples included Whites and African Americans as well as both males and females. A user interface (UI) was created to allow raters to view each photograph individually in a randomized order. The 5-point wrinkle severity scale was displayed above the photograph being assessed, and observers were prompted to rate the photograph from 0 to 4, using the scale as a guide with a slider controlled by their mouse, able to be submitted in increments of 0.01. The scale used included text describing the different degrees of wrinkle severity named "Absent" (0), "Mild" (1), "Moderate" (2), "Severe" (3), and "Extreme" (4) (Fig. 1). After an evaluation of a photograph, the user interface then proceeded to a new page for evaluating subsequent photographs, one by one (Table 1).

Crowd Evaluations

Amazon Mechanical Turk is one of the most popular of the widely used crowdsourcing web platforms in use. This site allows users to submit Human Interface Tasks for crowdworkers to complete, which normally consist of some sort of a survey or to give an opinion after viewing a video. Using Mechanical Turk, a series of Human Interface Tasks were created for this study. Multiple evaluation sessions were chosen such that different groups of people could be surveyed to provide a more reliable assessment, as well as learning whether ratings varied throughout different time periods. An estimated 100 ratings per photograph were obtained for each of the 8 rating sessions to obtain a 95% confidence interval width of 0.5. The 4 trials chosen are displayed in Table 1. These 4 trials were repeated during the second week, to analyze both the variability from week to week, as well as variability between the time of week and day.

Statistical Analysis

For each photograph, we calculated a crowd-based mean score for each time of week and day. The consistency of the crowd scores across days and times was assessed using an intraclass correlation coefficient. Test-retest reliability of ratings collected at the same time from week 1 to week 2 was assessed using Pearson's correlation coefficient. Reliability above 0.9 was considered excellent. A Bland–Altman plot was used to establish levels of agreement and to diagnose systematic differences in ratings between weeks across the rating scale. We conducted an additional analysis for investigating the degradation in correlation, as rating sample size requirements reduced from 100 down to 10. Using the complete data set for each set of day-times in each week, we used the psych package from the statistical computer language R[14] to bootstrap resample (with replacement) smaller data sets of 10 photographs each and recalculate the test-retest reliability correlation coefficients as a function of sample size.

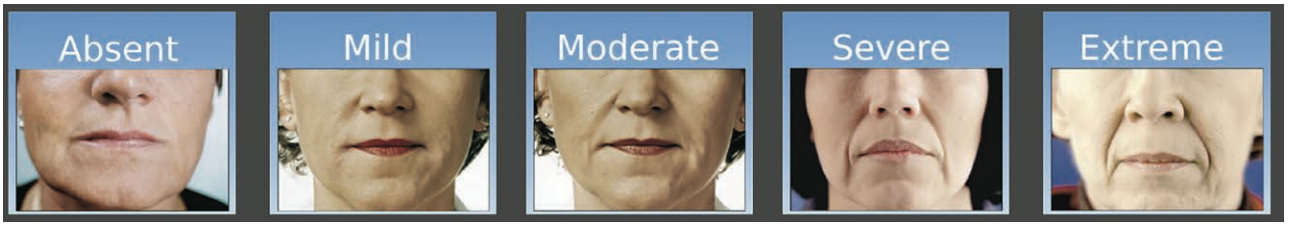

Fig. 1. Photographs displaying the numeric wrinkle severity scale. Each photograph corresponds to each rating of the relative presence of wrinkles in the midface.[6]

RESULTS

Figure 2 displays the mean scores for each of the 20 photographs, with a 95% confidence interval, for each of the 4 rating times in the first week. Ratings of photographs evaluated at different times of the week were highly concordant (ICC = 0.94; 95% CI = 0.89 to 0.97). Figure 3 displays the mean of each photograph's ratings for week 1 and week 2, with the strength of each correlation shown

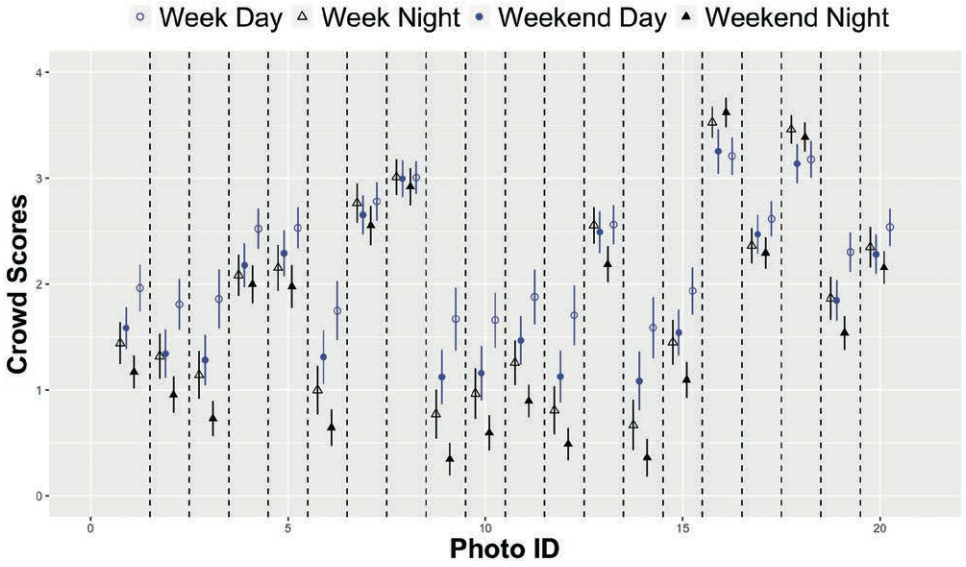

Fig. 2. Mean score (95% CI) of week 1 wrinkle severity ratings by photograph ID and rating time.

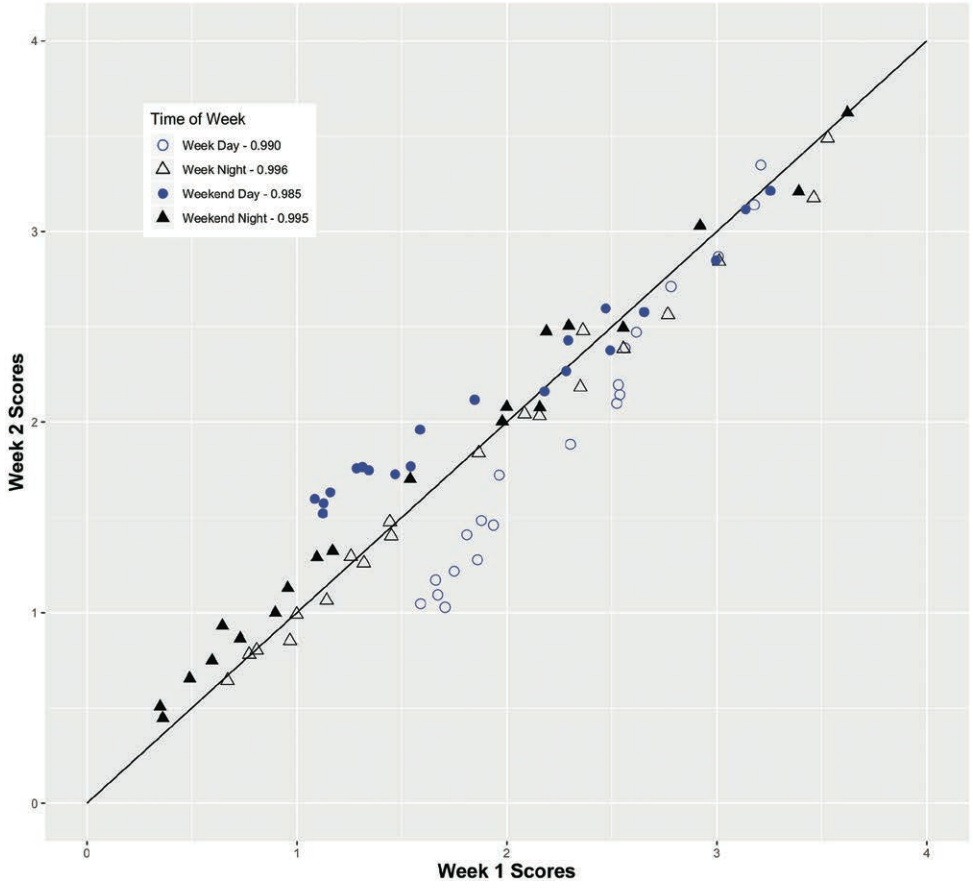

Fig. 3. Scatter plot of the midface wrinkle severity mean score for week 1 ratings compared with mean scores for week 2 ratings, at each of the 4 times of the week. The overall correlation was R = 0.946.

for each time of day. The overall correlation is R = 0.946, with the strongest correlation in photograph ratings being during the week, at night (R = 0.996), and the lowest being during the weekend in the day time (R = 0.985) (Figs. 2–5).

The Bland–Altman plot in Figure 4 illustrates the differences in scores from the mean for each of the 4 times of week. Week daytime ratings show 3 photographs with differences in ratings outside 1 SD from the average difference, with only 1 photograph being outside 1 SD for weekend daytime ratings. The test-retest reliability, shown in Figure 5, displays the test-retest reliability coefficient as a function of sample size. An inflection point occurs at roughly N = 40 ratings, in which the reliability continues increasing but to a diminished degree. For sample sizes above N = 40 ratings per photograph, the reliability coefficient is > 0.94.

DISCUSSION

Previous research on the uses of crowdsourcing in plastic surgery has generally fallen into one of the 2 broad categories. In the first group, investigators have surveyed crowd workers regarding their subjective preferences on a variety of topics. These articles generally group around preferences of beauty or opinions about the clinical delivery of various goods and services available in physicians' offices. Examples of studies of this type include Wu et al, who used crowdsourcing to learn a patient's preferences regarding which surgeon to choose, the use of before and after photographs, reputations, pricing, and experience.[15] Another study in the same realm used online worker assessments to gain knowledge of public perceptions toward plastic surgery, analyzing the existence of a gender bias in the field.[16]

Other studies in an opposing category (such as the Vartanian et al[17] study) collected assessments regarding the ideal thigh aesthetic based on thigh-to-buttock ratio and the buttock-thigh junction angle. Researchers have used various imaging modalities such as magnetic resonance imaging, and computed tomography to produce an objective measurement of facial anatomy.[18,19]

Studies in this category focus on analyzing the nature of specific pre or postoperative states. Tse et al notably used photographs of unilateral cleft lip patients to produce a ranked assessment of outcomes using the Asher-McDade rating system.[20,21] The research in this article is also more connected with this branch, as it analyzes the degrees of differing states in facial aging, as opposed to a completely subjective opinion of overall beauty by the same group of reviewers. The use of visual scales in this

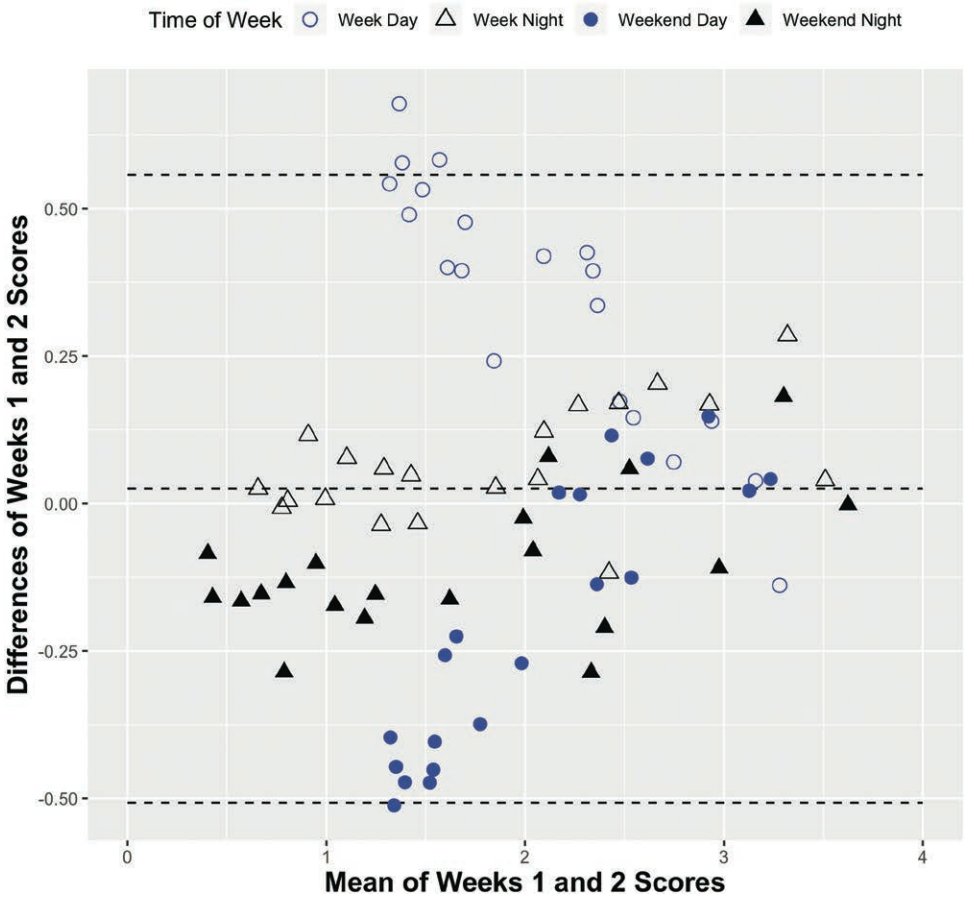

Fig. 4. Bland–Altman plot of the midface wrinkling scores for week 1 and 2. This illustrates the difference between mean ratings over the 2 weeks, with night time appearing to have smaller differences.

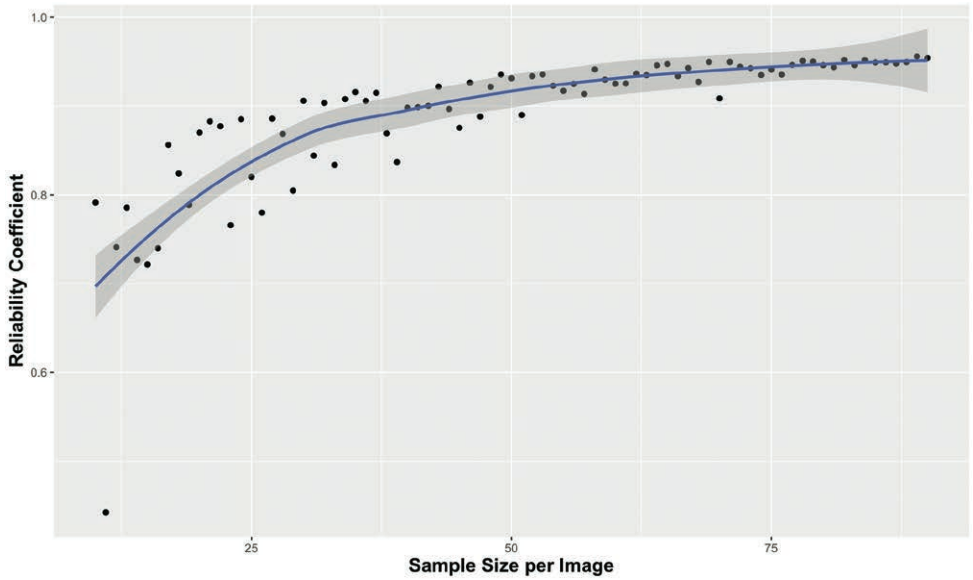

Fig. 5. The test-retest reliability of the midface wrinkle severity scores. Random groups of raters were selected from 10 to 90 raters, with the reliability coefficient calculated for each group selected. It appears that for around 40 raters, there is an inflection point at which the reliability was 0.935.

manner allows for the detachment of opinion from measuring incremental changes in visual states.

CONCLUSIONS

The high degree of correlation in assessments of facial aging gives merit to the technique of crowdworkers grading unknown photographs against photograph numeric scales. The small deviation from agreement in the change from night time to day time photograph ratings lends itself to deeper focus as to why this may be occurring, but overall the level of correlation in photograph ratings at all 4 times of the week were extremely high. Test-retest reliability of photograph ratings is encouraging, with an intraclass correlation coefficient of 0.94 after N > 40 ratings are obtained.

One limitation that will be taken into account in the future is that the scale used consisted only of White women, which could have created a bias in the ratings. Future research will aim to use more diverse groups of people in both rating scales and assessed photograph groups.

These results are promising, allowing the possibility of not only obtaining more objective evaluations of human aging, but additionally obtaining more precision than possible with a 5-point scale used by 1 person. This technique could make it possible to quantify granular changes in cosmetic procedure outcomes, as well as learning the level of improvement before and after receiving a cosmetic treatment, in a reliable manner. More work must be done to learn the limitations of photograph numeric scales on assessing the degree of laxity and wrinkling on human anatomy.

Jason D. Kelly, PhD
16 Hull St. Apt 3
Boston, MA 02113
E-mail: jkelly5207@gmail.com

REFERENCES

1. Pusic AL, Lemaine V, Klassen AF, et al. Patient-reported outcome measures in plastic surgery: Use and interpretation in evidence-based medicine. *Plast Reconstr Surg.* 2011;127:1361–1367.
2. Dittmann M. Plastic surgery: Beauty or beast? *Am Psychol Assoc.* 36;2005:30.
3. Honigman RJ, Phillips KA, Castle DJ. A review of psychosocial outcomes for patients seeking cosmetic surgery. *Plast Reconstr Surg.* 2004;113:1229–1237.
4. American Society of Plastic Surgeons. 2018 Plastic Surgery Statistics Report. 2018. Available at https://www.plasticsurgery.org/documents/News/Statistics/2018/plastic-surgery-statistics-full-report-2018.pdf. Accessed April 24, 2020.
5. Glogau Wrinkle Scale. Glogau dermatology. Published November 22, 2019. Available at https://sfderm.com/glogau-wrinkle-scale/. Accessed Oct 2, 2020.
6. Savoia A, Accardo C, Vannini F, et al. Outcomes in thread lift for facial rejuvenation: A study performed with happy lift revitalizing. *Dermatol Ther (Heidelb).* 2014;4:103–114.
7. Charalambous A, Adamakidou T. Risser patient satisfaction scale: A validation study in Greek cancer patients. *BMC Nurs.* 2012;11:27.
8. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. *Plast Reconstr Surg.* 2010;126:619–625.
9. American Academy of Facial Aesthetics. Wrinkles treatment. Available at: https://www.facialesthetics.org/patient-info/facial-esthetics/wrinkle-treatment/. Published 2015. Accessed April 24, 2020.
10. Day DJ, Littler CM, Swift RW, et al. The wrinkle severity rating scale: A validation study. *Am J Clin Dermatol.* 2004;5:49–52.
11. Brabham DC. *Crowdsourcing.* Cambridge, Mass.: The MIT Press; 2013.
12. Qarout RK, Checco A, Bontcheva K. Investigating stability and reliability of crowdsourcing output. Paper presented at: Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, July 5, 2018; 2018, Zurich, Switzerland.
13. Wagholikar K, Maclaughlin KL, Kastner TM, et al. Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. *J Am Med Inform Assoc.* 2013;20:749–57.

14. Revelle W. *psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.9.12.31.* Evanston, Ill.: Northwestern University; 2020.
15. Wu C, Hultman CS, Diegidio P, et al. What do our patients truly want? Conjoint analysis of an aesthetic plastic surgery practice using internet crowdsourcing. *Aesthet Surg J.*2017;37:105–118.
16. Bucknor A, Christensen J, Kamali P, et al. Crowdsourcing public perceptions of plastic surgeons: Is there a gender bias? *Plast Reconstr Surg Glob Open.* 2018;6:e1728.
17. Vartanian E, Gould DJ, Hammoudeh ZS, et al. The ideal thigh: A crowdsourcing-based assessment of ideal thigh aesthetic and implications for gluteal fat grafting. *Aesthet Surg J.* 2018;38:861–869.
18. Gosain AK, Amarante MT, Hyde JS, et al. A dynamic analysis of changes in the nasolabial fold using magnetic resonance imaging: Implications for facial rejuvenation and facial animation surgery. *Plast Reconstr Surg.* 1996;98:622–636.
19. Hutto JR, Vattoth S. A practical review of the muscles of facial mimicry with special emphasis on the superficial musculoaponeurotic system. *Am J Roentgenol.* 2015;204: W19–W26.
20. Tse RW, Oh E, Gruss JS, et al. Crowdsourcing as a novel method to evaluate aesthetic outcomes of treatment for unilateral cleft lip. *Plast Reconstr Surg.* 2016;138:864–874.
21. Mosmuller DGM, Bijnen CL, Kramer GJC, et al. The asher-mcDade aesthetic index in comparison with two scoring systems in nonsyndromic complete unilateral cleft lip and palate patients. *J Craniofac Surgery.* 2015;26:1242–1245.