Validation of the Unilateral Cleft Lip Severity Index for Surgeons and Laypersons

Alex Campbell, MD*†
Carolina Restrepo, MD*†
Gaurav Deshpande, MDS*†‡
Sarah M. Bernstein, MD, MHA§
Caroline Tredway, MD§
Lisa Wendby*
Bjorn Schonmeyr, MD, PhD*¶

Background: Severity of the primary unilateral cleft lip/nose deformity (UCL/N) is postulated to play a key role in postoperative complications, aesthetic result, and need for secondary surgery. There is no validated and widely accepted classification scheme of initial cleft severity. The purpose of this study was to validate the Unilateral Cleft Lip Cleft Severity Index as a reliable tool for evaluating presurgical UCL/N deformity by both surgeons and laypersons.

Methods: Twenty-five participants (10 surgeons and 15 laypeople) evaluated 25 sets of randomly selected presurgical standardized photographs of UCL/N patients. Each participant rated patients on a scale of 1–4 using the Cleft Severity Index. Interrater reliability for surgeons, laypersons, and all participants was determined using an intraclass correlation coefficient. Histograms and regression analysis were performed to compare average ratings between groups.

Results: Interrater reliability for all groups was classified as "very good" determined by intraclass correlation coefficients of 0.837 (laymen), 0.885 (surgeons), and 0.848 (all participants). These results indicate that there was a high degree of interrater across all 3 groups and that both surgeons and laypersons can reliability rate cleft severity using the Cleft Severity Index.

Conclusions: This study validates the use of the Cleft Severity Index by both surgeons and laypersons as a reliable tool for evaluating the degree of presurgical severity of patients with UCL/N. The Unilateral Cleft Lip Cleft Severity Index can thus serve as a reproducible and reliable grading system for primary UCL/N deformity and to categorize patients for future outcomes studies. *(Plast Reconstr Surg Glob Open 2017;5:e1479; doi: 10.1097/GOX.0000000000001479; Published online 13 September 2017.)*

INTRODUCTION

The ultimate outcome in cleft lip surgery has been shown to be influenced by surgeon experience, patient load, and organization of services.[1,2] Severity of the primary unilateral cleft lip deformity has also been accepted as a major determinant of ultimate appearance of the lip and nose after repair and has been postulated to play a key role in postoperative outcomes including aesthetic result, complications, and need for secondary surgery. In his elegant article on correlating objective measurements with severity of the cleft lip deformity, Fisher states that, "The ultimate appearance of the lip and nose is determined by a number of factors; however, the major determinant is the severity of the primary deformity. Thus, a permanent record of the primary deformity is necessary for any outcome study".[3]

Although classification schemes exist describing the degree of cleft involvement, there is no widely accepted measure of initial cleft severity. Numerous reports in the literature attempt to characterize the progressive phenotypic expression of cleft deformity, but no commonly used

Disclosure: *The authors have no financial interest to declare in relation to the content of this article. The Article Processing Charge was paid for by the authors.*

Supplemental digital content is available for this article. Clickable URL citations appear in the text.

classification scheme has emerged for grading preoperative severity and no known classification scheme that is useful for investigating outcomes.[4–7]

Development of a standardized tool would facilitate multicenter collaboration and comparison of outcomes. Although multiple attempts have been made to create a standardized assessment tool, the proposed methods to date vary widely in their study design, patient population, treatment stage, type of cleft, evaluators' education, and familiarity with the patients and techniques. Because of this, many of these tools have not resulted in reproducible results or high levels of interrater reliability. Thus, the potential influence of initial cleft severity on final results remains largely unmeasured.

The Unilateral Cleft Lip Cleft Severity Index is based on defined guidelines that evaluate the overall appearance of the deformity and separates patients into 4 categories according to the severity of their primary deformity. Grade I through Grade 4 cleft lip/nose deformities are defined according to the progressive degree of lip and nose involvement (Fig. 1).

1. Mild incomplete cleft lip
   A grade 1 defect is defined as a mild incomplete cleft lip where the cleft involves less than 50% lip height. There is usually a muscular depression above cleft and relatively mild nasal deformity.

2. More severe incomplete cleft lip
   A grade 2 defect describes a more severe incomplete cleft lip extending upward to more than 50% of lip height. The nasal deformity is more obvious. However, the nasal floor is intact.

3. Complete cleft lip
   A grade 3 defect describes a complete cleft lip with a nostril width ratio (NWR) less than 2 (Fig. 2 for calculation of NWR). Additional characteristics of the nasal deformity include a short hemicolumella, deviation of the columella and tip, posterior displacement of the alar base, and slumping of the lower lateral cartilage.

4. Severe complete cleft lip
   A grade 4 defect describes a severe complete cleft lip with a NWR greater than 2, meaning that the cleft side nostril width is more than double that of the noncleft side. There is wide separation between the medial and lateral elements that easily accommodates the tongue or endotracheal tube. There is also a severe nasal deformity with the alar is completely splayed across the cleft, often with complete distortion of normal alar curvature.

The purpose of this blinded, prospective study was to validate the Cleft Severity Index as a reliable tool for assessing the presurgical severity of unilateral cleft lip deformity through 3 specific aims:
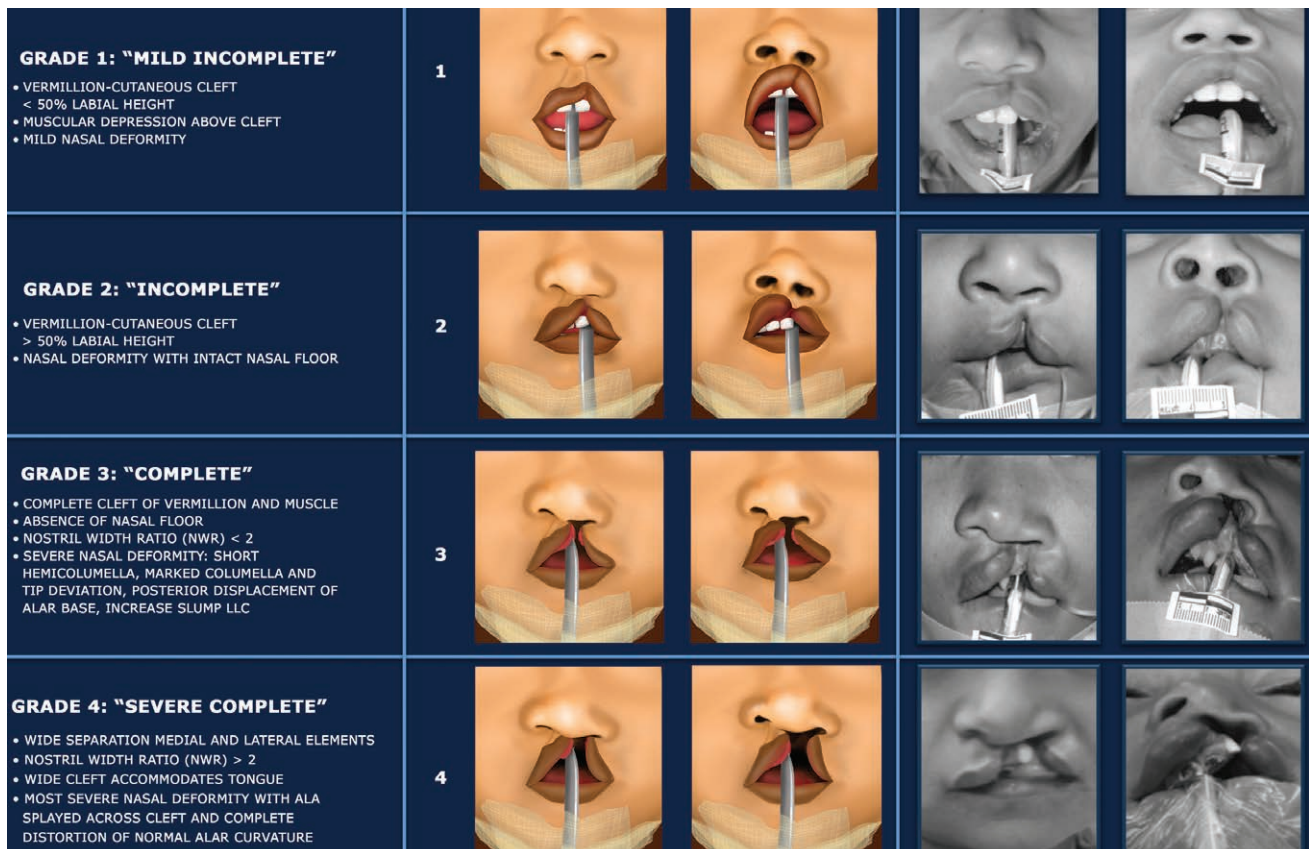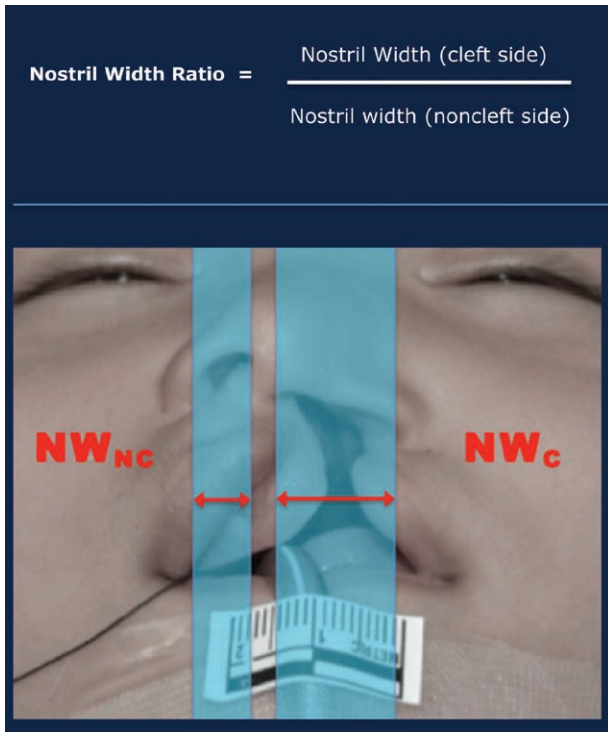

Fig. 1. Criteria and examples demonstrating each of the 4 grades of the Cleft Severity Index.

Fig. 2. Calculation of NWR. A caliper can be very helpful.

1. To determine whether experienced plastic surgeons can reliably grade the severity of cleft lip deformity using the Cleft Severity Index.
2. To determine whether inexperienced laypersons can reliably grade the severity of cleft lip deformity using the Cleft Severity Index.
3. To determine how similarly surgeons and laypersons grade the severity of cleft lip and deformity when using the Cleft Severity Index.

MATERIALS AND METHODS

Subjects

All study subjects were drawn from patients at the Guwahati Comprehensive Cleft Care Center in Assam, India, between 2011 and 2014 admitted for primary cheiloplasty. All patients' parents or guardians signed an informed consent allowing for the use of their medical records and photographs for research. All required forms and signatures were submitted to the institutional review board and ethical approval was granted.

Standardized basal and frontal presurgical photographs were obtained by 2 full-time photographers using a Nikon SLR digital camera. Patients ranged in age from 6–24 months and exclusion criteria included prior surgeries, known congenital syndromes, or other craniofacial abnormalities. From this group, 25 sets of photographs were randomly selected and deidentified. Photographs were cropped to minimize the portion of the face or body not affected by the cleft. All photographs were formatted to be of uniform length and width.

Evaluators

Two sets of evaluators were recruited to participate in the study. The first set consisted of 15 experienced senior cleft surgeons. All surgeons hold current plastic surgery board certification, international credentialing with Operation Smile, and have extensive experience in cleft lip and palate repair. Five surgeons failed to return their scores, so scores from a total of 10 plastic surgeons were collected and analyzed. The second set of 15 laypersons with no experience operating on cleft lip and palate were recruited from nonsurgical staff at the Guwahati Comprehensive Cleft Care Center and medical students at Emory University. Panel members ranged in age from 24 to 50 years.

Assessment

All participants were e-mailed a file containing the set of 25 standardized photographs, written and video instructions on the Cleft Severity Index, the Cleft Severity Index Reviewers Guide, and a flow chart to assist in grading (Fig. 3). Layperson volunteers also participated in a 1-hour informational training session on using the scale including practice cases. Participants graded each set of photographs using the Cleft Severity Index and recorded their answers in the table next to each set of photos (**see figure, Supplemental Digital Content 1**, with example of photos to evaluate, ***http://links.lww.com/PRSGO/A519***).

Statistical Analysis

All statistical analyses were performed using the Real Statistics Data Analysis Tool set in Excel (Microsoft). Statistical significance was defined as $P < 0.5$. Interrater reliability was determined by calculating the intraclass correlation coefficients (ICCs) for all surgeons, laypersons, and both groups combined. ICCs were calculated from a 2-factor analysis of variance (ANOVA) without replication using the following formula:

$$\frac{\left(MS_{Row} - MS_E\right)/k}{\left(MS_{Row} - MS_E\right)/k + MS_E} = \frac{MS_{Row} - MS_E}{MS_{Row} + \left(k-1\right)MS_E}$$

ICCs were then evaluated using the scale developed by Bland and Altman, which categorizes ICCs of 0.0–0.20 as poor, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good, and 0.81–1 as very good. The criteria for validation
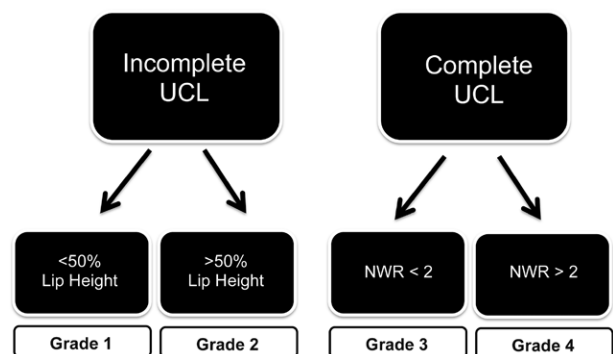
Fig. 3. Flow chart to grade severity according to the Unilateral Cleft Lip Severity Index.

of the scale was an interrater reliability of > 80% as measured by the intraclass correlation coefficient.[8]

The mean scores given by surgeon and layperson groups for each set of photographs were calculated and used to create a linear regression of the dataset. Histograms for surgeons and laypersons were also created by calculating the total number of times all evaluators from each group assigned a photograph as 1, 2, 3, or 4 across all 25 photographs. Finally, a Wilcoxon rank test was performed to test for differences in ratings between physicians and laypersons. A P value of < 0.5 was used to determine statistical significance.

RESULTS

In this study, plastic surgeons and laypersons graded the severity of cleft lip deformities using the Cleft Severity Index. ICCs are a measure of interrater reliability or the degree of agreement between evaluators when rating each subject. Two-way ANOVAs without replication were calculated and used to determine interrater reliability (**see table, Supplemental Digital Content 2** for 2-factor ANOVA without replication calculated from physician responses,

Table 1. ICCs for Physician Group, Layperson Group, and Both Groups Combined Calculated from Two Factor ANOVA without Replication

Group	ICC
Laymen	0.837
Physicians	0.885
Total	0.848

http://links.lww.com/PRSGO/A520; **see table, Supplemental Digital Content 3** for 2-factor ANOVA without replication calculated from layman responses, *http://links.lww.com/PRSGO/A521*; **see table, Supplemental Digital Content 4** for 2-factor ANOVA without replication calculated from physician and layman responses combined, *http://links.lww.com/PRSGO/A522*).

When 10 surgeons rated pictures of 25 clefts with varying degrees of severity, their ICC was found to be 0.885. When 15 laypersons evaluated the same 25 pictures, their ICC was 0.837. When all 25 participants (physician and laypersons) were considered a single group, their ICC was 0.848 (Table 1). As mentioned earlier, an ICC > 0.80 is considered very good. Because the ICCs for all groups were above 0.80, our calculations indicate that both laypersons and surgeons, independently and as a group, can reliably evaluate the severity of presurgical cleft lips and palates.

Next, the mean responses from the layperson and physician groups for each set of photographs were calculated (**see table, Supplemental Digital Content 5** for 2-factor ANOVA without replication calculated from physician responses, *http://links.lww.com/PRSGO/A523*). These averages were then used to create a linear regression shown in Figure 4. The equation for the best-fit line of this regression is $y = 1.0172x + 0.1192$ with an R^2 value 0.954. A slope of 1 and y intercept of 0 would indicate perfect agreement between groups. Thus, the slope of 1.0172 and y intercept of 0.1192 indicates that there is a high degree of correlation between the groups for each set of photographs, that is, surgeons and laypersons graded sets similarly. R^2 values range
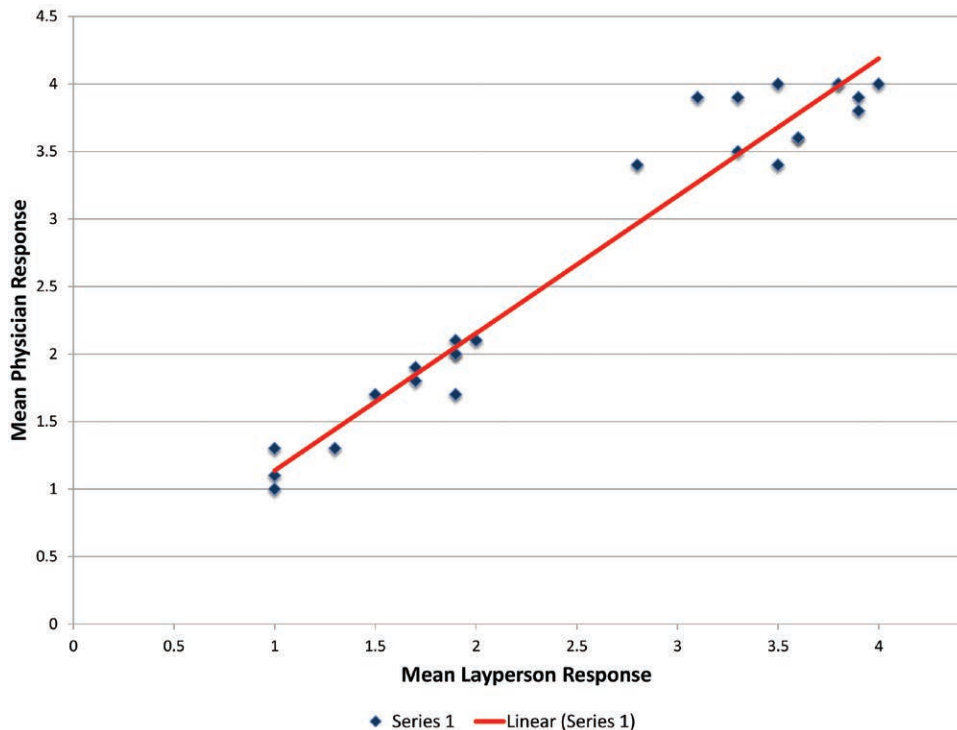
Fig. 4. Linear regression of mean scores given by physicians vs. laypersons across 25 patient photographs with $y = 1.0172x + 0.1192$.

from 0 and 1 and indicate how tightly the data points lie along the best-fit line. Because a score of 1 would indicate a perfect fit, the R^2 value of 0.954 adds further support to this conclusion.

Next, bin numbers showing the total number of times 10 responders from each group (physician and laypersons) assigned a score of 1, 2, 3, or 4 in the dataset was compiled (Table 2). For example, a score of 1 was assigned by 10 physicians a total of 43 times across all 25 photographs, whereas a score of 1 was assigned a total of 50 times by the same number of laypersons. To directly compare the frequency of specific scores between groups, 10 physician responses and the first 10 layperson responses were used to determine bin numbers.

Differences in physician and layperson scoring patterns is visually represented by the histogram in Figure 5. This illustrates that when the data are divided into complete and incomplete clefts (scores of 1 or 2 versus 3 or 4, respectively), physician responders are slightly more likely to assign a score of 2 or 4, compared with their layperson counterparts presenting an interesting finding explored further in the discussion section. Finally, a Wilcoxon rank test was performed showed that there were no significant differences ($P = 0.99$) between the plastic surgeon and layperson ratings (**see table, Supplemental Digital Content 6** for mean layperson and surgeon responses, rank order, and results of Wilcoxon rank test using 2 samples of n = 25, ***http://links.lww.com/PRSGO/A524***). When looked at individually and as a group, the results of the ICCs, regression, histograms, and Wilcoxon rank test all indicate that surgeons and layper-

Table 2. Bin Number Indicating the Total Number of Times 10 Physician and 10 Layperson Respondents Assigned Each Score for 25 Patients

Bin Number	Total Times Scores Assigned by Physicians	Total Times Scores Assigned by Laymen
1	43	50
2	76	68
3	29	49
4	102	83

Fig. 5. Histogram showing the total number of times 10 physicians and 10 laypersons assigned a particular grade photographs for all 25 patients.

DISCUSSION

The ultimate outcome of a surgical intervention is the determinant of success, and measuring and comparing outcomes is critical to the development of best practices. The ultimate goal of multidisciplinary cleft care is complete rehabilitation and normalization of a child. This takes a dedicated and experienced team of multidisciplinary specialists to restore central components of the form and function of a cleft patient including appearance, speech, hearing, eating, and self-perception. Successful rehabilitation in all of these areas liberates that child from the burdens of the cleft, freeing them to pursue life without disability and to achieve their full potential.

Cleft surgery has been and remains the largest, most sustained humanitarian global surgical effort. Charities treating cleft lip and cleft palate perform 150,000 operations per year, and evaluating outcomes is a stated goal of leaders such as Operation Smile, Smile Train, Interplast, and Shriners Hospital for Children.[9–13]

A central element of the overall outcome is the aesthetic appearance of the lip and nose after cleft lip repair. Surgical outcomes after cleft lip surgery are influenced by numerous factors, and many argue that severity of the primary deformity is perhaps the most important. Thus, postoperative results should be analyzed by taking into account the primary deformity. Unfortunately, no commonly accepted classification exists for cleft severity and thus comparison of postoperative outcomes remains ambiguous. Thus, development of a standardized tool to measure preoperative cleft severity is necessary to facilitate comparison of outcomes between surgeons, techniques, and institutions.
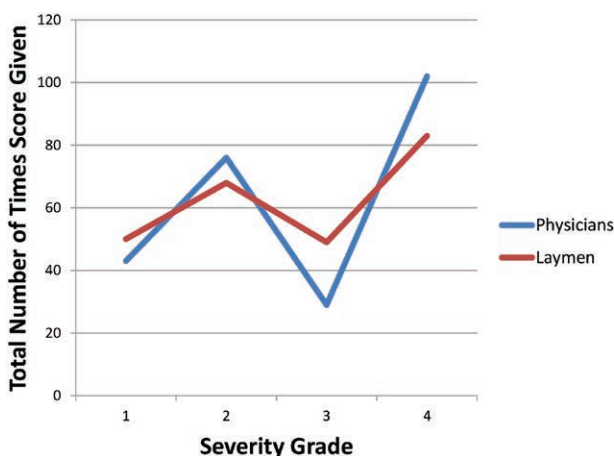
Numerous reports in the literature attempt to characterize the progressive phenotypic expression of cleft deformity, but there is no widely accepted measure of initial cleft severity and no known classification scheme that is useful for investigating outcomes. Millard described the "classification dilemma" of cleft lip noting that, "Many systems have been offered but none has been universally accepted because of language differences, inaccuracies, omissions, and lack of simplicity."[14] Thus, the potential influence of initial cleft severity on final results remains unmeasured and largely undocumented. Likewise, the degree of deformity has been postulated to play a role in surgical complications, though this is controversial and not substantiated in the literature.[15–18]

The Cleft Severity Index

An ideal classification of unilateral cleft lip would accurately characterize the severity of deformity in a manner that is simple, reproducible, relevant, and surgically applicable. Such a system would also allow for relevant inquiry into results and outcomes, comparing different maneuvers on similar patient sample groups. This would also allow for development of an algorithmic approach to

repair, employing specific series of maneuvers appropriate to correct anatomic irregularities characteristic of given deformities.

In unilateral cleft lip patients, varying degrees of displacement and hypoplasia of the lip, nose, and skeletal base produce a graduated severity of dysmorphology. Increasing degrees' deformity and separation of skeletal elements cause increasing malposition of the alar base and septum, resulting in progressive nasal deformity from flattening and elongation of the nostril with downward deflection of the alar cartilage.

Fisher's article in 2008 provided excellent evidence demonstrating that expert surgeons are able to reliably rank patients according to their subjective assessment and that these subjective assessments correlate with objective anthropometric measurements. He defined the NWR and demonstrated that this varies linearly with unilateral cleft lip nasal deformity and may act as an independent and objective indication of severity.[4] However, previous studies comparing the abilities of laypersons and surgeons to serve as evaluators have been mixed—while some studies have shown that there is a high degree of agreement between clinicians and laypeople, others have shown various degrees of disagreement.[19,20]

The Cleft Severity Index builds on the concepts and work of prior authors and provides a tool that quickly and easily grades the severity of a unilateral cleft lip/nose deformity. The Cleft Severity Index provides defined guidelines that evaluate the overall appearance of the deformity and separates patients into 4 categories according to the severity of their primary deformity. Grade I through Grade 4 cleft lip/nose deformities are defined according to the progressive degree of lip and nose involvement. The first step in grading shown by the flow chart was to determine whether a patient's cleft was complete or incomplete. This part seems to be the most objective as it only asks the evaluator to determine whether the nasal floor was intact. They have 2 options: present or absent. The next 2 steps are somewhat more subjective in that they ask the evaluator to determine "how much" of something exists. For example, whether the cleft > or < 50% of labial height or whether the NWR is > or < than 2.

The most significant results of this study were the ICCs for each group. It is only possible to scale cleft severity if people consistently agree on how they rate severity. The ICCs were 0.885 for surgeons, 0.837 for laypersons, and 0.848 for both groups combined. Our results indicate that there was a "very good" degree of interrater reliability indicated by ICCs > 0.80 across all 3 and that both surgeons and laypersons can reliability rate cleft severity when using the Cleft Severity Index.

When looked at individually and as a group, the results of this study indicate that surgeons and laypersons rate photographs similarly, showing a high degree of correlation between responders both within a group and between groups. The ICCs, regression analysis, histograms, and Wilcox rank test of this study all indicate high levels of interrater reliability between individuals and among groups. This successfully validates the Cleft Severity Index

as a reliable tool for grading initial cleft severity in patients born with unilateral cleft lip. This is particularly important, as experienced surgeons are often not readily available to analyze initial cleft severity for research purposes, particularly in resource-limited areas. We have learned this lesson time and time again, as reliance on surgeon evaluations has proven to be a central limiting factor in the development and implementation of a successful surgical outcomes program.

In summary, the results of this study validate the Cleft Severity Index as a reliable tool for evaluating presurgical unilateral cleft lip/nose deformity by both surgeons and laypersons. Although this particular study focuses on initial cleft severity and how it contributes to postsurgical outcomes, future research is needed to assess the scale's reliability across different age groups and how initial severity influences postsurgical outcomes in areas of appearance, language development, confidence, perceived social isolation, and overall well-being.

Alex Campbell, MD
Operation Smile, Inc.
3641 Faculty Boulevard
Virginia Beach, VA 23453
E-mail: alexcampbellmd@gmail.com

PATIENT CONSENT

Parents or guardians provided written consent for the use of the patients' image.

REFERENCES

1. Shaw WC, Dahl E, Asher-McDade C, et al. A six center international study of treatment outcome in patients with cleft of the lip and palate. *Cleft Palate Craniofac J.* 1992;32:434–441.
2. Sandy JR, Williams A, Mildinhall S, et al. The Clinical Standards Advisory group (CSAG) cleft lip and cleft palate study. *Brit J Orthod.* 1998;25:21–30.
3. Fisher DM, Tse R, Marcus JR. Objective measurements for grading the primary unilateral cleft lip nasal deformity. *Plast Reconstr Surg.* 2008;122:874–880.
4. Mosmuller DG, Griot JP, Bijnen CL, et al. Scoring systems of cleft-related facial deformities: a review of literature. *Cleft Palate Craniofac J.* 2013;50:286–296.
5. Tobiasen JM, Hiebert JM. Facial impairment scales for clefts. *Plast Reconstr Surg.* 1994;93:31–41; discussion 42.
6. Rajanikanth BR, Rao KS, Sharma SM, et al. Assessment of deformities of the lip and nose in cleft lip alveolus and palate patients by a rating scale. *J Maxillofac Oral Surg.* 2012;11:38–46.
7. Al-Omari M, Millet DT, Ayoub A. Methods of assessment of cleft-related facial deformity: a review cleft palate. *Craniofac J.* 2005;42:145–156.
8. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet.* 1995;346:1085–1087.
9. Bermudez L, Trost K, Ayala R. Investing in a surgical outcomes auditing system. *Plast Surg Int.* 2013;2013:671786.
10. Bermudez L, Carter V, Magee W, Jr, et al. Surgical outcomes auditing systems in humanitarian organizations. *World J Surg.* 2010;34:403–410.
11. Smile Train, ensuring safety and quality. Available at https://www.smiletrainindia.org/ensuring-safety-quality.php. Accessed March 16, 2017.

12. Shriners Hospital for Children. Outcomes. Available at https://www.shrinershospitalsforchildren.org/Locations/philadelphia/About/Outcomes. Accessed March 16, 2017.

13. Interplast. Volunteer manual 2016. Available at https://www.interplast.org.au/index.php. Accessed March 16, 2017.

14. Millard DR, Jr. The unilateral deformity. In: Millard DR, ed. *Cleft Craft*. Boston, Mass.: Little Brown; 1976. pp. 43–44.

15. Witt PD, Marsh JL. Advances in assessing outcome of surgical repair of cleft lip and cleft palate. *Plast Reconstr Surg*. 1997;100:1907–1917.

16. Johnson N, Williams A, Singer S, et al. Initial cleft size does not correlate with outcome in unilateral cleft lip and palate. *Eur J Orthod*. 2000;22:93–100.

17. Holtman B, Wray RC. A randomized comparison of triangular and rotation-advancement unilateral cleft lip repairs. *Plast Reconstr Surg*. 1983;71:72.

18. Tanikawa DY, Alonso N, Rocha DL. Evaluation of primary cleft nose repair: severity of the cleft versus final position of the nose. *J Craniofac Surg*. 2010;21:1519–1524.

19. Eliason MJ, Hardin MA, Olin WH. Factors that influence ratings of facial appearance for children with cleft lip and palate. *Cleft Palate Craniofac J*. 1991;28:190–193; 193.

20. Tobiasen JM, Hiebert JM, Boraz RA. Development of scales of severity of facial cleft impairment. *Cleft Palate Craniofac J*. 1991;28:419–424.