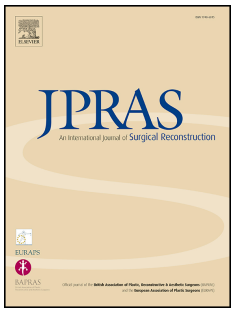
Facial aesthetic outcome analysis in unilateral cleft lip and palate surgery using web-based extended panel assessment

Bella, H.; Kornmann, N.s.s.; Hardwicke, Joseph; Wallis, K.l.; Wearn, C.; Su, T.-l.; Richard, B.m.

DOI:
10.1016/j.bjps.2016.05.006

Accepted Manuscript



Facial aesthetic outcome analysis in unilateral cleft lip and palate surgery using web-based extended panel assessment

H. Bella, N.S.S. Kornmann, J.T. Hardwicke, K.L. Wallis, C. Wearn, T.-L. Su, B.M. Richard

PII: S1748-6815(16)30081-X

DOI: 10.1016/j.bjps.2016.05.006

Reference: PRAS 4988

To appear in: *Journal of Plastic, Reconstructive & Aesthetic Surgery*

Received Date: 9 December 2015

Revised Date: 3 March 2016

Accepted Date: 17 May 2016

Please cite this article as: Bella H, Kornmann NSS, Hardwicke JT, Wallis KL, Wearn C, Su T-L, Richard BM, Facial aesthetic outcome analysis in unilateral cleft lip and palate surgery using web-based extended panel assessment, *British Journal of Plastic Surgery* (2016), doi: 10.1016/j.bjps.2016.05.006.

FACIAL AESTHETIC OUTCOME ANALYSIS IN UNILATERAL CLEFT LIP AND PALATE SURGERY USING WEB-BASED EXTENDED PANEL ASSESSMENT

H. Bella[1 *],

N.S.S. Kornmann[1 *],

J.T. Hardwicke[1,2,]

K.L. Wallis[1,]

C. Wearn[1,2],

T-L. Su[3],

B.M. Richard[1 †]

[1] Birmingham Institute for Paediatric Plastic Surgery (BIPPS) at the,

Birmingham Children's Hospital NHS Foundation Trust,

Steelhouse Lane,

Birmingham,

B4 6NH, UK.

[2] School of Clinical and Experimental Medicine,

University of Birmingham,

Edgbaston,

Birmingham,

B15 2TT, UK

[3] School of Dentistry,

University of Manchester,

Manchester,

M13 9PL,

UK.

* Joint first author

[†] **CORRESPONDING AUTHOR:**

Birmingham Children's Hospital NHS Foundation Trust,

Steelhouse Lane,

Birmingham,

B4 6NH,

UK.

KEYWORDS: cleft lip; aesthetic outcome; scoring system; random effects; web-based

This study was partly supported by the Healing Foundation (Registered Charity number 1078666).

ABSTRACT

Background: The reproducible measurement of aesthetic outcomes after cleft lip and palate (CLP) surgery remains elusive and there is no internationally recognised system. The aim of this pilot study was to better understand how humans rate post-operative aesthetic outcome after UCLP repair using a novel web-based rating platform with an extended panel of surgeon-raters.

Methods: Cropped images of five-year old UCLP patients were arranged in a randomly generated sequence within a web-based aesthetic scoring tool as part of an agreement/reliability study. Assessors rated the appearances of patients using a five-point Likert-type scale on two occasions. A mixed-effect statistical model was adopted to analyse the effects of rater, image and timing.

Results: Images of 76 patients were scored by 29 UK-based cleft surgeons. Intra-rater variability was found and the linear weighted Kappa was 0.56. This allowed identification of most and least consistent raters. The random image effect ($p<0.001$) suggested a broad range of aesthetic outcomes were included in the current study. Surgeon-raters in this study were likely to score the images more preferably at the second assessment.

Conclusions: A web-based scoring system provides extended data capture and mixed effects statistical modelling reveals the effect that time, image and rater has on the scorings. The selection and training of raters, in combination with an exemplary yardstick, might improve inter- and intra-rater agreement. There is a role for the development of objective measures based upon digital facial recognition to replace the highly variable subjective human influence on rating the aesthetic outcome.

INTRODUCTION

The measurement of aesthetic outcomes after cleft lip and palate (CLP) surgery remains elusive and despite numerous attempts at devising a scoring method there is still no internationally recognised system[1]. There is a need for a simple and reliable method of rating photographs that manages the intrinsic subjective nature of human assessment and produces a valid and reproducible result. In order to establish a valid measure, understanding what fluctuates the system is essential: what is it that the individual rater sees and how this is interpreted? These variables may include *rater-related factors* including the number of raters, timing of rating, and profession; *image-related factors* including whole or cropped aesthetic units, types of views, and two- or three-dimensional (2D or 3D) formats; *subject-related factors* including number, ethnicity, and laterality of the cleft; or *scoring-related factors* including five or seven-point Likert scales, visual analogue scales, or the use of exemplary comparators[2-6].

The validated measurement of outcomes has become an important factor in the evolution of current clinical practice: In 1998 the UK Clinical Standards Advisory Group on Cleft Lip and Palate (CSAG) recommended a centralisation of service provision for CLP patients to allow protocol driven management strategies[7]. These recommendations were based on the findings of long-term studies based upon outcome scoring systems for facial growth (using the Goslon Yardstick)[8] and speech (using the Cleft Audit Protocol for Speech – Augmented)[9]. With the addition of a scoring system for facial aesthetic outcome, optimised cleft management protocols could be further developed to allow the standardisation of best practice. Several large studies including the CSAG study, the Eurocleft and Americleft studies, have used Asher-McDade's system to assess facial

aesthetics[10-13]. Whilst many studies using this system state that it is a reliable assessment of the aesthetic outcome, they quote relatively low agreement between raters and use small numbers of mixed-professional raters, usually between four and six[10-14].

Whilst there is the expectation that computerised 3D imaging modalities might produce a valid outcome measure for cleft aesthetics[15-16], no such mechanism exists to date for either 2D or 3D images. In 2010 Pigott and Pigott introduced SymNose, a computer program designed to analyse clinical photographs by measurement of the symmetry of the lip and nose, as a surrogate for aesthetic outcome in UCLP patients[17]. Although this computer program enables rapid semi-objective comparison of these features, it remains unclear to what extent the symmetry corresponds with a subjective aesthetic result.

For the past eight years the Tri-centre Group in the UK (West Midlands, South West and Wales Regional Cleft Centres) have used the Asher-McDade-style system to evaluate cropped photographs for internal audit of practice. The aim of this pilot study was to better understand how humans rate post-operative aesthetic outcome after UCLP repair: Specifically, we would like to study inter- and intra-rater variability for an extended group of professional human raters; to characterise the images in term of their relationship to the five point Likert scale; and to study the side-cleft effect on an image being rated. All information was stored and carried out on a novel custom web-based rating portal.

MATERIALS AND METHODS

A retrospective analysis of 2D clinical photographs was undertaken and is presented according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS)[18]. Standardised anteroposterior (AP) images taken at five years of age were obtained from the Tri-Centre Cleft database of patients with UCLP born between January 1st 2000 and December 31st 2005. Exclusion criteria were: patients with any type of incomplete cleft lip and palate, bilateral cleft lip and palate, or a visible Simonart's Band on their pre-operative photographs. All images were screened and poor quality images, which could confound aesthetic scoring, were rejected: Quality was considered poor when the image resolution was less than 100 dots per inch, when saliva or mucous was obstructing view of the scar, nose or lip, when the patient was smiling or if there was no true AP view photographed.

Image processing

All AP images were cropped with a polygonal lasso to trapezoid-shape using Photoshop Elements software (Adobe Systems Incorporated, San Jose, CA). In summary, the images were initially rotated and levelled to the pupils. Horizontals were approximated to both the superior corneal limbi and the mental crease, with verticals set at both pupils. The trapezoidal crop was completed from the inferior transection of the horizontal and vertical lines, to the superior horizontal at a point corresponding to the medial canthus (Figures 1a and 1b). This technique was expanded from previously published data[3]. Hair, ears and irises were excluded from the assessment of photographs as they may influence the rating[3,19-20].

Web-based aesthetic scoring

The cropped images were arranged in a randomly generated sequence within a web-based aesthetic scoring portal on the Birmingham Institute of Paediatric Plastic Surgery secure website (Figure 1c). Invited assessors were given a personalised secure logon to access the scoring exercise and they proceeded to rate the aesthetic appearances of patients using a categorical five-point Likert-type scale (Table 1)[3]. This was done sequentially, one image at a time, as generated by the system and assessors were disabled to go back in the system and change their answers. The images were repeat scored again using the same method, after two to three weeks, by the same assessors.

SymNose analysis

The SymNose program (version 6.22; © Brian Pigott 2007-2015) provides a semi-objective measure of symmetry. Users trace the lower border of the nose and an outline of the upper lip using a digital trackpad or stylus. A vertical axis is created by bisecting a line joining the medial canthi. For the nose the axis of reflection is drawn parallel to this and equidistant from the widest points. For the lip the axis of reflection is drawn through the midpoint of the lip. The program then reflects the left side over the right. The total areas where left and right sides do not overlap (percentage mismatch), measured in pixels as a percentage of the traced area of the upper lip represent a surrogate measure of symmetry. Perfect symmetry would result in 0% mismatch. The average score from two users were given to each image.

Statistical analysis

Intra-rater agreement was studied using a weighted Kappa statistic for each rater as well as modelling in a mixed effect model (a random rater by time interaction): Two types of different weights (linear and squared) were used for Kappa statistics to identify the extreme cases with high and low intra-rater agreement. Notice that we were not to interpret the Kappa values themselves but examined the trend and identify the extreme cases.

To study the inter-rater and image variabilities of ratings, a proportional odds mixed-effect statistical model was adopted[21]. The fixed and random effects that underwent analysis were the rater, the image and the time point of the rating (first or second assessment). Likelihood ratio test was used to compare nested models for model selection purpose. To display the results of random effect, the conditional mode and variance were used to create a 95% confidence interval for each surgeon-rater and image. Details usage of this model is given in Supplement 1.

The aesthetic ranks assigned to each image from the random effect model (random image effect) were used to investigate if left- and right-sided cleft images were scored differently. The SymNose mismatch scores were also ranked to study such group difference. The Wilcoxon Rank Sum test was used to compare these ranks from each groups for human ratings and for SymNose results. All analyses were conducted using R (version 3.0.1; https://www.r-project.org/). R Packages *irr* (http://CRAN.R-project.org/package=irr) and *ordinal* (http://www.cran.r-project.org/package=ordinal) were used to calculate kappa statistics and for mixed effect modelling respectively.

RESULTS

A total of 76 patient images fulfilling the inclusion criteria were selected, cropped and uploaded onto the secure online portal. The images were scored by 29 UK-based consultant cleft surgeons on one occasion with 25 able to repeat the scoring on a second occasion between two and three weeks later. Fifteen surgeons rated all 76 images; seven surgeons rated 75 images; one surgeon rated 74 images; and two surgeons rated 73 images on both settings. Four surgeons who were only involved in one assessment had been excluded from from the intra-rater kappa analysis, but were included in mixed effect modelling. A total of 4,088 individual assessor scores were obtained from a maximum possible 4,408 scores (92.7 percent). After the model selection process, a proportional odds mixed effect model with "time" as a fixed effect, "image" and "surgeon-rater" as random effects and an "image by rater" random interaction term was judged to be suitable for the data.

Intra-rater agreement

Individual reproducibility of ratings: The linear weighted Kappa had a median of 0.56 and ranged between 0.33 and 0.67. The squared weighted Kappa had a median of 0.72 and ranged between 0.65 and 0.82. Although the median of the weighted Kappa ranged between moderate and good, there was still inconsistency in ratings at two time points for some raters. The same phenomenon was discovered by the mixed effect model approach that both methods had identified a group of the most inconsistent surgeon-raters. Interrogation of the data showed that three surgeons with the lowest agreement repeat-rated the images with the same scores at two occasions in only 34.2 percent, 43.4 percent

and 32.9 percent of cases respectively, whereas the three surgeons with the highest agreement repeat-rated the images consistently in 56.6 percent, 57.9 percent and 65.8 percent of cases respectively.

Time effect: From the mixed effect model analysis it suggested that the panel's rating behaviour changed with time: study the estimated Time effect coefficient indicating that the surgeons in this study were likely to rate the images more preferably at the second assessment.

Source of variability

In this mixed effects modelling approach the source of variation was considered after the time effect has been removed. Then the largest source of variation came from the image effect and the second largest source of variation came from the surgeon-rater effect.

Image effect: The random image effect was modelled and a large variation in ratings among these images was identified (Figure 2). This variation is desirable and reflects the intrinsic property of the images: that is a broad spectrum of aesthetic outcomes were included into the study and some are more aesthetically pleasing than others. A library of 23 exemplar images corresponding to each five-point Likert scale was constructed from the results of Figure 2, together with investigating raw image data with the most rater agreement and with clinical judgement; exemplar images of each category are shown in Figure 3.

Surgeon-rater effect: The random effect of the surgeon-rater was modelled and variation among these surgeon-raters was identified (Figure 4). Although a wide variety of images were presented, some surgeons tended towards low scores whereas some other surgeons had a tendency towards high scores. That is the five-point likert scale was not used fully by a small group of raters in that some surgeon-raters failed to recognise even one of the 76 images as “very poor” or as “excellent”. Thus the current scoring system reflected a rater’s personal views towards aesthetic outcomes. In this dataset, ten of the 29 surgeons showed statistically significant different ratings from the rest of the cohort: Five surgeon-raters tended to give low scores (towards aesthetic outcome marks as excellent) and five had a tendency to give high scores (towards aesthetic outcome marks as very poor).

SymNose analysis

From the 76 images, 51 images contained a left-sided cleft and 25 images had a right-sided cleft. Images with a left-sided cleft received significantly more favourable Likert scores than images with a right-sided cleft from the human-rater ($p = 0.02$). The asymmetry, as measured by SymNose as a percentage mismatch of lip and nose, was ranked and the distribution of the images’ ranks showed no evidence that any difference exists when between left- and right sided clefts with respect to lip mismatch ($p = 0.66$) or nose mismatch ($p = 0.69$).

DISCUSSION

A web-based aesthetic outcome assessment tool can be used to recruit an extended panel of assessors and allow them to rate freely at their own pace in their chosen environment. In this study, some surgeon-raters scored consistently well and some scored erratically and unreliably. The linear weighted Kappa in this study was slightly lower than the intra-rater reliability reported in other studies[5,10-11], yet all of these studies were carried out by only four to six mixed-professional raters compared to 29 cleft surgeon-raters in this study. The Kappa was utilised in the present study to identify the raters who were outliers from the norm, rather than for comparison with previous studies. Raters varied quite widely in their judgments, with some clearly tending to rate generously ("doves") and others having a more unfavorable approach to scoring aesthetic outcomes ("hawks"). One important finding was the effect of time on the rating, with surgeon-raters giving more preferable scores on the second assessment.

The five-point Likert scale as proposed by Asher-McDade[3] is the most commonly used method to assess the aesthetic nasolabial outcome after cleft surgery[1]. However, from the raw data in this study it can be seen that even the images receiving the highest agreement could not be categorised unanimously. The statistical robustness of the intra-rater reliability originally reported by Asher-McDade may have been compromised given the small number of raters involved[3]. Conversely, even with a larger cohort of professional raters, as in this study, the intra-rater reliability remains low. This would suggest that it is the subjective variance in perception and scoring of each rater that is the cause, rather than the scoring scale or number of raters.

Some images were marked more variably than others and it may be easier to mark an extreme result (Likert one or five) than images in the middle categories. In order to circumvent this problem Kuijpers-Jagtman *et al* suggested the need for reference photographs to produce a "yardstick"[4]. As each category represents a range of possibilities, Mercado *et al* suggested that the reliability of the assessment would improve with more than one exemplar image per category[5]. They identified four exemplar images per category and found that with the expanded nasolabial yardstick of reference the intra-rater reliability became very good. As such, a library of 23 exemplar images has been constructed from the present study. More rigorous training, selection of "reliable" raters and use of these exemplary images could produce a cohort of raters with a proven track record of steady judgments.

Regarding the laterality of the cleft, this study has found that surgeon-raters score the aesthetic outcome of right-sided clefts more severely than those on the left side. An objective difference was not evidenced by SymNose analysis. A similar conclusion has been demonstrated by mirror reversing right-sided cleft images and rescoring them[22]. There is a possibility that humans have a perceptual view of right-sided cleft repairs as being less aesthetically pleasing than left-sided.

In order to develop a reliable and objective assessment tool, the understanding of subjective assessments should be investigated further: It remains unclear why some images provided consistent ratings and why the majority of photos in this study led to widely variable judgments. Furthermore it is unknown if the appearance of the nose, the appearance of the lip or the appearance of the scar are equally contributing to a final rating. Mosmuller *et al* investigated if separate assessment of the nose and lip was more

reliable than the overall scorings[14]. They found that in assessments where lip and nose were scored together, the lip dominated the rating. The on-going development of computer programs as SymNose to perform objective quantitative measurements of lip and nose symmetry, as a surrogate measure of a good aesthetic outcome, is essential. It might be possible to identify better exemplary images or even abandon human-rater scorings and use digital facial pattern recognition to perform the assessments autonomously. We feel that, in the first instance, standardised AP photographs should be analysed prior to moving onto the worm's-eye view photograph, or more complex static or dynamic three-dimensional imaging modalities.

CONCLUSION

A web-based scoring system can provide large data capture and provides a platform for future studies. The use of a mixed effects statistical model to interrogate a large data pool has revealed the effect that time has on the ratings, as well as the effect of the images and the raters. Rating two to three weeks later tends to shift the scores globally towards the better end of the range, and a larger number of raters did not improve the statistical validity of any score. The selection and training of specific raters, with the use of an exemplary yardstick, might improve the inter- and intra-rater agreement. There is a role for the development of objective measures based upon digital facial recognition to replace the highly variable subjective human influence on rating the aesthetic outcome. However, to study the subjective human rating behaviour will help to identify the key features for such future digital automation operation.

CONFLICT OF INTEREST: None

FUNDING: None

REFERENCES

[1] Sharma VP, Bella H, Cadier MM, Pigott RW, Goodacre TE, Richard BM. Outcomes in facial aesthetics in cleft lip and palate surgery: A systematic review. J Plast Reconstr Aesthet Surg. 2012;65:1233-1245.

[2] Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;140:55.

[3] Asher-McDade C, Roberts C, Shaw WC, Gallager C. Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate. Cleft Palate Craniofac J. 1991;28:385-90.

[4] Kuijpers-Jagtman AM, Nollet PJ, Semb G, Bronkhorst EM, Shaw WC, Katsaros C. Reference photographs for nasolabial appearance rating in unilateral cleft lip and palate. J Craniofac Surg. 2009;20 Suppl 2:1683-1686.

[5] Mercado AM, Russell KA, Daskalogiannakis J, et al. The americleft project: A proposed expanded nasolabial appearance yardstick for 5- to 7-year-old patients with complete unilateral cleft lip and palate (CUCLP). Cleft Palate Craniofac J. 2015 [Epub ahead of print].

[6] Stebel A, Desmedt D, Bronkhorst E, Kuijpers MA, Fudalej PS. Rating nasolabial appearance on three-dimensional images in cleft lip and palate: a comparison with standard photographs. Eur J Orthod. 2015 [Epub ahead of print].

[7] Di Biase D, Markus A. Cleft lip and palate care in the UK: The CSAG report. Br Dent J. 1998;185:320-321.

[8] Mars M, Plint DA, Houston WJ, Bergland O, Semb G. The goslon yardstick: A new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. Cleft Palate J. 1987;24:314-322.

[9] Sell D, Grunwell P, Mildinhall S *et al*. Cleft Lip and Palate Care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech Outcomes. Cleft Palate J. 2001;38:30-37.

[10] Asher-McDade C, Brattstrom V, Dahl E, et al. A six-center international study of treatment outcome in patients with clefts of the lip and palate: Part 4. assessment of nasolabial appearance. Cleft Palate Craniofac J. 1992;29:409-412.

[11] Mercado A, Russell K, Hathaway R, et al. The americleft study: An inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 4. nasolabial aesthetics. Cleft Palate Craniofac J. 2011;48:259-264.

[12] Williams AC, Bearn D, Mildinhall S, et al. Cleft lip and palate care in the United Kingdom--the clinical standards advisory group (CSAG) study. part 2: Dentofacial outcomes and patient satisfaction. Cleft Palate Craniofac J. 2001;38:24-29.

[13] Kim JB, Strike P, Cadier MC. A simple assessment method for auditing multi-centre unilateral cleft lip repairs. J Plast Reconstr Aesthet Surg. 2011;64:195-200.

[14] Mosmuller DG, Bijnen CL, Don Griot JP, et al. Comparison of two scoring systems in the assessment of nasolabial appearance in cleft lip and palate patients. J Craniofac Surg. 2014;25:1222-1225.

[15] Trotman CA, Faraway JJ, Essick GK. Three-dimensional nasolabial displacement during movement in repaired cleft lip and palate patients. Plast Reconstr Surg. 2000;105:1273-1283.

[16] Al-Omari I, Millett DT, Ayoub A, et al. An appraisal of three methods of rating facial deformity in patients with repaired complete unilateral cleft lip and palate. Cleft Palate Craniofac J. 2003;40:530-537.

[17] Pigott RW, Pigott BB. Quantitative measurement of symmetry from photographs following surgery for unilateral cleft lip and palate. Cleft Palate Craniofac J. 2010;47:363-367.

[18] Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64:96-106.

[19] Bongaarts CA, Prahl-Andersen B, Bronkhorst EM, et al. Effect of infant orthopedics on facial appearance of toddlers with complete unilateral cleft lip and palate (dutchcleft). Cleft Palate Craniofac J. 2008;45:407-413.

[20] Prahl C, Prahl-Andersen B, van't Hof MA, Kuijpers-Jagtman AM. Infant orthopedics and facial appearance: A randomized clinical trial (dutchcleft). Cleft Palate Craniofac J. 2006;43:659-664.

[21] McCulloch CE, Searle SR, Neuhaus JM. *Generalized, linear, and mixed models. 2nd Edition*. New Jersey: John Wiley & Sons, Inc; 2001.

[22] Feragen KJ, Semb G, Magnussen S. Asymmetry of left versus right unilateral cleft impairments: An experimental study of face perception. Cleft Palate Craniofac J. 1999;36:527-532.

FIGURE LEGENDS

Figure 1 Preparation and cropping of digital photographs for web-based assessment extended panel assessment: (a) horizontal, vertical and oblique lines are set at the described anatomical landmarks on a non-cleft patient; (b) a trapezoidal crop is produced and (c) presented for rating via the secure web portal.

Figure 2 The image random effect mode with 95% confidence interval based on the conditional variance. The image identity is given for the 76 included photographs with exemplar image codes in red text (see Figure 3). Images to the left of scale tended towards excellent ratings for aesthetic outcome, and images to the right of the scale tended towards poorer ratings.

Figure 3 Five typical images for each Asher-McDade category from the library of 23 exemplar images: The images correspond to rating (1) Excellent (image ID: 41); (2) Good (image ID: 3); (3) Fair (image ID: 15); (4) Poor (image ID: 4); and (5) Very poor (image ID: 50).

Figure 4 The surgeon-rater random effect mode with 95% confidence interval based on the conditional variance. Surgeon-raters towards the left of the scale tend to rate generously (“doves”) and to the right have a more unfavorable approach to scoring aesthetic outcomes (“hawks”).

TABLES

Score	Description
1	Excellent
2	Good
3	Fair
4	Poor
5	Very poor

Table 1 The Likert-type scale with values and descriptors utilised in the web-based extended panel assessment.

Figure 1

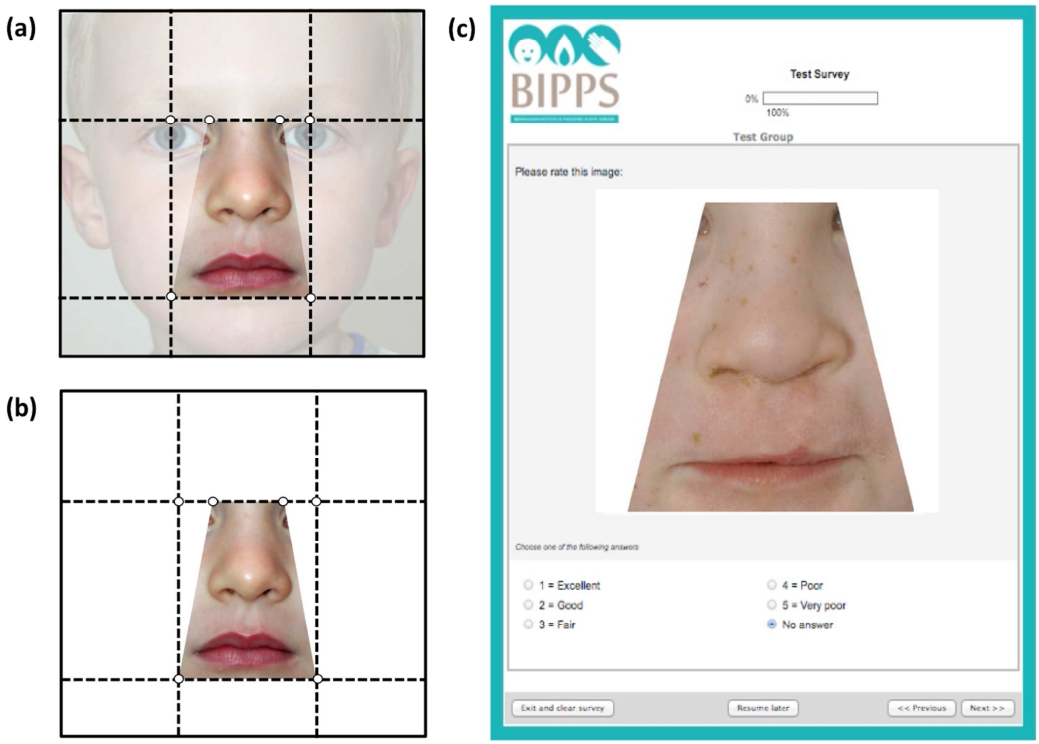


Figure 2

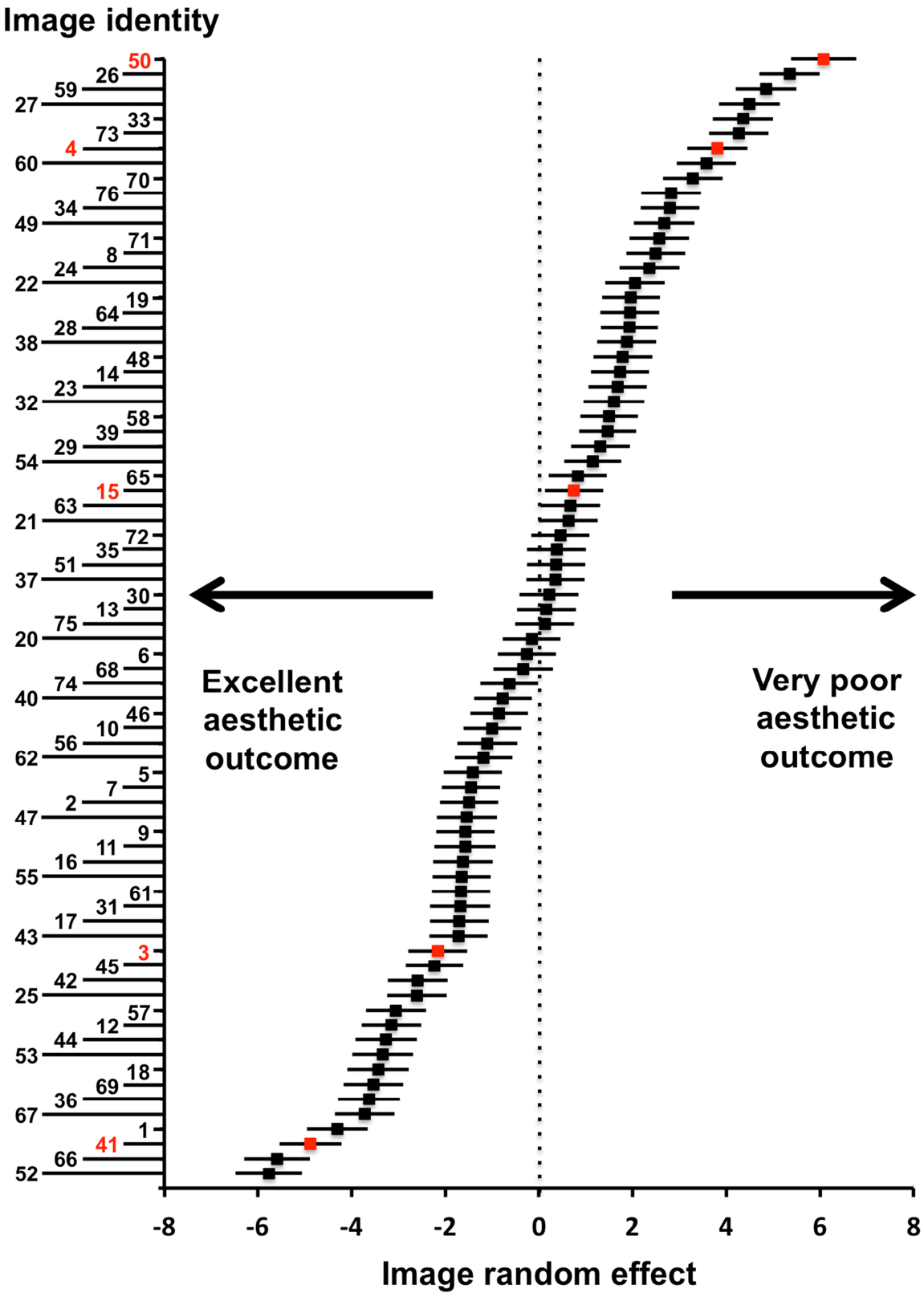
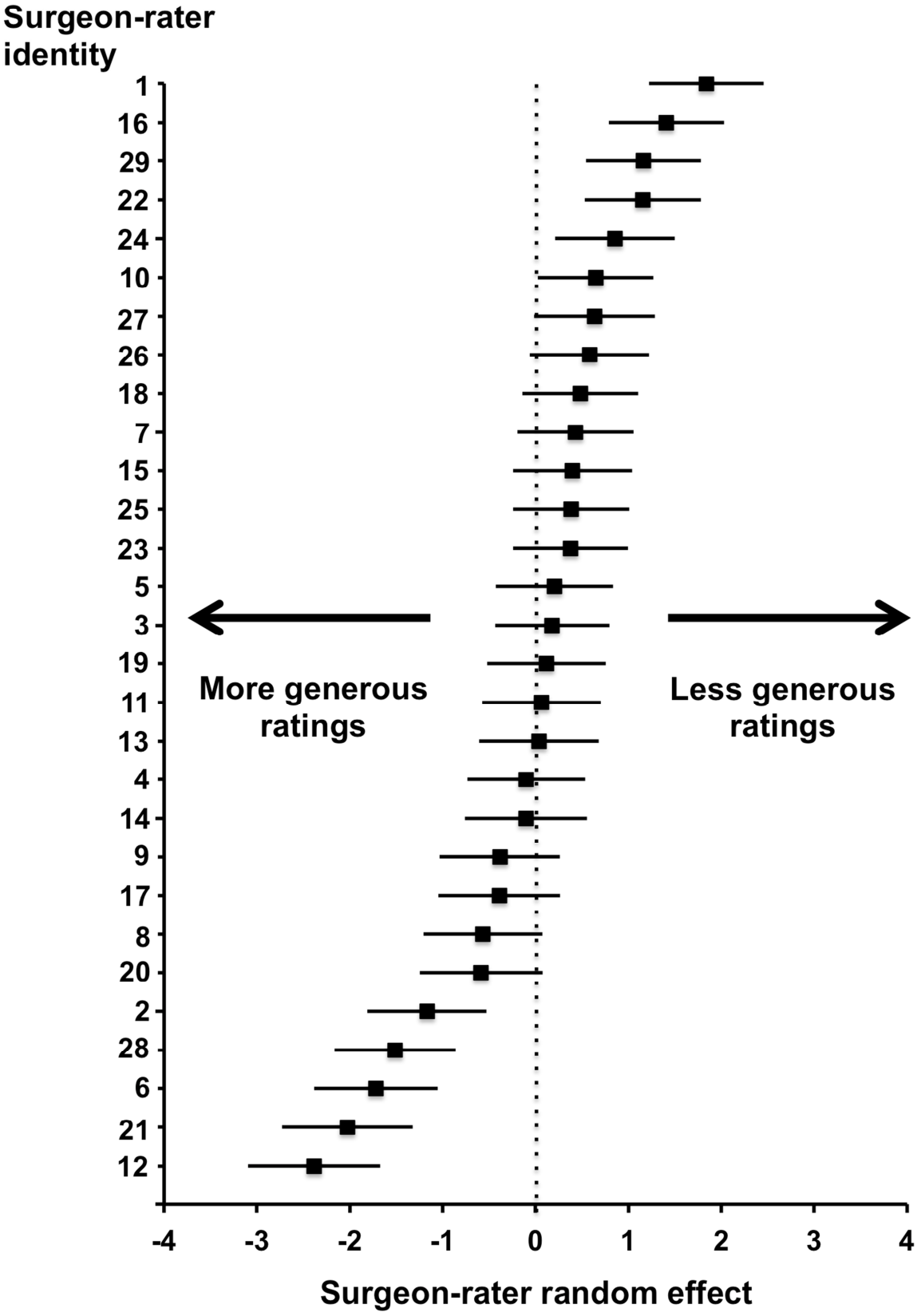


Image identity
50
26
59
27
33
73
4
60
70
76
34
49
71
8
24
22
19
64
28
38
48
14
23
32
58
39
29
54
65
15
63
21
72
35
51
37
30
13
75
20
6
68
74
40
46
10
56
62
5
7
2
47
9
11
16
55
61
31
17
43
3
45
42
25
57
12
44
53
18
69
36
67
1
41
66
52
Excellent aesthetic outcome
Very poor aesthetic outcome
-8
-6
-4
-2
0
2
4
6
8
Image random effect

Figure 3

Figure 4


Surgeon-rater identity
1
16
29
22
24
10
27
26
18
7
15
25
23
5
3
19
11
13
4
14
9
17
8
20
2
28
6
21
12
More generous ratings
Less generous ratings
-4
-3
-2
-1
0
1
2
3
4
Surgeon-rater random effect