**Introduction**

Historically, outcome analysis of facial plastic surgery has been complicated due to the subjective nature of aesthetic surgery and difficulty creating quantifiable metrics. Outcome analysis has typically been separated into two categories: patient-reported outcomes (e.g., FACE-Q)1 and clinical efficacy outcome measures (e.g., scar grading).2 However, there remains a paucity of well-designed objective outcome studies in the facial plastic surgery field.3

More recently, emotional expression has become an increasingly popular metric for outcome analysis for procedures such as brow lift surgery4, face transplant5, and facial reanimation surgery.6 Emotional expression has been a fundamental aspect of human communication and social connection throughout our species’ evolution.7 The practical importance of emotional expression for human interaction is accompanied by the aesthetic value humans place on these emotions; as seen by the implementation of surgeries such as facial rejuvenation.8 Universal emotional expression by means of facial movements have been subdivided into seven categories: sadness, happiness, anger, neutrality, surprise, fear, and disgust.9,10 These emotions have been systematically linked to the functioning of facial muscle action units through the Facial Action Coding System (FACS) originally developed by Ekman and Freisen in 1978.9,10 This system has been used as a means for analyzing facial expressions and the emotions they emit.10-12 However, utilizing FACS remains incredibly labor intensive: coding a one minute video can take up to two hours and FACS certification takes approximately 100 hours of training.13 Therefore it is rarely used in larger sample sizes.

Artificial intelligence technology has become an increasingly popular tool in academic research to expedite, scale, and decrease the cost of facial emotion analysis. FaceReader is a commercially available, validated software package that was trained and validated using the Amsterdam Dynamic Facial Expression Set14 and utilizes computer vision technology for facial expression recognition. Previous studies have validated FaceReader against the Facial Action Coding System (FACS) and found FaceReader to classify ~80% of images correctly.13 FaceReader has been studied as a potential means for creating an objective measure of facial surgery outcomes, such as facial rejuvenation and brow lift surgery.15,16

In addition to comparison with tools such as FACS, it is also important to understand how FaceReader analysis compares to lay-person interpretation of facial expressions, as this more accurately reflects patients’ and the general public’s perception of their emotional expression. Therefore, the purpose of this study is to compare FaceReader analysis to human evaluator interpretation of emotional expression to define standard values for the software output.

**Methods**

*Image Generation*

Randomly generated facial images were obtained from the website, thispersondoesnotexist.com. This website utilizes a class of machine learning frameworks called generative adversarial networks (GAN) to generate fake portraits from a large dataset of real images. Images were selected based on neutrality of expression and perceived age (20 children, 20 young adult, 20 middle aged, and 20 elderly). 38 male and 42 female images were selected. Images with beards, glasses, or hats were not included.

*Survey Analysis*    
The 80 images were analyzed by human evaluators through a survey. Participants were sourced using Amazon Mechanical Turk, a crowdsourcing marketplace utilized to outsource virtual tasks to a distributed workforce. Survey participants were offered $0.50 to complete out survey. A total of 496 responses were solicited due to budget constraints.

The 496 participants who each evaluated 20 randomly selected images. The number of evaluators per image ranged from 118 to 130, with an average of 124 evaluators per image. Survey participants were asked for basic demographic information including race (318 White, 94 Asian, 48 Black or African American, 22 Latino or Hispanic, 10 Native American, 4 other), gender identity (345 male, 149 female, 2 prefer not to say), and highest level of education (354 Bachelor’s, 75 Graduate, 27 some college but no degree, 21 High school degree or equivalent, and 19 Associate degree). Average survey participant age was 33.85 years old. The age, primary facial emotion (neutral, happy, sad, angry, surprised, scared, disgusted), and intensity of emotion (1=not intense to 3=very intense) determined by each evaluator was recorded. The results of the survey were averaged across participants.

*FaceReader Image Analysis*The 80 images were analyzed using a commercially available facial expression recognition software package (FaceReader™, Noldus Information Technology BV, Wageningen, Netherlands). The software generates data on the proportion of each emotion expressed for any given facial movement and the associated action units. The software’s capability to classify facial expressions was achieved by training an artificial neural network using more than 10,000 images that were manually annotated by trained experts.17-19 The system assesses the movements of more than 500 facial landmarks on each face to perform the classification. The age, facial emotion, and intensity of facial action units (0=no action unit detected to 4=the most intense action unit detected) generated by the software were recorded.

Statistical analysis was performed using JMP (SAS Institute Inc.) and RStudio(RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA). Cohen’s kappa coefficient was used to test correlation for categorical variables. Intraclass correlation coefficient was used to test correlation for continuous variables. Correlation interpretation was based on established categorizations: poor (κ = 0), slight (κ =0.01-0.20), fair (κ =0.21-0.40), moderate (κ =0.41-0.60), strong (κ =0.61-0.80), near perfect (κ =0.81-0.99), and perfect (κ =1). A value of p < 0.05 was considered statistically significant.

**Results**

*FaceReader Analysis*The most common expressions identified by FaceReader were happy (40.5%) and neutral (31.2%).

The most common emotion detected across age groups (child, adolescent, middle aged, and elderly) was also Happy followed by Neutral. There was no significant difference between primary emotion detected across age groups.

FaceReader also estimates the age range of images. FaceReader age range analysis is included in Table 1.

*Survey Analysis*In the survey, the most common primary emotions detected were also happy (63.8%), neutral (33.8%), and surprise (2.5%). Happiness was the most common primary emotion across all age ranges.

Interpreters also agreed more on the primary emotion if it was happiness (75% agreement vs. 65% agreement on average), as compared to neutral (49% agreement) and surprise (53% agreement). (Chart 1) For images where the primary emotion was detected as happiness, average intensity of emotion as determined by survey participants was 2.29 out of 3 (1=not intense to 3=very intense). For images where the primary emotion was neutral, average intensity of emotion was 2.09/3; for surprise it was 2.37/3.

Participants were also asked to estimate the age range of the images. Survey participant age range analysis is included in Table 2.

*FaceReader vs. survey evaluator comparison*For images where survey participants identified happiness as the primary emotion, FaceReader also rated happiness as the most common emotion (75.4%), as compared to surprised (22%) or neutral (5.87%) (Chart 2)*.* These images were also associated with a higher cheek raiser and lip corner puller intensity, as compared to those identified as neutral or surprised (0.27 vs. 0.02 vs. 0 and 0.64 vs 0.07 vs. 0.01, respectively). This aligns with our primary emotion analysis, as FaceReader correlates the emotion of happiness with the cheek raiser and lip corner puller action units.

Similarly, for images where survey participants identified surprise as the primary emotion, FaceReader detected surprise (14.6%) more often than happiness (0.35%) or neutral (2%). These images were associated with higher inner brow raiser intensity as compared to those identified as happy or neutral (0.017 vs. 0.001 vs. 0.013, respectively, as well as outer brow raiser intensity (0.09 vs. 0 vs. 0.02, respectively) and jaw drop intensity (0.68 vs. 0.02 vs. 0.06), respectively. This aligns with our primary emotion analysis, as surprise is correlated with the inner brow raiser, outer brow raiser, upper lid raiser and jaw drop action units. Interestingly, these images were not correlated with a higher upper lid raiser intensity (0 vs. 0.002 vs. 0.001, respectively).

Finally, for images where survey participants identified neutral as the primary emotion, FaceReader also rated neutral as the most common emotion (82.7%), as compared to happiness (16.2%) or surprise (32.2%).

Analysis of correlation between most common expression identified by FaceReader and the primary emotion detected by surveyors showed strong correlation (κ = 0.77, 95% CI = 0.64 – 0.91).

On analyzing this correlation by age group, there was fair correlation in children (κ = 0.40, 95% CI = 0.078 - 0.72), perfect correlation in young adults(κ = 1.0, 95% CI = 1.0 -1.0), strong correlation in middle aged adults(κ = 0.79, 95% CI = 0.53 - 1) and near perfect in elderly adults(κ = 0.9 , 95% CI = 0.7 - 1.0).

Survey participant and FaceReader age ranges were averaged to facilitate comparison. Age detection between FaceReader and survey responders had strong correlation (ICC = 0.886, 95% CI – 0.828 – 0.925, p<0.001).

**Discussion**

In this study, we demonstrated that human interpreters and FaceReader output had substantial agreement when interpreting facial emotional expression (kappa coefficient = 0.77, 95% CI = 0.64 – 0.91). Correlation was strongest in images of adults (young adults: κ = 1.0, 95% CI = 1.0 -1.0; middle aged adults, κ = 0.79, 95% CI = 0.53 – 1; elderly adults, κ = 0.9 , 95% CI = 0.7 - 1.0), whereas there was only fair agreement in images of children (κ = 0.40, 95% CI = 0.078 - 0.72). Furthermore, we found that human interpreters and FaceReader output had almost perfect agreement when detecting age (ICC = 0.886, 95% CI – 0.828 – 0.925, p<0.001). Therefore, out study found that FaceReader can be used as an approximate measure of lay person interpretation of emotional expression. To our knowledge, this is the first study creating standard values for FaceReader.

Emotional expression is becoming an increasingly important outcome measure in facial plastic surgery. Negative expressions of tiredness, sadness, and anger are a significant reason why many patients seek facial plastic surgery.20 Knoll et al. utilized human subjects to evaluate edited images to study influence of eyebrow, forehead, and periorbital aesthetics on the perception of tiredness, finding modifications in brow contour and length of pretarsal lid height had significant impacts of perceived emotion.4 Our group expanded on this study, utilizing artificial intelligence to compare emotional expression in pre-operative vs. post-operative brow lift patients; our findings included a significant increase in perceived happiness and decrease in perceived sadness in post-operative brow lift patients.16 Furthermore, a similar study on facial reanimation found that FaceReader detected a significantly greater happy signal in post-operative vs. pre-operative facial palsy patients after facial reanimation, as well as reduced sad signal and neutral signal.21 Finally, our analysis of facial rejuvenation surgery found an increased detection in happy emotion in post-operative facial rejuvenation patients, as well as significant decreased in angry emotion detection.15 This method of study has also been utilized for face transplant surgery; Dorante et. al utilized FaceReader to assess facial emotional expression after face transplant, finding a significant restoration of the happiness in post-transplant patients 1-year after transplant.5

While these studies have demonstrated the efficacy of utilizing AI as an objective means of evaluating facial expression in facial plastic surgery patients, it is important correlate lay person ratings and FaceReader generated output to properly inform patients. As the usage of artificial intelligence in surgery continues to grow22,23, the necessity of AI assurance continues to increase to identify potential biases and to ensure that generated analysis is relevant and trustworthy.24 Previous studies have used real world validation to assess artificial intelligence algorithms for ophthalmic imaging.25 This paper aims to serves to validate the usage of FaceReader for facial plastic surgery outcome analysis, providing a foundation for future papers. Furthermore, our findings suggest a strong correlation between detection of happiness and neutrality in FaceReader analysis when compared to survey evaluator ratings, providing a link between software outcomes and real-life perception. Clinicians can use this link as a foundation in patient discussions, specifically when discussing how others may perceive them pre- vs. post- surgery.

This study also provides a novel solution to the continuing problem of objective outcome analysis in aesthetic surgery. Current models largely rely on patient-reported outcomes in the format of time-consuming and laborious surveys. For example, a comprehensive study of cosmetic surgery outcome tools conducted by Ching et al. found body-image and quality-of-life measures, such as the FAST and DAS59 questionnaires, to be most useful; however the questionnaires have 18 and 59 questions, respectively.26 Similarly, FACE-Q, a validated tool that measures concepts and symptoms important to facial aesthetic patients, utilizes 40 independently functioning scales and checklists.27 While patient-reported outcomes remains an important tool for aesthetic surgery, artificial intelligence software such as FaceReader provide more objective and efficient analysis of both images and videos that can be conducted real-time at a large scale.

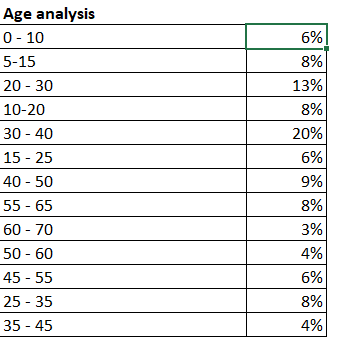
*Limitations*

This paper has several limitations. The overall sample size of respondents was relatively small, potentially skewing the survey results. Furthermore, only images with neutral expressions were chosen limiting FaceReader and human interpretation of additional emotions, such as sadness or anger. Finally, the age ranges outputted by FaceReader and survey participants were different; therefore, age ranges had to be averaged to accurately compare age detection.

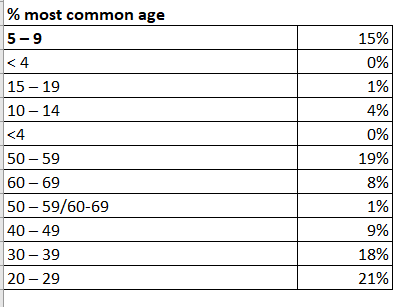
**Conclusion**We provided the first study defining the expected average values generated by FaceReader in generally smiling images. These values include emotion percentage and action unit intensity. This can be used as a standard in future studies and can also provide a correlation between the software and human raters. Furthermore, clinicians can use this information when discussing potential post-surgical outcomes with their patients in terms of public perception. Future studies will focus on FaceReader vs. human interpretation of images expressing various emotions, as well as video analysis.

**Figures and Tables**

**Table 1: FaceReader Age analysis**



**Table 2: Survey participant Age range analysis**



**Chart 1**

**Chart, bar chart

Description automatically generated**

**Chart 2**

**Chart, bar chart, waterfall chart

Description automatically generated**

1. Klassen AF, Cano SJ, Scott A, Snell L, Pusic AL. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. *Facial Plast Surg.* 2010;26(4):303-309.

2. Rhee JS, McMullin BT. Measuring outcomes in facial plastic surgery: a decade of progress. *Curr Opin Otolaryngol Head Neck Surg.* 2008;16(4):387-393.

3. Most SP, Alsarraf R, Larrabee WF, Jr. Outcomes of facial cosmetic procedures. *Facial Plast Surg.* 2002;18(2):119-124.

4. Knoll BI, Attkiss KJ, Persing JA. The influence of forehead, brow, and periorbital aesthetics on perceived expression in the youthful face. *Plast Reconstr Surg.* 2008;121(5):1793-1802.

5. Dorante MI, Kollar B, Obed D, Haug V, Fischer S, Pomahac B. Recognizing Emotional Expression as an Outcome Measure After Face Transplant. *JAMA Netw Open.* 2020;3(1):e1919247.

6. Dusseldorp JR, Hadlock TA. Using Artificial Intelligence to Measure Facial Expression following Facial Reanimation Surgery. *Plastic and Reconstructive Surgery.* 2022;149(2):343e-345e.

7. Jablonka E, Ginsburg S, Dor D. The co-evolution of language and emotions. *Philos Trans R Soc Lond B Biol Sci.* 2012;367(1599):2152-2159.

8. Barrett DM, Gerecci D, Wang TD. Facelift Controversies. *Facial Plast Surg Clin North Am.* 2016;24(3):357-366.

9. Ekman P, Friesen WV. Facial action coding system: a technique for the measurement of facial movement. 1978.

10. Clark EA, Kessinger JN, Duncan SE, et al. The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review. *Front Psychol.* 2020;11:920. doi:10.3389/fpsyg.2020.00920. Accessed 2020.

11. Asaad M, Dey JK, Al-Mouakeh A, et al. Eye-Tracking Technology in Plastic and Reconstructive Surgery: A Systematic Review. *Aesthet Surg J.* 2020;40(9):1022-1034.

12. Charafeddine AH, Drake R, McBride J, Zins JE. Facelift: History and Anatomy. *Clin Plast Surg.* 2019;46(4):505-513.

13. Skiendziel T, Rösch AG, Schultheiss OC. Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. *PLoS One.* 2019;14(10):e0223905.

14. Schalk J, Hawk S, Fischer A, Doosje B. Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion (Washington, DC).* 2011;11:907-920.

15. Boonipat T, Hebel N, Zhu A, Lin J, Shapiro D. Using artificial intelligence to analyze emotion and facial action units following facial rejuvenation surgery. *J Plast Reconstr Aesthet Surg.* 2022.

16. Boonipat T, Lin J, Bite U. Detection of Baseline Emotion in Brow Lift Patients Using Artificial Intelligence. *Aesthetic Plast Surg.* 2021;45(6):2742-2748.

17. Bishop CM. *Neural networks for pattern recognition.* Oxford university press; 1995.

18. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence.* 2001;23(6):681-685.

19. Van Der Schalk J, Hawk ST, Fischer AH, Doosje B. Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion.* 2011;11(4):907.

20. Matarasso A, Terino EO. Forehead-brow rhytidoplasty: reassessing the goals. *Plast Reconstr Surg.* 1994;93(7):1378-1389; discussion 1390-1371.

21. Boonipat T, Asaad M, Lin J, Glass GE, Mardini S, Stotland M. Using Artificial Intelligence to Measure Facial Expression following Facial Reanimation Surgery. *Plastic and Reconstructive Surgery.* 2020;146(5):1147-1150.

22. Dagli MM, Rajesh A, Asaad M, Butler CE. The Use of Artificial Intelligence and Machine Learning in Surgery: A Comprehensive Literature Review. *Am Surg.* 2021:31348211065101.

23. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg.* 2018;268(1):70-76.

24. Batarseh FA, Freeman L, Huang C-H. A survey on artificial intelligence assurance. *Journal of Big Data.* 2021;8(1):60.

25. Domalpally A, Channa R. Real-world validation of artificial intelligence algorithms for ophthalmic imaging. *The Lancet Digital Health.* 2021;3(8):e463-e464.

26. Ching S, Thoma A, McCabe RE, Antony MM. Measuring outcomes in aesthetic surgery: a comprehensive review of the literature. *Plast Reconstr Surg.* 2003;111(1):469-480; discussion 481-462.

27. Klassen AF, Cano SJ, Schwitzer JA, Scott AM, Pusic AL. FACE-Q scales for health-related quality of life, early life impact, satisfaction with outcomes, and decision to have treatment: development and validation. *Plast Reconstr Surg.* 2015;135(2):375-386.