2019

Let's Face It: The effect of orthognathic surgery on facial recognition algorithm analysis

Carolyn Bradford Dragon
Department of orthodontics

Let's Face It: The effect of orthognathic surgery on facial recognition algorithm analysis

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Dentistry at Virginia Commonwealth University.

By

Carolyn B. Dragon, D.M.D.
B.S., Biological Sciences, Cornell University, 2013
D.M.D., University of Connecticut, 2017

Thesis Director: Bhavna Shroff, D.D.S., M.Dent.Sc., M.P.A.
POSTGRADUATE PROGRAM DIRECTOR, DEPARTMENT OF ORTHODONTICS

Virginia Commonwealth University
Richmond, Virginia
April 2019

ACKNOWLEDGEMENT

First, I would like to thank my wonderful husband, Tom, for sticking with me through six years of post-graduate education and never once questioning my life choices. I cannot even begin to express my gratitude to Dr. Steven Lindauer, Dr. Bhavna Shroff, and Dr. Eser Tüfekçi for fostering my orthodontic education at Virginia Commonwealth University. I would especially like to thank Dr. Shroff, my thesis advisor, for her fearless guidance and encouragement as we delved into the heretofore unexplored trenches of facial recognition in the world of orthodontics. In addition, I would like to thank Dr. Caroline Carrico for her help with data analysis and interpretation. I would like to extend my sincere thanks to Andrew Arias and Spiro Stilianoudakis for their assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Aim: To evaluate the ability of a publicly available facial recognition application program interface (API) to calculate similarity scores for pre- and post-surgical photographs of patients undergoing orthognathic surgeries. Our primary objective was to identify which surgical procedure(s) had the greatest effect(s) on similarity score.

Methods: Standard treatment progress photographs for 25 retrospectively identified, orthodontic-orthognathic patients were analyzed using the API to calculate similarity scores between the pre- and post-surgical photographs. Photographs from two pre-surgical timepoints were compared as controls. Both relaxed and smiling photographs were included in the study to assess for the added impact of facial pose on similarity score. Surgical procedure(s) performed on each patient, gender, age at time of surgery, and ethnicity were recorded for statistical analysis. Nonparametric Kruskal-Wallis Rank Sum Tests were performed to univariately analyze the relationship between each categorical patient characteristic and each recognition score. Multiple comparison Wilcoxon Rank Sum Tests were performed on the subsequent statistically significant characteristics. P-Values were adjusted for using the Bonferroni correction technique.

Results: Patients that had surgery on both jaws had a lower median similarity score, when comparing relaxed expressions before and after surgery, compared to those that had surgery only on the mandible ($p = 0.014$). It was also found that patients receiving LeFort and bilateral sagittal split osteotomies (BSSO) surgeries had a lower median similarity score compared to those that received only BSSO ($p = 0.009$). For the score comparing relaxed expressions before surgery versus smiling expressions after surgery,

patients receiving two-jaw surgeries had lower scores than those that had surgery on only the mandible ($p = 0.028$). Patients that received LeFort and BSSO surgeries were also found to have lower similarity scores compared to patients that received only BSSO when comparing pre-surgical relaxed photographs to post-surgical smiling photographs ($p = 0.036$).

Conclusions: Two-jaw surgeries were associated with a statistically significant decrease in similarity score when compared to one-jaw procedures. Pose was also found to be a factor influencing similarity scores, especially when comparing pre-surgical relaxed photographs to post-surgical smiling photographs.

INTRODUCTION

Biometric technology, the conversion of physiologic or behavioral human characteristics into usable data sets, has been steadily gaining traction in both public and private security sectors. Biometric systems essentially consist of a *training* portion, in which physiognomic data is entered into the system and converted into a storable format, and an i*dentification* portion, in which new input data are converted into the same format as the stored data already on record and compared to find a match. Some authors argue that facial recognition serves as a biometric "sweet-spot" between fingerprint identification, which relies on voluntary participation, and iris scans, which can be expensive and considered too invasive by the general public.[1] A relatively recent article outlined efforts taken by some international cities to implement facial recognition for law enforcement purposes via closed-circuit television (CCTV). These places include Logan City, Queensland, which was reported to have embarked upon implementation of a CCTV system employing facial recognition capabilities and Bogota, Columbia, which tested a biometric-based video surveillance program in one of the city's transportation systems. This same article also reported a recommendation from London, England, which advised placement of CCTV cameras at face-level to permit effective use of facial recognition technologies.[2]

Modern society finds itself facing a culture increasingly reliant upon technologies, especially those created to increase efficiency and productivity. These technologies have infiltrated many aspects of everyday life, everything from the cars we drive, to the

security systems that guard our homes, to the phones we use for almost everything we do. Facial recognition technology is yet another example of the types of automation trickling into routine use.

The essence of facial recognition technology is the assessment of new images with respect to ones already stored in a searchable database. Matches and potential matches are identified by comparison of specific facial characteristics between two images. Early facial recognition was limited to two-dimensional analysis; thus, successful identification of face matches required nearly identical lighting, pose, etc. The advent of three-dimensional face analysis has improved facial recognition, relying more on the physical make-up of the face under analysis, with less dependence on similarity of image capture. Three-dimensional systems use specific algorithms to create a digital representation of the face, based on locations of facial landmarks; it is this digital construct that is then compared between images.[3]

Unlike humans, who rely on neural encoding from multiple exposures to a face and correct recall of “invariant features” to recognize a face, automated face recognition programs have much greater memory capacity than humans, and the quality of the encoded data does not degrade with time as human memories can. Both humans and automated face recognition programs improve recognition when more information is obtained (i.e. multiple views of a face); however, human recognition is frequently impaired when changes are made to “peripheral features,” such as hairstyle, facial hair, presence of accessories.[4] While automated facial recognition has attempted to combat confounding due to peripheral feature changes, the effect on automated face

recognition following changes to invariant features occurring with orthognathic jaw surgeries has not been explored.

Orthognathic surgery has both functional and esthetic utility. Jaw movements can be made to correct skeletal malocclusions and malformations, and these bony changes often come with concomitant changes to the soft tissue of the face. Thus, these changes may pose difficulties to facial recognition algorithms relying on the interrelationships of different facial points.[5] Previous studies examining the effect of plastic surgeries on older iterations of facial recognition technology suggest that those programs were not sufficiently powerful to make correct identifications; specifically, changes in facial geometry, texture, and features, all to varying degrees, combine to render automated recognition difficult.[6] While orthognathic surgeries can have similar effects on the human face, the impact of these procedures on facial recognition success appears unexplored as of now. As facial recognition technology becomes progressively more mainstream, the need for reliability becomes more and more critical.

Different facial recognition programs currently exist, from those utilized by the federal government to maintain national security, to the one powering the Apple iPhone security feature, Face ID. Retail giant Amazon markets its own facial recognition program for both public and private consumption. Some examples of widely known companies making use of Amazon's Rekognition software include: Scripps Networks for data sorting and search optimization for customers and employees, the Washington County Sheriff's Office for record organization, and C-SPAN for indexing content to facilitate searches.[3]

The aim of the present study was to evaluate the ability of a publicly available facial recognition application program interface (API) to calculate similarity scores for pre- and post-surgical photographs of patients undergoing orthognathic surgeries. Our primary objective was to identify which surgical procedure(s) had the greatest effect(s) on similarity score.

MATERIALS & METHODS

The charts of twenty-five patients were identified via a retrospective chart review of the Virginia Commonwealth University (VCU) Orthodontics Clinic electronic health record. All patients whose records were included in this study had previously undergone orthognathic surgery in one or both jaws. Pre- and post-surgical photographs and cephalograms were verified to be present in the clinic's Dolphin Imaging database. Exclusion criteria comprised age of less than eighteen years at the time of surgery and the presence of partial or complete cleft lip and palate.

For each of the twenty-five individuals whose records were used in this study, a unique numerical identifier was assigned, and gender, age at time of surgery, and type(s) of surgical procedure(s) performed were recorded. Frontal relaxed and frontal smiling photographs from one pre-surgical and one post-surgical timepoint were extracted from the standard set of orthodontic records photographs. An additional frontal relaxed photograph (termed "pre-pre-surgical" in this study) from an earlier pre-surgical timepoint was also identified to serve as a control comparison. For these same twenty-five cases, existing pre- and post-surgical lateral cephalograms were downloaded. Only photographs of patients with full fixed appliances were used in an effort to minimize confounding due to the presence/absence of metal braces. A password-protected Excel file using the assigned numerical identifiers was created to record all study data. At this point in the study, all personal identifiers (names, chart numbers, etc.) were dissociated from the photographs/cephalograms.

The selected records photographs were used to test Amazon Web Service's (AWS) Rekognition program. Of note, per the software Developer Guide, "Amazon Rekognition does not persist any information discovered about the input image" and "no input image bytes are persisted by non-storage API operations." In other words, these "non-storage" functions do not store uploaded photographs in the cloud-based service. The *Compare Faces* function within Amazon Rekognition's "Non-storage API Operations" (API = Application Program Interface) was used to measure similarity scores between the test images. The Compare Faces function was first used for two control pairings for each subject:

Control #1: pre-pre-surgical frontal relaxed vs. pre-surgical frontal relaxed i.e. two separate presurgical timepoint photographs, both of which were taken with fixed appliances in place

Control #2: pre-surgical frontal relaxed vs. pre-surgical frontal smiling i.e. relaxed and smiling photographs from the same timepoint to test for the effect of pose

The Compare Faces function was then used for four test pairings for each subject:

Test #1: pre-surgical frontal relaxed vs. post-surgical frontal relaxed

Test #2: pre-surgical frontal smiling vs. post-surgical frontal smiling

Test #3: pre-surgical frontal smiling vs. post-surgical frontal relaxed

Test #4: pre-surgical frontal relaxed vs. post-surgical frontal smiling

To quantify the magnitude of surgical movements, one examiner traced each of the 50 cephalograms, one pre-surgical cephalogram and one post-surgical cephalogram for each patient, using Dolphin Imaging. To ensure intraexaminer reliability, a random sampling of 10 cephalograms were re-traced at least one week after the original tracing was completed.

The following pre- and post-surgical lateral cephalogram metrics were recorded in the Excel file:

1. *SNA* - the angle formed between the cranial base and most upper jaw
2. *SNB* - the angle formed between the cranial base and the lower jaw
3. *ANB* - the angle formed between the upper jaw and lower jaw
4. *Upper incisor inclination*
5. *Lower incisor inclination*
6. *Upper lip to E-plane* - measure of upper lip protrusion
7. *Lower lip to E-plane* - measure of lower lip protrusion
8. *Facial convexity*

Wilcoxon Rank Sum tests were performed on the control measurements to compare facial poses before surgeries. These were used to confirm the validity of the application. The null hypothesis was that the comparison scores of relaxed and smiling facial expressions, before and after surgery, was equal to 99. A rejection of this hypothesis at the 0.05 alpha level would imply that the application was inherently flawed and that any further differences in scores could be attributed to the application itself. Additional Wilcoxon Rank Sum tests were used to assess if there were differences in angular measurements between the 8 cephalometric measures. The relationship between the

difference (in absolute value) between each cephalometric measure and the facial expression comparisons was then evaluated using Spearman Rank Correlation Coefficients. Additional Spearman Rank Correlation Coefficients were used to assess the pairwise relationship between each of the 4 test pairings. One-way Kruskal-Wallis Rank Sum Tests were then used to assess any differences in median scores for specific categorical variables, across each test pairing. Lastly, multiple comparison Wilcoxon Rank Sum tests were performed to confirm any within group differences. Type I error inflation was adjusted for using the Bonferroni correction procedure. All statistical analysis was performed in R version 3.5.1.

RESULTS

Records for a total of 25 patients who underwent orthognathic surgeries in the Virginia Commonwealth University Oral & Maxillofacial Surgery Clinic were identified. Table 1 presents the demographic characteristics of those included in the analysis. The average age of the patients was 24.5 years (SD 9.2 years). The distribution of sex was relatively even with 13 males (52%) and 12 females (48%). The majority of patients were classified as white, for a total of 14 (56%), compared to 8 African American patients (32%), and 3 recorded as other (12%).

Table 1. Demographic characteristics for patients whose records were included in this sample.

Variable	**N = 25**
Age (SD)	24.5 (9.2)
Gender (%)	
Male	13 (52%)
Female	12 (48%)
Ethnicity (%)	
African American	8 (32%)
White	14 (56%)
Other	3 (12%)

Table 2 shows that 13 patients had surgery performed on both jaws (52%), while 8 had surgery on only the maxilla (28%), and 5 had surgery on only the mandible (20%). The type of surgical procedure was grouped into 1 of 3 categories: bilateral sagittal split osteotomy (BSSO), LeFort I osteotomy (LeFort I), or both. A total of 12 patients were classified as having received both the BSSO and LeFort I procedures combined (48%). Eight of the 25 patients received the LeFort I, only (32%), while 5 patients received the BSSO, only (20%).

Table 2. Surgical characteristics.

Variable	N = 25
Jaw (%)	
Maxilla	7 (28%)
Mandible	5 (20%)
Both	13 (52%)
Type of Surgery	
BSSO	5 (20%)
LeFort I	8 (32%)
Both	12 (48%)

Similarity scores (0-100) used by the facial recognition application program interface (API), were evaluated for 4 combinations of facial compositions including: Pre-surgical Relaxed vs Post-surgical Relaxed, Pre-surgical Smiling vs Post-surgical Smiling, Pre-surgical Relaxed vs Post-surgical Smiling, and Pre-surgical Smiling vs Post-surgical Relaxed. Two control measurements were recorded comparing each subject's facial similarity score: comparison of relaxed vs smiling at the pre-surgical timepoint, and comparison of relaxed vs relaxed at two different pre-surgical timepoints. Two Wilcoxon Rank Sum Tests were used to confirm the validity of the application, in that there was not a significantly different score from 99. That is, when comparing the images with different expressions before surgery and then again after surgery, the application did not detect a difference, as would be expected.

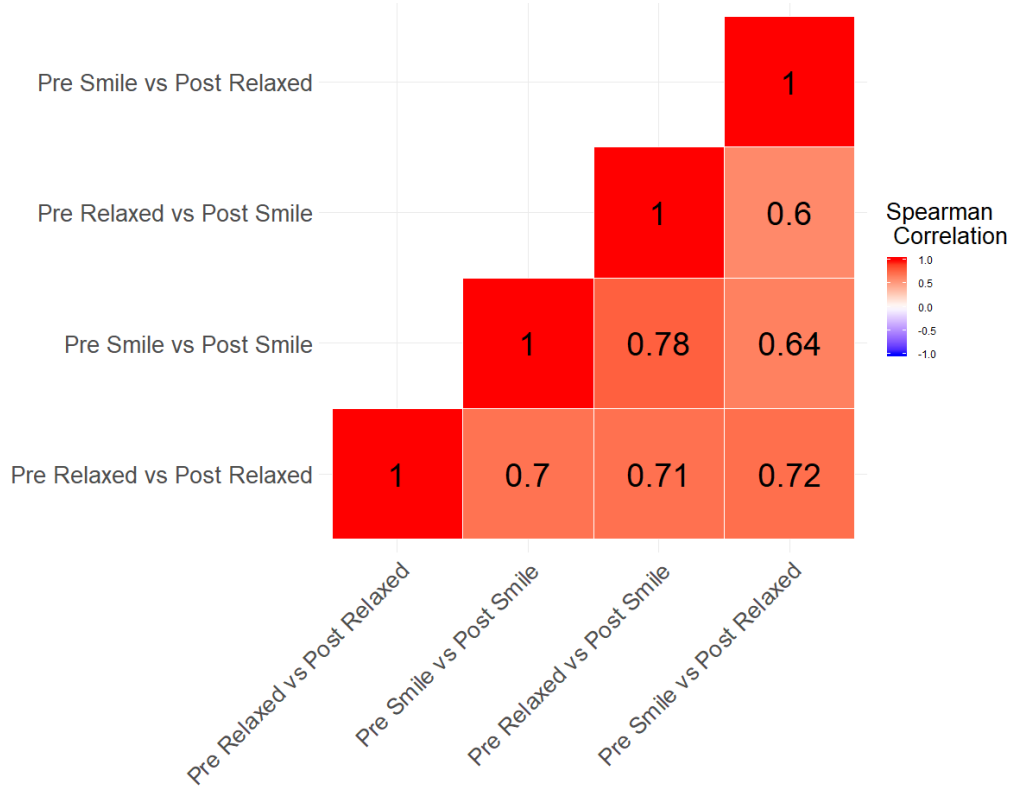
Table 3 presents the summaries of the similarity scores. Median values were used instead of means since the data were not normally distributed. When relaxed photographs taken before and after surgery were compared, the median similarity score was 98 (IQR: 97-99). Evaluations of smiling photographs before and after surgery also resulted in a median of 98 (IQR: 96-99). Similarity scores were also used to compare

the two facial expressions before and after surgery to account for the effect of facial pose – a known challenge for facial recognition programs. Pre-surgical Relaxed photographs tested against Post-surgical Smiling photographs showed a median score of 96 (IQR: 94-97). Lastly, the median scores comparing patients' smiling expressions before surgery and relaxed expressions after surgery was found to also be 96 (IQR: 94-98). When evaluating the sets of comparisons between expressions, we found that each comparison exhibited high correlations (Figure 1). The Spearman Rank correlation between the Pre-surgical Relaxed vs Post-surgical Relaxed and Pre-surgical Smiling vs Post-surgical Smiling was found to be 0.70 ($p < 0.001$), indicating that, when the similarity scores were compared between the tests of like facial expressions (i.e. relaxed vs relaxed and smiling vs smiling), there was a significantly high correlation between the similarity scores. In other words, the API's scoring was consistent for like facial expression comparisons in the same individual. Similarly, the correlation between the Pre-surgical Relaxed vs Post-surgical Smiling and Pre-surgical Smiling vs Post-surgical Relaxed was also relatively high with a value of 0.60 ($p=0.002$). This indicated that the ordering of relaxed and smiling in the pictures had little effect on the API's scoring.

Table 3. Similarity scores across the 2 different expressions, before and after orthognathic surgery. Scores were obtained via the Rekognition application program interface (API).

Pre- vs Post-surgical Similarity Score Medians (IQR)	
Relaxed Pre vs Post	98 (97-99)
Smiling Pre vs Post	98 (96-99)
Relaxed Pre vs Smiling Post	96 (94-97)
Smiling Pre vs Relaxed Post	96 (94-98)

Figure 1. Spearman Rank Correlation Coefficients of the 4 similarity scores used to measure facial recognition before and after surgery. Higher correlations are colored in darker red.



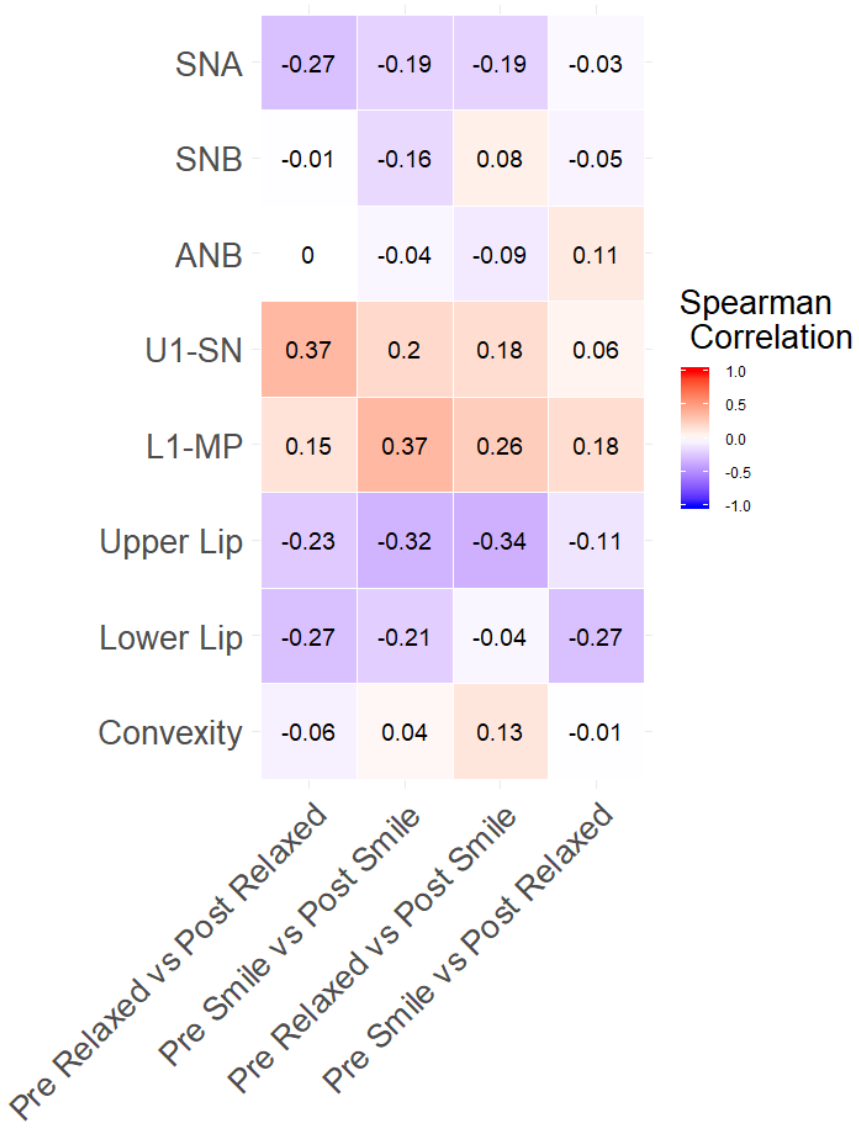
Eight different cephalometric points were traced on the pre- and post-surgical lateral cephalograms then subtracted to quantify the surgical movements. The metrics of interest were: Sella - Nasion - A point (*SNA*), Sella - Nasion - B point (*SNB*), A point - Nasion - B point (*ANB*), Maxillary Incisor to Sella – Nasion (*U1-SN*), Mandibular Incisor to Mandibular Point (*L1-MP*), *Upper lip point*, *Lower lip point*, and *Facial Convexity*. The intraclass correlation coefficient (ICC) for the 10 randomly-selected, retraced cephalograms was 0.998, indicating very strong calibration of the measures. The differences in angular measurements were tested against the null hypothesis of no difference using Wilcoxon Signed Rank Tests (Table 4). It was found that the differences for both *SNA* and the *lower lip* measurements were significantly different

from 0 at the 0.05 level. We then evaluated the relationship between the absolute value of the difference in each cephalometric measure and the facial expression comparisons (Figure 2). Using Spearman's Rank correlation, we found that most absolute difference measures were negatively correlated across each facial comparison. However, the *U1-SN* and *L1-MP* absolute difference measures were positively correlated with all facial comparison scores. none of the correlations were found to be significantly different from 0 at the 0.05 level.

Table 4. Results from univariate Wilcoxon Ranks Sum tests for each of the differences in cephalometric measures. The null hypothesis was that the differences between measures before and after surgery were 0.

Cephalometric Measure	Median Angular Difference (Range)	P-Value
SNA	1.8 (0.2-5.1)	0.002
SNB	0.7 (-0.3-4.0)	0.076
ANB	0.0 (-1.3-3.1)	0.399
U1-SN	-0.5 (-7.0-4.1)	0.258
L1-MP	-0.1 (-3.9-4.0)	0.808
Upper lip	0.7 (-0.3-3.3)	0.065
Lower lip	-0.3 (-1.8-0.6)	0.048
Convexity	1.8 (-2.6-5.3)	0.277

Figure 2. Spearman Rank Correlation Coefficients between the differences in cephalometric measures in absolute value with the facial expression comparison scores. Negative correlations (blue), positive correlations (red). None of the correlations were significant



Univariate comparison for each categorical patient characteristic – surgerized jaw(s), gender, surgery type, and ethnicity – to each of the 4 recognition scores – Pre-surgical Relaxed vs Post-surgical Relaxed, Pre-surgical Smiling vs Post-surgical Smiling, Pre-surgical Relaxed vs Post-surgical Smiling, and Pre-surgical Smiling vs Post-surgical Relaxed – was performed using nonparametric Kruskal-Wallis Rank Sum Tests. Table 5

presents the results from the Kruskal-Wallis tests for each recognition score. The jaw in which the surgery was performed was found to be significantly associated with the recognition score when comparing relaxed expressions before and after surgery ($p = 0.013$). The type of surgery was also found to be significantly associated with this specific recognition score ($p = 0.006$). None of the patient characteristics were significantly associated with the smiling expression score before and after surgery, at the 0.05 level. However, both jaw and surgery type were significantly associated with the recognition score for the Pre-surgical Relaxed versus Post-surgical Smiling comparison ($p = 0.041$ and $p = 0.040$ respectively). When comparing patient characteristics to the recognition score for the Pre-surgical Smiling versus Post-surgical Relaxed test set, it was found that none of the characteristics were significantly associated with the score.

Table 5. Results from the Kruskal-Wallis Rank Sum tests.

	Frontal Facial Agreement Scores							
	Relaxed: Pre vs Post		**Smiling: Pre vs Post**		**Relaxed Pre vs Smiling Post**		**Smiling Pre vs Relaxed Post**	
Variable	**Kruskal-Wallis Chi-Squared Value**	**P-Value**	**Kruskal-Wallis Chi-Squared Value**	**P-Value**	**Kruskal-Wallis Chi-Squared Value**	**P-Value**	**Kruskal-Wallis Chi-Squared Value**	**P-Value**
Jaw	8.687	0.013	2.518	0.284	6.417	0.041	4.154	0.125
Gender	0.0197	0.888	0.607	0.436	0.17	0.68	0.019	0.891
Surgery Type	10.333	0.006	4.025	0.134	6.42	0.040	3.878	0.144
Ethnicity	3.504	0.173	0.068	0.967	0.32	0.852	1.859	0.397

Table 6 presents the results of multiple comparison Wilcoxon Rank Sum Tests for the two recognition scores that yielded significant results for both jaw and surgery type: Pre-surgical Relaxed vs Post-surgical Relaxed and Pre-surgical Relaxed vs Post-

surgical Smiling. For the Pre-surgical Relaxed vs Post-surgical Relaxed comparison, it was found that patients that had surgery performed on both jaws had significantly different scores than those that had surgery performed on only the mandible (p = 0.014). Specifically, the median score for patients that received surgery on both jaws was 97 (IQR = 96-98), while those that had surgery on only the mandible had a median score of 99 (IQR = 99-99). Likewise, patients that received both types of surgery had significantly different scores compared to those that received the BSSO only (p = 0.009). Patients that received both surgical procedures had a median score of 97 (IQR = 96-98), while patients with the BSSO only had a median score of 99 (IQR = 99-99). Similarly, for the score comparing relaxed expressions before surgery versus smiling expressions after surgery, it was again found that patients that received the surgery on both jaws had significantly different scores from those that had surgery on only the mandible (p = 0.028; M = 95 (IQR = 92-96) vs M = 97 (IQR = (97-99), respectively). Finally, patients that received both surgery types were also found to have significantly different scores compared to patients that only received BSSO for this facial comparison (p = 0.036; M = 94.5 (IQR = 91.5-96.3) vs M = 97 (IQR = (97-99), respectively).

Table 6. Multiple comparison Wilcoxon Rank Sum tests for jaw and surgical type patient characteristics. Type I error inflation was adjusted for using the Bonferroni correction procedure.

		Facial Recognition Score	
		Relaxed: Pre vs Post	**Relaxed Pre vs Smiling Post**
Characteristic	**Comparison**	**P-Value**	**P-Value**
Jaw	Both vs Maxillary	0.432	0.795
	Both vs Mandible	0.014	0.028
	Maxillary vs Mandible	0.711	1
Surgical Type	Both vs BSSO	0.009	0.036
	Both vs LeFort	0.158	0.689
	BSSO vs LeFort	0.723	0.876

DISCUSSION

Successful automated facial recognition depends upon accurate pattern recognition during analysis of facial geometries. Face comparison is performed by either testing two images directly against each other or by testing a probe image against a database of face images. Features from the probe image are extracted, and the geometric relationship between the constructed points is compared to an existing database in search of the closest match, with similarity recorded as a percentage.[7] We designed this study to compare similarity scores for pairs of photos known to contain the same person, with the primary changes between photographs being facial pose and/or pre- vs post-surgical status. Validity of the API selected for this study was confirmed through comparison of photographs taken at a single timepoint, varying only facial expression between the two images.

One of the primary difficulties with automated facial recognition is the nonlinear nature of the human face – in other words, the three-dimensional spatial relationships between the various points of interest on the human face. Different filters exist within these programs to process and code images of faces into data amenable to automated comparison. These filters attempt to control for differences in facial expression or ambient light between images; however, the addition of such filters drastically increases the complexity of these programs. A literature review of face recognition discussed the difficulties raised by changes in illumination and pose with respect to automated face analysis. Changes in lighting between images are almost a guarantee when the images are obtained in unconstrained environments – indoor versus outdoor or incandescent versus fluorescent lighting, for example. Changes in pose likewise create difficulty for

automated analysis. Unless the reference gallery is composed of faces photographed from every possible angle, the algorithm running the program must possess the ability to identify a face that may be viewed from a different angle than what is on record. There are some programs sufficiently robust to recognize faces from various three-dimensional viewpoints interpolated from known gallery images; however, the galleries for these sophisticated programs are small.[1] In order to minimize confounding from the known facial recognition challenges of illumination and pose variables, only standard orthodontic records photographs were analyzed in this study. All photographs used in this analysis were taken in the Virginia Commonwealth University Orthodontics Clinic, using the two single-lens reflex (SLR) clinic cameras used for all initial, progress, and final orthodontic records. Additionally, both relaxed and smiling frontal photographs were studied, and comparisons between like facial poses at different timepoints and different facial poses at the same timepoint were tested to assess the ability of the Rekognition program to control for pose when comparing images. All photographs were taken in front of a light box, using standardized camera settings.

In order to calculate similarity scores. Amazon Rekognition generates a “feature vector” for each face, and it is this vector that is used for comparison, not the actual image itself. The Developer Guide for the Rekognition program recommends a 99% similarity threshold when a highly accurate identification is needed.[8] In our study, we found control comparisons to have a mean similarity score of 99%, which is what we would expect, given that the only differences between the two photographs were the change in facial expression and the amount of time needed for the subject to change from a relaxed posed to a smiling pose and for the photograph to be taken. Thus, merely

changing facial expression from relaxed to smiling or vice versa, should not significantly decrease similarity score below the recommended threshold (i.e. 99%) for recognition for situations in which accurate recognition is considered important. When like facial expressions were compared to each other before and after surgery, the median similarity scores were found to be 98% for each expression. When different facial expressions were compared to each other before and after surgery, the median similarity scores dropped to 96%. Given that all pre- versus post-surgical image comparison medians were found to be below the recommended 99% threshold, it is reasonable to conclude that these surgical changes could result in inaccurate automated face recognition outside of the constrained clinic environment as well. However, with technology constantly evolving, it would not be fair to conclude that this identified inadequacy is an insurmountable problem.

Feature vectors and other automated constructs used for face recognition are typically based on the idea that there should not be too much change in facial landmark positions for the same person between two given images, though different facial expressions may temporarily alter the locations of these landmarks. In theory, the greater the number of landmarks used for comparison, the greater the confidence in the comparison itself. Orthognathic surgery, performed on one or both jaws, inherently changes the locations of various facial landmarks and their relationships to each other. A recent study examined 25 consecutive patients with mandibular prognathism, who were treated with either mandibular surgery or bimaxillary surgery. The changes in soft tissue landmarks were measured three-dimensionally using cone-beam computed tomography (CBCT) scans. When comparing changes observed in 1-jaw versus 2-jaw surgeries, soft tissue

changes were found to be more pronounced in the midface of the 2-jaw surgery group than in the 1-jaw surgery group.[9] It is reasonable to assume that such changes in the midface soft tissue could present challenges to a facial recognition algorithm relying on landmark locations for identification. Consistent with these findings, our study found a significant decrease in similarity score between those receiving 2-jaw surgeries and those receiving surgery in the mandible only, though we did not subdivide the group by initial malocclusion, given the relatively small sample size. We found that the 2-jaw group had a median similarity score of 95 (IQR = 92-96), while the mandible only, group had a median similarity score of 97 (IQR = (97-99). Thus, 2-jaw orthognathic surgery seems to pose a significant challenge to automated face recognition, consistent with existing literature reporting more significant soft tissue changes with 2-jaw versus 1-jaw surgeries.

In our study, cephalometric analysis was also performed to determine the effect of the extent of surgical movement on similarity score. We did not find a significant association between changes in the eight cephalometric values measured on pre- and post-surgical lateral cephalograms. Though the pre- and post-surgical cephalometric values did change, as was expected, since all patients underwent surgery, the cephalometric changes were not significantly associated with the similarity scores. One possible explanation for the lack of significant correlation between surgical movements and similarity scores is that ratios of hard tissue to soft tissue changes with orthognathic surgery are known to vary. Three-dimensional analysis has been used to calculate ratios of soft tissue movement to bony movement to improve the quality of surgical predictions. Surgical changes in the maxilla result in significantly lower degrees of soft

tissue change, while mandibular surgical changes tend to correlate more closely with resultant soft tissue changes.[10] Though not statistically significant in this study, most of the absolute value changes in hard and soft tissue cephalometric values were found to negatively correlate with similarity score. We would expect this finding since more surgical change would logically lead to bigger changes in facial appearance. The variability of soft tissue change predictability based on hard tissue changes has been described often in the oral surgery literature[9–11] and could at least partially explain the lack of significance that we found in our study since the API only evaluates facial soft tissues from the frontal view. The only positive correlations with similarity score found in our cephalometric analysis were the upper and lower incisor inclinations. The positive correlations make sense since orthodontic decompensation prior to orthognathic surgery often inclines the incisors in an anteroposterior direction opposite that of the surgical movement of that jaw

Though no other studies were found examining the effect of orthognathic surgery on automated facial recognition, studies have been performed to explore the effect of plastic surgery on facial recognition algorithm analysis. Singh et al. created a 900-person database of individuals who underwent plastic surgery procedures for either reconstructive or cosmetic purposes. Some of the procedures studied included rhinoplasty, brow lift, genioplasty, skin resurfacing, and lip reshaping. Baseline performance of the six facial recognition algorithms used was assessed on a group of 900 subjects who did not undergo plastic surgery from a publicly available face database. The algorithms were then used to test the images of the 900 individuals who had undergone plastic surgery. The six algorithms were shown to function at 18-54%

when tested on the images of those who underwent plastic surgery, far below the threshold of what is acceptable, with performance rates lowest for global procedures, including facelifts and facial resurfacing. They concluded that more research is needed in order to "teach" facial recognition algorithms to recognize *faces* that have had plastic surgery procedures performed and/or to recognize *when* a plastic surgery procedure has been performed.[6] Given our findings, it is fair to extend these conclusions to include orthognathic surgery procedures as well. If these algorithms could recognize certain hallmarks of orthognathic surgery in an individual's face – identification of such "hallmarks" is beyond the scope of the present study – then the algorithm could then analyze the face perhaps with a different similarity threshold in order to increase the likelihood of a correct match. To date, such advanced technology is not widely available.

These demonstrated barriers to successful automated identification following facial surgical procedures brings to light a few considerations. While inaccurate recognition following a facial surgical procedure could pose as an inconvenience for someone relying on facial recognition to unlock their smartphone, more nefarious sequelae could be surmised as well. It has been proposed that the inadequacy of current facial recognition programs could enable intentional surgical alterations of one's face to either deliberately evade facial recognition *or* to impersonate another individual.[6,12] From a national security standpoint, it might seem reasonable to require those who undergo these face-altering procedures to register this activity; however, matters of personal privacy, especially with respect to health-related medical procedures, significantly complicate such a proposal.[6]

With the effects of routine orthognathic surgeries on facial recognition found in this study, it can be concluded that these procedures do in fact pose challenges to automated face recognition, especially when surgery is performed on both jaws. As such, orthodontists and oral and maxillofacial surgeons should consider advising their double-jaw orthognathic patients appropriately. Those patients who make use of biometric passports should be advised to proactively update their information following surgery. Patients who use facial recognition to secure their smartphones should similarly be advised that they may need to re-program their security settings after the surgery (and likely again once complete healing has occurred).

This study serves as an excellent foundation for future research in this area. Strengths of this study include its novel application of an emerging technology that is gaining rapid and widespread use and the standardized nature of the photographs used. Opportunities for further research include use of a larger sample size to permit analysis of subgroups based on initial malocclusion and surgical procedure type, evaluation of multiple facial recognition algorithms, and measurement of facial landmark changes using three-dimensional imaging technology. Use of a three-dimensional facial recognition program, such as the software used in smartphones for biometric face analysis, could potentially provide more information about spatial changes between which facial points pose the biggest challenge to automated face analysis.

CONCLUSIONS

1. Orthognathic surgery can cause enough facial change to significantly reduce similarity scores calculated by an automated face recognition program below the recommended threshold of acceptability
2. Two-jaw surgeries – specifically LeFort I with BSSO – have more significant negative effects on similarity score than one-jaw surgeries
3. Facial pose variation appears to compound the negative impact of orthognathic surgery on similarity score

REFERENCES

1. Abate AF, Nappi M, Riccio D, Sabatino G. 2D and 3D face recognition : A survey. *Pattern Recognit. Lett.* 2007;28:1885–906.

2. Anon. CCTV feeds facial recognition systems for law enforcement. *Biometric Technol. Today* 2015;(4):3.

3. Anon. Three-Dimensional Facial Recognition. *U.S. Dep. Homel. Secur. Syst. Assess. Valid. Emerg. Responders* 2008;(May):1–2.

4. Devue C, Wride A, Grimshaw GM. New Insights on Real-World Human Face Recognition. *J. Exp. Psychol. Gen.* 2018;(Advance online publication).

5. Loon B Van, Heerbeek N Van, Bierenbroodspot F, Verhamme L, Xi T. Three-dimensional changes in nose and upper lip volume after orthognathic surgery. *Int. J. Oral Maxillofac. Surg.* 2015;44:83–9.

6. Singh R, Vatsa M, Bhatt HS, Bharadwaj S, Noore A, Nooreyezdan SS. Plastic surgery: A new dimension to face recognition. *IEEE Trans. Inf. Forensics Secur.* 2010;5(3):441–8.

7. Kumar A, Narain Y. Evaluation of Face Recognition Methods in Unconstrained Environments. *Procedia - Procedia Comput. Sci.* 2015;48:644–51. Available at: http://dx.doi.org/10.1016/j.procs.2015.04.147.

8. Amazon Web Services. *Amazon Rekognition Developer Guide*.; 2019.

9. Kim B, Oh K, Cevidanes LHS, et al. Analysis of 3D Soft Tissue Changes After 1- and 2-Jaw Orthognathic Surgery in Mandibular Prognathism Patients. *J. Oral Maxillofac. Surg.* 2013:151–61.

10. Ruellas CO, Yatabe MS, Westgate PM, Cevidanes LHS, Huja SS. Soft Tissue Changes Measured With Three-Dimensional Software Provides New Insights for Surgical Predictions. *J Neurol Surg Rep* 2017;75:2191–201.

11. Rupperti S, Winterhalder P, Rudzki I, Mast G, Holberg C. Changes in the facial soft-tissue profile after mandibular orthognathic surgery. *Clin. Oral Investig.* 2018:15–20.

12. Nappi M, Ricciardi S, Tistarelli M. Deceiving faces: When plastic surgery challenges face recognition. *Image Vis. Comput.* 2016;54:71–82. Available at: http://dx.doi.org/10.1016/j.imavis.2016.08.012.