medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Machine Learning Methods to Track Dynamic Facial Function in Facial Palsy

Akshita A. Rao[1], Jacqueline J. Greene[2], and Todd P. Coleman*[1]

[1] Stanford University, Department of Bioengineering, Stanford, CA, USA
[2] University of California San Diego, Department of Otolaryngology – Head & Neck Surgery, San Diego, CA, USA
*Correspondence: Todd P. Coleman, toddcol@stanford.edu

Abstract

For patients with facial paralysis, the wait for return of facial function and the resulting vision risk from poor eye closure, difficulty speaking and eating from flaccid oral sphincter muscles, as well as the psychological morbidity from the inability to smile or express emotions through facial movement can be devastating. There are limited methods to assess ongoing facial nerve regeneration: clinicians rely on subjective descriptions, imprecise scales, and static photographs to evaluate facial functional recovery and thus facial nerve regeneration remains poorly understood. We propose a more precise evaluation of dynamic facial function through video-based machine learning analysis which would facilitate a better understanding of the sometimes subtle onset of facial nerve recovery and improve guidance for facial reanimation surgery. Specifically, we here present machine learning methods employing likelihood ratio tests, optimal transport theory, and Mahalanobis distances to: 1) assess the use of defined facial landmarks for binary classification of different types of facial palsy; 2) identify regions of asymmetry and potential paralysis during specific facial cues; and 3) determining severity of abnormal facial function when compared to a reference class of normal facial function. Our work presents promising results of utilizing videos, rather than static photographs, to provide robust quantitative analyses of dynamic properties for various facial movements without requiring manual assessment. The long-term potential of this project is to enable clinicians to have more accurate and timely information to make decisions for facial reanimation surgery which will have drastic consequences on quality of life for affected patients.

I. Introduction

There are an estimated over 500,000 cases of peripheral nerve transection requiring repair annually in the United States [1] and peripheral nerve injuries cost the US healthcare system $150 billion annually [2]. Despite this, the current understanding of the spectrum of nerve injury and regenerative outcomes is based on clinical observational data alone with unfortunately little objective data to guide surgical intervention [3]. Acquired facial paralysis occurs in over 150,000 Americans every year from a variety of causes: Bell's Palsy, benign or malignant tumors, trauma, infections and neurologic and iatrogenic injuries [4].

While facial nerve (FN) recovery is robust in Bell's Palsy and should be managed conservatively, for many other conditions, the prognosis is much more uncertain and may take months to years to complete. In such cases, a critical time window for successful facial reanimation surgery is passing (typically after 18-24 months of facial paralysis muscle reinnervation through a nerve transfer is unlikely to be successful due to muscle degeneration and atrophy [5]). As detecting the onset of nerve regeneration and muscle reinnervation can be subtle and difficult to assess clinically, determining prognosis for FN recovery represents a black-box scenario with limited data to guide clinicians and patients. In the setting of no overt visible facial movement, the choice to proceed with facial reanimation surgery (such as a cranial nerve V-VII transfer) is fraught with difficulty [6] [7] [8] [9].

There are limited objective methods to quantify FN function; clinicians rely on subjective descriptions, imprecise scales, and static photographs to evaluate facial functional recovery [10] [11] [12] [13] [14]. There are no current imaging modalities to detect the onset, progression (or failure) of nerve regeneration [3]. The lack of objective FN outcomes data and lengthy recovery times has hampered the field of facial reanimation surgery for some time [15]; for example, only 24% of patients requiring FN sacrifice during parotid tumor excision receive FN repair [16]. For patients with facial paralysis, the wait for return of facial function and the resulting vision risk from poor eye closure, difficulty speaking and eating from flaccid oral sphincter muscles, as well as the psychological morbidity from the inability to smile or express emotions through facial movement, can be devastating.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A. Current Methods to Quantify and Track Facial Nerve Function

The current methods for quantifying facial nerve (FN) function primarily include clinical grading systems such as House-Brackmann (HB) score and the Sunnybrook Facial Grading System, and collection of facial electrophysiology signals. While these tools have been widely used, they do have limitations in capturing the full spectrum of FN palsy.

House-Brackmann (HB) Score: The HB system was approved by the American Academy of Otolaryngology-Head and Neck Surgery FN dysfunction committee as the reference standard for grading facial palsy [17]. The system uses a six-point scale, where grade I corresponds to normal and grade VI to complete flaccid paralysis. However, it relies on subjective clinical judgment, leading to potential interobserver variability. The discrete grading may not capture subtle changes, making it less sensitive to nuanced improvements or deteriorations.

Sunnybrook Facial Grading System: The Sunnybrook facial grading system comprises a regional scale involving facial symmetry at rest, voluntary movements, and synkinesis, or unwanted contractions of muscles during attempted movement. The composite score ranges from 0 to 100, where 100 corresponds to normal facial function and 0 corresponds to complete paralysis. The Sunnybrook classification system has been reported to assess facial synkinesis, involuntary movements, better than the HB systems [18]. Like the HB score, it involves subjective clinical assessment, and the scores may not fully reflect patients' perceptions of their facial function.

Electrophysiological methods: Objective measures that collect facial electrophysiology, such as needle or evoked electromyography (EMG), are invasive, painful, and may cause patient discomfort. EMG measurements vary based on practitioner experience by up to 20% [19], are logistically difficult to arrange and track over time, may be impacted by edema in post-surgical or trauma patients, and do not inform prognosis of recovery in if complete facial paralysis is present (HB VI). Furthermore, EMG may be effective in detecting significant changes in facial muscle activity, but there have not yet been studies that show it is sensitive enough to track subtle or gradual onset of facial palsy. Therefore, electrophysiological methods are less suitable for monitoring early stages of FN dysfunction or recovery.

B. Applications of Machine Learning in Analysis of Facial Palsy

Machine learning (ML) algorithms for facial analysis have advanced significantly in the past few decades [20] [21] but are based on training datasets of intact facial function. Application of these tools to facial paralysis patients photographs was reported to cause significant landmark inaccuracy that precludes clinical usage [22] [23]. The most advanced method of automated facial analysis in facial paralysis patients, Emotrics, is limited to static photos only and requires significant manual landmark adjustments [23]. Other attempts to quantify facial function have incorporated proprietary marketing software to estimate facial emotional expressions [24] [25] [26] [27] or were limited to a single surgical case report [28] following surgical interventions. An objective, open-source, rigorous quantification of dynamic facial function in videos has remained elusive [28] [29] [30], as there are multiple challenges to accumulating a sufficiently large training dataset to improve landmark accuracy in the facial paralysis population.

There is an unmet need to develop a more precise evaluation of dynamic facial function to facilitate a better understanding of the spectrum of facial palsy, improve guidance for and outcomes following facial reanimation surgery, and potentially detect the subtle onset of FN recovery (for which there currently is no reliable test [15]). Contrary to contemporary beliefs regarding the limitations of ML algorithms, the latest open-source computer vision and ML algorithms available through Python (OpenCV, dlib) [21] [31] on facial palsy patients recently revealed surprisingly robust landmark accuracy despite significant facial asymmetry [32]. Specifically, a recent finding demonstrated a significant reduction in landmark error rate in videos compared to photos in a standardized data set across all facial palsy severity types, in part because the volume of data available from a video of a standard FN exam is orders of magnitude larger than static photography (from 8 photos to 3600 frames for a 60 second video recorded at 60fps) [32].

Building upon recent accomplishments in reducing landmark error rates in videos compared to photos, we here present machine learning methods employing likelihood ratio tests, optimal transport theory, and Mahalanobis distances to: 1) assess the use of defined facial landmarks for binary classification of different types of facial palsy; 2) identify regions of asymmetry and potential paralysis during specific facial cues; and 3) determining severity of

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

abnormal facial function when compared to a reference class of normal facial function. The paper concludes with a discussion of the results and future directions of the presented work.

II. Materials and Methods

A. Data Collection

Videos of normal subjects and patients clinically diagnosed with facial palsy were prospectively gathered. The Massachusetts Eye and Ear Infirmary (MEEI) Facial Palsy Photo and Video Standard Set [33] consists of videos from 50 subjects and the University of California, San Diego (UCSD) Facial Nerve Database consists of videos from 15 patients clinically diagnosed with facial palsy. Thus, a total of 65 patients diagnosed with having abnormal facial function, from the combination of these two datasets, was used in this study. Additionally, videos of 50 healthy controls were gathered at Stanford University. Here, we provide a brief description of the two databases used, collection of healthy control data, experimental setup, and tasks. Figure 1a summarizes the data used in this study.

MEEI Facial Palsy Photo and Video Standard Set: This dataset is an open-source, hosted on the Sir Charles Bell Society website, standardized set of facial photographs and videos that captures the spectrum of flaccid and nonflaccid facial palsy [33]. Normal intake protocol was followed, including a clinician assessment of facial function (eFACE) as well as a full set of photographs and a video. Exclusion criteria included prior facial reanimation surgery or extensive facial scarring, currently active chemodenervation, and bilateral facial palsy. Cases from both flaccid and nonflaccid (aberrantly regenerated or synkinetic) states were included resulting in 25 flaccid palsy subjects and 25 synkinetic palsy subjects.

Subjects who consented to enroll in the standard set were categorized by their eFACE score into the following: normal, near-normal, mild, moderate, severe and complete flaccid or nonflaccid facial palsy. The degree of palsy was quantified using eFACE, House-Brackmann (HB), and Sunnybrook scales by two expert clinicians in the field. The fifty facial palsy subjects from this database were defined as having abnormal facial function and the remaining 10 subjects defined as having normal facial function for this study.

UCSD Facial Nerve Database: IRB approval was obtained from the Office of IRB Administration at UCSD prior to beginning this study. Patients who have facial palsy greater than a HB score of 1 are currently being recruited from the UCSD Facial Nerve Clinic. A total of 15 subjects with known facial palsy were used in this study, of which 11 have flaccid palsy and 4 have synkinetic palsy.

Healthy Controls: A total of 50 healthy controls were used in this study. Ten healthy controls (5 male and 5 female) that were included in the MEEI Facial Palsy Photo and Video Standard Set were used. The remaining 40 individuals (20 male and 20 female) consists of healthy volunteers, who have no facial movement disorders, recruited.

The Facial Nerve Exam: All the subjects included in this study performed the same standard facial nerve (FN) exam, in which the subject was requested to perform the following facial cue for a few seconds before returning back to face at rest.

1) Face at rest: used as baseline
2) Raise their eyebrows
3) Gently close their eyes
4) Forcefully close their eyes
5) Gentle smile
6) Full effort smile
7) Pucker their lip
8) Show their bottom teeth

Figure 1b presents the intended facial muscle activation for each of the given cues. The eyebrow raise cue aims to target frontalis muscle activation, the eye closure cue targets orbicularis oculi activation, the smile cue targets zygomaticus major and minor, levator anguli oris, and risorius activation, the lip pucker cue targets orbicularis oris activation, and the bottom teeth cue targets depressor inferioris and depressor anguli oris activation.

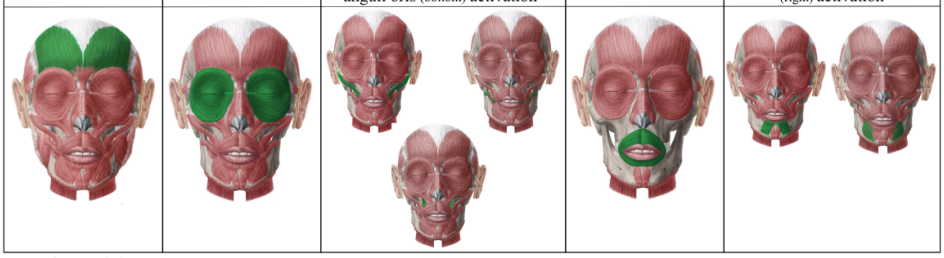
medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

a.

Category	eFACE	Palsy Type	# of Subjects	Database	Total Count
Normal	96-100	none	10	MEEI	Total Normal = 50 subjects
			40		
Palsy	<96	Flaccid	11	UCSD	Total Flaccid = 36 subjects
		Synkinetic	4		
Near-Normal Palsy	91-95	Flaccid	5	MEEI	Total Synkinetic = 29 subjects
		Synkinetic	5		
Moderate Palsy	80-90	Flaccid	5	MEEI	Total Abnormal = 65 subjects
		Synkinetic	5		
Severe Palsy	70-79	Flaccid	5	MEEI	
		Synkinetic	5		
Complete Palsy	<60	Flaccid	5	MEEI	
		Synkinetic	5		

b.

Eyebrow raise: Frontalis muscle activation	Eye closure: Orbicularis oculi activation	Smile: Zygomaticus major & minor (top left), Risorius (top right), Levator anguli oris (bottom) activation	Lip pucker: Orbicularis oris activation	Show bottom teeth: Depressor labii inferioris (left), Depressor anguli oris (right) activation



Images from Kenhub Anatomy.

Fig. 1. a: Summary of all the subjects (n=115) used in this study. Information includes the eFACE score assigned to them (if applicable), whether their palsy was flaccid or synkinetic (if applicable), and which database they belong to. b: The five facial cues given during the facial nerve (FN) exam and their intended facial muscle activation, where the region of interest is highlighted in dark green.

B. Data Processing

Facial Landmarks Extraction: The process to extract 38 facial landmarks from a face within one input image is given as follows [34]:

1) Extract the number of frames in the input video.
2) For each frame in the video:
   a) Convert the color input image to gray scale.
   b) Detect the face on the image using the open-source and publicly available *dlib* and *OpenCV* Python libraries.
   c) Store the coordinates in the selected frame.
   d) Save coordinates for all the frames in the video as a .csv file for future data processing.

Note that the MEE shape predictor is trained to detect 68 points, but only 38 of them were used in this work. These 38 points were reorganized, as seen in Figure 2, to proceed with the computational measures.

Segmenting Data for Specific Cues: The time points of when each FN exam cue started and ended in a subject's video was observed and recorded. These time points were converted to specific frame numbers based on the frames per second of the inputted video. With these frame points indicating the start and end of each cue, the landmark coordinates could be segmented as such and stored as a Python dataframe for future processing.

Pre-Processing Steps: In order to account for camera drift:

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

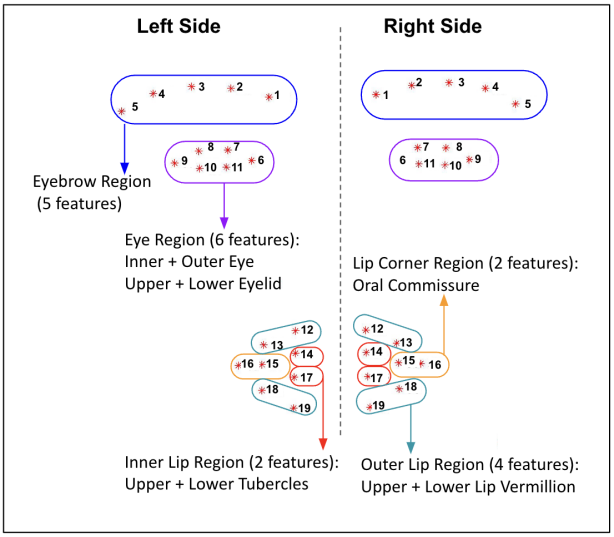


Fig. 2. Diagram of the facial features that were extracted from each side of the patient's face during the recorded FN exam. Difference between the x- or y-coordinate from each side of the face for each feature was calculated after pre-processing steps were completed.

1) The x and y coordinates of each facial landmark was subtracted from the average coordinates during the face at rest cue.
2) The x and y coordinates of every facial landmark was then subtracted from the midpoint of the subject's face as identified by dlib.

C. Implementing supervised learning for hypothesis testing

Feature selection: In order to choose features that can probe aspects of facial symmetry, we extracted the x,y positional coordinates from the 38 landmarks and computed their difference relative to the midline of the face. An analogous procedure was done for the left and right y-position coordinates. Therefore, each patient would have a $\underline{Y} \in \mathbb{R}^{n \times d}$, where n = 38 and d represents the number of frames in the patient video. We observe statistically significant differences ($p < 0.001$) in some key facial regions, such as the lip corner (oral commissure) and outer and inner lip features, between the average y-position difference of the abnormal patient class and the healthy controls, as shown in Figure 3a.

A similar process can be applied to the landmark coordinate data segmented at selected cues, resulting in $\underline{Y}_{smile} \in \mathbb{R}^{n \times d_{smile}}, \ldots, \underline{Y}_{eyeclose} \in \mathbb{R}^{n \times d_{eyeclose}}$. During each cue, we observe statistically significant differences in varying facial regions, with highest significance in context-specific regions of interest. For example, during the eyebrow raise cue, there is a significant difference between the two subject groups for the eyebrow features ($p < 0.01$) and upper eye features ($p < 0.05$) (Figure 3b). Whereas, during the smile cue, there is a significant difference between the patient groups for the outer and inner lip features ($p < 0.001$) and lip corner (oral commissure) features ($p < 0.001$) (Figure 3c).

1) Setup for binary classification: Our setup for binary classification can be described as having labeled data ($\underline{Y}^{(1)}$,..., $\underline{Y}^{(k)}$), which are our arbitrary feature vectors for every subject in $\mathbb{R}^k$, and $(H_1, \ldots, H_k)$ such that $H_i \in \{0, 1\}$, where 0 is our label for normal facial function and 1 for abnormal. The goal of binary classification is to find a decision function that allows us to use the input features $\underline{Y}^{(i)}$ that provide the corresponding class labels H_i. This can be viewed as a hypothesis testing problem, where it is assumed that under H_0, $\underline{Y}$ has a joint density $f_{\underline{Y}}(\underline{y}; \theta_0)$ and under H_1, $\underline{Y}$ has a joint density $f_{\underline{Y}}(\underline{y}; \theta_1)$.

Both hypothesis classes were modeled as a multivariate Gaussian distribution, with mean $\underline{\mu}$ and covariance matrix Σ, namely $\underline{Y} \sim \mathcal{N}(\underline{\mu}, \Sigma)$. Then, note that its density is given by

$$f_{\underline{Y}}(\underline{y}; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right) \tag{1}$$

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

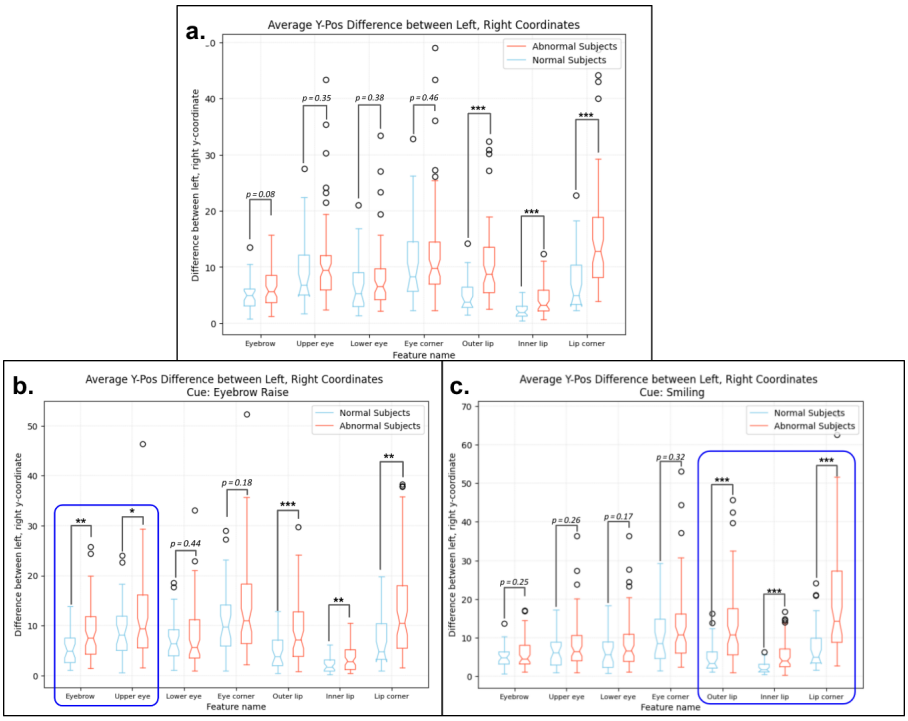


Fig. 3. a: Average difference of y-coordinates on the left and right side of the face for all the abnormal patients and normal subjects over the entire course of the FN exam. b: Average difference of y-coordinates on the left and right side of the face for all abnormal and normal subjects during the eyebrow raise cue. There is a significant difference between the two patient groups for the eyebrow features ($p < 0.01$) and upper eye features ($p < 0.05$). c: Average difference of y-coordinates on the left and right side of the face for all abnormal and normal subjects during the smile cue. There is a significant difference between the two patient groups for the outer and inner lip features ($p < 0.001$) and lip corner (oral commissure) features ($p < 0.001$).

We can utilize supervised learning to partition the samples for each class, where we use 70% of the dataset for training and the remaining 30% for testing. The training dataset has the labels for their associated classes and so we can calculate distribution parameters, θ_i, for every training sample under H_0 and under H_1.

2) Clustering probability distributions using Wasserstein barycenters: Optimal transport theory is a geometrically meaningful way of measuring distances between probability distributions, and it has recently become important in applications in data science and machine learning [35]. In this work, we utilize the concept of barycenters to find the "center of mass" of a collection of probability measures, where distances between each are determined by the optimal transport Wasserstein distance.

Consider a set $P_1, \ldots, P_k$ of elements, with their associated weights $\lambda_1, \ldots, \lambda_k$, satisfying $\lambda_i > 0$ and $\sum_{i=1}^{k} \lambda_i = 1$, belonging to a metric space M. The barycenter P^* on this metric space (M, d_M) is defined as [36]:

$$P^* := \sum_{i=1}^{k} \lambda_i d_M^2(P_i, P^*) = \min\left\{\sum_{i=1}^{k} \lambda_i d_M^2(P_i, P), P \in M\right\} \tag{2}$$

If we were to consider the set $P_1, \ldots, P_k$ of elements to be probability distributions, where the metric space (M, d_M) corresponds to the 2-Wasserstein distance, then P^* becomes the Wasserstein barycenter [37]. when $P_1, \ldots, P_k$ are all multivariate Gaussian distributions with parameters (μ_i, Σ_i), then the mean of the Wasserstein barycenter is simply $\mu^* = \sum_{i=1}^{k} \lambda_i \mu_i$ and the covariance Σ^* is the only positive definite matrix Σ satisfying the equation [38]:

$$\Sigma = \sum_{i=1}^{k} \lambda_i (\Sigma^{1/2} \Sigma_i \Sigma^{1/2})^{1/2} \tag{3}$$

Using the probability distributions of each training samples in either hypothesis class, the Wasserstein barycenter was calculated to define distribution parameters, $\hat{\theta}_0$ and $\hat{\theta}_1$, with $\lambda_j \frac{1}{k}$ for each distribution. Then, with the test

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

dataset, where the label of the subject was unknown, a likelihood ratio test was performed on each sample, using the learned parameters, $\hat{\theta}_0$ and $\hat{\theta}_1$ from the training dataset. A likelihood ratio test was performed under the two Gaussian Wasserstein barycenters:

$$\log \frac{f_{\underline{Y}}(\underline{y};\theta_1)}{f_{\underline{Y}}(\underline{y};\theta_0)} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\lessgtr}} \tau. \tag{4}$$

This supervised learning method for hypothesis testing was used to classify between: 1) normal and abnormal subjects, 2) normal and synkinetic subjects, and 3) normal and flaccid subjects.

D. Identifying regions of dynamic facial asymmetry using specific facial cues

1) Pearson correlation coefficient to define facial symmetry: Pearson correlation coefficients are well-suited for capturing linear relationships between facial landmarks. The coefficients quantify the degree of linear dependence between two variables, making them effective in identifying proportional changes in facial features during dynamic expressions. Furthermore, the magnitude of the Pearson coefficient reflects the strength of the correlation, allowing for a quantitative assessment of how facial features move in tandem. The sign of the coefficient indicates the direction of the relationship, where a positive correlation implies a simultaneous increase or decrease in facial features, and a negative correlation suggests an inverse relationship.

From the 38 facial landmarks described in Figure 2 Pearson correlation coefficients were calculated for each of the corresponding right and left landmark, resulting in a coefficient vector, $\underline{v} \in \mathbb{R}^{19}$. These coefficient vectors can either be calculated to capture correlation across features for the entire duration of the FN exam or for the duration of a specific facial cue that was executed during the exam.

2) Mahalanobis distance as a metric to determine context specific asymmetry: To compute how much a patient deviates from normal facial function, the normalized Mahalanobis distance (MD) were calculated for each FN exam cue. This metric identifies which of the cues has the most significant dynamic asymmetry when compared to normal function and gives insight into the specific region and context, or facial cue, of asymmetry.

Though seven cues were provided in the FN exam, the two eye closure cues and the two smile cues were grouped together. Therefore, the five facial cues of interest include eyebrow raise, eye closure, smile, lip pucker, and showing bottom teeth. For each facial cue, the MD was used to measure the degree of error from normal facial function based on the 19 Pearson coefficients of the symmetric facial landmarks. The MD takes into account the variability and correlation structure of the data, making it suitable for multivariate analysis, since the Pearson coefficient vectors, $\underline{v} \in \mathbb{R}^{19}$. The following is how MD was computed for each cue:

1) The Pearson correlation coefficient vectors was calculated for all 50 normal subjects, $\underline{V} \in \mathbb{R}^{19\times 50}$
2) The mean of the Pearson coefficients for all normal subjects was calculated, $\underline{\mu} \in \mathbb{R}^{19}$. This represents the average correlation values across all facial regions.
3) The covariance matrix of the Pearson coefficients for all normal subjects was determined, $\Sigma \in \mathbb{R}^{19\times 19}$. This describes the relationships and variability among the different facial regions.
4) For the patient of interest, the MD was calculated at each of the 5 cues:

$$D = \sqrt{(\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu})} \tag{5}$$

   where $\underline{X}$ is the vector of Pearson coefficients for the patient of interest, $\underline{\mu}$ is the mean vector of all the normal subjects, and Σ is the covariance matrix of all the normal subjects.
5) Once all the distances were calculated for every cue $\overline{D} = [D_{eyebrow-raise}, D_{smile}, \ldots, D_{pucker}]$, the five distances were normalized allowing us to determine which cue resulted in maximum deviation from normal function.

The MD will provide a measure of how far away the abnormal patient's data is from the average normal facial symmetry, considering the multivariate distribution of the normal subjects Pearson coefficients. A higher MD indicates that the abnormal patient's facial function is more divergent from the normal range.

E. Using multivariate outlier detection to identify abnormal facial function

1) Wasserstein barycenter distance as metric to define abnormal facial function: The Wasserstein distance quantifies the minimum "work" required to transform one probability distribution into another. As it defines a metric space

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

within the space of probability measures, it can be applied to compare the distribution of facial feature positions during expressions in individuals with varying degrees of facial palsy to a reference distribution representing normal facial function. To calculate the Wasserstein distance of one multivariate Gaussian distribution to another Gaussian, $d := W_2(\mathcal{N}(\underline{\mu}_1, \Sigma_1); \mathcal{N}(\underline{\mu}_2, \Sigma_2))$, the following closed-form equation can be used [35, eq 2.41]:

$$d^2 = ||\underline{\mu}_1 - \underline{\mu}_2||_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}). \tag{6}$$

To extend the notion of the MD that was used for the $\underline{v} \in \mathbb{R}^{19}$ of Pearson correlation coefficients to the patient specific y-position coordinates, Wasserstein distances were calculated. This distance to the cue-specific barycenter of all healthy controls was calculated for each subject in this study (n = 115):

1) The cue in which a subject deviated most from the average normal function was identified by using normalized MDs. The cue with the largest MD was identified as having maximum deviation.
2) For this cue, the Wasserstein barycenter that summarizes the Gaussian distribution for all healthy controls was calculated. See (2), (3).
3) The Wasserstein distance between the individual subject's Gaussian distribution at the cue and the Wasserstein barycenter for all normal subjects was calculated.

This distance serves as a representative measure of the dissimilarity between an individual's facial movements and the typical facial expressions of the healthy population. We hypothesized that patients with abnormal facial function would have noticeably larger Wasserstein distances from the averaged normal barycenter at that cue. As such, the Wasserstein barycenter distance can be useful for outlier detection because it provides a measure of how far a particular distribution is from the central tendency represented by the barycenter.

To account for outliers in the healthy controls, Wasserstein distances that were above two standard deviations from the mean were not used for the following linear regression models. As such, seven healthy controls were removed for the model selections to follow.

2) Linear regression to predict clinical scores of facial palsy: Linear regression can be employed as a valuable tool for predicting clinical scores related to facial palsy, such as the HB and Sunnybrook scales. In this context, the Wasserstein distance from the barycenter of normal facial function serves as a unique and innovative feature that captures the dissimilarity between an individual's facial movements and a reference distribution of normal facial expressions. Both the MEEI Facial Palsy Photo and Video Standard Set dataset and the UCSD Facial Nerve Database includes clinical scores, specifically the HB score, for individuals with facial palsy. The healthy controls were assigned a HB score of I since they exhibit normal facial function in all areas. As such, the calculated Wasserstein distances were used as the independent variable and the clinical scores are the dependent variable.

III. Results

A. Performance of binary classification of facial palsy

The receiver operating characteristic (ROC) curve offers one way to measure effectiveness of predicting by calculating the area under the curve (AUC), which can be interpreted as the probability that the test result from a randomly chosen abnormal individual is more indicative of paralysis than that from a randomly chosen normal individual [39]. To determine the reliability and performance of the classification, bootstrapping (n = *500)* was used to provide a range of ROC curves, helping us understand how robust the model evaluation is to variations in the dataset. The average AUC, with a 95% confidence interval, was calculated for the three binary classification scenarios described above. Figure 4 presents the ROC curves along with the average AUC values for each scenario: 1) normal and abnormal subjects: 0.9262, 2) normal and synkinetic subjects: 0.9192 and 3) normal and flaccid subjects: 0.9419.

B. Patient-specific assessment of dynamic facial asymmetry is more robust during specific contexts

The Pearson coefficients were calculated at the five facial cues of interest for all 50 healthy controls, and the average values of all 19 coefficients that correspond to the y-position of the landmarks in Figure 2 are presented in Figure 5a. It is observed that all 19 features reveal high positive correlations between the right and left side y-coordinates during the full duration of the FN exam. Heatmaps were created to spatially localize levels of synchrony for the 19 pairs of facial landmarks across the face for each of the five cues executed. The average Pearson coefficients for all healthy controls during the five cues are presented in Figure 5b, where dark blue corresponds to a positive correlation of 1 and dark red corresponds to a negative correlation of -1. During all the cues, there are strong

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

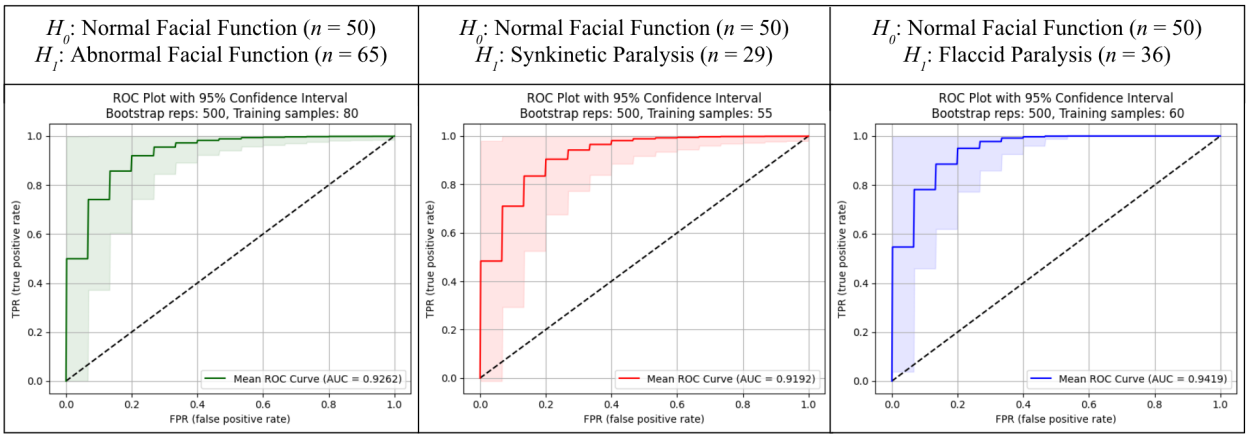


Fig. 4. Presents ROC curves of 500 bootstrap samples for the log-likelihood ratio test used to perform binary classification between (left) abnormal (*n=65*) and normal (*n=50*) subjects, where the mean ROC AUC was 0.9262, (middle) synkinetic (*n=29*) and normal subjects, where the mean ROC AUC was 0.9192, and (right) flaccid (*n=36*) and normal subjects, the mean ROC AUC was 0.9419.

positive correlations, above 0.90 for each cue's regions of interest: $r_{y-eyebrow} = 0.911 \pm 0.052$ during the eyebrow cue, $r_{y-eyeclosure} = 0.952 \pm 0.022$ during the eye closure cue, $r_{y-lipregion} = 0.934 \pm 0.056$ during the smile cue, $r_{y-lipregion} = 0.900 \pm 0.043$ during the lip pucker cue, and $r_{y-lipregion} = 0.911 \pm 0.066$ during the show bottom teeth cue.

Figure 6a presents four sample subjects from the MEEI Facial Palsy Photo and Video Standard Set with varying types of facial palsy and their corresponding heatmaps during specific FN exam cues (please view section V regarding images of patients in figures). For example, the complete synkinetic palsy patient (SP1) (left in Figure 6a) presents palsy in the eyebrow region and upper lip vermilion and oral commissure, and this is captured in both the magnitude and direction of the Pearson coefficients in these regions: $r_{y-eyebrow} = -0.280 \pm 0.187$, $r_{y-uppervermilion} = -0.228 \pm 0.327$, and $r_{y-oralcommissure} = 0.55$. Similarly, the complete flaccid palsy patient (FP1) (second from right in Figure 6a) presents palsy in the oral commissure and upper lip vermilion, and this is captured in the Pearson coefficients of these regions: $r_{y-oralcommissure} = -0.54$ and $r_{y-uppervermilion} = -0.525 \pm 0.009$.

When comparing the above results to the Pearson coefficients of the whole time series, rather than specific cues, we lose some of the spatial specificity and sensitivity of asymmetry but the algorithm can still identify key regions of abnormal function. Figure 6b presents the heatmaps of the same two patients described above but for the full duration of the FN exam. Over the whole time course, SP1 has the following Pearson coefficients for the initially identified regions of asymmetry: $r_{y-eyebrow} = 0.032 \pm 0.0488$, $r_{y-uppervermilion} = 0.685 \pm 0.007$, and $r_{y-oralcommissure} = 0.67$. While FP1 has the following Pearson coefficients for the initially identified regions of asymmetry: $r_{y-oralcommissure} = -0.58$ and $r_{y-uppervermilion} = -0.515 \pm 0.077$.

In the case of FP1 the regions of asymmetry are apparent during the duration of the whole FN exam, but as evident in SP1, there are specific contexts or facial cues in which their abnormal motor movement is most evident. To better determine under which context a patient's palsy is most noticeable in, the MD was used to measure the degree of error or how off a patient is from normal facial function. Figure 6c presents a patient with moderate flaccid palsy (FP2) performing all the five facial cues, with the normalized MD values for each 19 symmetric landmarks spatially plotted on top. It can be observed that in some cues, such as the smile and lip pucker cue, FP2's palsy is not as evident as it is during the eye closure and eyebrow raise cue. A higher MD indicates that landmarks at that specific cue are more divergent from the normal range, allowing us to identify at which context patients differ most from the norm. In FP2's case, the highest MD is attributed to the eye closure cue, indicating that this is when the patient is the most abnormal from the normal function of all healthy controls defined in Figure 5b.

By creating separate heatmaps for each facial feature, clinicians can focus on specific regions of interest. This feature-specific analysis helps identify asymmetries or patterns related to particular facial landmarks, aiding in the assessment of facial expressions or cues.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

a.

Average of 19 Pearson correlation coefficients, y-position, for all healthy controls (*n=50*)			
Feature	**Pearson Coefficient (*r*)**	**Feature**	**Pearson Coefficient (*r*)**
Eyebrow 1	0.985 ± 0.015	Outer Eye Corner	0.936 ± 0.065
Eyebrow 2	0.978 ± 0.024	Upper Lip Vermillion 1	0.996 ± 0.004
Eyebrow 3	0.965 ± 0.037	Upper Lip Vermillion 2	0.987 ± 0.011
Eyebrow 4	0.939 ± 0.059	Upper Tubercles	0.996 ± 0.003
Eyebrow 5	0.905 ± 0.075	Oral Commissure 1	0.979 ± 0.017
Inner Eye Corner	0.974 ± 0.027	Oral Commissure 2	0.995 ± 0.003
Upper Eyelid 1	0.983 ± 0.019	Lower Tubercles	0.975 ± 0.023
Upper Eyelid 2	0.976 ± 0.025	Lower Lip Vermillion 1	0.930 ± 0.068
Lower Eyelid 1	0.963 ± 0.040	Lower Lip Vermillion 2	0.936 ± 0.060
Lower Eyelid 2	0.948 ± 0.055		

b.

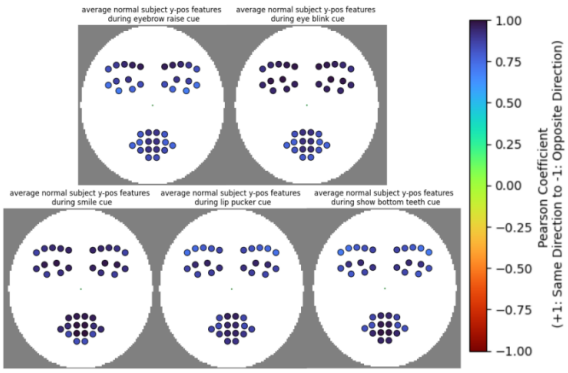


Fig. 5. a: The average and standard deviation of the 19 Pearson correlation coefficients for all healthy controls used in this study (n=50). These are the correlation coefficients calculated for the entire course of the FN exam. b: Plotting the average Pearson coefficients for the 19 features during specific cues for all healthy controls. These heatmaps represent the range of normal dynamic symmetry when comparing the changes in y-coordinates of the left side and right side of their face.

C. Performance of using Wasserstein distances to predict clinical scores

While the heatmaps provide insight to visualize specific regions of palsy, we wanted to associate the facial landmarks to clinically validated scores used in the field of Otolaryngology–Head and Neck Surgery, such as the House-Brackmann and Sunnybrook scale.

Wasserstein distances were calculated for all the subjects (n= 108 after 7 healthy control outliers removed). To determine the effectiveness of using Wasserstein distances to distinguish between normal and abnormal subjects, an ROC curve was plotted for 1000 iterations using 80% as the training data and 20% as testing. The mean AUC from all the iterations was 0.9152 (Figure 7a). Linear regression models between a patient's Wasserstein distance and given HB score was determined, with a regression coefficient of $r = 0.64$. A similar process was executed on a patient's Wasserstein distance and given Sunnybrook score, with a regression coefficient of $r = -0.68$. As seen in Figure 7b, c, we observe similar performances in models between the two clinician scales and the calculated distance.

To determine the efficacy of using Wasserstein distances as an objective metric to classify patients into their respective clinician scores, a linear regression model was trained under a subset of data and evaluated. The data was split among each of the HB score categories using a 80%-20% stratified train-test split to preserve the proportions of subjects in each categeory. The linear regression model was trained, where the Wasserstein distance was the independent variable and the HB scores were the dependent variable. Chance-level accuracy for this six-class classification was 1/6 = 16.67%, and when all distances were trained and tested accordingly for 1000 iterations, the mean accuracy was 20.55%. As seen in Figure 8a, the mean true positive values for HB I was 0.0. To observe the performance of a linear regression model trained on only the HB I and HB II subjects, a similar model was trained using a 80%-20% stratified train-test split, resulting in a mean accuracy of 85.64% for 1000 iterations (Figure 8b).

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

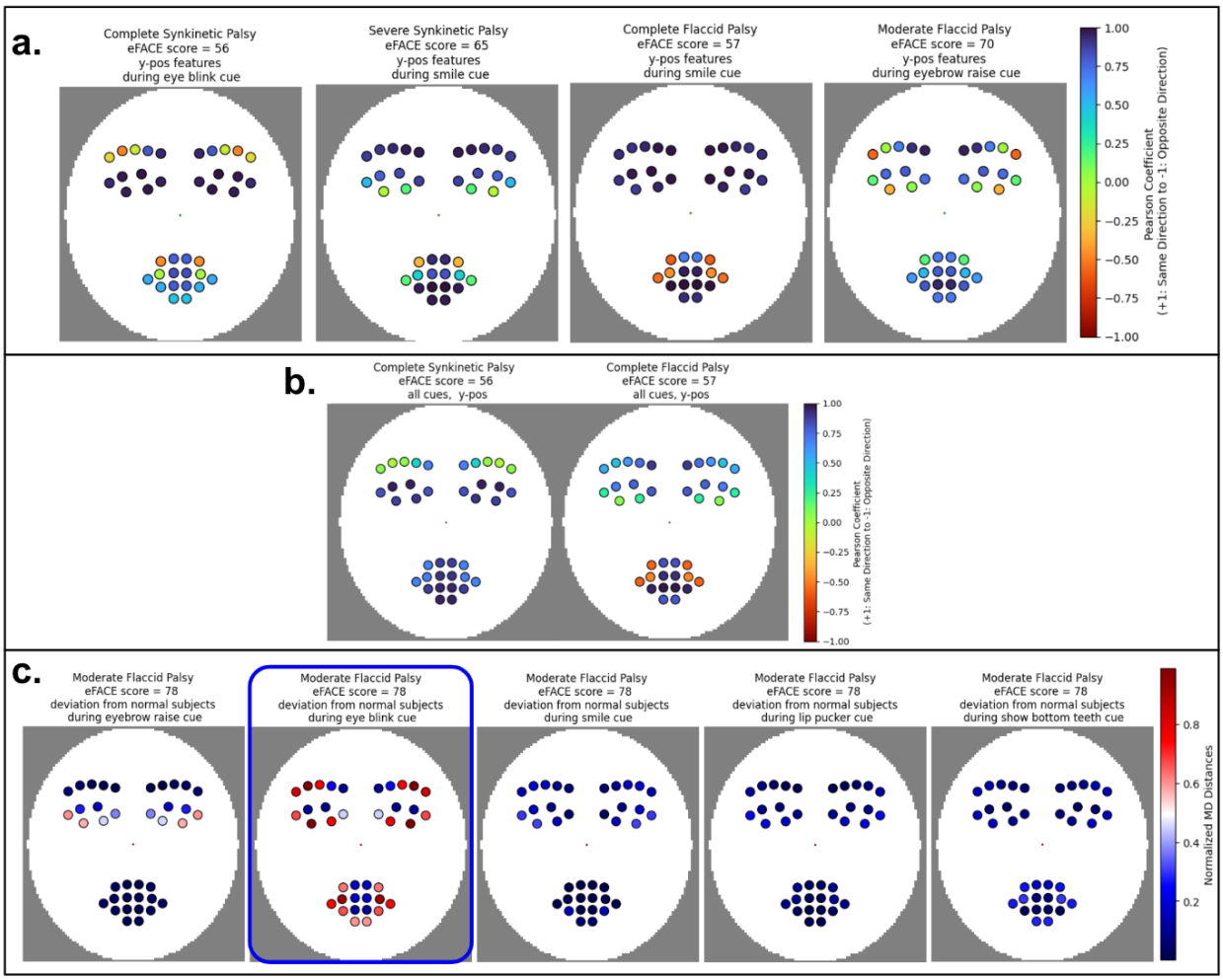


Fig. 6. a: Plotting Pearson coefficients for the 19 features during specific cues for each patient. These heatmaps represent individual patient dynamic asymmetry when comparing the changes in y-coordinates of the left side and right side of their face. In the order of left to right presents a: complete synkinetic palsy patient (SP1) (eFACE = 56), severe synkinetic palsy patient (eFACE =65), complete flaccid palsy patient (FP1) (eFACE = 57), and moderate flaccid palsy patient (eFACE = 70). b: Plotting Pearson coefficients for the 19 features during the whole time series for SP1 and FP1. c: Using normalized Mahalanobis distances (MDs) to compute how much a patient deviates from normal facial function (by comparing them to the average of all the controls) during each cue. This metric identifies which of the cues, and prospectively which regions of interest, has the most significant dynamic asymmetry when compared to normal function. Presented is a patient with moderate flaccid palsy (FP2), eFACE score = 78.

From this, a nested regression model was implemented, such that the subset of predictor variables that were initially given a score of HB I or II from the six-class classification model underwent another two-class classification to better separate the two categories. As a result, the mean accuracy for this six-class classification nested regression model increased approximately 3 times to 59.67% for 1000 iterations (Figure 8c).

D. Application to Surgical Patient Cases

Due to the severe morbidity of chronic facial palsy, decisions in facial nerve (FN) management can carry high stakes, particularly when the degree of injury to the FN is uncertain, as a stretch or compression injury can mimic a complete transection in the early stages of recovery. Here we present two cases of facial palsy, the first due to an iatrogenic injury and the second due to head and neck cancer, that was tracked over the course of time. We present how our proposed algorithms in these two cases could help inform clinicians of potential onsets of FN recovery or worsening paralysis, that can guide their decisions to improve patient outcomes.

1) Case 1: Tracking of subtle onset facial nerve recovery for an iatrogenic patient: Iatrogenic FN injury has a reported incidence of 11% to 40% [40] , where rates can vary depending on the type of surgical procedure done and

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

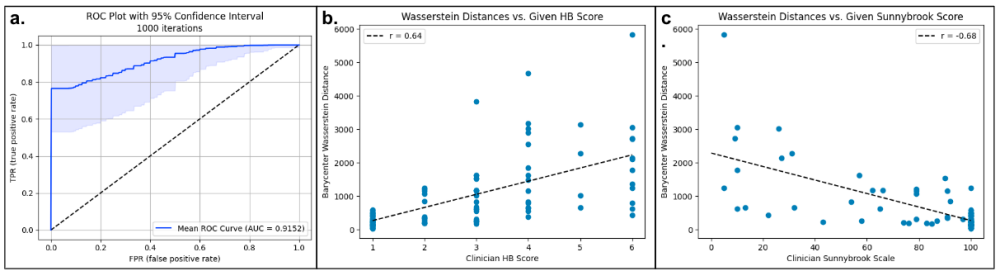


Fig. 7. a: Presents ROC curve of 1000 iterations for using Wasserstein distances to perform binary classification between abnormal and normal subjects. The mean AUC of the ROC curves was 0.9152. Linear regression models between the Wasserstein distance and (b) the given clinician HB score, with $r = 0.64$ and (c) the given clinical Sunnybrook score, with $r = -0.68$.

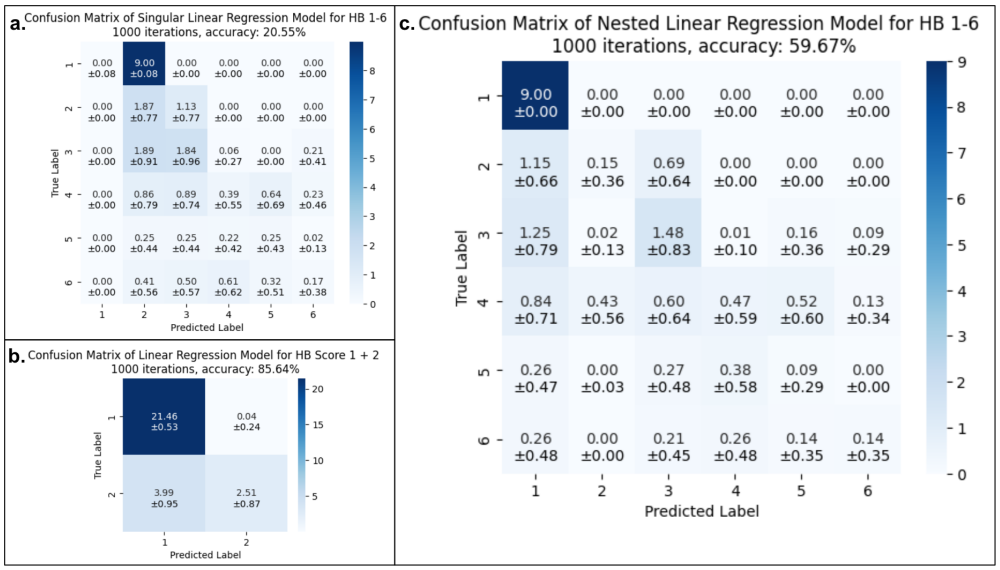


Fig. 8. Confusion matrix that evaluates the performance of a linear regression model that uses Wasserstein distances to predict HB scores. a: Presents the results of the initial linear regression model that was trained on a 80%-20% stratified train-test split of the dataset. The mean accuracy was 20.55% for *n=1000* iterations, compared to chance-level accuracy of 16.67% . b: Presents the results of the two-class linear regression model trained on the HB I and II data. The mean accuracy was 85.64% for *n=1000* iterations. c: Presents results of the nested regression model that implements a combination of the models describes in a and b. The mean accuracy was 59.67% for *n=1000* iterations, compared to chance-level accuracy of 16.67%.

expertise of the surgeon themselves. There is consensus amongst peripheral nerve surgeons that the critical window following acute injury during which nerve repair is feasible before scar formation is up to 72 hours—although with contemporary instruments, this timeline may be extended several weeks [41]. It is recommended that reconstruction of FN damage be done during this acute phase to avoid atrophy of the target facial muscles [41]. However, subtle changes in FN performance are difficult to evaluate both from an observational manner and from current measurements, such as scoring scales like HB or EMG, which only yields 33% accuracy in traumatic injury patients [42].

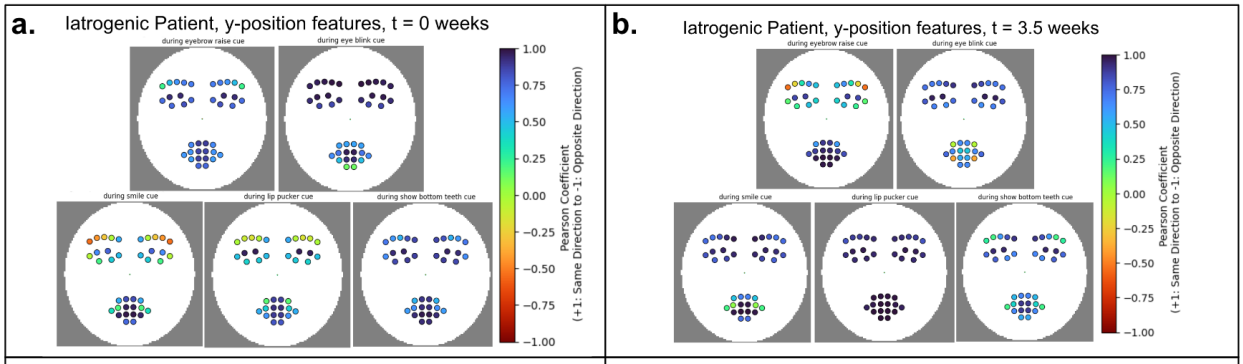
This case presents an iatrogenic patient (IP), Figure 9, who underwent a tumor excision where the surgeon was not certain whether a cranial nerve was damaged or not. A FN exam was performed the week following injury and then again 3.5 weeks after (please view section V regarding images of patients in figures). Figure 9 present the Pearson correlation coefficients during both trials, Figure 9a at 0 weeks and Figure 9b at 3.5 weeks. Differences in Pearson coefficient values are observed specifically during the smile cue and the lip pucker cue; the increased magnitudes of the eyebrow $(p < 0.01)$ and eye $(p < 0.05)$ features increase after 3.5 weeks are statistically significant (Figure 9c), indicating increased facial dynamic symmetry. However, since the magnitudes of coefficients for the lip region features significantly decreased $(p < 0.05)$, IP still has not complete normal facial function. Furthermore, Figure

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

9d presents significant increases in the eyebrow feature coefficients $(p < 0.01)$ during both the lip pucker and smile cues.

The MD was used to identify the context under which palsy was most apparent for IP, and Figure 10a presents the smile cue being the most deviation from normal function. With this context, the Wasserstein distances were calculated for IP and the linear regression model was used to predict HB scores. The model predicts IP to have a HB score of V at week 0 and a HB score of III at week 3.5, as seen in Figure 10b. These predicted scores match the clinician assigned scores, which were HB V at week 0 and HB III at week 3.5, with a mean square error (MSE) of 0.5 for the two predictions.

Due to the signs of improvement over this time duration, IP's clinician suggested to hold off on surgical intervention and continue monitoring her recovery over the next few months. The proposed visualization algorithm and metrics are able to guide clinicians of IP's overall FN recovery over the 3.5 weeks. Therefore, these tools can be used to supplement a clinician's decision on whether reanimation surgery is necessary or not, proving to be useful when a time-sensitive decision needs to be made.



c.

IP's Pearson correlation coefficients at regions of interest during smile cue			
Feature	**Pearson Coefficient at 0 weeks**	**Pearson Coefficient at 3.5 weeks**	**Significance Value**
Eyebrow 1	0.911	0.983	$p \leq 0.01$ (**)
Eyebrow 2	0.799	0.951	
Eyebrow 3	0.729	0.913	
Eyebrow 4	0.705	0.904	
Eyebrow 5	0.703	0.888	
Upper Eyelid 1	0.938	0.980	$p \leq 0.05$ (*)
Upper Eyelid 2	0.891	0.969	
Outer Eye Corner	0.838	0.957	
Oral Commissure 1	0.929	0.759	$p \leq 0.05$ (*)
Lower Tubercles	0.905	0.705	

d.

IP's Pearson correlation coefficients at eyebrow region during smile + lip pucker cue			
Feature	**Pearson Coefficient at 0 weeks**	**Pearson Coefficient at 3.5 weeks**	**Significance Value**
Eyebrow 1	0.939	0.985	$p \leq 0.01$ (**)
Eyebrow 2	0.858	0.971	
Eyebrow 3	0.812	0.950	
Eyebrow 4	0.787	0.946	
Eyebrow 5	0.751	0.923	

Fig. 9. a: Plotting Pearson coefficients for the 19 features during specific cues for the iatrogenic patient (IP) at 0 weeks post-injury. b: Plotting Pearson coefficients for the 19 features during specific cues for IP at 3.5 weeks post-injury, where there are improvements in values during the smile and lip pucker cues. c: Table of Pearson coefficients at the regions of interest during the smile cue. We observe statistically significantly increases in coefficient values in the eyebrow $(p < 0.01)$and eye region $(p < 0.05)$, and statistically significantly decreases in the lip region $(p < 0.05)$. d: Table of Pearson coefficients at eyebrow region during both the smile and lip pucker cue. We observe statistically significantly increases in coefficient values $(p < 0.01)$.

2) Case 2: Tracking facial nerve function pre- and post-facial nerve sacrifice and repair of a head and neck cancer patient: Subtle, progressive and non-resolving facial palsy can be the sole initial clinical manifestation of perineural spread of head and neck cancer occurring along the facial nerve (CN VII), and be misdiagnosed as a benign process such as Bell's Palsy in cancer survivors, particularly in the absence of a discrete tumor. Facial paralysis continues to pose a significant concern during cancer resection and facial nerve interventions are often not prioritized amidst urgent oncologic treatments [43]. Detection of subtle changes in facial nerve function preoperatively could greatly

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

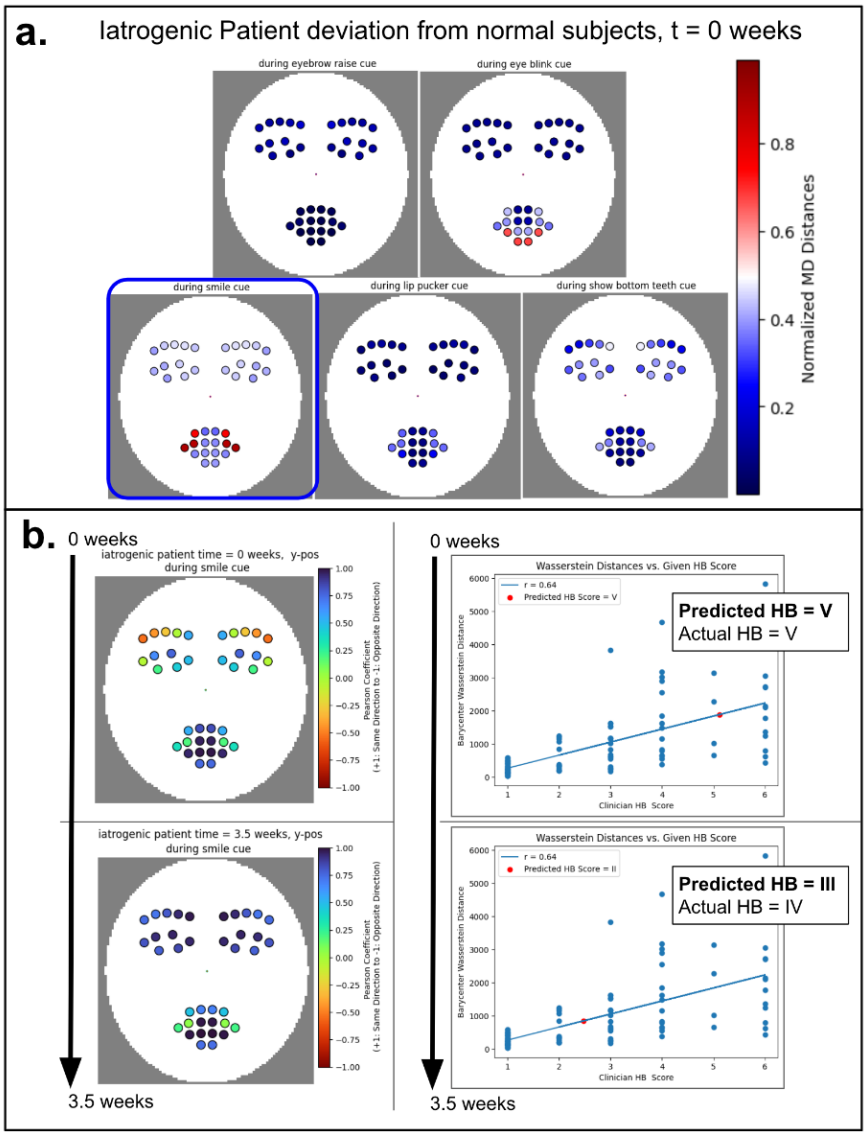


Fig. 10. a: Using normalized MDs to compute how much IP deviates from normal facial function during each cue. Algorithm identifies the smile cue to have the most deviation. b: (left) Tracking Pearson coefficients during the cue that exhibited maximum deviation from normal function, smiling. Increases in Pearson coefficient magnitudes are observed over time, indicating FN recovery. (right) Using the linear regression model to predict IP's HB score over the course of the 3.5 weeks. IP's score drops from an initial HB V to a HB III and these scores match the clinician given score with an MSE of 0.5.

aid surgical planning as the optimal time for facial nerve reconstruction is at the time of tumor resection. For patients who develop subtle facial weakness following cancer treatments with an intact facial nerve, additional metrics tracking function could potentially detect cancer recurrence at an earlier date and lead to more effective, earlier treatments. Monitoring outcomes after facial reanimation surgery could also greatly advance with more robust, objective metrics that what are currently utilized.

This case presents a patient with a right parotid malignant tumor (CP), Figure 11, who underwent a radical right parotidectomy with a FN sacrifice, lateral temporal bone resection, and neck dissection to remove the tumor. Other reanimation procedures were performed on CP such as a fascia lata and suture static suspension to the nasolabial fold, nerve harvests and grafting for FN repair, and placement of a eyelid weight in their right eyelid (please view section V regarding images of patients in figures). Videos of CP performing the FN exam were taking 3 weeks prior to surgery when they started developing FN weakness, 8 months after surgery to remove the tumor and repair

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

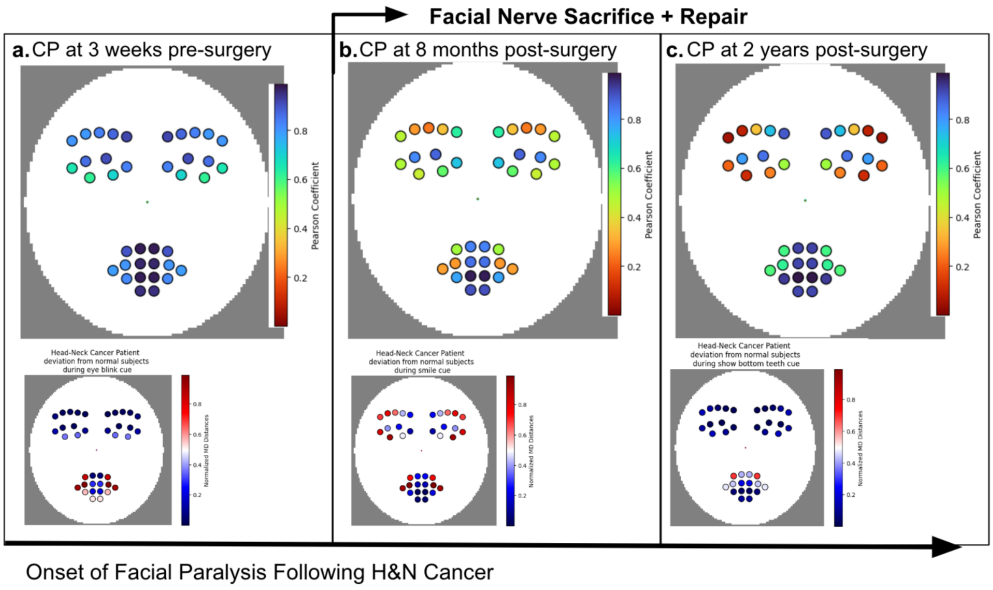


Fig. 11. Tracking CP's FN recovery over the course of when they developed FN weakness 3 weeks prior to surgery (a), 8 months after their tumor removal and facial reanimation surgery (b), and 2 years after surgery (c). The cue that deviates most from normal function, as identified by MD algorithm, is presented below the Pearson coefficient heatmaps.

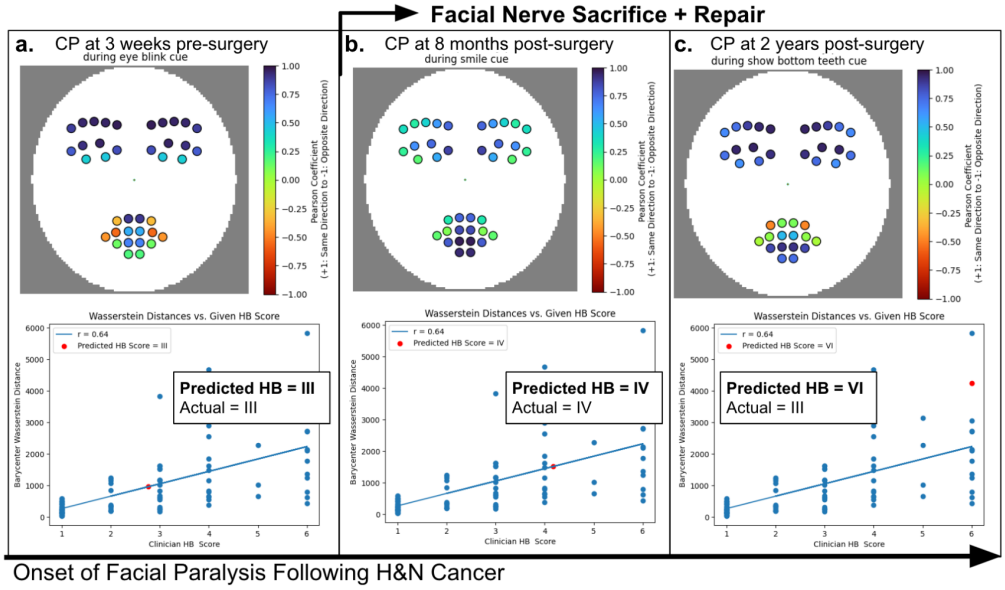


Fig. 12. Plotting the Pearson coefficients at CP's cues of maximum deviation at every time point (as selected by MDs in Figure 11). Using the linear regression model to predict CP's HB score over the course of their clinic visits. a: The model predicts an initially HB score of III when they first develop FN weakness 3 weeks prior to surgery. b: The model predicts a HB score of IV 8 months post-surgery, when CP developed mild synkinesis. c: The model predicts a HB score of VI 2 years post-surgery, when CP developed moderate synkinesis. These three predictions match the clinician given score with an MSE of 1.0.

FNs, and 2 years after the surgery.

Figure 11 tracks CP's initial onset of FN weakness and FN function 8 months and 2 years after surgery. The initial FN weakness was initially most apparent under the eye closure cue, where slight weakness is observed in the lower eye regions and oral commissure areas (Figure 11a). Figures 11b, c present worsening paralysis in the eyebrow region after surgery with the maximum deviation contexts varying from the smile cue at 8 months to the bottom teeth cue at 2 years.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

With the maximum deviation contexts determined by the MD algorithm (bottom of Figure 11), CP's Wasserstein distances were calculated at each time point. The linear regression model was used to predict HB scores both before and after surgery. The model predicts an initially HB score of III when they first develop FN weakness 3 weeks prior to surgery (Figure 12a). Once CP had their FN sacrifice surgery, CP developed complete flaccid paralysis and was diagnosed with a HB score of VI after surgery. CP visited the clinic at 4 months post-surgery, and static photographs of the FN exam were taken at to capture the extent of CP's flaccid paralysis. However, the study's analyses could not be applied as a video, rather than photos are required. The model predicts a HB score of IV 8 months post-surgery, when CP developed mild synkinesis. (middle of Figure 12). the score increases to a VI 2 years after surgery (right of Figure 12). These predicted match the clinician assigned scores, which were HB III 3 weeks post-surgery, HB IV 8 months post-surgery, and HB III 2 years post-surgery, with a MSE of 1.0 for the three predictions.

This case successfully captures the trends of CP's initial FN weakness before surgery and the worsening of paralysis after the tumor was removed, and the regression model matches clincian HB scores fairly well. These visualization algorithms and metrics can therefore be applied to various causes that may lead to onsets of FN weakness. Furthermore, implementation of this algorithm in similar cases could provide guidance to surgeons on whether tumor resection needs to be performed earlier, leading to better patient outcomes post-surgery.

IV. Discussion

Current methods of facial nerve (FN) assessment rely on clinician experience and subjective scales such as House-Brackman (HB) or Sunnybrook [44], [45] which can miss subtle changes to a patient's FN exam such as interval asymmetry or segmental weakness. Clinician-based evaluation of facial function is inherently subjective with a reported interobserver variability up to 64% for severe facial palsy (HB III or worse) [46]. Machine learning (ML) algorithms for facial analysis have advanced significantly in the past few decades [20] [21] but are based on training datasets of intact facial function; application of these tools to facial paralysis patients photographs was reported to cause significant landmark inaccuracy that precludes clinical usage [22] [23]. An objective, rigorous quantification of dynamic facial function in videos has remained elusive [29] [30] [28] as there are multiple challenges to accumulating a sufficiently large training dataset to improve landmark accuracy in the facial paralysis population. Recent work has demonstrated that the latest the open-source computer vision and machine learning facial analysis [34] on videos of facial paralysis patients reduced landmark errors significantly compared to static image analysis [32]. In this study we report for the first time a novel technique of using barycenters and wasserstein distributions to develop robust ML techniques to classify between normal and abnormal facial function in videos of facial palsy patients and one of the few papers that uses robust metrics to predict commonly used clinical grading scales like HB and Sunnybrook.

Although our ability to read and interpret facial expressions is an evolutionary trait learned at a young age, our understanding of normal facial movements [47] and ability to describe it mathematically is limited, particularly in the asymmetric and asynchronous movements of facial palsy patients. In our study, we demonstrate Pearson correlation coefficients as a valuable metric for quantifying the degree of symmetry in facial movement over the duration of a video. By leveraging the coordinates of facial landmarks across multiple frames for certain cues, these correlation coefficients provide a robust statistical measure of the linear relationship between facial movements, while also allowing us to identify regions of asymmetry. This video ML analysis could for the first time characterize the dynamic properties of various facial movements, like smiling versus raising your eyebrows, without requiring manual assessment, and significantly advance our understanding of facial function. This context specific facial function is what drove the final analyses of determining the Wasserstein distances that aided in developing prediction models to map back to the clinical HB score. Focusing on specific facial cues where paralysis is most pronounced in palsy patients enhances the accuracy of ML models by reducing generalizability and honing in on critical regions of dysfunction. Additionally, it offers clinicians valuable context regarding which facial regions require particular attention, optimizing their diagnostic and surgical strategies for patients.

There are several limitations of this work. Firstly, the gold standard for facial analysis remains the human eye and landmark errors still occur during videos of facial palsy patients. We have found particular facial features and certain cues that decrease landmark accuracy such as eyelids during forceful blinking, as well as facial rhytids, facial hair, increased BMI, and patient movement are some of the suspected sources of error. Patient effort can also affect the range of facial movement, particularly in the facial palsy patient population, who often consciously

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

or subconsciously avoid large movements like smiles to mask their asymmetry. Other errors in our study could be attributed to human error during the FN exam resulting from unstable camera setups and other noise artifacts. We hope to continue developing stronger pre-processing steps after facial landmark extraction to increase the accuracy of our presented clinical score prediction models.

Uncertainty analysis has been previously investigated in sleep tests for obstructive sleep apnea with the primary advantage of automated identification identification of difficult to analyze data that is flagged for ”clinician-in-the-loop” review [48]. One important finding that is likely reflected in other clinical datasets analyzed by ML tools, is the notable lack of perfect concordance among the “gold-standard” clinician review. The elements of doubt, uncertainty or gray areas are well recognized as part of the art of medicine; providing uncertainty metrics allows for this information to be automatically reported for additional review. We hope to conduct future investigations in developing insight on uncertainty metrics that could trigger additional clinician review and augment the accuracy of our facial functional rating.

Furthermore, increasing the database of both healthy controls and facial palsy subjects in a controlled setting will allows us to better understand the spectrum of normal dynamic facial function, particularly with the limitations of available facial palsy datasets. In the case of the surgical head and neck cancer patient (Figure 12), the prediction model had difficulties identifying the patient’s transition from complete flaccid paralysis to moderate synkinesis post-surgery. Increasing facial palsy databases will facilitate a comprehensive analysis of facial dysfunction, enabling a detailed comparison between synkinetic and flaccid patients. By discerning the distinctive patterns and nuances in each type of palsy, we can develop more accurate prediction models tailored to differentiate between them, thus advancing personalized treatment strategies and improving outcomes for individuals affected by facial paralysis.

Future applications of dynamic facial analysis are manifold. Recent advances in telemedicine technology have enabled secure, patient-driven, HIPAA compliant, iOS telemedicine applications that have applied advanced computer vision and artificial intelligence techniques for diverse purposes such as tracking rehabilitation outcomes after knee arthroplasty [49], epilepsy monitoring [50], to assessing liver steatosis from donors [51]. There are currently half a million survivors of Head and Neck Cancer in the US. Development of this technology could empower patients to track their own facial function after Head and Neck cancer therapy with a telemedicine tool that employs Machine-Learning algorithms for facial analysis. Providing patients with this tool to monitor their facial function remotely could potentially reduce diagnostic delays, facilitate objective data that can be shared with clinicians through telemedicine, and empower patients with ownership of their own data. Most critically, this tool will be available across socioeconomic and insurance statuses as the vast majority of Americans (97%) own a cellphone and nine-in-ten own a smartphone [52]. Advances in technology and a shift towards more patient-centered assessments may contribute to improving the accuracy and sensitivity of FN function evaluations and facilitate an understanding of dynamic facial function in a manner not previously possible.

V. Addendum

Due to pre-print screening requirements, images of the facial palsy subjects discussed in this study will only be released in the final publication. Readers may contact the corresponding author prior to final publication to request access to these materials.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

REFERENCES

[1] K. Brattain, "Analysis of the Peripheral Nerve Repair Market in the United States," *Magellan Medical Technology Consultants, Inc.*, 2013.

[2] D. Grinsell and C. P. Keating, "Peripheral nerve reconstruction after injury: a review of clinical and experimental therapies," *Biomed Res Int*, vol. 2014, p. 698256, 2014.

[3] R. Gupta, J. P. Chan, J. Uong, W. A. Palispis, D. J. Wright, S. B. Shah, S. R. Ward, T. Q. Lee, and O. Steward, "Human motor endplate remodeling after traumatic nerve injury," *J Neurosurg*, vol. 135, pp. 220–227, Sept. 2020.

[4] J. N. Bleicher, S. Hamiel, J. S. Gengler, and J. Antimarino, "A survey of facial paralysis: etiology and incidence," *Ear Nose Throat J*, vol. 75, pp. 355–358, June 1996.

[5] N. Jowett and T. A. Hadlock, "An Evidence-Based Approach to Facial Reanimation," *Facial Plast Surg Clin North Am*, vol. 23, pp. 313–334, Aug. 2015.

[6] S. C. Prasad, K. Balasubramanian, E. Piccirillo, A. Taibah, A. Russo, J. He, and M. Sanna, "Surgical technique and results of cable graft interpositioning of the facial nerve in lateral skull base surgeries: experience with 213 consecutive cases," *J Neurosurg*, vol. 128, pp. 631–638, Feb. 2018.

[7] J. J. Greene, J. Tavares, S. Mohan, N. Jowett, and T. Hadlock, "Long-Term Outcomes of Free Gracilis Muscle Transfer for Smile Reanimation in Children," *J Pediatr*, vol. 202, pp. 279–284.e2, Nov. 2018.

[8] A. Kochhar, M. Albathi, J. D. Sharon, L. E. Ishii, P. Byrne, and K. D. Boahene, "Transposition of the Intratemporal Facial to Hypoglossal Nerve for Reanimation of the Paralyzed Face: The VII to XII TranspositionTechnique," *JAMA Facial Plast Surg*, vol. 18, pp. 370–378, Sept. 2016.

[9] P. D. Knott, "A Facial Nerve Anniversary–Twelve Months of Treatment Time Saved," *JAMA Facial Plast Surg*, vol. 18, no. 1, pp. 60–61, 2016.

[10] C. A. Banks, P. K. Bhama, J. Park, C. R. Hadlock, and T. A. Hadlock, "Clinician-Graded Electronic Facial Paralysis Assessment: The eFACE," *Plast Reconstr Surg*, vol. 136, pp. 223e–230e, Aug. 2015.

[11] J. H. Ng and R. Y. S. Ngo, "The use of the facial clinimetric evaluation scale as a patient-based grading system in bell's palsy," *The Laryngoscope*, vol. 123, no. 5, pp. 1256–1260, 2013.

[12] R. A. Gaudin, M. Robinson, C. A. Banks, J. Baiungo, N. Jowett, and T. A. Hadlock, "Emerging vs Time-Tested Methods of Facial Grading Among Patients With Facial Paralysis," *JAMA Facial Plast Surg*, vol. 18, pp. 251–257, July 2016.

[13] A. Y. Fattah, A. D. R. Gurusinghe, J. Gavilan, T. A. Hadlock, J. R. Marcus, H. Marres, C. C. Nduka, W. H. Slattery, A. K. Snyder-Warwick, and Sir Charles Bell Society, "Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity," *Plast Reconstr Surg*, vol. 135, pp. 569–579, Feb. 2015.

[14] P. C. Revenaugh, R. M. Smith, M. A. Plitt, L. Ishii, K. Boahene, and P. J. Byrne, "Use of Objective Metrics in Dynamic Facial Reanimation: A Systematic Review," *JAMA Facial Plast Surg*, vol. 20, pp. 501–508, Dec. 2018.

[15] T. Hadlock, "Standard Outcome Measures in Facial Paralysis," *JAMA Facial Plast Surg*, vol. 18, pp. 85–86, Mar. 2016. Publisher: AMA - American Medical Association.

[16] C. D. Bovenzi, P. Ciolek, M. Crippen, J. M. Curry, H. Krein, and R. Heffelfinger, "Reconstructive trends and complications following parotidectomy: Incidence and predictors in 11,057 cases," *Journal of Otolaryngology - Head & Neck Surgery*, vol. 48, p. 64, Jan. 2019. Publisher: SAGE Publications.

[17] N. Mat Lazin, H. Ismail, S. Abdul Halim, N. A. Nik Othman, and A. Haron, "Comparison of 3 Grading Systems (House-Brackmann, Sunnybrook, Sydney) for the Assessment of Facial Nerve Paralysis and Prediction of Neural Recovery," *Medeni Med J*, vol. 38, pp. 111–119, June 2023.

[18] J. E. Berner, P. Kamalathevan, I. Kyriazidis, and C. Nduka, "Facial synkinesis outcome measures: A systematic review of the available grading systems and a Delphi study to identify the steps towards a consensus," *J Plast Reconstr Aesthet Surg*, vol. 72, pp. 946–963, June 2019.

[19] M. Sonoo, D. L. Menkes, J. D. P. Bland, and D. Burke, "Nerve conduction studies and EMG in carpal tunnel syndrome: Do they add value?," *Clinical Neurophysiology Practice*, vol. 3, pp. 78–88, Jan. 2018.

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge," in *2013 IEEE International Conference on Computer Vision Workshops*, (Sydney, Australia), pp. 397–403, IEEE, Dec. 2013.

[21] D. E. King, "Dlib-ml: A Machine Learning Toolkit,"

[22] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett, "Toward an Automatic System for Computer-Aided Assessment in Facial Palsy," *Facial Plast Surg Aesthet Med*, vol. 22, no. 1, pp. 42–49, 2020.

[23] D. L. Guarin, J. Dusseldorp, T. A. Hadlock, and N. Jowett, "A Machine Learning Approach for Automated Facial Measurements in Facial Palsy," *JAMA Facial Plast Surg*, vol. 20, pp. 335–337, July 2018.

[24] J. R. Dusseldorp, D. L. Guarin, M. M. van Veen, N. Jowett, and T. A. Hadlock, "In the Eye of the Beholder: Changes in Perceived Emotion Expression after Smile Reanimation," *Plastic and Reconstructive Surgery*, vol. 144, p. 457, Aug. 2019.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[25] J. R. Dusseldorp, M. M. van Veen, D. L. Guarin, O. Quatela, N. Jowett, and T. A. Hadlock, "Spontaneity Assessment in Dually Innervated Gracilis Smile Reanimation Surgery," *JAMA Facial Plast Surg*, vol. 21, pp. 551–557, Dec. 2019.

[26] T. Boonipat, M. Asaad, J. Lin, G. E. Glass, S. Mardini, and M. Stotland, "Using Artificial Intelligence to Measure Facial Expression following Facial Reanimation Surgery," *Plast Reconstr Surg*, vol. 146, pp. 1147–1150, Nov. 2020.

[27] J. Fuzi, C. Meller, S. Ch'ng, T. M. Hadlock, and J. Dusseldorp, "Voluntary and Spontaneous Smile Quantification in Facial Palsy Patients: Validation of a Novel Mobile Application," *Facial Plast Surg Aesthet Med*, vol. 25, no. 4, pp. 312–317, 2023.

[28] T. Hidaka, M. Kurita, K. Ogawa, Y. Tomioka, and M. Okazaki, "Application of Artificial Intelligence for Real-Time Facial Asymmetry Analysis," *Plast Reconstr Surg*, vol. 146, pp. 243e–245e, Aug. 2020.

[29] D. L. Guarin, A. Bandini, A. Dempster, H. Wang, S. Rezaei, B. Taati, and Y. Yunusova, "The Effect of Improving Facial Alignment Accuracy on the Video-based Detection of Neurological Diseases," preprint, Sept. 2020.

[30] B. Johnston and P. d. Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP Journal on Image and Video Processing*, vol. 2018, p. 86, Sept. 2018.

[31] C. Meijerink, "Facial landmark detection under challenging conditions," July 2021.

[32] S. K. Kalavacherla, M. Davis, and J. J. Greene, "Learning from machine learning: advancing from static to dynamic facial function quantification," *arXiv preprint arXiv:2024/584911*, 2024.

[33] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock, "The spectrum of facial palsy: The MEEI facial palsy photo and video standard set," *The Laryngoscope*, vol. 130, no. 1, pp. 32–37, 2020.

[34] J. C. Martinez, "Detecting Face Features with Python. Live Code Stream.," July 2020. https:// livecodestream.dev/post/detecting-face-features- with-python/

[35] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[36] P. C. Álvarez Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in Wasserstein space," *Journal of Mathematical Analysis and Applications*, vol. 441, pp. 744–762, Sept. 2016.

[37] C. Villani *et al.*, *Optimal transport: old and new*, vol. 338. Springer, 2009.

[38] M. Agueh and G. Carlier, "Barycenters in the wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, pp. 29–36, Apr. 1982. Publisher: Radiological Society of North America.

[40] M. H. Hohman, P. K. Bhama, and T. A. Hadlock, "Epidemiology of iatrogenic facial nerve injury: A decade of experience," *The Laryngoscope*, vol. 124, no. 1, pp. 260–265, 2014.

[41] E. Fliss, R. Yanko, A. Zaretski, R. Tulchinsky, E. Arad, D. J. Kedar, D. M. Fliss, and E. Gur, "Facial Nerve Repair following Acute Nerve Injury," *Arch Plast Surg*, vol. 49, pp. 501–509, July 2022.

[42] G. Mannarelli, G. R. Griffin, P. Kileny, and B. Edwards, "Electrophysiological measures in facial paresis and paralysis," *Operative Techniques in Otolaryngology-Head and Neck Surgery*, vol. 23, pp. 236–247, Dec. 2012.

[43] K. L. Crawford, J. A. Stramiello, R. K. Orosco, and J. J. Greene, "Advances in facial nerve management in the head and neck cancer patient," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 28, p. 235, Aug. 2020.

[44] S. E. Coulson, N. J. O'dwyer, R. D. Adams, and G. R. Croxson, "Expression of emotion and quality of life after facial nerve paralysis," *Otol Neurotol*, vol. 25, pp. 1014–1019, Nov. 2004.

[45] S. D. Reitzen, J. S. Babb, and A. K. Lalwani, "Significance and reliability of the House-Brackmann grading system for regional facial nerve function," *Otolaryngology–Head and Neck Surgery*, vol. 140, no. 2, pp. 154–158, 2009.

[46] C. Scheller, A. Wienke, M. Tatagiba, A. Gharabaghi, K. F. Ramina, K. Scheller, J. Prell, J. Zenk, O. Ganslandt, B. Bischoff, C. Matthies, T. Westermaier, G. Antoniadis, M. T. Pedro, V. Rohde, K. von Eckardstein, T. Kretschmer, M. Kornhuber, F. G. Barker, and C. Strauss, "Interobserver variability of the House-Brackmann facial nerve grading system for the analysis of a randomized multi-center phase III trial," *Acta Neurochir*, vol. 159, pp. 733–738, Apr. 2017.

[47] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Annual Review of Psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[48] D. Y. Kang, P. N. DeYoung, J. Tantiongloc, T. P. Coleman, and R. L. Owens, "Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine," *npj Digit. Med.*, vol. 4, p. 142, Sept. 2021.

[49] S. Bini and J. Mahajan, "Clinical outcomes of remote asynchronous telerehabilitation are equivalent to traditional therapy following total knee arthroplasty: A randomized control study," *J Telemed Telecare*, vol. 23, pp. 239–247, Feb. 2017. Publisher: SAGE Publications.

[50] W. O. Tatum, L. J. Hirsch, M. A. Gelfand, E. K. Acton, W. C. LaFrance, Jr, R. B. Duckrow, D. K. Chen, A. S. Blum, J. D. Hixson, J. F. Drazkowski, S. R. Benbadis, G. D. Cascino, and for the OSmartViE Investigators, "Assessment of the Predictive Value of Outpatient Smartphone Videos for Diagnosis of Epileptic Seizures," *JAMA Neurology*, vol. 77, pp. 593–600, May 2020.

[51] M. Cesaretti, N. Poté, F. Cauchy, F. Dondero, S. Dokmak, A. Sepulveda, A. S. Schneck, C. Francoz, F. Durand, V. Paradis, and O. Soubrane, "Noninvasive assessment of liver steatosis in deceased donors: A pilot study," *Liver Transpl*, vol. 24, pp. 551–556, Apr. 2018.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.16.24304357; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[52] P. Center, "Mobile Fact Sheet: Mobile phone ownership over time," 2024. Numbers, Facts and Trends Shaping Your World. https://www.pewresearch.org/internet/fact-sheet/mobile/.