CLINIFACE: PHENOTYPIC VISUALISATION AND ANALYSIS USING NON-RIGID REGISTRATION OF 3D FACIAL IMAGES

R. L. Palmer[1,*], P. Helmholz[1], G. Baynam[1,2,3]

[1] School of Earth and Planetary Sciences, Faculty of Science and Engineering, Curtin University, Perth, WA 6845 Australia
(r.l.palmer, petra.helmholz)@curtin.edu.au
[2] Western Australian Register of Developmental Anomalies and Genetic Services of Western Australia,
King Edward Memorial Hospital, Dept. of Health, Gov. of Western Australia, Perth WA 6008 Australia
[3] Telethon Kids Institute and Division of Paediatrics, Faculty of Health and Medical Sciences, UWA, Perth, WA 6008 Australia
gareth.baynam@health.wa.gov.au

Commission II, WG II/3

KEY WORDS: 3D Image, Anthropometrics, Cliniface, Dysmorphology, Facial Landmarks, Non-Rigid Registration, Phenotype

ABSTRACT:

Facial appearance has long been understood to offer insight into a person's health. To an experienced clinician, atypical facial features may signify the presence of an underlying rare or genetic disease. Clinicians use their knowledge of how disease affects facial appearance along with the patient's physiological and behavioural traits, and their medical history, to determine a diagnosis. Specialist expertise and experience is needed to make a dysmorphological facial analysis. Key to this is accurately assessing how a face is significantly different in shape and/or growth compared to expected norms. Modern photogrammetric systems can acquire detailed 3D images of the face which can be used to conduct a facial analysis in software with greater precision than can be obtained in person. Measurements from 3D facial images are already used as an alternative to direct measurement using instruments such as tape measures, rulers, or callipers. However, the ability to take accurate measurements – whether virtual or not – presupposes the assessor's facility to accurately place the endpoints of the measuring tool at the positions of standardised anatomical facial landmarks. In this paper, we formally introduce *Cliniface* – a free and open source application that uses a recently published highly precise method of detecting facial landmarks from 3D facial images by non-rigidly transforming an anthropometric mask (AM) to the target face. Inter-landmark measurements are then used to automatically identify facial traits that may be of clinical significance. Herein, we show how non-experts with minimal guidance can use *Cliniface* to extract facial anthropometrics from a 3D facial image at a level of accuracy comparable to an expert. We further show that *Cliniface* itself is able to extract the same measurements at a similar level of accuracy – completely automatically.

1. INTRODUCTION

The human phenotype – including our apparent physiological form – is modified by our genotype and the expression of our genes which is affected by developmental and environmental factors. Many rare diseases are caused by genetic variations which can result in perturbation of normal growth and functioning. While the variety of different genetic diseases is individually rare, it is estimated that cumulatively between 6–8 percent of the human population is affected by a rare disease (Nguengang Wakap et al., 2020). Rare diseases are frequently characterised by medical complexity, pain, suffering, disability, and premature death. Up to 30% of children affected by a rare disease die before their fifth birthday (Global Genes, 2020).

Assessing the phenotype (phenotyping) is a clinician's daily work, including searching for clues to disease diagnosis. The shape and growth of the face is a focus of phenotyping because characteristic facial variations are found in approximately 1 in 3 rare diseases (Ferry et al., 2014). By measuring the face and head and comparing these measurements against statistical norms and expected patterns of growth, the clinician can make inferences about possible diagnoses. With a large and accurate enough set of observations, the scope of possible diseases is reduced to support diagnosis, including guiding genetic testing where relevant. Unfortunately, due to variation in facial appearance and the subtle nature of many clinically salient phenotypic traits, the ability to perform a dysmorphological analysis of the human face requires specialist training and experience that is uncommon to general medical practitioners.

Clinicians are now beginning to use 3D facial imaging to support the measurement and interpretation of patients' facial features (Poulton et al., 2018, Baynam et al., 2016, Baynam et al., 2013, Hammond et al., 2012), and to explore the genetic basis for the phenotypic expression of facial features (Shaffer et al., 2016, Hammond, Suttie, 2012). Two-dimensional photographs have been used for some time to help clinicians assess patients (Ferry et al., 2014), but the advent of photogrammetric technologies that can capture very detailed and spatially accurate 3D representations of the face and head gives clinicians the ability to take real world measurements that can be compared against existing norms to identify unusual and clinically relevant traits. A standardised ontology of phenotypic terms known as the Human Phenotype Ontology (HPO) (https://hpo.jax.org) (Köhler et al., 2018) has also been developed to better define and communicate the nature of these traits and more effectively share knowledge about (facial) phenotypic variation.

Being able to visualise a patient's face in 3D is itself useful to clinicians due to the extra spatial dimension (over 2D photographs) and the ability to see a face from different viewpoints;

* Corresponding author

such information can help to make better subjective clinical assessments. The image can also be revisited later on if reassessment is needed or new information comes to light. However, taking accurate measurements directly from the 3D facial image itself is not possible or accessible for most clinicians. Given a 3D image capture system, producing a 3D image of a subject's face and then viewing it usually entails using the hardware vendor's software which is tied by licensing arrangements to the system's owner. Images can usually be converted into non-proprietary formats for viewing in other applications, but measurement data (if it can be taken) cannot be stored with the image or communicated to others, except by external means. Third-party applications that allow 3D facial images to be imported for viewing may also suffer from user interface complexity and are not generally tailored to the specific use cases of clinicians interested in facial anthropometrics and dysmorphology supporting rare disease diagnosis.

Using 3D facial images in a clinical setting, allowances must be made for how clinicians take facial measurements which are defined in terms of standardised anthropometric landmarks. Some of these are difficult to localise – especially on severely dysmorphic faces. Several landmarks are surface approximations for the position of cephalometric points and a clinician would normally palpate the face to find cranial ridges that help to position these landmarks. Since such practices are not possible on 3D images, algorithms must be able to accurately position the landmarks. This enables the extraction of facial measurements if the clinician has access to suitably tailored software.

2. BACKGROUND AND CONTRIBUTION

Several algorithms have been developed to approach the challenge of accurate anthropometric landmark placement on 3D facial images (Gilani et al., 2015, Yang, 2011, Myronenko, Song, 2010, Chui, Rangarajan, 2003). Recently, a new method was developed using the approach of mapping a 3D anthropometric mask (AM) (Claes et al., 2012) through non-rigid deformation to a target face (White et al., 2019). This method uses an affinity matrix of symmetrically weighted nearest neighbour correspondences between the AM and the target surface. Over several iterations, the correspondences are adjusted by gradually transforming and relaxing the mask toward and away from the target surface in a way that evens out possibly erroneous transformative steps, taking a less local approach to the mask's deformation (the authors call this the *Visco-Elastic* step). The algorithm's parameters can be changed to weight the affinity matrix to include or exclude certain parts of the mask or target surface, to change the number of points used in K-Nearest Neighbour correspondence finding, and to change the size of the local smoothing region after each iteration. These changes result in greater mapping detail or improved anthropometric correspondence (accuracy of these concerns is traded against one another).

Landmark detection is performed by establishing correspondences between anthropometrically similar regions on faces. A simple barycentric coordinate mapping between the corresponding triangles of the two surfaces is used to transfer landmarks from the mask to the target face. The geometric surface of the same AM supplies a common reference between heterogenous faces. This means the algorithm can be used for generic landmark placement and it is especially suited to localising landmarks at positions where the local geometry is relatively uniform unlike methods that depend upon the presence of regions having curvature of a certain form *e.g.,* spline-patch based methods (Gilani et al., 2015). The algorithm has been shown to out-perform the automated landmark placement accuracy of other state-of-the-art methods (White et al., 2019).

The original implementation is available as an open source toolbox for MATLAB™called *MeshMonk*. This greatly improves its utility – especially among other researchers and those seeking to understand the statistical norms of the human face. However, software such as MATLAB™is too generic and unwieldy for most occasional users. The toolbox functionality is also targeted to supporting the non-rigid correspondence algorithm and its use in registering many faces against one another. As such there is a clear need to incorporate this algorithm into a platform that has been designed from the ground up to support facial anthropometrics and dysmorphological analysis.

2.1 Cliniface

We formally introduce *Cliniface*: sofware for interactively visualising, analysing, investigating anthropometrics, and detecting dysmorphological traits from 3D facial images. We have developed *Cliniface* through international collaboration with researchers and clinicians to provide them with a suite of tools to easily visualise and interrogate the 3D facial images of their patients or study subjects, irrespective of how the images are generated. With *Cliniface*, a user can interactively view a 3D facial image, take more than 50 different measurements, and view a report on whether any of the more than 40 different specific phenotypic facial traits of potential clinical significance listed in *Cliniface*'s database are present.

A reimplementation of the *MeshMonk* non-rigid correspondence algorithm is used in *Cliniface* to perform facial landmark detection. Some library dependencies are removed and processing speed is increased by more than 25% but it is in essence the same algorithm which is explained in detail in the originally published paper (White et al., 2019). In *Cliniface*, the algorithm's parameter tuning is modified to allow for acceptable accuracy (especially on very dysmorphic faces) but faster speeds befitting user expectations of an interactive application. The parameters remain tunable via *Cliniface*'s preferences to allow adjustment for research purposes or for improved registration accuracy on a case-by-case basis. *Cliniface* includes a suitable bilaterally symmetric AM for generic facial registration and landmarking (Ekrami et al., 2018) (used with permission) but *Cliniface* also allows researchers to easily incorporate and use their own AMs. Figure 1 shows left to right the non-rigid deformation of the AM (left) to an input target face (centre) and the result of transferring the landmarks from the deformed AM to the target face's original surface.

Cliniface includes curvature and facial asymmetry visualisations and two or more images can be viewed simultaneously to directly compare anthropometrics between faces, whether of the same or different individuals (contingent on memory limitations). Users can also take investigative measurements between any two points on the face (landmarks or not) including angles, depth, and both surface and straight-line distances. All analytic results including landmark positions and detections of atypical facial traits are saved alongside the original 3D image within the same file archive (a compressed format called a 3DF). Results can also be exported to XML or JSON formats for further analysis outside of *Cliniface* and PDF reports of the dysmorphological analysis can be produced containing a fully manip-

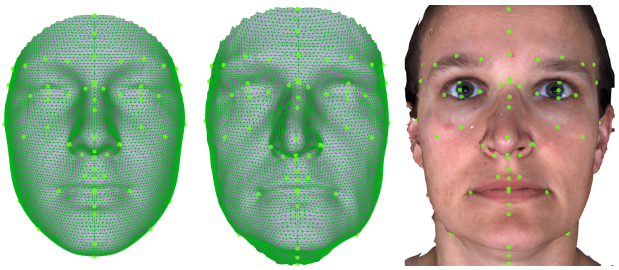

Figure 1. Non-rigid transformation of the AM with template landmarks (left) to a target face showing the resulting deformation (centre), followed by landmark transfer to the target face using barycentric coordinates (right).

ulable 3D model of the face (viewable using Adobe's Acrobat Reader™).

Cliniface has been built to support ongoing research in facial anthropometrics and dysmorphology using 3D images, so it has been designed around a centralised plugin architecture to simplify its continuing extensibility to new features. *Cliniface* does not send data offsite – all processing and analysis is performed client side – so it can be used in situations where privacy is of concern or in research without unduly complicating ethical considerations. The solution can travel to the data, rather than the data to the solution.

Cliniface is free and open source to help drive the uptake and utility of 3D imagery in the clinical setting and accelerate research into effective techniques of deriving diagnostic criteria from the facial phenotype. *Cliniface* is being continually improved upon to incorporate feedback and new features suggested by users; the latest version can be downloaded for Windows and Linux from https://cliniface.org.

2.2 Aims and Hypotheses

Automatically generated straight-line distance measures using *Cliniface* ought to be comparable in accuracy to those of an expert clinician because landmark positioning is handled by the non-rigid registration algorithm. The measurement endpoints depend upon the accuracy of landmark placement, however it is not necessarily the case that inaccurate automatic landmark placement by the non-rigid registration algorithm will entail inaccuracy of any given inter-landmark distance measure. This is because the error vector of the positions of both landmarks in the inter-landmark pair may be equivalent. Therefore, the accuracy of straight-line distance measures generated by *Cliniface* should be more tolerant to errors in landmark positioning and *Cliniface*'s measurement accuracy should be tested independently of the how accurately landmark positions are automatically placed.

We test empirically against 3D facial images taken from a diverse range of subjects generated by two different hardware capture systems. Measurement accuracy should not be impacted adversely because of the different 3D image generation systems used (assuming the systems meet an acceptable minimum standard of accuracy and detail in 3D image generation).

The measurement tools built-in to *Cliniface* and the static nature of the 3D facial image allow for a degree of finesse in measurement placement that is not possible when taking measurements from subjects' faces in real life. We test the accuracy of measurements ascertained through *Cliniface* if the automatic landmark registration process offers subjectively poor enough results that the user is willing to manually override the detected landmark positions so that measurement accuracy depends upon the user's placement of the landmarks.

In this study, manual measurements are taken from participants' faces by an experienced clinical geneticist. These are compared against the same measurements from 3D facial images of the same participants taken using *Cliniface*'s built-in measuring tool by non-expert assessors, and by measurements generated automatically by *Cliniface* using non-rigid registration. Four hypotheses are tested:

1. Measurements taken manually by the expert are not significantly affected by the choice of measuring device.
2. Measurements by non-experts using *Cliniface* are not significantly affected by how the images are generated.
3. Measurements by non-experts using *Cliniface* are not significantly different from expert manual measurements.
4. Measurements automatically made by *Cliniface* using non-rigid registration are not significantly different from expert manual measurements.

3. METHODOLOGY

Ten different inter-landmark straight-line distances involving ten different facial landmarks were selected for measurement. The landmarks were Nasion (N), Exocanthion (EX), Endocanthion (EN), Pronasale (PRN), Alar Curvature Point (AC), Subalare (SBAL), Subnasale (SN), Cheilion (CH), and Labrale Superius (LS). The standard anatomical positions of these landmarks are shown by the yellow dots next to their abbreviations on the example face in figure 2.

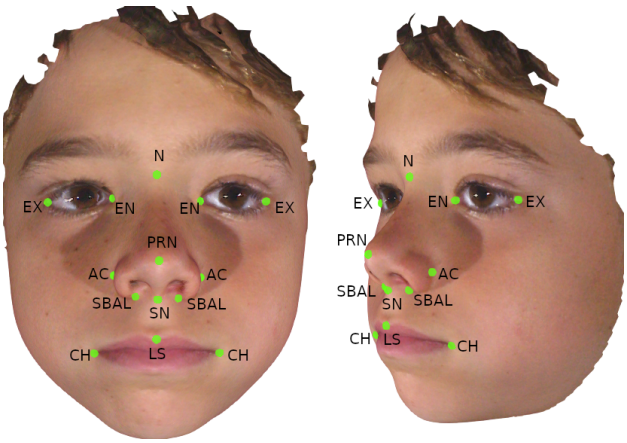

Figure 2. The standard anatomical positions of the landmarks used to define the distance metrics.

The list of measures used is shown in table 1 which shows the two landmarks defining the endpoints of the measure (L_1 and L_2), the measure name, and if it is bilateral (*i.e.,* measured on both sides of the face).

The landmarks and their associated measures were chosen for their relative ease of identification by the non-expert assessors

L_1	L_2	Measure Name	Bilateral?
EN	EN	Intercanthal Distance	-
EX	EX	Outercanthal Distance	-
EX	EN	Palpebral Fissure Length	Yes
CH	CH	Labial Fissure Length	-
SN	LS	Philtral Length	-
SBAL	SBAL	Subnasal Width	-
N	PRN	Nasal Bridge Length	-
AC	PRN	Nasal Ala Length	Yes

Table 1. The inter-landmark distance measures.

and for the range of practical difficulty in measuring between the landmark pairs.

Four non-experts (P1–P4) and one experienced clinical geneticist (CG) were tasked as assessors. To establish the baseline level of accuracy, the experienced assessor took the above ten measures directly from the faces of 25 different adult participants ranging in sex, age, and ethnicity using both a pair of Vernier callipers and a standard tape measure. The aim of using both devices was to identify the degree (if any) of measurement bias due to the use of different measuring devices.

The 25 subjects had 3D images of their faces taken using two different 3D image capture systems: the 3dMD static system in the dual camera module configuration (3dMD Ltd, London, UK), and the Vectra H1 handheld imaging system (Canfield Scientific, New Jersey, USA). Each non-expert assessor was tasked with successively viewing each of the 50 facial images produced by these systems in *Cliniface* and using its inbuilt measurement tool to interactively drag the endpoints of each measure into position. Figure 3 shows *Cliniface*'s interface while using its "virtual callipers" to place the endpoints of a measure between two arbitrary points on an example face.

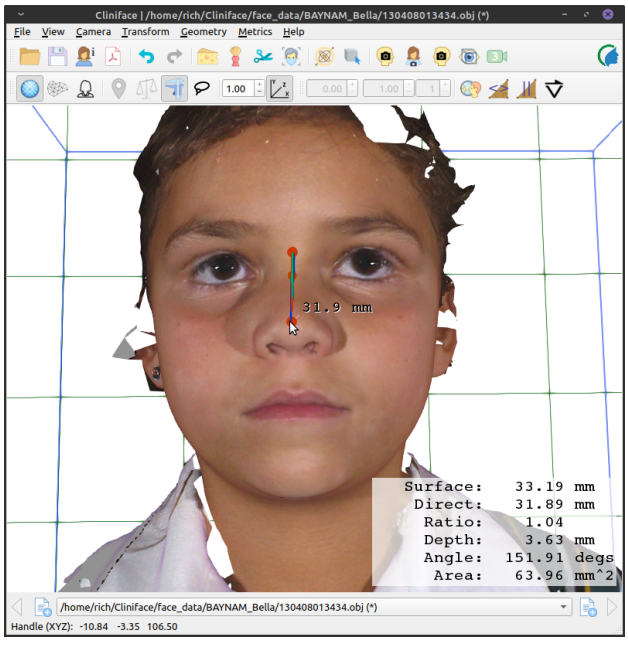

Figure 3. *Cliniface*'s user interface during interactive placement of the nasal bridge measure.

Prior to this, each assessor was briefed on the nature of the task and shown images (similar to figure 2) to advise them of how to identify and localise the standard anatomical positions of the landmarks. Assessors interpreted how best to place the inter-landmark measurement endpoints in accordance with this guidance for each of the 3D facial images.

Finally, the 50 facial images were sequentially loaded into *Cliniface* and the automated landmark detection algorithm was used to record the same set of measurements from each face without any user intervention or readjustment of the detected landmark positions.

The same 3D images were used across all assessments by the non-experts and in *Cliniface* to undertake automated landmark positioning and measurement. After taking the images of the participants and before their use in the experiments, some of the facial models in the images were repositioned slightly to ensure proper centring within *Cliniface*'s viewer. Aside from this, no other pre-processing of the images was performed; the surface geometries were left as captured by the respective 3D image capture systems. This was not ideal for the automated landmark detection algorithm as non-face areas such as the neck or shoulders can potentially confuse the registration algorithm, but it was deemed acceptable to standardise and simplify the protocol for this study. Figure 1 shows *Cliniface*'s facial registration and unadjusted landmark mapping on one of the authors. This shows the full set of landmarks detected by *Cliniface* but only those shown in figure 2 were used in this study.

4. RESULTS AND ANALYSIS

The first two hypotheses in section 2.2 concern how the facial measures are taken (*i.e.,* the mode of measurement). The second two hypotheses concern how accurately non-experts and *Cliniface* itself can take the facial measures. To test the second two hypotheses, the baseline set of measurements for comparison against first had to be established. The analysis for this determination is given in section 4.1.1. The second two hypotheses are tested with reference to this baseline in section 4.2.

4.1 Measurement Mode

The violin plots in figure 4 show the distributions of measurement differences over all participants due to the use of different measuring devices by the expert assessor (EX) shown in red, or due to the use of differently generated 3D images by the non-expert assessors (P1–P4) shown in blue, and by *Cliniface* (CF) shown in green. Plot EX shows measurement differences by the expert as callipers *minus* tape measure. Plots P1–P4, and CF show measurement differences as 3dMD *minus* Vectra H1. The plots show the median value (centre vertical bar), the interquartile range (thick black bar), and the extreme values (end vertical bars).

4.1.1 Callipers versus Tape Measure The use of two different devices by the expert assessor resulted in large and consistent measurement differences. The RMSE in measurement over all participants, was found to be 3.82 mm. Paired samples t-testing confirmed clear rejection of the null hypothesis at $\alpha = 0.05$ *i.e.,* accuracy was significantly affected by the choice of measuring device.

The expert assessor reported that of the two devices, they were more comfortable using the tape measure. For any particular measure, discomfort in using one device over the other might appear as increased variance over all participants for one of the devices. F-testing was performed over all participants for each

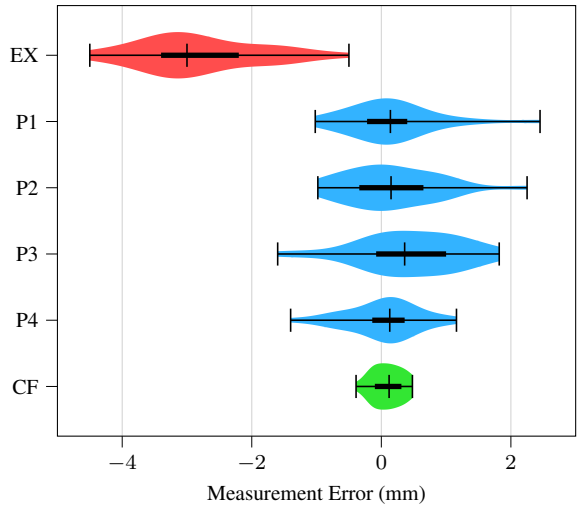

Figure 4. Distributions of measurement differences over participants due to the expert's use of measuring device (red), 3D image source by the non-experts using *Cliniface* for manual measurement (blue), and 3D image source by *Cliniface* using non-rigid registration to derive measurements (green).

of the expert's measurements. The null hypothesis was formed as: for a given measure, measurement variance between the two devices is not significantly different. Rejection of the null hypothesis would imply decreased precision for some reason – possibly due to a lack of familiarity with the device. However, at $\alpha = 0.05$, this hypothesis was not rejected for any of the measures. Paired samples t-testing for the *individual* measures was also performed, and together with the f-testing results this confirmed that for every one of the ten measures, accuracy was significantly affected by device selection while precision was not.

Since the nature of the error was found to be one of accuracy rather than precision, the expert determined using his experience that the tape measure derived measurements were in fact a better reflection of the true distances and should be used as the baseline set for comparison against.

4.1.2 3dMD versus Vectra H1 In figure 4, plots P1–P4, and CF show differences as the measurements obtained from the 3dMD images subtracted by the those same measurements obtained from the Vectra H1 images for the four non-experts and *Cliniface* respectively. Across all assessors (including *Cliniface*), there is a small but consistent bias resulting in the distances obtained from the 3dMD images to be slightly longer. The RMS errors for assessors P1–P4, and *Cliniface* are respectively: 1.77, 2.05, 2.11, 1.50, and 0.90 [mm].

To test the second hypothesis, paired samples t-tests were performed. At $\alpha = 0.05$, this hypothesis was rejected by assessor P3 with $P = 0.001 < \alpha$. When measurement differences were averaged together over assessors P1–P4, the hypothesis was rejected with $P = 0.025 < \alpha$. Upon further investigation, it was found that human assessors had more trouble placing the labial fissure distance measure than any other. The mean RMS error over all measures was found as 1.30 mm with standard deviation of 0.57 mm. The labial fissure measure with RMSE of 2.86 mm was found to be an outlier with sigma of 2.7. Figure 5 shows the most severe example of this issue where the labial fissure measure (*i.e.,* between the mouth corners) is shorter in the Vectra H1 image on the right (and incorrectly placed) than in the 3dMD image on the left. Note the apparent difference in skin tone in the two images which may have had some bearing on the assessors' ability to determine the correct position of the mouth corners. Note also that the participant maintained a perfectly neutral mouth position in both images.

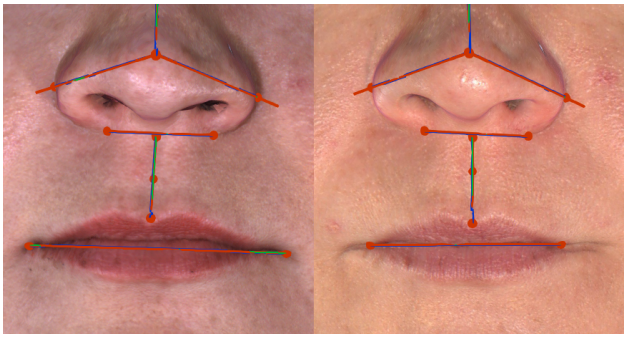

Figure 5. Inaccurate placement by an assessor of the labial fissue measure in the Vectra H1 image (right) while placement is accurate in the 3dMD image (left).

Once the most severe example of this issue was removed from the analysis (the participant's images shown in figure 5), even though the issue was still present in other participants' images, it was found that paired samples t-testing over the averaged together measurement differences of all four non-expert assessors no longer resulted in rejection of the hypothesis with $P = 0.123$. That is, in aggregate, it was found that there were no significant differences in measurement accuracy when switching between the 3dMD and Vectra H1 generated images. However, the hypothesis was still rejected in the case of assessor P3. This appears to be a human assessor dependent issue however because difficulties placing particular measures were not observed in the automatically obtained *Cliniface* measurements. In the case of *Cliniface*, the RMSE for measurement of the labial fissure was not significantly different from the other individual measurement RMS errors, and measurement variance in general due to image source difference was much narrower as seen in the bottom plot of figure 4.

For the remainder of the experiments, the problematic participant's images were retained because while posing a problem to some assessors when measuring certain features, it was felt that the images were still typical of what might be generated under normal circumstances.

4.2 Measurement Accuracy

In this section, the facial measurements taken by the non-expert assessors P1–P4, and the automatically generated measurements of *Cliniface* are compared against the expert assessor's baseline measurements decided upon in section 4.1.1. Due to the similarity of measurements obtained from the 3dMD and Vectra generated images for each participant and there being no particular reason to favour one 3D image type over the other, each assessor's measurements from both sets of images were averaged together for comparison against the expert assessor's.

The violin plots in figure 6 show the distributions of measurement differences from the expert's over all participants for assessors (P1–P4) shown in blue, and *Cliniface* (CF) using non-rigid registration shown in green.

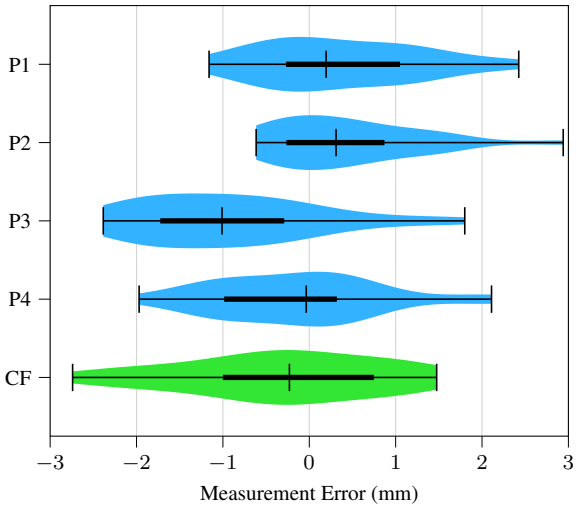

Figure 6. Distributions of measurement differences from the expert assessor's over participants for the non-experts using *Cliniface* for manual measurement (blue), and *Cliniface* using non-rigid registration to derive measurements (green).

Paired samples t-testing was performed with $\alpha = 0.05$ to evaluate the third null hypothesis for each non-expert assessor *i.e.,* that the measurements taken by the assessor using *Cliniface* are not significantly different in accuracy to those taken by the expert. For assessors P1, P2, and P4, P values of 0.139, 0.087, and 0.564 respectively were obtained – failing to reject the null hypothesis. For assessor P3, a P value of 0.001 was obtained and the null hypothesis was rejected. This means that the accuracy of measurements taken using *Cliniface* when *not* relying upon its automated landmark detection, is still very dependent upon the assessor's understanding of how to position the landmarks on the given face.

Paired samples t-testing with $\alpha = 0.05$ was also used to evaluate the fourth and final hypothesis *i.e.,* that the automatically generated measurements of *Cliniface* are not significantly different in accuracy to those taken by the expert. A P value of 0.282 was obtained – failing to reject the null hypothesis and providing support to the conclusion that *Cliniface* is able to automatically generate measurements of facial features at a level of accuracy comparable to an expert.

RMS errors for assessors P1–P4 were calculated as 4.02, 4.10, 4.20, and 3.50 [mm]. The RMSE of *Cliniface*'s measurements was calculated as 4.87 mm. F-testing on *Cliniface*'s measurement errors versus an average of the errors over assessors P1, P2, and P4 (not P3 due to the lack of accuracy) revealed that the variance in measurement error by *Cliniface* was not significantly different to the variance in error by the three human assessors for every individual measure except subnasal width ($P = 0.02 < \alpha = 0.05$).

5. DISCUSSION

We found that the majority of non-expert assessors using *Cliniface* had comparable measurement accuracy to the expert using direct manual measurement. Additionally, we demonstrated that the accuracy of measurements generated by *Cliniface* was similar to the expert's measurements, and that the precision of *Cliniface*'s automatically generated measurements was similar to those of the accurate non-expert human assessors.

It should be noted that the results of comparisons between the accuracy of measurement by non-expert assessors and *Cliniface* depend upon the objective veracity of the measurements obtained by the expert. The comparison against a single expert is a study limitation which could be addressed in further studies, including to further assess interrater reliability, accuracy, and precision of manual measurements and for comparisons of multiple measurement approaches – both manual and virtual.

The analysis in section 4.1.1 did not demonstrate an improved precision of measurement by the expert when using callipers over the tape measure. This is interesting because intuitively the callipers offer improved accuracy due to their more reliable construction and the fact that, unlike the flexible tape measure, the callipers will always report the distance of a straight line. Since it is also the case that callipers are typically indicated as the preferred method for taking many anthropometric measurements (Karen W. Gripp, Allanson, 2013), this aspect of the study – though unrelated to *Cliniface* – should be investigated further.

The analysis in section 4.1.2 found that averaged over assessors there was no significant difference due to the images ascertained by different image capture systems. However, for an individual assessor (P3) the choice of image mattered. This may be a stochastic effect, or it may point to a meaningful difference when assessing selected facial measurements. Human qualitative assessment of the captured images may remain an important complementary part of the workflow. Other methods of generating 3D images are available and will be developed in the future and the results of this study do not extend to 3D images in general.

In section 4.2, the analysis found that for the facial images tested, *Cliniface*'s automatically generated measurements were comparable in their accuracy to those taken by the expert. It was also found that suitably guided human assessors could take measurements using *Cliniface* at the expert level of accuracy (excepting assessor P3). The precision in *Cliniface*'s measurements was also found to be similar in all but one measure to those human assessors who matched in accuracy.

In total, it took under 45 minutes for *Cliniface* to sequentially process and annotate the 50 3D images of the participants – less than one minute per image. For the human assessors, the task took at least four times longer. In addition, some of the human assessors needed to review their assessments later due to accidentally skipping an image, or failing to place a measurement. *Cliniface* is more reliable in its ability to take a comprehensive set of measurements.

It is important to note that the observed accuracy in the distance measurements automatically generated by *Cliniface* does not imply accuracy of the landmark endpoints for each of the measures due to possible equivalent errors for inter-landmark pairs. Figure 7 shows an example of this where the mouth/lip landmarks are placed lower than they should be but the nasal landmarks are placed correctly on the participant's face (left). This results in the horizontal labial fissure length remaining reasonably accurate while the accuracy of the vertical philtral length measurement is reduced. This is due to the poor non-rigid registration of the AM to the participant's mouth (right most image).

In normal use, *Cliniface*'s workflow addresses the need to improve the accuracy of automatic landmark placement by prompt-

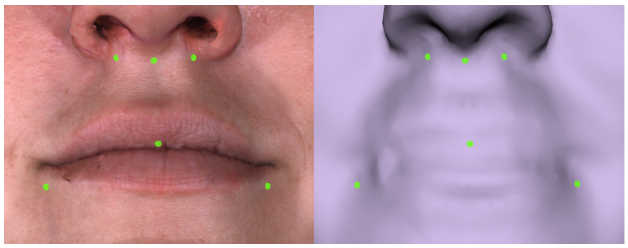

Figure 7. Incorrect placement of mouth landmarks on the participant's face (left) due to poor non-rigid registration of the mouth (right) without causing measurement error in the labial fissure length.

ing the user after automated facial registration and landmarking to confirm landmark positions using a dialog that describes the standard anatomical landmark positions. For this study, prompting the assessors to check landmark placement on a per image basis was not undertaken and so it was possible for the assessors to make mistakes in measurement placement. Errors of the kind shown in figure 7 would be easier to avoid using *Cliniface* under normal circumstances.

Finally, it is likely that *Cliniface* can obtain more accurate measurements if the user has the expertise to fine-tune the parameters of the non-rigid registration algorithm. The version of the algorithm used in *Cliniface* is tuned to trade some accuracy for speed.

6. CONCLUSION

In this study, the accuracy of measures placed on 3D facial images by both non-expert human assessors and an automated algorithm were evaluated against a baseline set of expert measurements. *Cliniface*: a novel interactive software application for the visualisation, extraction, and analysis of measurements from 3D facial images was introduced which provided both the means for the non-expert assessors to take measurements, and also the means to automatically generate measurements from 3D facial images without user intervention.

We found that three out of four of the non-expert assessors were able to take measurements using *Cliniface* at a level of accuracy comparable to the expert for the measures under evaluation. Also that the measurements automatically generated by *Cliniface* were similar in accuracy to the measurements obtained by the expert, and that the precision of *Cliniface*'s automatically generated measurements were similar to those of non-expert human assessors.

This study's conclusions are tempered by the lack of certainty concerning the objective veracity of the expert assessor's measurements; future research should first try to more accurately ascertain the true objective measurements from the 3D facial images under evaluation. This can be achieved by using more expert assessors so that confidence intervals of accuracy can be obtained. This will also allow for a study of measurement reliability to be undertaken which should be prioritised to properly evaluate the use of *Cliniface* going foward for investigative research into new and diagnostically useful 3D facial features.

ACKNOWLEDGEMENTS

We are extremely grateful to all the participants who volunteered their time so willingly for this research: Aasta Kelly, Alexis Hunt, Anne Harvey, Ben Kamien, Bhavya Vora, Brenely Vargas, Cassie Dowson, Cassie Greer, Cathy Kirali-Borri, Dian Karina, Fiona Baldacchino, Fiona McKenzie, Helen Mountain, Isabelle Dinu, Jackie Soraru, Jennifer Reemeijer, Karen Harrop, Kate Mountain, Lauren Dreyer, Mandy McShane, Marnie Russell, Michelle Ward, Sarah Long, Sharron Townshend, and Sophia Skoglei.

Special thanks to Lyn Schofield and Dylan Gration for providing essential organisational support. We also acknowledge and thank other *Cliniface* team members past and current including Paula Fievez, Cathryn Poulton, Yarlalu Thomas, Stefanie Kung, Tracey Tsang, and Hedwig Verhoef.

Due to this being the first academic publication focusing on *Cliniface*, we would like to use this opportunity to thank the following groups who have either helped fund *Cliniface*'s development, or have provided valuable consultation and feedback: Angela Wright Bennet Foundation, Curtin University, FASD Research Australia Centre for Research Excellence, FrontierSI (formerly Cooperative Research Centre for Spatial Information), Genetic Services of Western Australia, Genome.One, Linear Clinical Research, McCusker Charitable Foundation, Murdoch Children's Research Institute, Patches Paediatric Clinics, Perth Children's Hospital Foundation, RD-Connect, Roy Hill Community Foundation, Telethon Kids Institute Genetic and Rare Diseases Program, the Western Australian Register of Developmental Anomalies, and the Western Australian Department of Health.

Cliniface is free and open source software. It depends upon the innovations and efforts of the many hundreds of researchers, developers, and maintainers of other open source software products and technologies. We salute your benevolence.

REFERENCES

Baynam, G., Pachter, N., McKenzie, F., Townshend, S., Slee, J., Kiraly-Borri, C., Vasudevan, A., Hawkins, A., Broley, S., Schofield, L., Verhoef, H., Walker, C. E., Molster, C., Blackwell, J. M., Jamieson, S., Tang, D., Lassmann, T., Mina, K., Beilby, J., Davis, M., Laing, N., Murphy, L., Weeramanthri, T., Dawkins, H., Goldblatt, J., 2016. The rare and undiagnosed diseases diagnostic service – application of massively parallel sequencing in a state-wide clinical service. *Orphanet Journal of Rare Diseases*, 11(1), 77. https://doi.org/10.1186/s13023-016-0462-7.

Baynam, G., Walters, M., Claes, P., Kung, S., LeSouef, P., Dawkins, H., Gillett, D., Goldblatt, J., 2013. The Facial Evolution: Looking Backward and Moving Forward. *Human Mutation*, 34(1), 14-22. https://doi.org/10.1002/humu.22219.

Chui, H., Rangarajan, A., 2003. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2), 114 - 141. https://doi.org/10.1016/S1077-3142(03)00009-2. Nonrigid Image Registration.

Claes, P., Walters, M., Clement, J., 2012. Improved facial outcome assessment using a 3D anthropometric mask. *International Journal of Oral and Maxillofacial Surgery*, 41(3), 324 - 330. https://doi.org/10.1016/j.ijom.2011.10.019.

Ekrami, O., Claes, P., White, J. D., Zaidi, A. A., Shriver, M. D., Van Dongen, S., 2018. Measuring asymmetry from high-density 3D surface scans: An application to human faces. *PLOS ONE*, 13(12), 1-17. https://doi.org/10.1371/journal.pone.0207895.

Ferry, Q., Steinberg, J., Webber, C., FitzPatrick, D. R., Ponting, C. P., Zisserman, A., Nellåker, C., 2014. Diagnostically relevant facial gestalt information from ordinary photos. *eLife*, 3, e02020-e02020. https://doi.org/10.7554/eLife.02020.

Gilani, S. Z., Shafait, F., Mian, A., 2015. Shape-based automatic detection of a large number of 3d facial landmarks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4639–4648.

Global Genes, 2020. RARE Facts. Global Genes website. https://globalgenes.org/rare-facts (30 April 2020).

Hammond, P., Suttie, M., 2012. Large-scale objective phenotyping of 3D facial morphology. *Human Mutation*, 33(5), 817-825. https://doi.org/10.1002/humu.22054.

Hammond, P., Suttie, M., Hennekam, R. C., Allanson, J., Shore, E. M., Kaplan, F. S., 2012. The face signature of fibrodysplasia ossificans progressiva. *American Journal of Medical Genetics Part A*, 158A(6), 1368-1380. https://doi.org/10.1002/ajmg.a.35346.

Karen W. Gripp, Anne M. Slavotinek, J. G. H., Allanson, J. E., 2013. *Handbook of Physical Measurements 3rd Edition*. Oxford University Press, 198 Madison Avenue, New York, NY 10016 USA.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Lauledeerkind, S. J., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Lourghi, H., Della Rocca, M. G., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., Robinson, P. N., 2018. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1), D1018-D1027. https://doi.org/10.1093/nar/gky1105.

Myronenko, A., Song, X., 2010. Point Set Registration: Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12), 2262-2275. https://doi.org/10.1109/TPAMI.2010.46.

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., Rath, A., 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2), 165-173. https://doi.org/10.1038/s41431-019-0508-0.

Poulton, C., Azmanov, D., Atkinson, V., Beilby, J., Ewans, L., Gration, D., Dreyer, L., Shetty, V., Peake, C., McCormack, E., Palmer, R., Lewis, B., Dawkins, H., Broley, S., Baynam, G., 2018. Silver Russel syndrome in an aboriginal patient from Australia. *American Journal of Medical Genetics Part A*, 176(12), 2561-2563. https://dot.org/10.1002/ajmg.a.40502.

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., Cunningham, M. L., Hecht, J. T., Kau, C. H., Nidey, N. L., Moreno, L. M., Wehby, G. L., Murray, J. C., Laurie, C. A., Laurie, C. C., Cole, J., Ferrara, T., Santorico, S., Klein, O., Mio, W., Feingold, E., Hallgrimsson, B., Spritz, R. A., Marazita, M. L., Weinberg, S. M., 2016. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLOS Genetics*, 12(8), 1-21. https://doi.org/10.1371/journal.pgen.1006149.

White, J. D., Ortega-Castrillón, A., Matthews, H., Zaidi, A. A., Ekrami, O., Snyders, J., Fan, Y., Penington, T., Van Dongen, S., Shriver, M. D., Claes, P., 2019. MeshMonk: Open-source large-scale intensive 3D phenotyping. *Scientific Reports*, 9(1), 6085. https://doi.org/10.1038/s41598-019-42533-y.

Yang, J., 2011. The thin plate spline robust point matching (TPS-RPM) algorithm: A revisit. *Pattern Recognition Letters*, 32(7), 910 - 918. https://doi.org/10.1016/j.patrec.2011.01.01.