A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery

Paul G. M. Knoops1,2,3, Athanasios Papaioannou1,2,4, Alessandro Borghi1,2, Richard W. F. Breakey1,2, Alexander T. Wilson5, Owase Jeelani1,2, Stefanos Zaferiou4, Derek Steinbacher5, Bonnie L. Padwa3, David J. Dunaway1,2 & Silvia Schievano1,2

Current computational tools for planning and simulation in plastic and reconstructive surgery lack sufficient precision and are time-consuming, thus resulting in limited adoption. Although computer-assisted surgical planning systems help to improve clinical outcomes, shorten operation time and reduce cost, they are often too complex and require extensive manual input, which ultimately limits their use in doctor-patient communication and clinical decision making. Here, we present the first large-scale clinical 3D morphable model, a machine-learning-based framework involving supervised learning for diagnostics, risk stratification, and treatment simulation. The model, trained and validated with 4,261 faces of healthy volunteers and orthognathic (jaw) surgery patients, diagnoses patients with 95.5% sensitivity and 95.2% specificity, and simulates surgical outcomes with a mean accuracy of $1.1 \pm 0.3$ mm. We demonstrate how this model could fully-automatically aid diagnosis and provide patient-specific treatment plans from a 3D scan alone, to help efficient clinical decision making and improve clinical understanding of face shape as a marker for primary and secondary surgery.

Over 200,000 maxillofacial procedures, including orthognathic (jaw) surgery, are performed in the USA every year to treat a range of diseases, defects and injuries in the head, neck and face1. For these operations, vast quantities of patient data are collected2, thus providing a great opportunity for the development of machine-learning-based methods, for use in clinical decision-making and to enable automated personalised medicine approaches3,4. Although the application of machine learning in plastic and reconstructive surgery is not new – in orthognathic surgery it has been used to elucidate how syndromes affect skull growth5, to quantify6 or to predict7 the corrective effect of surgical techniques on skull deformities, and for outcome evaluation8 – its clinical usefulness has been limited due to the low number of samples, absence of automated processing methods, and lack of state-of-the-art mathematical models. Therefore, we propose a machine-learning-based framework involving a large number of data points and fully automated processing for diagnosis and clinical decision making in orthognathic surgery.

Medical imaging and computer-assisted surgical planning form an integral part of the preoperative workup9,10, as exploring various operative approaches in a virtual environment can reduce operation time11 and cost12, and facilitate more consistent and optimised surgery13. A digital patient model, reconstructed from computed tomography (CT) or magnetic resonance (MR) images14,15, can be manipulated by the surgeon to determine the optimal bone cuts (osteotomies) and bone position, and to simulate changes in face shape16,17. Based on this virtual patient-specific treatment plan, custom-made surgical wafers, cutting guides, plates and implants can be manufactured to deliver the plan18,19. Although 3D computer-assisted surgical planning has been around for over 30 years20, and there are known benefits over traditional 2D planning21, implementation in clinical practice has been

1 UCL Great Ormond Street Institute of Child Health, London, UK. 2 Craniofacial Unit, Great Ormond Street Hospital for Children, London, UK. 3 Department of Plastic and Oral Surgery, Boston Children’s Hospital & Harvard School of Dental Medicine, Boston, MA, USA. 4 Department of Computing, Imperial College London, London, UK. 5 Department of Plastic and Reconstructive Surgery, Yale University School of Medicine, New Haven, CT, USA. Paul G. M. Knoops and Athanasios Papaioannou contributed equally. Correspondence and requests for materials should be addressed to S.S. (email: s.schievano@ucl.ac.uk)
Our models were characterised and validated with the following intrinsic metrics: compactness, generalisation, and specificity (see Methods). Additionally, we benchmarked our models to the LSFM database (see Methods) which contains high-dimensional patient and volunteer shape data; (3) a manifold visualisation to compare high-dimensional patient and volunteer shape data; (4) a classification for automated diagnosis of faces with orthognathic features as an indicator for orthognathic surgery, and (5) an analysis of different regression techniques to simulate patient faces for automated patient-specific surgical planning.

**Results**

In this section, to demonstrate the power of our large-scale clinical 3DMM, we present the following results: (1) a description of how the models were built and the 3D face databases used, including intrinsic statistical validation metrics; (2) an evaluation of mean face shape to compare, quantitatively and qualitatively, how patient faces differ from volunteer faces preoperatively and postoperatively; (3) a manifold visualisation to compare high-dimensional patient and volunteer shape data; (4) a classification for automated diagnosis of faces with orthognathic features as an indicator for orthognathic surgery, and (5) an analysis of different regression techniques to simulate patient faces for automated patient-specific surgical planning.

**Model construction and validation.** Two databases were used to build the 3DMMs, comprising volunteer and patient 3D face scans. For the volunteer faces, we used the LSFM database (see Methods) which contains 9,663 3D face scans from the general public with mean age 24.5 ± 14.7 years, 52% female, and 82% white heritage (Table 1). For the patient faces, 274 3D scans were retrospectively selected from a database of 151 patients who underwent orthognathic procedures at Boston Children’s Hospital and Yale-New Haven Hospital, with mean age at surgery 18.4 ± 2.4 years, 52% female, and 76% white heritage (Table 1). Additional patient demographics are summarised in Supplementary Fig. 1.

| Characteristics                  | Dataset                        | LSFM dataset (all ages) | LSFM subset (14–28 yr) |
|----------------------------------|--------------------------------|-------------------------|------------------------|
| Number of subjects               | Patient                        | 151                     | 9,663                  | 3,943                  |
| Number of images                 |                                | 273                     | 9,663                  | 3,943                  |
| Age: mean (s.d.), years          | 18.4 (2.4), n = 151            | 24.5 (14.7), n = 9,460  | 22.2 (3.7), n = 3,943  |
| Age: range, years                | 14–28, n = 151                 | 0–85, n = 9,460         | 14–28, n = 3,943       |
| Gender (% male/female)           | 44%/56%, n = 151               | 47.7%/52.2%, n = 9,582  | 55.5%/44.5%, n = 3,942 |
| Ethnicity                        | 72% White, 10% Asian, 10% Mixed Heritage/Other, 8% Black, n = 151 | 82% White, 9% Asian, 5% Mixed Heritage, 3% Black, 1% Other, n = 9,554 | 83% White, 8% Asian, 4% Mixed Heritage, 3% Black, 2% Other, n = 3,928 |

Table 1. Characteristics of patient and volunteer faces in the databases. n is the number of individuals for whom that measurement was available. s.d. indicates the standard deviation.

limited to highly specialised hospitals, mainly due to the complexity of commercial software and the contested planning accuracy.

Machine-learning-based models, including statistical shape models, have been proposed to streamline and automate processes in computer-assisted surgical planning, thereby making this a more accessible technology. However, accurate statistical modelling of face shape features is a challenging task due to the large anatomical variation in the human population; and to build a statistical model that can truthfully represent each given face, a large collection of high-quality 3D images is required from a population diverse in age, gender, and ethnicity. State-of-the-art computer vision algorithms are required to process these 3D images and construct a high-dimensional statistical model. A popular machine learning approach, originally used to reconstruct accurate and complete 3D representations from single 2D images and for photo-realistic manipulation, involves 3D morphable models (3DMM) – statistical models of face shape and texture. Current applications of 3DMM include facial recognition, expression normalisation, and face reconstruction from videos.

Here, we present the first fully-automated large-scale clinical 3DMM involving supervised learning for diagnostics, risk stratification, and treatment simulation. Using databases comprising 10,000 3D face scans of healthy volunteers and patients admitted for orthognathic surgery, we trained and validated a 3DMM, and demonstrated its potential for clinical decision making, including fully-automated diagnosis and surgery simulation. We believe our proposed model is an important step towards making computer-assisted surgical planning cheaper, and more accessible for surgeons and patients. This model could potentially transform patient-specific clinical decision making in orthognathic surgery and other fields of plastic and reconstructive surgery.

In this section, to demonstrate the power of our large-scale clinical 3DMM, we present the following results: (1) a description of how the models were built and the 3D face databases used, including intrinsic statistical validation metrics; (2) an evaluation of mean face shape to compare, quantitatively and qualitatively, how patient faces differ from volunteer faces preoperatively and postoperatively; (3) a manifold visualisation to compare high-dimensional patient and volunteer shape data; (4) a classification for automated diagnosis of faces with orthognathic features as an indicator for orthognathic surgery, and (5) an analysis of different regression techniques to simulate patient faces for automated patient-specific surgical planning.

**Model construction and validation.** Two databases were used to build the 3DMMs, comprising volunteer and patient 3D face scans. For the volunteer faces, we used the LSFM database (see Methods) which contains 9,663 3D face scans from the general public with mean age 24.5 ± 14.7 years, 52% female, and 82% white heritage (Table 1). For the patient faces, 274 3D scans were retrospectively selected from a database of 151 patients who underwent orthognathic procedures at Boston Children’s Hospital and Yale-New Haven Hospital, with mean age at surgery 18.4 ± 2.4 years, 52% female, and 76% white heritage (Table 1). Additional patient demographics are summarised in Supplementary Fig. 1.

We trained three 3DMM (see Methods): a global model, a bespoke preoperative model, and a bespoke postoperative model. The global model (n = 4,216) comprised all patient scans as well as volunteer scans from the same age range (Table 1). The bespoke preoperative (n = 119) and postoperative (n = 127) models were made exclusively with patient scans – ‘bespoke’ refers to the fact that these models are custom made to represent preoperative or postoperative patient faces.

Our models were characterised and validated with the following intrinsic metrics: compactness, generalisation, and specificity (see Methods). Additionally, we benchmarked the performance of our models to the large-scale facial model (LSFM), a state-of-the-art 3DMM constructed with 9,663 scans. Compactness showed that 81.8% and 91.6% of the variance are respectively described by the first 10 and 20 principal components for the bespoke preoperative model, 79.6% and 89.3% for the global model, and 79.9% and 89.3% for LSFM (Fig. 1a). The generalisation error demonstrated the ability to describe patient faces that were not used for training. Using leave-one-out cross-validation, at 100 components, we found that the global (0.3 mm) and bespoke preoperative (0.4 mm) models outperformed LSFM (1.4 mm), due to lack of patient data in the latter (Fig. 1b). Moreover, the bespoke preoperative model initially outperformed the global model, but after 48 components this trend reversed, as the bespoke model ran out of statistical variance sooner due to a lower number of samples. For specificity,
we synthesised faces \((n = 10,000)\) and compared them to their closest real neighbour. Values in the range of 0.3–0.4 mm quantitatively indicated good agreement with real faces (Fig. 1c,d).

**Qualitative and quantitative shape evaluation.** To investigate how the three models differed, we qualitatively and quantitatively evaluated the shape and variance. Specifically, the mean shape and first five principal components with standard deviation \((\sigma_i)\) of \(+3\sigma_i\) and \(-3\sigma_i\) were computed (see Methods), and differences between the mean shapes were calculated. In the average volunteer face (Fig. 2a) and in the postoperative face (Fig. 2c), lengthening-widening (component 1) and concavity-convexity (component 2) captured most variance, whilst in the mean preoperative face (Fig. 2b), a component of under-overdevelopment of the upper and lower jaw (component 2) was present. These differences were confirmed by a direct comparison of the mean preoperative face to the mean volunteer face (Fig. 3), revealing maxillary hypoplasia (underdevelopment of the upper jaw) and mandibular hyperplasia (overdevelopment of the lower jaw) preoperatively in our patient cohort. The operation successfully ameliorated the jaw discrepancy but some nose orthognathic shape features remained postoperatively (Fig. 3).

**Manifold visualisation.** To test the diagnostic potential, we used t-SNE for dimensionality reduction of the high dimensional shape vectors in order to visualise the global manifold in two dimensions with different hyper-parameters (see Methods, Supplementary Fig. 2). With labels for volunteer, preoperative, and postoperative, no distinct groups were uncovered (Fig. 4a). To elucidate similarity amongst neighbouring faces, we display...
a patient’s face (Fig. 4b) that is close to two volunteer faces (Fig. 4c,d) in the t-SNE embedding, showing resemblance in the facial profile and particularly in the upper lip area. Patient faces that often appeared to be close to the average volunteer face in the classification experiment, as detailed in the next paragraph, are also displayed (Fig. 4e–g). Although the majority of patient faces, preoperatively and postoperatively, appeared to populate the perimeter of the t-SNE embedding, no distinct groups were observed which suggests that patient faces and average volunteer faces, overall, demonstrated substantial shape similarity.

**Classification for diagnosis.** Classification was performed with all preoperative patient scans (n = 119) and randomly sampled subsets of volunteer face scans in the 14–28 age range (see Methods). Three different splits for training and testing were investigated for 1,000 iterations: a split of 80–20% between training and testing data provided overall classification accuracy of 95.4% (Fig. 5a). Patient faces were diagnosed with 95.5% sensitivity and 95.2% specificity, and a positive and negative predictive value of 87.5% and 98.3%, respectively (Fig. 5b). False negatives – patient faces incorrectly labelled as being from the volunteer sample (Figs 4e–g, 5c) – were observed.
with a frequency of 12.5% and 4 patients were incorrectly classified in more than 200 out of 1,000 iterations. False positives – volunteer faces incorrectly labelled as patients (Fig. 5d) – were less common with an occurrence of 1.7%, and 3 patients were incorrectly classified in more than 150 out of 1,000 iterations.

**Regression for surgery simulation.** To demonstrate the automated simulation of the postoperative face shape, we tested linear regression (LR), ridge regression (RR), least-angle regression (LARS), and least absolute shrinkage and selection operator regression (LASSO) on our global model (see Methods). The average
error between the predicted shape and the ground-truth postoperative shape, at 100 components, was lowest with LARS (1.1 ± 0.3 mm) and RR (1.1 ± 0.3 mm), followed by LASSO (1.3 ± 0.3 mm) and LR (3.0 ± 1.2 mm) (Fig. 6a), which is as accurate as traditional computer-assisted surgical planning methods23,24. Using more than 40

Figure 5. Classification of preoperative patient and non-patient faces. (a) A split of 80–20% provides 95.4% classification accuracy at 96 principal components, superior to other splits. (b) Average confusion matrix, obtained from classification using preoperative patient scans (n = 140) and randomly selected non-patient scans (n = 280), representing the average of 1,000 iterations. With an 80–20% split, patient (n = 112) and non-patient (n = 224) scans were used for training and patient (n = 28) and non-patient (n = 56) for testing. Orthognathic shape features were diagnosed with 95.5% sensitivity and 95.2% specificity, and with a positive and negative predictive value of 87.5% and 98.3%, respectively. (c) For 1,000 iterations, 4 unique patient scans were classified as false positive in more than 200 out of 1,000 iterations, and (d) 3 unique volunteer scans were classified as false negative in more than 150 out of 1,000 iterations, suggestive of untreated patients in our volunteer sample.
components, LR exhibited overfitting which reduced its generalization beyond the training data. To demonstrate the quality of our patient-specific predictions, the differences between preoperative, postoperative and simulated, were visualised and quantified (Fig. 6b). To check that predictions were indeed patient-specific rather than mimicking the population mean, all simulated faces ($n = 113$) were additionally compared to the mean global face and mean bespoke postoperative face (Supplementary Fig. 3). At 100 components, the difference between RR simulations is much smaller compared to the postoperative 3D scan (1.1 mm, see above) than compared to the mean global face (1.8 mm) and the mean bespoke postoperative face (1.6 mm).

Discussion

Although there has been great interest in the use of machine learning in plastic and reconstructive surgery, a lack of data and complex interpretability currently limit its adoption in routine clinical practice\(^{36}\). In this study, we have introduced a novel approach involving 3DMM trained with 4,216 3D. Using a state-of-the-art computer vision framework\(^{37}\), we designed our model that comprehensively integrates high-quality 3D scans to automatically classify orthognathic patient faces and faces from volunteers – as an indication if someone should be seen by a specialist based on their aesthetics – and to automatically predict the patient-specific postoperative outcome. Our model can help objective assessment of preoperative and postoperative face shape, which may help inform patients better during a medical consultation. Additionally, this approach provides a goal-driven surgical planning approach for the surgeon.

Figure 6. Regression for postoperative face shape simulation. (a) Overall error of the ground-truth postoperative face shape compared to the simulated shape using ridge regression (RR), linear regression (LR), least-angle regression (LARS), and least absolute shrinkage and selection operator regression (LASSO). (b) Mean and standard deviation for ridge regression: at 100 components, average Euclidean distance = 1.1 mm, s.d. = 0.3 mm. (c) The best simulated face, with an error of 0.6 mm between the simulated shape and the postoperative face shape (d). The worst simulated face, with an error of 2.4 mm as compared to the postoperative face shape.
Our machine learning approach has several important advantages over computer-assisted surgical planning with traditional software. Conventional surgical simulation is a time-consuming explorative process in which the surgeon manually tests various procedural approaches and assesses the optimal osteotomy and bone position. Our approach accurately and automatically predicts the postoperative face shape and reduces the planning process to a single step. However, it does leave the surgeon to decide on the appropriate surgical procedure that delivers the simulated face shape since the surface scans do not contain volumetric bone data. To automatically project the necessary bone movements, a mathematical method that can inversely deduce the modification to the skull to achieve a soft tissue shape, or a combined soft tissue-skeletal model will need to be implemented. However, these models would require a large number of head CT or MR images, thus renouncing the advantages that 3D surface imaging has over volumetric imaging methods. Moreover, large CT and MR image databases are currently not available.

Considering our results, the average models showed an interesting difference between the average volunteer and the postoperative face shape. Whilst the operation successfully ameliorated the jaw discrepancy, some preoperative nose shape features postoperatively which is in line with known shape effects of Le Fort I advancement. Looking at the classification results, the false negative rate of 12.5% is undesirable as real patients would be missed by the model. We partly attribute this to multifactorial indications for surgery, as previous reports showed that aesthetics is the primary driver for surgery in only 71% of patients. Thus, shape alone may not be the main consideration for at least 12.5% of orthognathic surgery patients. To further improve on the performance of the model, additional scans should be collected, and stricter selection criteria should be adopted to exclude volunteers with mild functional and aesthetic indications for orthognathic surgery.

Indeed, we propose that our model should be used as a machine-learning-based support tool in clinical decision-making, not to fully replace human assessment, and further improvements to the model can be made. Whilst the false positive rate of 1.7% is low, it is substantially larger than the incidence of craniofacial anomalies, including jaw malformation (1 in 1,600 live births) and the incidence of cleft lip and palate (1 in 700 live births, with about 20% requiring an operation later in life). Faces in the LSFM database were collected from the general population where subjects were not excluded for facial anomalies, or untreated functional issues volunteers’ jaws – as this information was not available. It should also be considered that orthognathic surgery has become a routine practice that some subjects may choose to undergo surgery for mild anomalies that are present within the general population, and no perfect binary classifier exists. The reported accuracy of computer-assisted orthognathic surgery simulation ranges from 0.5 mm to 2.0 mm, depending on the software used. Clinically meaningful predictions can be obtained with most commercial software, but intrinsic limitations restrict its use in doctor-patient communication. We demonstrated that our model performs within the range of traditional programs whilst being fully-automated. Although our model shows high sensitivity and specificity, the results also suggest there is scope for further improvements, for example by employing increasingly powerful algorithms and by having access to larger numbers of data.

Recently, in computer vision and machine learning, new modelling frameworks and algorithms have been developed that achieve remarkable success in various applications. Deep Neural Networks (DNNs) include Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have greatly impacted and increased the performance of automatic systems designed for speech recognition, visual object detection, scene recognition, and face recognition. Whilst it is likely that these models will be used for clinical application in the near future, including computer-assisted surgical planning, their big data requirement limits their current use, and approaches that handle 3D data are still relatively poor compared to more traditional methods like 3DMM. Therefore, the first step to improve further the performance of our models is to increase the number of scans. Generally, machine learning and artificial intelligence rely on big data for their success, but for rare diseases there are limited resources and it is often difficult to obtain access to high-quality, standardised data. Cloud-based platforms have been proposed to integrate data collection, ultimately to improve the quality of care for rare diseases. Alternatively, unsupervised methods have been tested, although their use has not been demonstrated on patient populations. In addition, with the projection of 7.2 billion smartphone subscriptions globally in 2023, and the potential of using smartphones to capture high-quality photos and 3D scans, mobile devices equipped with diagnostic algorithms will play an increasingly important role in low-cost universal care. Our approach, relying on non-ionising 3D scans, can help to accelerate this development and pave the way for shape analysis in other parts of surgery, including craniofacial and aesthetic surgery, and to replace applications that rely on CT scans. A second way of increasing the performance of our model would be the integration of shape data with traditional software. Conventional surgical simulation is a time-consuming explorative process in which the surgeon manually tests various procedural approaches and assesses the optimal osteotomy and bone position. Our approach accurately and automatically predicts the postoperative face shape and reduces the planning process to a single step. The performance of both classification and regression supports the paradigm of using machine-learning-based tools in clinical decision making and specifically computer-assisted surgical planning. Future validation of the model in larger patient cohorts and diverse surgical specialisms, or multimodal models where shape models are combined with electronic medical data...
Records, may lead to valuable diagnostic and planning tools, ultimately facilitating low-cost care, objective treatment planning and evaluation, and safer and more precise surgery.

Methods

Institutional review board statement. Patients provided informed consent and data for were retrospectively retrieved from electronic medical records after receiving approval from the Institutional Review Board at Boston Children’s Hospital (#00019505) and the Human Investigations Committee at Yale-New Haven Hospital (HIC #110100793). Data was analysed in accordance with the guidelines laid out in the Declaration of Helsinki.

Data sources. Two face databases were used, one database containing faces from the general public and one propriety patient database. Non-patient face scans were collected from the Large Scale Facial Model (LSFM) database which is available under a non-commercial licence for academic use. LSFM comprises 9,663 3D scans from volunteers taken with a 3DMD face system (3DMD LLC, Atlanta, GA, USA) under standardised imaging conditions at the Science Museum in London with the following demographics (Table 1): mean age = 24.5 ± 14.7 years, 48% male and 52% female, and ethnicity = 82% White, 9% Asian, 5% mixed heritage, 3% Black, and 1% other. The patient database includes 274 3D surface scans taken with the Vectra M3 system (Canfield Scientific, Parsippany, NJ, USA) under standardised conditions, collected from 151 patients who underwent orthognathic surgery at Boston Children’s Hospital and Yale-New Haven Hospital between December 2010 and September 2017 (Table 1), with the following demographics: mean age = 18.4 ± 2.4 years, 44% male and 56% female, and ethnicity = 76% White, 10% Asian, 10% Mixed Heritage/Other, and 8% Black. Non-syndromic patients who had a Le Fort I or bimax osteotomy were included in this retrospective study providing they had complete medical records including preoperative and long-term postoperative 3D surface scans. Additional information that was extracted from the electronic medical records including scan date, operation date, surgical procedure, indication for surgery, and syndromic diagnosis. The patient dataset analysed in this study is not publicly available due to ethical restrictions.

3D morphable model construction. Our 3DMM training pipeline, based on the approach proposed in, operates in four main functional blocks:

1. Automatic annotation – each 3D mesh was rendered from a number of virtual cameras positioned around the subject into 2D images. A vector defines the geometry of each 3D facial mesh:
   \[ X = [x_1^T, x_2^T, \ldots, x_n^T]^T \in \mathbb{R}^{3n}, \]
   where \( n \) is the number of vertices and \( x_i = [x_i^x, x_i^y, x_i^z]^T \in \mathbb{R}^3 \) describes the \( X \), \( Y \) and \( Z \) coordinates of the \( i \)-th vertex. A landmark localisation algorithm – an active appearance model (AAM) – was applied to find the 2D landmarks on the rendered images, and each 2D landmark set was projected onto the 3D surface, rendering the 3D landmarks.

2. Alignment and statistical modelling – the collection of scans was brought into the same space by removing similarity effects (rotation, translation, scale) via generalised Procrustes analysis (GPA); leaving only shape information.

3. Dense correspondence – the aligned collection of 3D scans was registered into a form where each scan had the same number of points joined into a triangulation shared across all scans. Dense correspondence was accomplished via non-rigid ICP (NICP), using the LSEM mean face as a template. This process deforms the template mesh to the shape of each patient face to obtain a set of deformed templates.

4. Statistical analysis – the 3DMM model was built by applying principal component analysis (PCA) on the corresponding meshes and finding the eigenvectors-bases with the greatest variance. Any 3D face shape can be defined as a linear combination of these bases, which makes up the 3DMM as follows:
   \[ X^* = M + \sum_{i=1}^{d} \alpha_i U_i = M + U\alpha \]  
   (1)
   Where \( M \in \mathbb{R}^{3n} \) represents the mean shape, \( U = [U_1 \ldots U_d] \in \mathbb{R}^{3n \times d} \) is the orthonormal basis matrix with columns containing the shape eigenvectors \( U_i \), \( \alpha = [\alpha_1, \ldots, \alpha_d] \in \mathbb{R}^d \) is the shape vector containing the coefficients that define a specific shape instance for a given deformable shape model. Any given input face mesh \( X \) can be projected on the model subspace by identifying the shape vector \( \alpha \) that produces a shape as close as possible to \( X \). The optimal shape vector and projection \( P(X) \) are given by:
   \[ \alpha = U^T(X - M), \quad P(X) = M + UU^T(X - M) \]  
   (2)

Model characteristics. The intrinsic characteristics of each 3DMM were evaluated, including compactness, generalisation, and specificity. Compactness is a measure of the cumulative variance in the data that is retained with a certain number of principal components and was extracted from the model construction. Generalisation describes how well a face unknown to the 3DMM can be approximated by the existing model. Specifically, we used leave-one-out cross-validation for all patient faces in all three models (LSFM, global, bespoke preoperative): a model was constructed for all faces but one, and then fitted to the excluded face. The error between the excluded face and the model was quantified using the average Euclidean distance, with a large generalisation error suggesting overfitting, the inability of a model to represent previously unseen faces. We repeated this for all patient faces. Specificity measures how well synthesised faces can be approximated by ground-truth images. Specifically, faces
(n = 10,000) were randomly synthesised for each model, and the specificity error was computed as the lowest average Euclidean distance of all vertices between a synthesised face and the closest ground-truth neighbour.

**Error quantification.** The error was quantified using the average Euclidean distance (AED), calculated from the per-vertex distance between two meshes:

\[
AED = \frac{\sum_{i=1}^{n} (x_{i,A} - x_{i,B})^2 + (y_{i,A} - y_{i,B})^2 + (z_{i,A} - z_{i,B})^2}{n}
\]

where \( x, y, \) and \( z \) corresponded to the Cartesian coordinates of mesh \( A \) and \( B \), and \( n \) was the number of vertices per mesh. Each mesh was in dense correspondence and therefore vertex \( i \) represented the same anatomical location in both meshes. To compute signed errors (such as shown in Figs 3 and 6), the error was positive if the reference mesh had a larger value on the \( z \)-axis.

**Manifold visualisation.** T-distributed stochastic neighbour embedding (t-SNE) was used as a dimensionality reduction technique\(^{48}\) to visualise a high-dimensional manifold onto a 2-dimensional space. We tested various hyper-parameters (perplexity = 2 to 100; iterations = 1,000 to 5,000, Supplementary Fig. 2) as well as different numbers of randomly sampled non-patient faces (\( n = 200 \) to \( n = 3,000 \)) together with all preoperative patient faces (\( n = 119 \)) and postoperative patient faces (\( n = 127 \)).

**Classification.** Classification was performed using a subgroup of randomly selected volunteer faces (\( n = 300 \)) and faces from pre-operative patients (\( n = 119 \)). We split the whole dataset in a stratified manner with various proportions between training and test set (80–20%, 60–40%, and 50–50%). Thus, for the 80–20% case, we used patient (\( n = 95 \)) and non-patient (\( n = 240 \)) faces for the training set and patient (\( n = 24 \)) and non-patient (\( n = 60 \)) faces for the test set. For the classifier, we chose a Support Vector Machine (SVM)\(^{49}\) with linear kernel \( x, x' \) as SVMs are powerful tools for small sample size problems. We employed the scikit\(^{50}\) implementation of SVM with “one-vs-the-rest” multi-class strategy with default values for the penalty parameter (\( C = 1.0 \)) and gamma. To calculate the mean accuracy, training and test sets were created according to a Monte-Carlo cross-validation scheme by randomly selecting the training and test set 1,000 times.

**Regression.** To automatically predict face shape outcomes based on the preoperative scan, linear regression (LR), ridge regression (RR), least-angle regression (LARS), and least absolute shrinkage and selection operator regression (LASSO) were tested, using their scikit\(^{50}\) implementation. For RR and LASSO, the alpha parameter that defines the strength of regularization term was set to 0.1 and 0.1, respectively. For LARS, the number of nonzero coefficients was set to 1. We kept the default values for all the other parameters. We used the global model to perform regression, which included volunteer (\( n = 3,664 \)), preoperative (\( n = 113 \)) and for the same unique patients their postoperative scans postoperative (\( n = 113 \)) faces. For our experiment, we used the leave-one-out scheme. A design matrix was learnt between the components of the preoperative and postoperative patients, which was then used to map the preoperative components to the postoperative components. We repeated the experiment leaving out each patient and for various components (113 times). All regression methods but LR penalised the weight of the components with a regulizer which made them more robust to overfitting.

**References**

1. American Society of Plastic Surgeons. *2017 Plastic Surgery Statistics Report*. (2018).
2. Kanevsky, J. et al. Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation. *Plast. Reconstr. Surg. 137*, 896–7e (2016).
3. Bennett, C. C. & Hauser, K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artif. Intell. Med. 57*, 9–19 (2013).
4. Mirnezami, R. & Ahmed, A. Surgery 3.0, artificial intelligence and the next-generation surgeon. *Br. J. Surg. 105*, 463–465 (2018).
5. Maas, B. D. P. J. et al. Using principal component analysis to describe the midfacial deformities in patients with craniofacial microsomia. *J. Cranio-Maxillofacial Surg., https://doi.org/10.1016/j.jcms.2018.09.019 (2018).*
6. Rodriguez-Florez, N. et al. Quantifying the effect of corrective surgery for trigonocephaly: A non-invasive, non-ionizing method using three-dimensional handheld scanning and statistical shape modelling. *J. Cranio-Maxillofacial Surg. 45*, 387–394 (2017).
7. Crombag, G. A. J. C. et al. Assessing the corrective effects of facial bipartition distraction in Apert syndrome using geometric morphometrics. *J. Plast. Reconstr. Aesthetic Surg. 67*, e151–e161 (2014).
8. Meulstee, J. et al. A new 3D approach to evaluate facial profile changes following BSSO. *J. Cranio-Maxillofacial Surg. 43*, 1994–9 (2015).
9. Steinbacher, D. M. Three-Dimensional Analysis and Surgical Planning in Craniofacial Surgery. *J. Oral Maxillofac. Surg. 73*, 540–56 (2015).
10. Pfaff, M. J. & Steinbacher, D. M. Plastic Surgery Applications Using Three-Dimensional Planning and Computer-Assisted Design and Manufacturing. *Plast. Reconstr. Surg. 137*, 603e–616e (2016).
11. Steinhufer, T. et al. Is Virtual Surgical Planning in Orthognathic Surgery Faster Than Conventional Planning? A Time and Workflow Analysis of an Office-Based Workflow for Single- and Double-Jaw Surgery. *J. Oral Maxillofac. Surg., https://doi.org/10.1016/j. joms.2017.07.162 (2018).*
12. Restnick, C. M. et al. Is There a Difference in Cost Between Standard and Virtual Surgical Planning for Orthognathic Surgery? *J. Oral Maxillofac. Surg. 74*, 1827–33 (2016).
13. Xia, J. J. et al. Accuracy of the Computer-Aided Surgical Simulation (CASS) System in the Treatment of Patients With Complex Craniofacial Deformity: A Pilot Study. *J. Oral Maxillofac. Surg. 65*, 248–54 (2007).
14. Eley, K. A., Watt-Smith, S. R., Sheerin, F. & Golding, S. J. “Black Bone” MRI: a potential alternative to CT with three-dimensional reconstruction of the craniofacial skeleton in the diagnosis of craniosynostosis. *Eur. Radiol. 24*, 2417–26 (2014).
15. Wippold, F. J. Head and neck imaging: The role of CT and MRI. *Magn. Reson. Imaging 25*, 453–65 (2007).
16. Knoops, P. G. M. et al. A novel soft tissue prediction methodology for orthognathic surgery based on probabilistic finite element modelling. *PLoS One 13*, e0197209 (2018).
46. Esteva, A.

43. Rivera, S. M.

42. Metzler, P., Geiger, E. J., Chang, C. C., Sirisoontorn, I. & Steinbacher, D. M. Assessment of three-dimensional nasolabial response to

45. Good, P. M., Mulliken, J. B. & Padwa, B. L. Frequency of Le Fort I osteotomy after repaired cleft lip and palate or cleft palate.

37. Booth, J., Roussos, A., Ponniah, A., Dunaway, D. & Zafeiriou, S. Large Scale 3D Morphable Models.

49. Genova, K.

35. Kittler, J., Huber, P., Feng, Z. H., Hu, G. & Christmas, W. 3D morphable face models and their applications. in

19. Mazzoni, S., Bianchi, A., Schiariti, G., Badiali, G. & Marchetti, C. Computer-aided design and computer-aided manufacturing
cutting guides and customized titanium plates are useful in upper maxilla waferless repositioning. J. Oral Maxillofac. Surg.,

21. Ho, C. T., Lin, H. H., Liou, E. J. W. & Lo, L. J. Three-dimensional surgical simulation improves the planning for correction of facial

32. Blanz, V. & Vetter, T. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer

graphics and interactive techniques - SIGGRAPH '99, https://doi.org/10.1109/31535.31556 (1999).

33. Blanz, V. Face recognition based on a 3D morphable model. In FGR 2006: Proceedings of the 7th International Conference on

Automatic Face and Gesture Recognition, https://doi.org/10.1109/FGR.2006.42 (2006).

Amberg, B., Knotho, R., Vettern, T. Expression invariant 3D face recognition with a morphable model. In 2008 IEEE
International Conference on Automatic Face and Gesture Recognition, FG 2008, https://doi.org/10.1109/AFGR.2008.4813376 (2008).

Küttner, I., Huber, P., Feng, Z. H., Hu, G. & Christmas. W. 3D morphable face models and their applications. in Lecture Notes in
Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), https://doi.org/10.1007/978-3-59-417783-19 (2016).

Hashimoto, D. A., Rosman, G., Rus, D. & Meireles, O. R. Artificial Intelligence in Surgery: Promises and Perils. Ann. Surg. 268, 70–6 (2018).

Booth, J., Roussos, A., Ponniah, A., Dunaway, D. & Zafeiriou, S. Large Scale 3D Morphable Models. Int. J. Comput. Vis. 126, 233–254 (2018).

Lubkoll, L., Schiela, A. & Weiser, M. An Optimal Control Problem in Polyconvex Hyperelasticity. SIAM J. Control Optim. 52, 1403–22 (2014).

Madsen, D., Lüthi, M., Schneider, A. & Vettern, T. Probabilistic Joint Face-Skull Modelling for Facial Reconstruction. cvpr, doi:
https://doi.org/10.1109/CVPR.2018.00535 (2018).

Sawh-Martinez, R. et al. Clockwise and Counterclockwise Le Fort I Movements Influence Nasolabial Morphology Differently. Plast.
Reconstr. Surg. 142, 1572–1581 (2018).

DeSesa, C. R., Metzler, P., Sawh-Martinez, R. & Steinbacher, D. M. Three-dimensional nasolabial morphologic alterations following
Le Fort I. Plast. Reconstr. Surg. Glob. Open 4, e848 (2016).

Metzler, P., Geiger, E. J., Chang, C. C., Sirisoontorn, I. & Steinbacher, D. M. Assessment of three-dimensional nasolabial response to
le Fort i advancement. J. Plast. Reconstr. Aesthetic Surg. 67, 756–63 (2014).

Rivera, S. M. et al. Patients’ own reasons and patient-perceived recommendations for orthognathic surgery. Am. J. Orthod. Dentofac.
Orthop., https://doi.org/10.1067/mod.2000.107010 (2000).

Mossey, P. & Castilla, E. In Global registry and database on craniofacial anomalies 15–33 (2001).

Good, P. M., Mulliken, J. B. & Padwa, B. L. Frequency of Le Fort I osteotomy after repaired cleft lip and palate or cleft palate. Cleft
Palate-Craniofacial J. 44, 396–401 (2007).

Estepa, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–8 (2017).

Tewari, A. et al. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In
Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017, https://doi.org/10.1109/ICCVW.2017.153 (2018).

Long, E. et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat.
Biomed. Eng. 1 (2017).

Genova, K. et al. Unsupervised training for 3D morphable model regression. in IEEE Conference on Computer Vision and Pattern
Recognition (CVPR) 8577–8386, https://doi.org/10.1109/CVPR.2018.00874 (2018).

Cervall, P., Lundwall, A., Jonsson, P., Carson, S. & Möller, R. Ericsson Mobility Report 2018. Available at, https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf (2018).

Muravyov, O. et al. 3D Capture: 3D Reconstruction for a Smartphone. In IEEE Computer Society Conference on Computer Vision and
Pattern Recognition Workshops, https://doi.org/10.1109/CVPRW.2016.116 (2016).

Mendoza, C. S. et al. Personalized assessment of craniosynostosis via statistical shape modeling. Med. Image Anal. 18, 635–46 (2014).

Porter, M. E., Larsson, S. & Lee, T. H. Standardizing Patient Outcomes Measurement. N. Engl. J. Med. 374, 504–6 (2016).

Tassabehji, M. et al. GTF2IRD1 in Craniofacial Development of Humans and Mice. Science (80-.). 310, 1184–7 (2005).

Zafeiriou, S. & Dunaway, D. Large Scale Facial Model (LSFM). Available at, https://sip.ucsb.edu/software/LSFM.html. (Accessed: 29th November 2018) (2018).

Amberg, B., Romdhani, S. & Vettern, T. Optimal step nonrigid ICP algorithms for surface registration. In Proceedings of the IEEE
Computer Society Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPR.2007.383165 (2007).

Styner, M. A. et al. In Information processing in medical imaging. IPMI 2003. Lecture Notes in Computer Science (eds Taylor, C. &
Noble, J. A.), https://doi.org/10.1007/b111820 (2003).

Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (2008).

Cortes, C. & Vapnik, V. Support-Vector Networks. Mach. Learn. 20, 273–297 (1995).

Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
Acknowledgements
This research was supported by Great Ormond Street Hospital Children's Charity through the FaceValue programme Grant (No. 508857), the Engineering and Physical Sciences Research Council (EP/N02124X/1) and the European Research Council (ERC-2017-StG-757923). This work was undertaken at GOSH/ICH, UCLH/UCL who received a proportion of funding from the United Kingdom Department of Health’s NIHR Biomedical Research Centre funding scheme. The funders had no role in study design, collection, analysis and interpretation of the data, decision to publish, or preparation of the manuscript.

Author Contributions
P.G.M.K., D.S., B.L.P and S.S. initiated and designed the study. P.G.M.K. and A.P. developed the algorithms, designed data analyses and performed experiments, and S.Z. validated the methodology. A.B., R.W.F.B., D.J.D. and O.J. helped design the study and helped with data interpretation. B.L.P. and D.S. provided the electronic medical record data, and P.G.M.K. and A.W. extracted the electronic medical record data. P.G.M.K., A.P. and S.S. wrote the paper in conjunction with D.S., B.L.P. and D.J.D. who wrote the clinical introduction and discussion. A.B., A.W., R.W.F.B., O.J. and S.Z. provided manuscript feedback.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49506-1.

Competing Interests: A.P. works as a consultant for FaceSoft Ltd and S.Z. works as CTO for FaceSoft Ltd. D.S. has received grant funding from Synthes and KLS Martin. The other authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019