Optimizing digitalization effort in morphometrics

Allowen Evin, Vincent Bonhomme, Julien Claude

BIOLOGY Methods & Protocols

Biology Methods and Protocols, 2020, 1–10

doi: 10.1093/biomethods/bpaa023
Advance Access Publication Date: 16 November 2020
Methods Manuscript

METHODS MANUSCRIPT

Optimizing digitalization effort in morphometrics

Allowen Evin,* Vincent Bonhomme, and Julien Claude

Institut des Sciences de l'Evolution-Montpellier, UMR 5554-ISEM, CNRS, Université de Montpellier, IRD, EPHE, 2 place Eugène Bataillon, CC065, 34095 Montpellier Cedex 5, France

*Correspondence address. Institut des Sciences de l'Evolution-Montpellier, UMR 5554-ISEM, CNRS, Université de Montpellier, IRD, EPHE, 2 place Eugène Bataillon, CC065, 34095 Montpellier Cedex 5, France. E-mail: allowen.evin@umontpellier.fr

Abstract

Quantifying phenotypes is a common practice for addressing questions regarding morphological variation. The time dedicated to data acquisition can vary greatly depending on methods and on the required quantity of information. Optimizing digitization effort can be done either by pooling datasets among users, by automatizing data collection, or by reducing the number of measurements. Pooling datasets among users is not without risk since potential errors arising from multiple operators in data acquisition prevent combining morphometric datasets. We present an analytical workflow to estimate within and among operator biases and to assess whether morphometric datasets can be pooled. We show that pooling and sharing data requires careful examination of the errors occurring during data acquisition, that the choice of morphometric approach influences amount of error, and that in some cases pooling data should be avoided. The demonstration is based on a worked example (*Sus scrofa* teeth) using a combinations of 18 morphometric approaches and datasets for which we identified and quantified several potential sources of errors in the workflow. We show that it is possible to estimate the analytical power of a study using a small subset of data to select the best morphometric protocol and to optimize the number of variables necessary for analysis. In particular, we focus on semi-landmarks, which often produce an inflation of variables in contrast to the number of available observations use in statistical testing. We show how the workflow can be used for optimizing digitization efforts and provide recommendations for best practices in error management.

Keywords: data sharing; geometric morphometrics; interoperability; measurement error

Introduction

Quantifying and analyzing phenotypic variation have greatly benefited from conceptual and analytical developments in morphometrics, and from the development of new acquisition methods and hardware. The large geometric morphometrics toolbox [1] offers the possibility to use different methods to estimate shape and size from a set of coordinates: e.g. various Procrustes methods based on landmarks or sliding semi-landmarks [2–4], or outline analyses on a collection of points digitized along an outline, commonly addressed using elliptic Fourier analysis [5, 6]. These geometric morphometric approaches complement 'traditional' morphometrics mainly based on collection of linear measurements. Not all these methods are similar in terms of data input and time spent on acquiring primary data. While a simple set of inter-landmark distance measurements or an outline mask allowing direct coordinates extraction can be rapidly obtained, digitizing dense configurations of points can be more time consuming. Recent years have seen an important revolution in the discipline with the possibility to analyze dense datasets made by a large number of points on surfaces or outlines (e.g. [7]). The inflation in the number of variables collected for geometric morphometric studies (e.g. [8, 9]) is often motivated by the necessity of capturing shape as accurately as possible in order to detect the most subtle

Received: 18 September 2020; **Revised:** 5 November 2020. **Accepted:** 13 November 2020

variation possible. But the resulting highly dimensional datasets may on the one hand lead to a dramatic increase in the digitization time, and on the other hand potentially lead to biologically inaccurate results and misleading interpretations [9]. In addition, this data inflation does not guarantee an increase of precision in inferences since highly dimensional datasets are not inherently devoid of error, and increasing the number of variables do not necessarily temper digitization errors.

The increase of time needed for acquiring data regarding phenotypic variation and the necessity of having a sufficiently large dataset could be partially mediated by the possibility of pooling and sharing data among users. Indeed, like many other fields, emerging repositories (e.g. https://www.morphosource.org, https://morphobank.org, and https://morphomuseum.com) facilitate data archiving and sharing. Pooling morphometric datasets can be desirable to scale up and generalize phenotypic studies leading to simultaneously larger scale analyses and reduced acquisition efforts. However, pooling datasets never comes without risk and morphometrics makes no exception.

In that context, an increasing number of studies have begun to pool datasets obtained from multiple operators and/or devices and even crowdsourcing phenotypic data acquisition for geometric morphometrics has been used [10]. However, several studies have highlighted the large influence of artificial, systematic, and sometimes directional, variation introduced by inter-operator (IO) errors on these measurements (e.g. [11–15]). While some studies have evidenced small variation between multiple operators when compared with the targeted biological signal (e.g. [16]) others have demonstrated that IO bias can lead to substantial variation on geometric morphometric analyses (e.g. [17]). IO measurement errors (MEs) can have dramatic impacts on morphometric studies, especially when the phenotypic variation under investigation is subtle or varies in the same direction as the explored biological variation [16]. This is true particularly as the methods such as geometric morphometrics (using coordinates of landmarks, curves, or surfaces) are especially used for detecting small-scale shape and size variation [18, 19]. Several approaches have been proposed to assess error in morphometrics based on replicated measurements [12, 20, 21]. Less studies have addressed the problem of pooling data from various sources [13, 14, 16].

Yezerinac et al. [20] identified multiple sources of imprecision in morphometric measurements including: not well-defined measurements, structure flexibility, operator experience, IO and repeatability variation, non-human sources precision, and lightning conditions. These can be summarized into the three categories of methodological, instrumental, and personal sources of error [14]. Unfortunately, MEs, both within and among operators, are rarely evaluated and even more rarely quantified and published.

When morphometric data are pooled, error likely increases with the acquisition workflow complexity. For instance, morphometric data can be acquired directly on the specimens using calipers or a 3-D digitizer, but very often, these data are obtained using an intermediate image, either in 2D or 3D obtained with a camera or a 3-D model acquisition devise (e.g. CT scan, photogrammetry). On these images, coordinates or measurements are later obtained with 2-D, or 3-D digitizing software. In the latter case, when an acquisition device is used, error will be the result of variations in the object preparation (if any), the device used, and the acquisition and post-processing of the data. Furthermore, error can increase when data are shared and multiple operators are involved, which is typical of pooling data from different studies (Fig. 1A).

Common error quantification relies on the comparison of the amount of variation among replicates of the same measurements and between higher categories (e.g. individuals, populations, or species), ensuring that the explored variation significantly exceeds the ME. Pooling data involves estimating variation introduced by multiple operators, insuring that the sum of errors (IO error and intra-operator MEs) does not alter the interpretation of the results. In the simplest and ideal case, pooling data among multiple operators introduces an excess of variation that goes beyond the variation introduced by a single operator. When operators introduce a systematic and directional bias, and when there is little overlap of common data acquisition among operators disentangling operator effects and true variation from pooled data can become difficult or impossible.

Indeed, even if there is high reproducibility within a single operator, the systematic error can be high by comparison to the research question. Such case of autocorrelated pattern due to user occurs, e.g. when an operator repeatably misplaces a specific landmark. One must therefore estimate whether ME introduced by various users is significantly greater than intra-operator based on a set of similar object. To do so we propose the workflow illustrated in Fig. 1B. Error can be sensitive to both the data acquisition procedure and the way that morphometric parameters of shape and size are extracted from the data. Different analytical approaches may lead to different amount of error within and among operators even if based on the exact same set of data (Fig. 1B). The proposed workflow represents different steps for validating both the data acquisition protocol, i.e. the choice of the position, number, and type of points (e.g. landmarks, or points along curves or outlines) and the choice of the analytic approach to be used (e.g. using or not sliding semi-landmarks; elliptic Fourier analysis or sliding semi-landmarks, and sliding semi-landmarks using the bending energy or not). Both aspects should be carefully examined prior to pooling datasets. This workflow is based on the comparison of the intra-operator MEs (one per operator) with the IO error. The proposed workflow comes in addition to already available workflows (e.g. [13]) by formalizing the specific context of pooled datasets obtained from multiple operators.

Here, we demonstrate how to choose morphometric approaches and datasets for conducting a morphometric analysis and assess whether morphometric datasets can be pooled without risk. We use an example based on the common problem of taxonomic identification based on morphometrics but the approach can be easily transposed to other classification questions, and more generally to many biological questions involving morphometric data. To do so we used the workflow presented in Fig. 1B to assess different source of error when pooling morphometric datasets such as the intra- and IO errors, the importance of the choice of the analytical method, and the impact of the number of variables in morphometric analyses. After a selection of best approaches for a given problem (here identifying two taxa), we show how to potentially reduce digitization efforts.

Material and methods

Worked example

We selected a dataset with relatively subtle morphometric variation and for which several morphometric approaches have

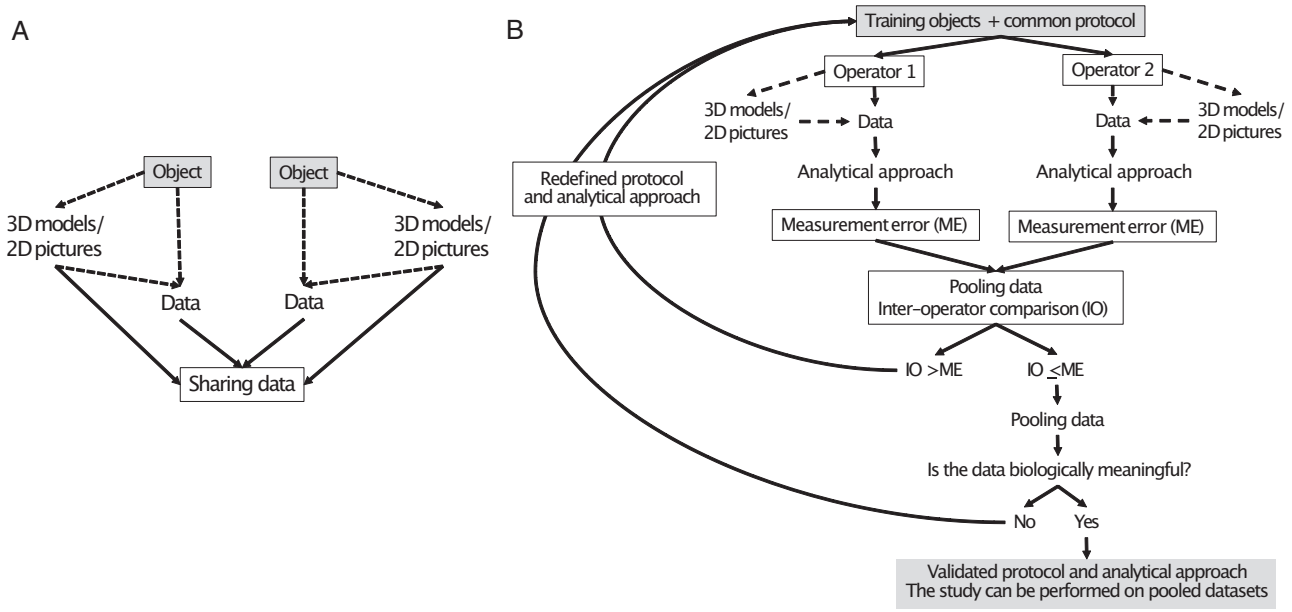

Figure 1: (**A**) Classical workflow when data from different studies are shared. The dotted lines represent steps including error. (**B**) Analytical workflow for validating the data acquisition protocol and the analytical approach to be used for pooling data obtained from multiple operators. We recommend to select a protocol for which the intra-operator MEs exceed the IO ME. The protocol and analytical approach can be redefined recursively to reach this goal.

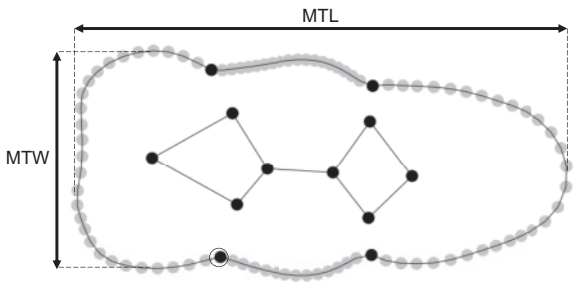

Figure 2: Data acquisition protocol for measuring pig lower third molar. Geometric morphometric data consists of 12 landmarks (in black) and 87 sliding semi-landmarks (in gray). Traditional metric data consists of two linear measurements; maximum tooth length (MTL), and width (MTW). The anterior part of the tooth is on the left.

been used in the past [22–24]. Distinguishing wild boar from domestic pigs in the archaeological record is particularly challenging due to bone fragmentation and the close morphological similarity between the two [25]. Traditional measurements [maximum tooth length and widths (Fig. 2), especially of the lower third molar] have been traditionally used to separate small domestic pigs from the large wild boar (e.g. [26]). More recently, 2-D landmark and sliding semi-landmark-based geometric morphometric approaches have been used for investigating shape variation between and within the two taxa (review in [27]). Pooling the data from these various datasets is the next step forward for future research.

Two datasets were used to apply the proposed analytical workflow. A first set of data was used to assess the intra and inter-operator effects. This dataset corresponds to repeated measurement on six third lower molars belonging to four adult wild boars (two specimens have left and right teeth). Photographs of the teeth were obtained using a Nikon D90 DSLR camera paired with a 60-mm micro lens (AF-S Micro Nikkor), and the 2-D coordinates of points (landmarks and sliding semi-landmarks) were acquired following the protocols detailed in [23, Fig. 2) using the tpsDig2 software (v2.18, [28]). Maximum tooth length and width of the same teeth were measured using a caliper (Mitutoyo). A second dataset was used to quantify the discrimination between wild and domestic pigs using various methods. Data were publicly available ([24], http://dx.doi.org/10.6070/H4ZK5DNC), measured with the same protocol (Fig. 2) and include measurements of the third lower molars of 42 domestic pigs and 129 wild boars. In order to have a protocol that will be suitable to identify the two taxa, this protocol should show reduced error variation and should discriminate between the two categories.

All analyses were performed following 18 analytical morphometric approaches (Table 1) including a single approach based on linear measurement and various geometric morphometric approaches (landmarks, sliding landmarks, and outlines) based on complete or subsamples of the original datasets. Analyses were performed for the two classical cases of data pooling: (i) in the case that primary data are shared (i.e. pictures) when only ME consist of multiple operators digitizing landmarks and (ii) in the case that different users obtained primary data independently (independent picture and landmark acquisition). We also investigated the effect of data inflation (i.e. number of points) in terms of relative quantity of error produced and of effort needed to discriminate between groups.

Assessing repeatability

First, ME of each operator (measurer measurement error, MME) due to digitization practice was assessed. Five people with varying amounts of experience in morphometrics and pig tooth anatomy (from novice to experienced) placed 12 landmarks on 6 specimen, 5 times each (Fig. 3). Five people also measured five times the maximum tooth length and width using a caliper.

Then, in order to estimate simultaneously the effects of the photography protocol and digitization, and to estimate total ME (TME); three operators were asked to take five independent pictures of six teeth and to digitize point coordinates on each of

Table 1: List and description of R packages and options used to establish the 18 analytical approaches used in the study

Approach	Description	R Package	Options
A1	12 landmarks	Morpho	Default
A2	8 landmarks	Morpho	Default
A3	91 Sliding semi-landmarks	geomorph	Procrustes distance
A4	91 Sliding semi-landmarks	geomorph	Bending
A5	91 Sliding semi-landmarks	Morpho	Procrustes distance, tol = 1e−5
A6	91 Sliding semi-landmarks	Morpho	Bending, tol = 1e−5
A7	91 Sliding semi-landmarks	Morpho	Procrustes distance, tol = 1e−7
A8	91 Sliding semi-landmarks	Morpho	Bending, tol = 1e−7
A9	12 landmarks + 87 Sliding semi-landmarks	geomorph	Procrustes distance
A10	12 landmarks + 87 Sliding semi-landmarks	geomorph	Bending
A11	12 landmarks + 87 Sliding semi-landmarks	Morpho	Procrustes distance, tol = 1e−5
A12	12 landmarks + 87 Sliding semi-landmarks	Morpho	Bending, tol = 1e−5
A13	12 landmarks + 87 Sliding semi-landmarks	Morpho	Procrustes distance, tol = 1e−7
A14	12 landmarks + 87 Sliding semi-landmarks	Morpho	Bending, tol = 1e−7
A15	Outline (91 coordinates)	Momocs	9 harmonics
A16	Outline (91 coordinates)	Momocs	6 harmonics
A17	Outline (91 coordinates)	Momocs	2 harmonics
A18	Traditional metrics (MTL, MTW)		

MTL and MTW, maximum tooth length and width; Tol, tolerance.

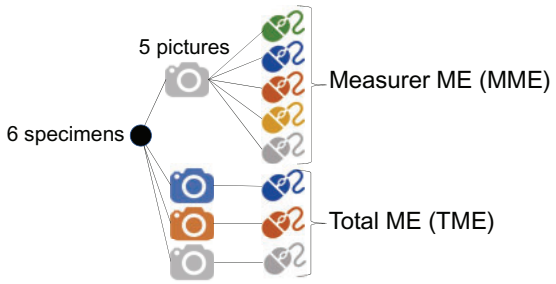

Figure 3: Protocol design for estimating the different sources of error when sharing morphometric data. We estimate the MME which correspond to the case when pictures or models are shared and only the error in landmarking is considered, as well as the TME combining error in acquiring the medium and in landmarking. For both protocol, we compare the individual ME for each operator with the IO ME.

these pictures. This allowed for the quantification of the relative importance in photograph positioning and landmarking (both included in TME). For both approaches, we quantified the error of each operator (ME) before quantifying the IO ME.

Morphometric methods, packages, and computation options

The 18 different possible combination of approaches and dataset are given in Table 1. The first two approaches are based on *true landmarks only* and include either the 12 landmarks (A1) described in Fig. 2, or only the 8 landmarks (A2) localized in the inner part of the tooth, excluding the 4 landmarks on the outline. Generalized Procrustes Analyses (GPA) was performed with the package Morpho (Schlager, 2017 [29]). Approaches (A3)–(A8) include the *sliding semi-landmarks only* with one true landmark (landmark 10) and 90 sliding semi-landmarks and were performed using Morpho (ProcSym function) and geomorph (gpagen function [30]). Approaches (A3)–(A8) were performed using the Procrustes and bending options of both functions and we also tested two levels of tolerance modified it in the ProcSym function of Morpho package (we first used the default threshold of 10^{-5} and then decreased that threshold to 10^{-7} while other parameters were set to default). Approaches (A9)–(A14) combined *landmarks and sliding landmarks* and used the same variation in options outlined previously. Approaches (A15)–(A17) correspond to elliptic Fourier approaches applied to the *outlines* (using the Momocs package, efourier function [31]) and vary in the number of Fourier harmonics (9, 6, and 2) kept as descriptive variables to capture the outline. Finally, approach (A18) corresponds to the *traditional metrics*, with maximum tooth length and width analyzed jointly using Mosimann's log shape ratio [32].

Subsampling

We used approach (A5) (sliding semi-landmarks analyzed with the package Morpho, the Procrustes distance sliding procedure, and the tolerance set to 10^{e-5}) to explore the effect of the number of sliding semi-landmarks on both the error and the discrimination power. Initial number of sliding semi-landmarks was progressively subsampled from 90 to 10 points.

Error quantification

Detailed quantification of among- and within-operator variation relies on repeated measurements of the same object. The percentage of error relative to inter-individual variation was computed following the ANOVA design presented in [20, 33]. In this approach, mean squares are used to estimate the respective proportion of variance associated with replication measurement and inter-individual variation. This percentage has the advantage of being comparable between different studies and approaches [20, 33]. Percentage of error is calculated as the ratio between the variance within a specimen (between the replicates, s^2within) and the sum of the within and among variances (s^2within + s^2among), expressed as a percentage (i.e. multiplied by 100). Our design was balanced and allowed the use of least squares estimates. For unbalanced designs, we remind the reader that estimation of variances can be done via the use of maximum likelihood as it is now routinely obtained with software or libraries designed for mixed effect models [34].

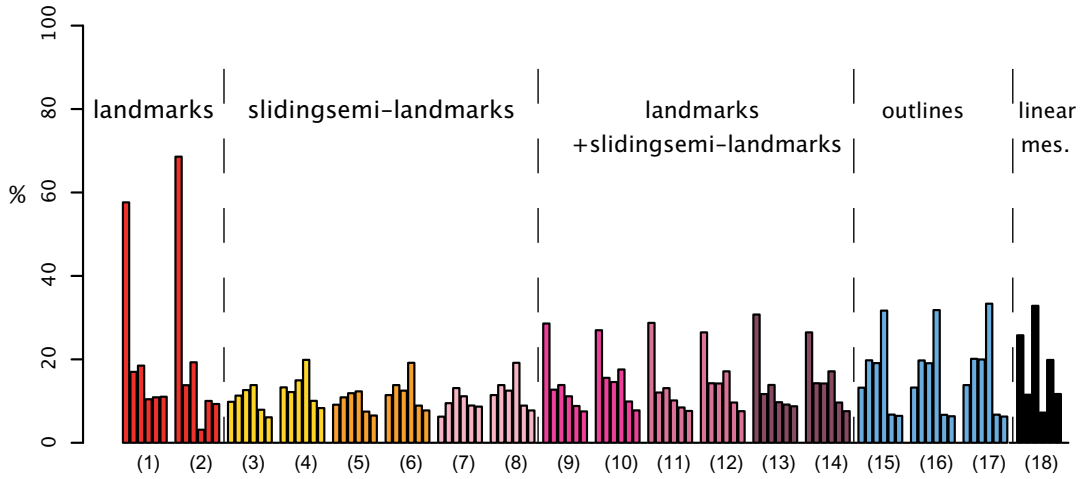

Figure 4: Level of MME (percentage of total variance) for the 18 explored approaches (numbers in brackets refer to Table 1). For each set of bars (separated by a space): the first bar on the left represent the IO error while the five next bars on the right represent the error of the five operators (ME). mes., measurements.

Amounts of error were obtained for each operator separately (with five replicates/photos measured per specimen) before quantifying the IO error by pooling the replicates of each operator resulting in five or three replicates (for MME and TME, respectively).

Quantification of discrimination power

Because one of the main objectives of the study was to pool multiple bioarchaeological datasets to identify wild and domestic pigs, we contrasted the error estimates with the discriminating power of the different approaches. Percentages of accuracy, i.e. correct cross-validation, were obtained from predictive Linear Discriminant Analyses (LDAs) on shape variables paired with a leave-one-out cross-validation procedure. When the number of variables exceed the number of specimens we applied the approach proposed in [35, 36] in which raw variables are replaced by the first principal component analysis (PCA) scores which number maximize the rate of leave-one-out cross validation [35]. Moreover, because the number of wild boars highly exceed the number of domestic pigs in our datasets we used the approach proposed in [22](Evin et al. 2013) providing a mean cross validation and a 90% confidence interval calculated from a distribution of cross validation percentages (CVP) based on 100 balanced resampled datasets.

Results

MME

Analysis of ME linked with digitalization show strong differences between approaches and datasets (Fig. 4). The two approaches including only landmark data (A1 and A2) have particularly high proportion of total variance accounted by IO error. Approaches A9–A14 combine landmarks and sliding-semi landmarks and show much higher percentages of error when multiple operators acquired the data. Outline approaches (A15–A17) show relatively low percentage of individual operator errors but high variation between operators. Linear measurements (A18) show an intermediate pattern with relatively a high percentage of IO error and variable amount of individual operator error. High heterogeneity between operators should be noted for some approaches, e.g. (A2), with one operator showing a particularly low level of ME. Such cases where intra-operator errors are relatively low compared with the IO amount of error likely reveal a systematic error for at least one of the operators. This is particularly noticeable for the two approaches including only landmarks data (A1 and A2). Finally, approaches based on sliding semi-landmarks (A3–A8) show low percentage of IO and intra-operator errors, with the IO ME being even smaller than any intra-operator error for the approach including the use of the package Morpho, the Procrustes distance criterium for the sliding procedure and the threshold set to 10^{-7} (A7).

TME

Analysis of the TME (Fig. 5) combining error in acquiring the pictures and the landmark coordinates reveals a very similar pattern to the MME. Approaches including landmarks (A1, A2, and A9–A14) and outlines (A15–A17) show high IO error, as well as the traditional metrics (A18). Again, on the contrary, approaches based on sliding semi-landmarks (A3–A8) show lower percentage of IO errors; however, none of the approach show smaller percentage of IO error than any intra-operator error.

Correlation between MME and TME

We explore the correlation between errors in landmarking only (MME) and the cumulative effect of landmarking and photographing (TME) for both the IO error and for the mean of the intra-operator errors. In both cases, MME and TME are highly correlated (IOs: adjusted $R^2 = 0.71$, $P = 7e^{-6}$, mean intra-operators: adjusted $R^2 = 0.85$, $P = 3e^{-8}$) (Fig. 6) highlighting again that the choice of analytical approach deeply influences error, and the importance of the point coordinate acquisition in error measurement. A large majority of tested approaches show higher TME than MME (Fig. 6), showing that the photography process also induces error measurement. This is especially true for the outline approaches (A15–A17, Fig. 6, left) when the IO error is assessed, while in other cases, error in point coordinate acquisition greatly influences intra-operator error (e.g. A2–8 landmarks, Fig. 6, right). Therefore, the relative importance of error in photographing and landmark coordinate acquisition varies depending on the analytical approach used.

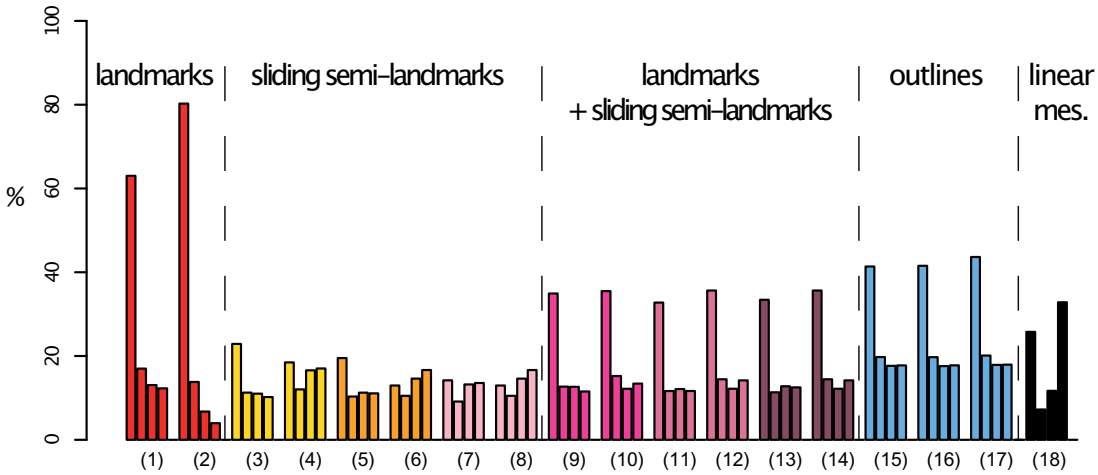

Figure 5: Level of TME (percentage of total variance) for the 18 explored approaches (numbers in brackets refer to Table 1). For each set of bars, the first one on the left represent the IO error while the three next on the right represent the error of the three operators (ME)., mes., measurements.

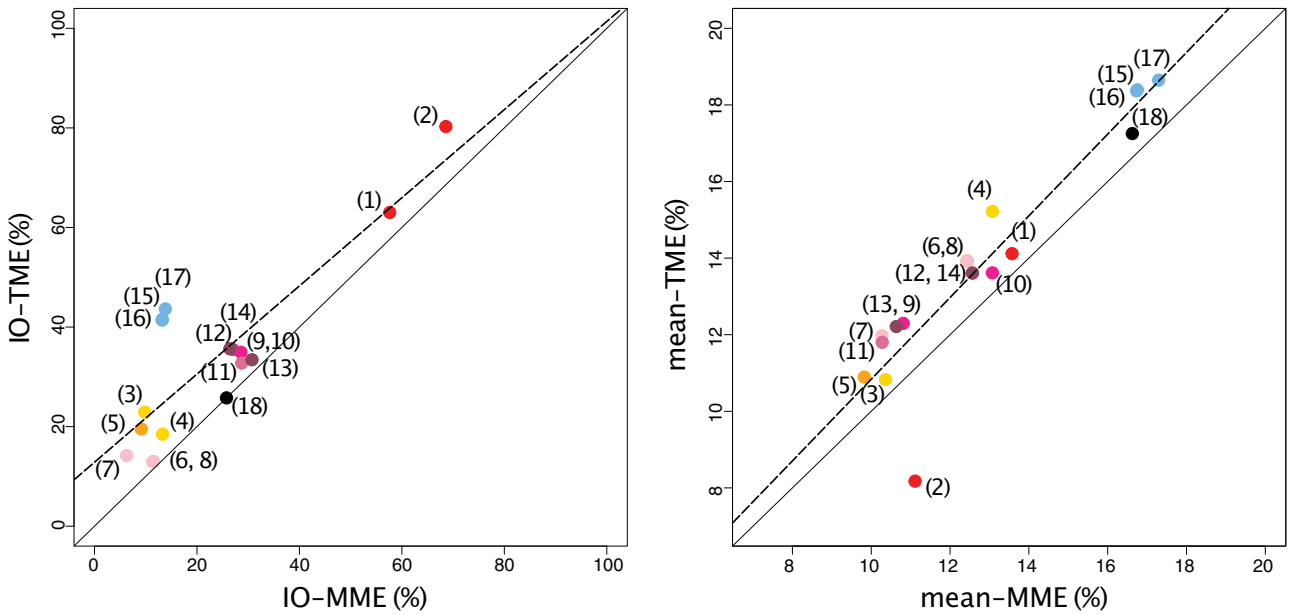

Figure 6: Relationship between (left) IO MME and TME and (right) mean MME and mean TME. Numbers in brackets refer to Table 1 and correspond to the different approaches. The dotted lines represent the regression line, while the plain lines represent a perfect regression between the two error measures (a slope of one and an intercept of zero).

Correlation between error and discriminant power

Though error and discriminant power correspond to two different aspects of the study we ultimately aim at identifying at the same time the approach minimizing error and maximizing the discrimination between two known groups. In our working example, the discriminant power varied from 64.7% (A19, traditional metrics) to 89.3% (A16, outline with six harmonics) in mean, depending on the approach used (Supplementary Fig. S1). The lowest CVPs were obtained for approaches including 8 (A1) and 12 (A2) landmarks only, outline analysis using the two first harmonics only (A17), and traditional metrics (A18). All other distributions largely overlap around 87.6% in mean (Supplementary Fig. S1).

Contrasting the discrimination power with the TME and the MME (Fig. 7) provide similar patterns with: approaches with high error and/or low CVP (A1, A2, A18, A19); approaches with moderate error and/or moderate discriminant power (A9–A17); and approaches with the lowest error and highest CVP (A3–A8). Both the MME and the TME slightly correlate with the mean CVP of the approaches used ($P = 0.01$ and adjusted $R^2 = 0.27$, $P = 0.04$ and adjusted $R^2 = 0.19$, respectively) revealing a relatively small, but significant, link between the discriminant power and the percentage of error in our worked example. These correlations became non-significant when landmarks approaches (A1 and A2) are discarded.

Impact of subsampling

Protocols including 90 sliding semi-landmarks were subsampled down to 10 points coordinates. The discriminant power correlates with error (adjusted $R^2 = 0.8$, $P = 0.0005$), with the smaller the number of coordinates included, the smaller the

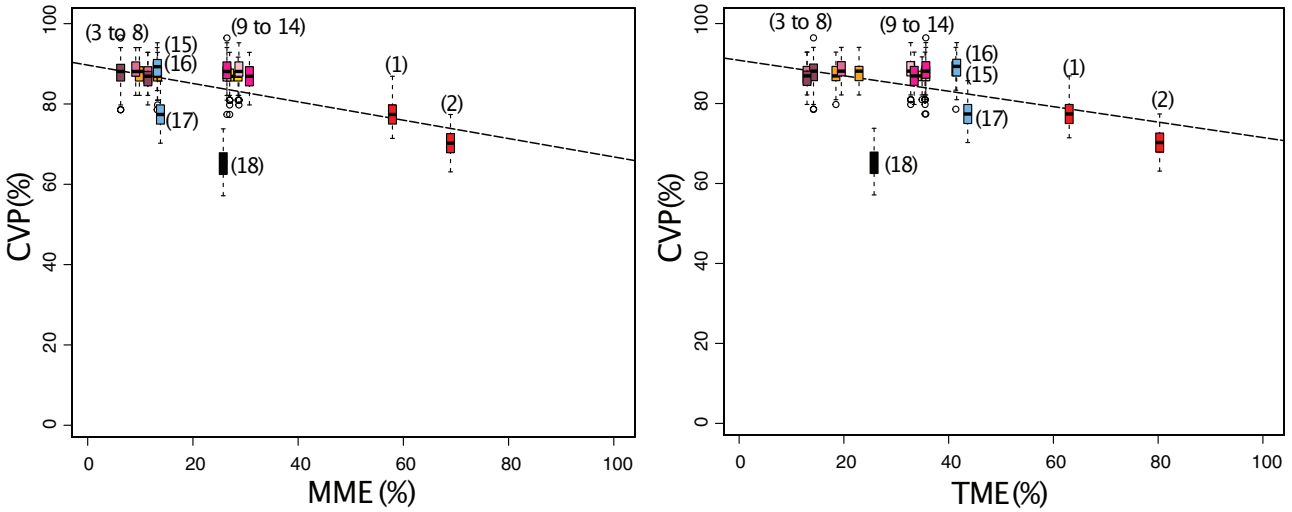

Figure 7: Relationship between discriminant power (CVPs) and IO error (TME on the left, MME on the right). Numbers in brackets refer to Table 1.

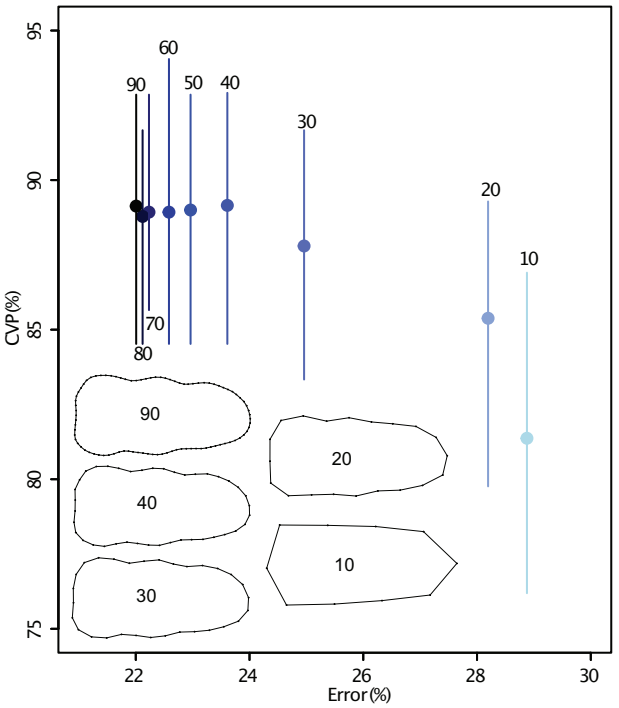

Figure 8: Evolution of CVPs and error (TME) in relation to the number of sliding semi-landmarks (down sampling from 90 to 10) used to measure the tooth shape. Data were obtained using approach A7. Dots represent the mean CVP while the vertical bars represent 90% of the CVP distributions. Visualization of the mean shapes is depicted for various number of sliding semi-landmarks (see Fig. 2 for the initial tooth shape).

discriminant power and the higher the error (Fig. 8). This is especially true when fewer than 40 points were used. From 90 to 40 points, the CVPs of the discrimination between the two known categories remain stable despite an increase in error rate. As expected, details in tooth shape decrease with the number of points along the external outline: angles become steeper and the lateral sides flatten (Fig. 8). In our worked example, we identify 40 points as a good compromise between landmarking effort, error rate, and discriminant power.

Discussion

The proposed workflow was used to determine an analytical approach offering the possibility of sharing morphometric data using an optimized protocol and minimizing acquisition effort by ensuring high repeatability, while preserving the biological significance of the results.

Worked example

In our worked example of wild and domestic pigs, the aim was to validate the possibility of sharing point coordinates and/or 2-D pictures. We aimed at identifying a morphometric protocol which would allow sharing data obtained by multi-operators or published by different studies. We tested 18 analytical approaches and demonstrated that not all of them can be used for data sharing. Pooling data from different operators always increase error rate and data should not be pooled in some cases. A rule of thumb could be to avoid pooling data when the IO error rate exceeds the error of any of the individual operators. In such instances, it could be wise to either exclude some of the operators or to keep the data separated. In our worked example, approaches based solely on landmark data could not be used for combining datasets obtained from various studies since the inter-operator error appears in some case at least four times higher than any of the individual operator error rates. On the contrary, other approaches, such as the ones based on sliding semi-landmarks only offer good repeatability and sometimes even buffer the inter-operator error rate. The approach used in previous published work (A11 with 12 landmarks and 87 sliding semi-landmarks) offer good intra-operator repeatability, but the high amount of IO repeatability due to the presence of landmark data prevent it from being confidently used on pooled data. We identified the best approach providing at the same time low intra- and IO errors, with the IO error being even smaller than any intra-operator errors in landmarking (i.e. MME) and similar in the TME combining the error in landmarking and picturing. For the case of pig teeth discrimination, we therefore recommend, if the data has to be shared, to use approach A7 based only on sliding semi-landmarks along the outline of the teeth, the package Morpho, and a threshold for GPA convergence fixed to 10^{-7}. In the specific case of pig lower third molar morphometric analysis, we also highlight the possibility of optimizing the

number of landmarks used in the protocol and recommend the use of 40 sliding semi-landmarks. This number offers a good trade-off between digitalization effort, error rate, and discrimination power. In this example, reducing the number of landmarks by more than half compared with the initial approach does not alter the biological information explored here, i.e. the discrimination between two taxonomical groups. This optimization can reduce the time dedicated to recording point coordinates. This protocol, applied only to a relatively small set of modern specimens (though the biggest available so far for these taxa and this methodology), has been identified specifically for our case study and may provide different results if applied to other datasets (either species or skeletal element). We highly recommend applying the same workflow before starting a study or pooling data between operators.

How can we pool morphometric data?

Error is ubiquitous

Error is ubiquitous in morphometric data acquisition, can never be completely avoided, and can be a major problem when the explored variation is relatively small and subtle [21]. Error cannot be estimated without replicates, and any error testing protocol has to be carefully thought out and implemented as the first step of any morphometric study. It could even be recommended to perform the tests at several stages of the study, especially if it is spread over a long period of time, since operators may progressively, and unconsciously, change the way they acquire data. If possible, tests should be performed using the exact same set of reference specimens. That being said, error should always be interpreted in the context of the biological question explored [14, 21, 37]. The amount of acceptable error will indeed depend on the data used and the comparisons being made [17, 21, 38, 39]. For example, error could be too important to compare populations but acceptable when higher taxonomical ranks are explored such as, e.g. species or genera. Though not limited to 2-D data, 3-D data either obtained directly on the objects or via 3-D models are not free from error and different approaches (e.g. photogrammetry versus CT-scans), different devices (e.g. different surface scans) [40], or different software's used for the model reconstructions can induce biases and should not be pooled without prior assessing of error rate. Reducing impact of error can be made by replicating multiple times (at least twice) the measurements and averaging those data prior to analysis [21, 39]. Finally, when there is a doubt on repeatability and when datasets do not overlap between users, it is also possible to remove the effect of the origin of the data by introducing this as a factor in the analysis (e.g. [16]) but this requires the critical assumption that the variation of the hypothetical bias is not in the same direction as the biological variation explored. This can however be done in trying to keep balanced data among data providers for avoiding violating "the marginality principle" in linear modeling. The ME performed in this study does not disentangle random from systematic errors, the two main types of error [40]. However, we detected a likely case of systematic error, with one operator who digitalized one of the landmarks in a systematic and different manner than the other operators. This could have been identified and mitigated through a careful examination of the position of the landmarks on the pictures prior to the analyses. We therefore recommend, when possible, a careful examination of the coordinate positions on the initial images or 3-D models.

Identifying and classifying sources of error

We explored and compared: (i) the error linked with landmarking only with several operators having digitalized several times the same set of pictures, which is typical when pictures are shared between studies, and (ii) the cumulative error linked with both 2-D picture acquisition and landmarking, which corresponds to the case where raw coordinates are published and pooled between studies.

We have found a general positive correlation between landmarking (MME) and total error (photographing and landmarking, TME) across morphometric methods but only when protocols comprising fuzzy points, the landmarks, were included in the analysis. Overall, our results reveal the importance of landmarking and the relatively limited effect of photographing in the amount of error. Here, though based on a very limited number of operators (not enough for statistical testing), "experts" show slightly lower amount of error than more "beginners" operators. As a consequence, if pig teeth data have to be shared, special attention has to be given to the coordinate acquisition step. Importantly, once identified, error can be reduced by training as demonstrated by noticed differences between expert and novice operators [42]. Operators must be familiar with both the specimens under study and the way the data are acquired [39]. Prior to performing a morphometric study, the operator has to have an idea of the range of variation that will be included in the study and sufficient practice in data acquisition so they will have no, or very limited, hesitation when performing the data acquisition. Here, all operators used the same photographic equipment excluding the inter-device bias that may occur when data from multiple studies are combined.

Optimizing data acquisition

Minimizing the impact of ME can be done in several ways including by; reducing the number of steps between the specimen and the data, thoroughly standardizing protocols of data acquisition across specimens (ideally the same equipment should be used throughout the data collection process within a single study), calibration, insuring quality of equipment, -adaptation of the dataset/protocols to the question, and repeated measurements [43].

A common question when starting a study is about how many landmarks or measurements should be used to accurately quantify variation. The answer is never straightforward and as this study shows, more is not always better [9, 37]. Measuring effort should be contrasted with the aims of the study and the number of coordinates acquired depends on the complexity of the structure measured. There is a trend for increasing number of variables, but adding too many can bias the results of the analyses [9, 44, 45]. While a high ratio between sample size (N) and number of variables (P) is often recommended this is not always the case in geometric morphometric studies (Cardini, 2020 and references there in). Between group CVPs were obtained from [22] which provide a careful examination of the sensitivity of discriminant analyses to both unequal sample size and number of shape predictors. Selecting the minimum number of principal component scores to be included in the discriminant analyses allows simultaneously maximizing the cross-validated accuracy of the classification but also reducing unwanted variation (noise) in the data while simultaneously removing, through the PCA, collinearity problems [13, 34]. As a consequence, including fewer or more principal component scores will result in lower, or at best similar, between group CVPs. Acquisition effort should be therefore contrasted with the complexity of the

studied object, the subtlety of the targeted differentiation, and the number of specimens available for the study. We show that adding sliding-semi landmarks, that capture variation in landmark-free regions, to "true" landmarks greatly improve classification accuracy while simultaneously mitigate the amount of error, a likely result of the generalized Procrustes superimposition that distributes error over all point coordinates. With the exception of the landmarks localized on the outline of the tooth, the landmarks used to capture pig molar shape correspond to type III landmarks on Bookstein typology [46]. We show here that while these landmarks can be used when a single measurer operate, they have to be avoided for multi-operator comparisons. Here, these landmarks provide information on shape changes on the occlusal surface of the tooth, not located on the outline and thus not captured by outline analyses. As a consequence, removing those points will allow to gain IO repeatability but at the same time result in a loss of biological information. One could decide to fix an error cut-off but it would require to compare the amount of error removing the landmarks one by one and contrast the gain in repeatability with the biological information carried by the data. Outlines analyses appeared more influenced by the picture on which the coordinates were obtained than the sliding semi-landmarks approaches suggesting important differences in data processing. Slight differences between pictures have more impact in the elliptic harmonic coefficients than on superimposed sliding semi-landmarks coordinates. We show that not all datasets are equal in terms of discrimination power nor in terms of error they embed. We show also that it is possible to simplify datasets without affecting discrimination up to a satisfying, optimized, level. For datasets purely made of landmark coordinates or linear measurements, it is also possible to iteratively jackknife variables by starting to remove points or measurement that increase overall amount of error while simultaneously not decreasing statistical power. One should keep in mind that more is not always better and that artificially inflating the number of variables by adding more points along curves or on surfaces can be in fact statistically counterproductive in terms of the statistical power researched.

Conclusion

Combining datasets requires simultaneously minimizing the impact of ME *and* maximizing the reproducibility under the universal constraint of minimizing data acquisition effort. Our worked example shows that not all morphometric approaches are suitable for data sharing, but the workflow presented here allows for carefully determining those that allow pooling datasets obtained by different operators. We therefore recommend to: (i) systematically measure error at the beginning of any morphometric study and throughout its duration; (ii) evaluate how various sets of alternative methodological approaches can help to provide good predictions; and (iii) optimize data acquisition by downgrading the number of variables to its minimal number while maintaining discriminatory power, to insure enough statistical power to solve the research question.

Supplementary data

Supplementary data are available at *Biology Methods and Protocols* online.

Data accessibility

MEs and a set of photos for assessing intra-operator ME are available from the LabArchives database (accession number: DOI:10.25833/g2bj-2d30).

Acknowledgements

The authors warmly thank the following persons who participated in the data acquisition: Laurent Bouby, Camille Martinand-Mari, Sergio Ferreira-Cardoso, Fabrice Lihoreau, Marine Durocher, Colline Brassard, and Fiona Laviano. They are grateful to Carmelo Fruciano and to one anonymous reviewer for their thoughtful comments that significantly improved the manuscript.

Authors' contributions

AE conceived the study. All authors analyzed and interpreted the data, and wrote the manuscript.

Funding

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [Grant Agreement No. 852573].

Conflicts of interest. The authors declare no conflict of interest.

References

1. Rohlf JF, Marcus LFLF. A revolution morphometrics. *Trends Ecol Evol* 1993;**8**:129–32.
2. Bookstein FL. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis* 1997;**1**:225–43.
3. Dryden IL, Mardia KV. *Statistical shape analysis*. New York, NY: John Wiley and Sons, 1996.
4. Perez SI, Bernal V, Gonzalez PN. Differences between sliding semi-landmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. *J Anatomy* 2006;**208**:769–84.
5. Kuhl FP, Giardina CR. Elliptic Fourier features of a closed contour. *Comput Graph Image Process* 1982;**18**:236–58.
6. Rohlf FJ, Archie JW. A comparison of fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). *Syst Zool* 1984;**33**:302.
7. Goswami A, Watanabe A, Felice RN *et al*. High-density morphometric analysis of shape and integration: the good, the bad, and the not-really-a-problem. *Integr Comp Biol* 2019;**59**:669–83.
8. Cornette R, Baylac M, Souter T, Herrel A. Does shape covariation between the skull and the mandible have functional consequences? A 3D approach for a 3D problem. *J Anat* 2013;**223**:329–36.
9. Cardini A. Less tautology, more biology? A comment on "high-density" morphometrics. *Zoomorphology* 2020a;**139**:513–29.
10. Chang J, Alfaro ME. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods Ecol Evol* 2016;**7**:472–82.

11. Trut LN, Prasolova LA, Kharlamova AV, Plyusnina IZ. Directional left-sided asymmetry of adrenals in experimentally domesticated animals. *Bull Exp Biol Med* 2002;**133**:506–9.
12. Fruciano C. Measurement error in geometric morphometrics. *Dev Genes Evol* 2016;**226**:139–58.
13. Fruciano C, Celik MA, Butler K *et al.* Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. *Ecol Evol* 2017;**7**:7034–46.
14. Fox NS, Veneracion JJ, Blois JL. Are geometric morphometric analyses replicable? Evaluating landmark measurement error and its impact on extant and fossil *Microtus* classification. *Ecol Evol* 2020;**10**:3260–75.
15. Vrdoljak J, Sanchez KI, Arreola-Ramos R *et al.* Testing repeatability, measurement error and species differentiation when using geometric morphometrics on complex shapes: a case study of Patagonian lizards of the genus Liolaemus (Squamata: liolaemini). *Biol J Linn Soc* 2020;**130**:800–12.
16. Fruciano C, Schmidt D, Ramírez Sanchez MM *et al.* Tissue preservation can affect geometric morphometric analyses: a case study using fish body shape. *Zool J Linn Soc* 2020;**188**: 148–62.
17. Daboul A, Ivanovska T, Bülow R *et al.* Procrustes-based geometric morphometrics on MRI images: an example of inter-operator bias in 3D landmarks and its impact on big datasets. *PLoS ONE* 2018;**13**:e0197675–20.
18. Adams DC, Rohlf FJ, Slice DE. Geometric morphometrics: ten years of progress following the 'revolution'. *Ital J Zool* 2004;**71**: 5–16.
19. Adams DC, Rohlf FJ, Slice DE. A field comes of age: geometric morphometrics in the 21st century. *Hystrix* 2013;**24**:7–14 (doi: 10.4404/hystrix-24.1-6283).
20. Yezerinac SM, Lougheed SC, Handford P. Measurement error and morphometric studies: statistical power and observer experience. *Syst Biol* 1992;**41**:471–82.
21. Arnqvist G, Mårtensson T. Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. *Acta Zool Acad Sci Hung* 1998;**44**: 73–96.
22. Evin A, Cucchi T, Cardini A *et al.* The long and winding road: identifying pig domestication through molar size and shape. *J Archaeol Sci* 2013;**40**:735–43.
23. Evin A, Cucchi T, Escarguel G *et al.* Using traditional biometrical data to distinguish West Palearctic wild boar and domestic pigs in the archaeological record: new methods and standards. *J Archaeol Sci* 2014;**43**:1–8.
24. Evin A, Dobney K, Schafberg R *et al.* Phenotype and animal domestication: a study of dental variation between domestic, wild, captive, hybrid and insular *Sus scrofa*. *BMC Evol Biol* 2015; **15**:6.
25. Rowley-Conwy P, Albarella U, Dobney K. Distinguishing wild boar from domestic pigs in prehistory: a review of approaches and recent results. *J World Prehist* 2012;**25**:1–44.
26. Vigne J-D, Peters J, Helmer D. *The First Steps of Animal Domestication: New Archaeozoological Techniques (Proceedings of the 9th ICAZ Conference)*. Oxford: Oxbow Books Limited, 2005.
27. Evin A, Dobney K, Cucchi T. A history of pig domestication: new ways of exploring a complex process. In: Melletti M, Meijaard E (eds), *Ecology, Conservation and Management of Wild Pigs and Peccaries*. Cambridge: Cambridge University Press, 2017, 39–48.
28. Rohlf FJ. The tps series of software. *Hystrix* 2015;**26**:1–4 (doi: 10.4404/hystrix-26.1-11264).
29. Schlager S. Morpho and Rvcg—shape analysis in {R}. In: Zheng G, Li S, Szekely G (eds), *Statistical Shape and Deformation Analysis*. San Diego, CA: Academic Press, 2017, 217–56.
30. Adams DC, Collyer ML, Kaliontzopoulou A. *Geomorph: Software for Geometric Morphometric Analyses*. R package version 3.1.0, 2019
31. Bonhomme V, Picq S, Gaucherel C, Claude J. Momocs: outline analysis using R. *J Stat Soft* 2014;**56**:1–24.
32. Mosimann JE. Size allometry: size and shape variables with of the lognormal characterizations and generalized gamma distributions. *J Amer Statist Assoc* 1970;**65**:930–45.
33. Claude J. *Morphometrics with R*. New York, NY: Springer, 2013.
34. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *J Stat Soft* 2015;**67**:1–48.
35. Baylac M, Friess M. Fourier descriptors, procrustes superimposition and data dimensionality: an example of cranial shape analysis in modern human population. In Slice DE (ed.), *Modern Morphometrics in Physical Anthropology, Part 1: Theory and Methods*. New York: Kluwer Academic/Plenum Publishers, 2005, 142–65.
36. Chiari Y, Claude J. Morphometric identification of individuals when there are more shape variables than reference specimens: a case study in Galápagos tortoises. *C R Biol* 2012;**335**: 62–8.
37. Cardini A. Modern morphometrics and the study of population differences: good data behind clever analyses and cool pictures? *Anat Rec* 2020b;February:1–19 (doi: 10.1002/ar.24397).
38. Kohn L, Cheverud J. Anthropometric imaging system repeatability. In: Vannier M, Yates R, Whitestone J (eds), *Proceedings of the Cooperative Working Group in Electronic Imaging of the Human Body*. Dayton (OH): CSERIAC, 1992, 114–23.
39. Lele SR, Richtsmeier JT. (2001). *An Invariant Approach to Statistical Analysis of Shapes*. Boca Raton (FL): Chapman and Hall/CRC (doi: 10.1201/9781420036176).
40. Rabinovich SG. *Measurement Errors: Theory and Practice*. New York (NY): American Institute of Physics, 1995.
41. Evin A, Souter T, Hulme-Beaman A *et al.* The use of close-range photogrammetry in zooarchaeology: creating accurate 3D models of wolf crania to study dog domestication. *J Archaeol Sci Rep* 2016;**9**:87–93.
42. Osis ST, Hettinga BA, Macdonald SL, Ferber R. A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. *Comput Methods Biomech Biomed Eng* 2015;**18**:1108–16.
43. Rabinovich SG. *Measurement Errors and Uncertainties*. 3rd edn. New York: Springer, 2005.
44. Bookstein FL. Pathologies of between-groups principal components analysis in geometric morphometrics. *Evol Biol* 2019; **46**:271–302.
45. Cardini A, O'Higgins P, Rohlf FJ. Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evol Biol* 2019;**46**:303–16.
46. Bookstein FL. *Morphometric Tools for Land- Mark Data: Geometry and Biology*. New York: Cambridge University Press, 1991