Cretaceous Research 55 (2015) 200–209

Contents lists available at ScienceDirect

Cretaceous Research

journal homepage: www.elsevier.com/locate/CretRes

Missing data estimation in tyrannosaurid dinosaurs: Can diameter take the place of circumference?

Gavin J. Bradley*, Michael E. Burns, Philip J. Currie

Department of Biological Sciences, University of Alberta, 11455 Saskatchewan Drive, Edmonton, Alberta T6G 2E9, Canada

A R T I C L E I N F O

Article history:
Received 6 October 2014
Accepted in revised form 24 February 2015
Available online 23 March 2015

Keywords:
Missing data
Femoral diameter
Tyrannosaurid
Femoral circumference

A B S T R A C T

Missing data, due to taphonomic deformation, inaccessibility of specimens, or human error in collecting, cataloguing and measuring features, is a formidable problem in current palaeontological studies. Missing values within a data set can undermine confidence in analyses, skew results in promoting analyses of small portions of a population, not necessarily representative of the entire data set, and drastically decrease sample sizes. Missing data estimation methods, however, may reduce the effects of these missing values and potentially boost sample sizes for palaeontological studies. Here, six missing data estimation models for the prediction of femoral circumferences in tyrannosaurids were statistically tested and their predictive success measured against true circumferences, and other models. The statistical analyses suggest that estimation models based on anteroposterior diameter values of tyrannosaurid femora were poor predictors of circumference, whereas those based on mediolateral diameters were much more successful. Three out of the six models, were presented as viable alternatives to missing measured circumferences and may be used to boost tyrannosaurid samples with significant levels of missing data.

1. Introduction

Missing data are a major limitation in palaeontological studies, decreasing confidence in analyses and drastically reducing sample sizes (Kearney & Clark, 2003). The problem of missing data is far reaching within many subsets of the discipline, and affects a host of palaeontological methodologies. In fossil reconstructions, missing data can limit the accuracy and extent to which an extinct organism can be recreated (Grillo and Azevedo, 2011). It can also impact the rate at, and specificity with, which new specimens may be described and assigned a phylogenetic position; missing data have prevented more exclusive phylogenetic resolution for the potential Chinese tyrannosauroid *Chingkangosaurus fragilis*, known only from fragmentary scapula remains (Brusatte et al., 2013), and delayed a comprehensive phylogenetic analysis of *Deinocheirus mirificus* for fifty years (Lee et al., 2014).

The potential impact of missing data in morphometric analyses is well known, and alternative missing data estimation techniques, such as 'Bayesian Principal Component Analysis' and 'mean substitution' have been proposed and compared (Campione and Evans, 2011; Brown et al., 2012). Phylogenetic analyses are also complicated by the effect of missing data (Wilkinson, 2003; Wiens, 2003; Wiens, 2006; Wiens and Morrill, 2011; Lemmon et al., 2009), and in this subfield of palaeontology, missing data estimation models have also been devised to reduce the effect of spaces in the sample (Norell & Wheeler, 2003).

The most common cause of missing data in a palaeontological sample is taphonomic deformation (Chapman, 1990; Dilkes, 2001; Miyashita et al., 2011; Arbour & Currie, 2012; Tschopp et al., 2013; Hedrick & Dodson, 2013). The geological pressures that distort and destroy fossils as they form are beyond the control of palaeontologists; however, paucity in sample size can also be compounded by anthropological factors. Human error might lead to the destruction of fossils due to accidental damage or vandalism (Lipps, 2009), the misplacement of specimens or accompanying information between excavation and preparation, or even inappropriate measurement of specimens. Missing data can also be generated by lack of access to the complete specimen due to backlogs in fossil preparation laboratories, its inclusion in a mounted museum or university display, or even poor relations between palaeontological institutions.

This investigation looks at femoral circumference in tyrannosaurid dinosaurs as a case study in the examination of missing data

* Corresponding author. Tel.: +1 7807072077.
E-mail address: gbradley@ualberta.ca (G.J. Bradley).

http://dx.doi.org/10.1016/j.cretres.2015.02.010
0195-6671/

sets within palaeontology. Load bearing bones like femora form an important component of the mammalian-like erect posture that allowed dinosaurs to attain great speeds and vast body sizes (Persons & Currie, 2011a; Persons & Currie, 2011b; Benton et al., 2014; Seymour et al., 2011; O'Gornan & Hone, 2012). More specifically, femoral circumferences are important in several subfields of palaeontology. Their proportions compared to other limb bones have been used in descriptive studies, such as that of *Gigantoraptor erlianensis* (Xu et al., 2007), and allometric scaling studies have also made use of femoral dimensions (Christiansen, 1998, 1999; Christiansen and Fariña, 2004; Carrano, 2001; Bybee et al., 2006; Kilbourne & Makovicky, 2010; Funston et al., 2015). Lee et al. (2008) also used femoral circumferences along with histological growth lines to create a life history model for *Hypacrosaurus*.

Missing femoral data are a problem with tyrannosaurids particularly because they are bipedal and thus have a reduced number of bones which can be used for bone loading (e.g. Farlow et al., 1995), locomotion (e.g. Heinrich et al., 1993; Persons & Currie, 2011a; Persons & Currie, 2011b) and body mass studies (e.g. Erickson et al., 2004; Campione & Evans, 2012; Campione et al., 2014; Benson et al., 2014). Tyrannosaurids face an additional problem in that they are also popular mounted exhibits in museums, and thus access to measure femoral circumference will often be impossible.

Herein, we used anteroposterior and mediolateral diameter measurements to predict femoral circumference in tyrannosaurids through six different missing data estimation models, which may be divided crudely into two main groups: generalised geometric equation models and regression based equation models. By comparing the predictive success of the model generated circumferences to a set of true circumferences, and the relative performance of each model by juxtaposing error values, it was hoped to assess whether or not tyrannosaurid femoral diameters may be used to predict circumferences in data sets with missing measurements. If so, this could increase the sample sizes for studies involving femoral circumference in tyrannosaurids and minimise the effects of missing data.

2. Materials and methods

Six models were tested in this study: (1) **CAP**, the circumference of a circle using the diameter of the anteroposterior shaft width (x); (2) **CML**, circumference of a circle using the diameter of the mediolateral shaft width (y); (3) **ELL**, circumference of an ellipse using both anteroposterior and mediolateral diameters; (4) **APR**, linear regression equation for the anteroposterior diameter of the shaft and true circumference; (5) **MLR**, linear regression equation for the mediolateral diameter of the shaft and true circumference; (6) **ELLR**, multiple linear regression equations for both the true anteroposterior and mediolateral diameters of the shaft and true circumference. The equation for the circular models, **CAP and CML** was:

$$C = \pi D$$

Estimating the femoral circumference using **ELL**, the anteroposterior and mediolateral diameters (x and y) were expressed as a radical fraction in the standard equation:

$$C \approx 2\pi \frac{\sqrt{x^2 + y^2}}{2}$$

To create predictive equations for **APR** and **MLR**, true diameters (x axis) were plotted against corresponding true circumferences (y axis) in bivariate linear regression graphs using Microsoft Excel 2007. The linear relationship between the true circumferences and their corresponding diameters was expressed in the standard regression format: $y = mx + b$, where b is the 'y axis' intercept, m is the slope of the line, and 'y' and 'x' represent the circumference and diameter values respectively (mm). These equations may be used by substituting a known diameter in place of 'x' and calculating a predicted circumference at 'y'. Because regressions of raw data introduced violations of the underlying assumptions of linear regressions (Seber and Lee, 2003), all data were log-transformed before regression. Log-transformation of data has been demonstrated to reduce the influence of outliers in previous palaeontological morphometric regression analyses (Cawley and Janacek, 2010; Campione and Evans, 2012). Because the experimental design was unbalanced, due to the range in sample sizes of each genus (McDonald, 2014), each raw, true diameter and circumference was appropriately weighted using the following formula in order to fulfil the statistical assumption of independence:

$$WD = R \cdot \frac{1}{n}$$

where: WD = Weighted Datum, R = Raw Datum and n = sample size of genus.

Paired two-tailed Student's t-tests were carried out between each model's set of predicted and true circumferences using Microsoft Excel 2007 and repeated using Graph Pad 'Quick Calcs'. Prior to carrying out these t-tests, data from the regression based formulae were required to be back-transformed and weighted back to the scale of the original raw values. The latter was performed using the inverse of the previous equation:

$$R = \frac{WD}{\left(\frac{1}{n}\right)}$$

Percent Prediction Error (PPE) and Standard Error of the Estimate (SEE) values were calculated for each set of predictions using the 'ppe' and 'see' functions in 'MASSTIMATE' (Campione, 2013; Campione et al., 2014): a package within 'R' (R Development Core Team, 2014). Unpaired two-tailed t-tests were carried out between the PPE values of all models, as well as between SEE values. For the latter set of t-tests, five SEE values, one from each genus of tyrannosaurid, were used for each model.

Due to the high number of t-tests carried out on PPE and SEE values, it was necessary to take into account the increased potential for false discoveries, or, the 'False Discovery Rate' (Benjamini and Hochberg, 1995). For this, adjusted p-values, or 'q-values' were returned using the 'p.adjust' function within the 'stats' package in 'R' (R Development Core Team, 2014). These values were then compared to a statistical significance cut-off value (c), calculated at $\alpha = 0.05$, using the formula suggested by Benjamini and Hochberg (1995), to identify false positives from the multiple t-tests.

Measurements were taken with a tape measure or digital callipers. The anteroposterior and mediolateral diameters were measured at the minimum shaft width of the femur, which in tyrannosaurids is distal to the mid-length of the femur. The original sample set consisted of 71 tyrannosaurid femora; however, only 51 had measured circumferences and therefore could be used as comparisons for predictive models (Appendix 1). The taphonomic alteration of each specimen dictated which diameters could be measured and hence, which models could be tested for which specimen. Thus, 45 anteroposterior diameters were available to test the **CAP** and **APR** models, 41 could be tested using **CML** and **MLR**, and only 35 were complete enough to be tested using **ELL** and **ELLR**. The femora represent five tyrannosaurid genera: *Albertosaurus, Daspletosaurus, Gorgosaurus, Tarbosaurus* and *Tyrannosaurus*. One specimen, BMNH 2002.004.001, previously designated