Received: 16 August 2016 | Revised: 14 January 2017 | Accepted: 25 February 2017

DOI 10.1002/ajpa.23206

TECHNICAL NOTE

WILEY

Comparisons of statistical techniques to assess age-related skeletal markers in bioarchaeology

Colleen M. Cheverko[1] | Mark Hubbe[1,2]

[1]Department of Anthropology, The Ohio State University, Columbus, Ohio

[2]Instituto de Investigaciones Arqueológicas y Museo, Universidad Católica del Norte, Chile

Correspondence
Colleen Cheverko, 4034 Smith Laboratory, 174 West 18th Avenue, Columbus, OH 43210.
Email: cheverko.1@osu.edu

Abstract

Objectives: Many authors argue that inconsistencies between studies of skeletal markers are based on different data collection protocols, especially when comparing age-related markers such as osteoarthritis. Less attention is given to the choice of statistical techniques that are used to test the hypotheses associated with the data. This paper addresses how different statistical techniques compare the prevalence of age-related skeletal indicators, specifically osteoarthritis.

Materials and methods: Osteoarthritis prevalence was scored in eight postcranial joints in 243 adult individuals from seven prehistoric archaeological sites in Central California, and data was compared between three time periods [Early (4800–2800 BP), Middle (2800–1200 BP), and Late (1200–250 BP)] using commonly used statistical tests: chi-square, Fisher's exact, and odds ratios. In addition, we analyzed the data with tests that are able to take into consideration the effect of age on osteoarthritis prevalence: ANCOVA and Factorial ANOVA. Finally, bootstraps were applied to the data to investigate how fluctuating frequencies, sample size, and age-at-death distributions affected the interpretations resulting from each test.

Results: The results demonstrate that the tests that consider age as a covariate (ANCOVA and Factorial ANOVA) are more efficient in rejecting the null hypothesis when smaller magnitudes of difference are observed between samples, irrespective of sample size, even though osteoarthritis prevalence fails to meet assumptions of normal distribution and homoscedasticity.

Discussion: ANCOVAs or Factorial ANOVAs that incorporate age as a covariate should be considered more often in studies that test different prevalences of age-related osteological markers among past populations.

KEYWORDS
ANCOVA, osteoarthritis, quantitative methods

1 | INTRODUCTION

In studies of age-related skeletal markers, especially osteoarthritis, authors are frequently concerned with how to compare their results to the published literature, since comparison between samples is one of the most straightforward ways to establish a context of relevance for the data collected. Most of these discussions focus on how conflicting data collection protocols complicate the ability to compare results between studies (Bridges, 1993; Waldron & Rogers, 1991). For example, studies that apply the criteria for scoring osteoarthritis prevalence proposed by Rogers and Waldron (1995) may not directly correlate to studies that apply the criteria for scoring osteoarthritis severity proposed by Jurmain (1999). Other authors suggest that differing sample sizes complicate their ability to compare results across studies, but this effect is less challenging to overcome than the first issue (e.g., Bartelink, 2001). Few authors, if any, discuss how choices of statistical tests can impact comparability between interpretations, despite how statistical choices may impact the final conclusions derived from comparisons between samples.

This question is important to the osteoarthritis literature because of the various statistical techniques applied in previous studies. A review of 33 bioarchaeological studies that use osteoarthritis demonstrates that 13 different statistical tests were used in the past to analyze the prevalence of this osteological marker of activity (Table 1).

TABLE 1 Literature review of statistical tests used to access osteoarthritis patterns

Test	N	Citations
Chi-square	12	Bartelink (2001), Bridges (1991), Cheverko (2013), Debono, Mafart, Jeusel, and Guipert (2004), Hemphill (1999), Hollimon (1988), Jurmain (1980), Novak and Slaus (2011), Smith (2004), Ullinger, Sheridan, and Ortner (2012), Williamson (2000), and Woo and Pak (2014)
Fisher's exact	8	Bartelink (2001), Cheverko (2013), Crubezy et al. (2002), Debono et al. (2004), Lieverse, Weber, Bazaliiskii, Goriunova, and Savel'ev (2007), Ortner (1968), Ullinger et al. (2012), and Woo and Sciulli (2011)
Spearman coefficient	6	Lieverse, Bazaliiskii, Goriunova, and Weber (2013), Molnar, Ahlstrom, and Leden (2011), Palmer, Hoogland, and Waters-Rist (2014), Schrader (2012), and Weiss (2006, 2007)
Frequency comparison	4	Bridges (1994), Hollimon (1992), Jurmain (1990), and Waldron (1995)
Mann–Whitney	4	Bridges (1991), Lieverse et al. (2013), Palmer et al. (2014), and Schrader (2012)
Kruskal–Wallis	4	Lieverse et al. (2013), Rando and Waldron (2012), Schrader (2012), and Walker and Hollimon (1989)
Odd's ratio	2	Klaus et al. (2009) and Vrezas, Elsner, Bolm-Audorff, Abolmaali, and Seidler (2010)
Rank order	2	Waldron (1995) and Weiss (2007)
t Test	1	Machicek and Beach (2013)
Linear regression	1	Hemphill (1999)
Logistic regression	1	Baker and Pearson (2006)
Pearson's correlation	2	Hemphill (1999) and Jurmain (1980)
Principal components analysis	1	Jurmain (1991)
Total	48	

Older articles present a high percentage of descriptive frequency comparisons (e.g., Hollimon, 1992; Jurmain, 1990). Chi-square and Fisher's exact tests were used most often in these prior studies (n = 21), followed by Spearman's correlation coefficient (n = 5). More recent papers applied odds ratios to investigate change through time in activity and mobility patterns (e.g., Klaus, Larsen, & Tam, 2009). Methods that correct for the effect of age, such as Analyses of Covariance (ANCOVAs) or Factorial Analyses of Variance (ANOVAs), were not found in any study surveyed for this paper.

The number of potential statistical tests is problematic for comparisons between studies, because they can report contrasting results to some extent. As Bridges (1993) notes, several practices complicate our ability to compare results between studies, including the varying use or lack of statistical tests. From her description, the author concludes that bioarchaeologists will have difficulty determining whether patterns of osteoarthritis, and therefore the changes in activity they aim to infer, are truly valid. Bioarchaeologists have partially addressed this concern over the past two decades by enacting more rigorous statistical testing (Jurmain, Cardosa, Henderson, & Villotte, 2012); however, the use of so many different tests to infer the same information may still be problematic.

Therefore, the goal of this technical note is to compare the results of various statistical techniques commonly employed by bioarchaeologists to test hypotheses about different prevalences between groups, with the ultimate goal of discussing the comparability of studies that use these common tests. Specifically, this paper will determine whether we can compare the results of chi-square, Fisher's exact, and odds ratios tests. ANCOVAs and Factorial ANOVAs, which are able to incorporate the effect of variation due to age (Hair et al., 2009) in the test of the Null Hypothesis, are also considered to determine whether they can be used to better address similar research questions. The question of whether ANCOVAs and Factorial ANOVAs can be used in this context is relevant to all studies that use age-related skeletal traits because the use of one test that addresses an age covariate may increase the statistical power of the analysis by simultaneously testing multiple independent variables at once, including any potential age-related biases, and as such are better alternatives to the traditional strategy of breaking down the sample into smaller sub-samples to restrict comparisons within specific age-categories. Combined, these goals represent a key step to any bioarchaeological project where investigators must decide which methods to use, including which statistical test is most appropriate to assess the hypotheses.

2 | MATERIALS AND METHODS

The skeletal collections used in this study include individuals from seven archaeological sites located within the Sacramento-San Joaquin Delta region (Table 2). The sites date to the Early (4800–2800 BP), Middle (2800–1200 BP), and Late (1200–250 BP) periods based on time-sensitive artifacts and chronometric dating techniques (Bennyhoff, 1994; Bouey, 1995; Heizer, 1949; Hoffman, 1987; Johnson, 1937; Lillard, Heizer, & Fenenga, 1939; Ragir, 1972; Rosenthal, 2009). For this analysis, the sites are aggregated by time period to increase the sample size within each period. The remains are curated at the Phoebe A. Hearst Museum of Anthropology at the University of California, Berkeley, where initial data collection occurred.

Osteoarthritis prevalence was scored in the hip, knee, shoulder, and elbow of each individual using criteria established by Rogers and Waldron (1995). Osteoarthritis was scored as present when eburnation was present or there was a combination of marginal lipping, new bone formation or pitting on the joint surface, or an alteration in the shape of the joint (Rogers & Waldron, 1995). Skeletal elements were scored when 50% or more of the joint surfaces were present (Jurmain, 1991). Age and sex of each individual were estimated using criteria summarized by Buikstra and Ubelaker (1994). Adults were assigned to one of three age categories: young adult (20–29), middle adult (30–39), and

TABLE 2 Central California Archaeological Sample sizes according to period and age categories

Period	Central California Archaeological Sites	20–29	30–39	40+	Total
Early	CA-SJO-68, CA-SJO-142	24	25	34	83
Middle	CA-SAC-43, CA-SAC-60, CA-SJO-154	29	17	18	64
Late	CA-SAC-06, CA-SAC-43, CA-SAC-60, CA-SJO-154, CA-CCO-138	30	28	38	96
Total		83	70	90	243

older adult (40+). A three-stage aging system is a clear limitation in the estimation of age-at-death, since it creates arbitrary categories with limits that do not represent the biological nature of the aging process, and also limits any interpretations about osteological markers that appear in the older age group (Milner & Boldsen, 2012; Milner, Wood, & Boldsen, 2008). This is even more problematic for osteological markers such as osteoarthritis, which have increased prevalence in older ages (Aspden, 2008; Bitton, 2009; Jurmain et al., 2012; Roach & Tilley, 2009; Sharma, 2007). Moreover, these age categories usually underrepresent potential variation within old adults in a population (Milner & Boldsen, 2012; Milner et al., 2008), which is also impactful on analyses of skeletal traits that are age dependent, especially if the traits tend to develop late in an individual's life, as is the case with most degenerative joint diseases. There is an increasing trend in the field of using more refined age estimation techniques, such as Transition Analysis (Boldsen & Milner, 2012; Boldsen, Milner, Konigsberg, & Wood, 2002), and we expect that the availability of data with better estimates of age in the future will be useful to correct the current limitation of this variable in the analyses of age-dependent markers (Knudson & Stojanowski, 2008). However, we chose to keep this three-stage system because it represents a common scoring system used by many contemporary bioarchaeologists (e.g., Fontanals-Coll, Subira, Diaz-Zorita Bonilla, & Gibaja, 2017; Kinaston, Roberts, Buckley, & Oxenham, 2016; Marklein, Leahy, & Crews, 2016; Scott, Choi, Mookherjee, Hoppa, & Larcombe, 2016; Yaussy, DeWitte, & Redfern, 2016; Yonemoto, 2016; Zampetti, Mariotti, Radi, & Belcastro, 2016). Juveniles were excluded because they are less likely to demonstrate osteoarthritis due to its age-progressive nature (e.g., Bridges, 1991; Jurmain, 1990, 1999; Walker & Hollimon, 1989; see Cheverko, 2013 for further details on data collection and comparisons between periods). It is important to state that the use of a limited number of age categories is a limitation in our analyses (see below), particularly for the ANCOVAs, since an ordinal variable with only three stages is not a proper covariate for a general linear model.

Three commonly applied statistical tests—chi-square, Fisher exact, and Odds Ratio—were chosen to evaluate their ability to identify significant differences in osteoarthritis prevalence between time periods based on their use in previous bioarchaeological literature (Table 1). Because age-at-death has a cumulative effect in osteoarthritis prevalence, comparisons of the prevalence between samples need to consider the progressive effect of age on the prevalence of osteoarthritis. To incorporate this source of variation in the analyses, the data were also analyzed within a set of General Linear Model analyses that are able to incorporate multiple independent variables to predict the behavior of a dependent variable, which allows researchers to test the effects of multiple independent categorical variables (e.g., period and sex) and their interactions simultaneously on the dependent variable (e.g., prevalence of osteoarthritis in this case), while taking into account the effect of a numeric covariate variable (e.g., age-at-death). Given that our hypothesis tests for differences in osteoarthritis prevalence between periods, which is defined as a categorical variable, alongside the idea that age can be seen as a numerical, progressive variable, an Analyses of Covariance (ANCOVA) using age as a covariate is a good design to test the null hypothesis of differences between periods. However, as discussed above, the way age was recorded in both our analyses and those of many other studies of this nature limits its use as a true covariate. Therefore, we repeated the analysis as a Factorial ANOVA, assuming both period and age as categorical independent variables. A full Factorial ANOVA requires that the interaction between the independent variables is explored, therefore we considered the interaction between age and period as a third independent variable. Since this model adds a new component in the decomposition of the variance of the dependent variable, its direct comparison with the results of the ANCOVA is not straightforward. Therefore, we also ran a third analysis consisting of the Factorial ANOVA without the inclusion of the interaction between the independent variables. These different analyses were implemented to illustrate the potential impact that violations of the assumptions of the ANCOVA (i.e., that the covariate is numeric, parametric, homoscedastic and linearly associated to the independent variable) have on the final decision of rejecting or supporting the Null hypothesis of the test.

In all these tests, osteoarthritis presence/absence is used as the dependent variable. Osteoarthritis scoring was kept as a binary variable instead of ordinal, because many of the studies that use this marker and other age-related skeletal stress markers are commonly represented that way. However, the power of either ANCOVA or Factorial ANOVA will increase if the dependent variable is not binary, so this consideration needs to be taken into account in the results presented here. Finally, sex was not considered as one of the independent variables in our analyses to facilitate the comparisons of the results between the univariate and the General Linear Model results and to limit the number of interactions considered within the Factorial ANOVAS (a three-way ANOVA would require us to handle four interaction components). However, sex is likely an important factor in many age-related skeletal indicators and should be considered as part of the model in future studies.

The first step of the analysis was to compare the initial *p*-values obtained between the four statistical tests in each of the eight joint

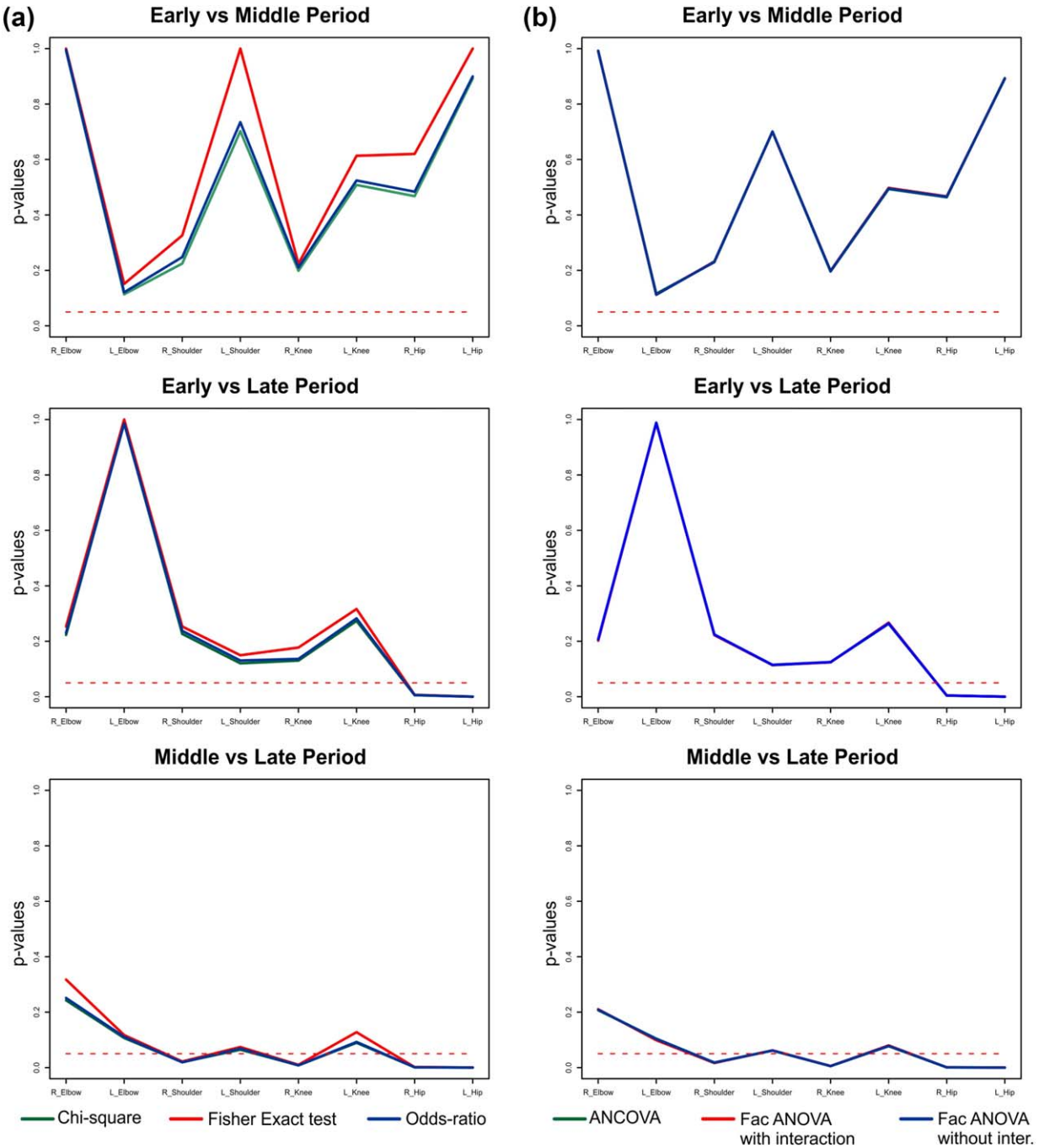

FIGURE 1 Comparisons of results obtained from chi-square, Fisher's exact, and odd's ratios tests and results obtained from ANCOVA and Factorial ANOVA (with and without the interactions of covariates) between the Early and Middle, Middle and Late, and Early and Late periods. Significance ($p = .05$) is marked as the dashed red line

complexes. All tests were performed between pairs of periods (e.g., Early vs. Middle, Early vs. Late, and Middle vs. Late), despite the fact that the tests could be used to test all periods simultaneously. No post-hoc correction was applied to the results obtained, so that the Type I error associated to each test is comparable. This step helps determine how the *p*-values obtained from one test compares with each of the other tests in this skeletal sample.

Following the initial comparisons of *p*-values, data were bootstrapped to explore the impact of varying frequencies, sample sizes, and age-at-death associations on the ability to reject the null hypothesis. In the first analysis, 1,000 permutations of the original data were created to estimate varying frequencies of osteoarthritis in each period, while the original sample size was held constant for each period and joint. The goal of this analysis is to investigate how variations in frequencies between groups affect the results of each test. The second analysis is designed to investigate how fluctuations in sample size and frequency affect the results of the various statistical tests. In this case, a progressive number of individuals were randomly selected 1,000

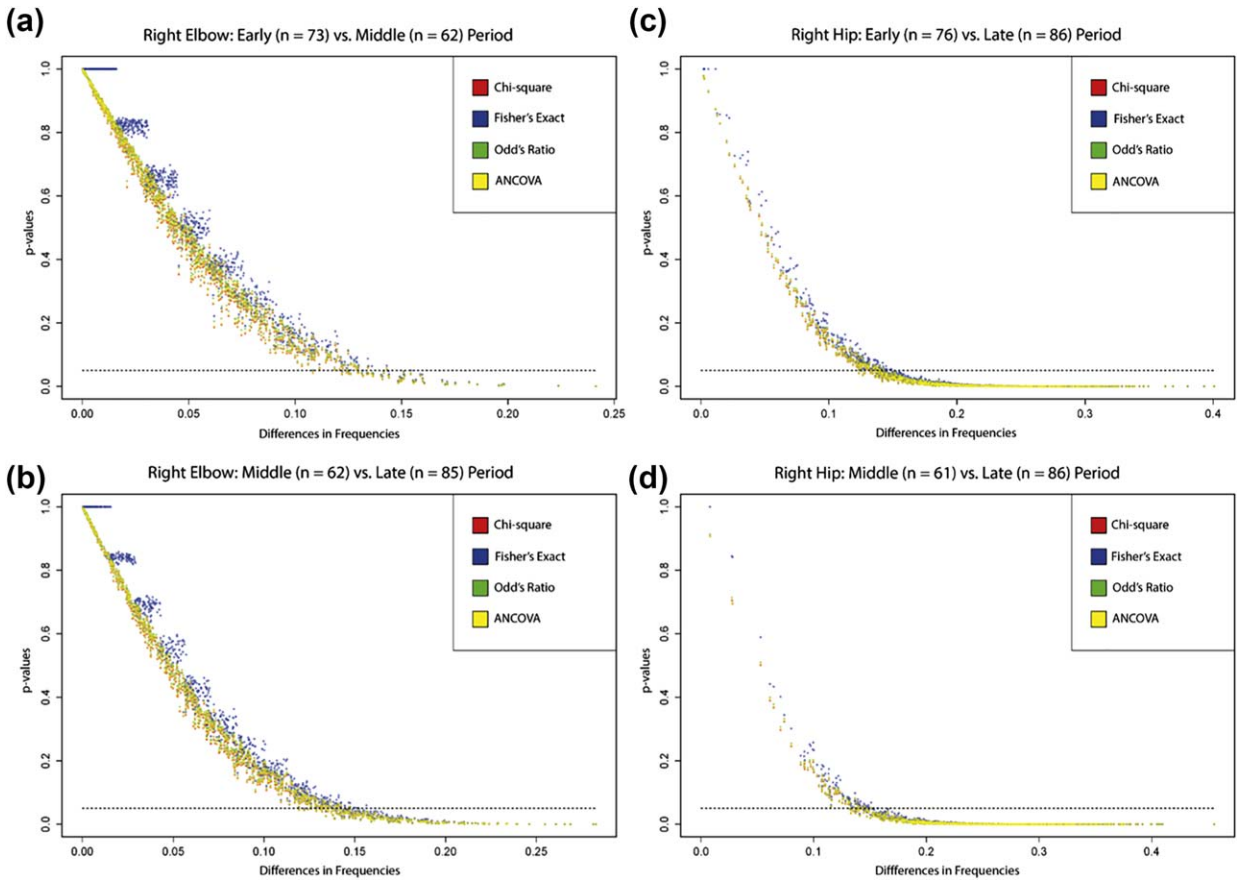

FIGURE 2 Probabilities of Type-I error based on different prevalence of osteoarthritis derived from 1,000 bootstraps of original data between sites for chi-square, Fisher's exact, odd's ratio, and ANCOVA tests. The original data for A and C show no significant differences between sites, while the original data for B and D show significant differences at $\alpha = 0.05$

times from each period and joint to explore the impact of variations of sample size and osteoarthritis prevalence. Sample sizes were changed from n = 5 to n = 100 in increments of five. This analysis was restricted to chi-square, ANCOVA, and Fisher's exact tests. Odds ratios were excluded from this step because of their inability to calculate *p*-values for observed frequencies of zero, which was a common outcome in bootstraps with low sample sizes. Factorial ANOVAs were not included because their results are undistinguishable from the results of the ANCOVAs. Finally, the third analysis represents a re-sampling of age of the individuals 1,000 times to investigate the degree to which *p*-values obtained in ANCOVAs fluctuate given different age-at-death distributions between samples and differing degrees of correlation between age-at-death and osteoarthritis prevalence. An alpha of $p = .05$ was assumed for all tests in this study. The analyses were performed in R (R Core Team, 2016), with functions written specifically for this study.

3 | RESULTS

3.1 | Comparison of *p*-values

While many researchers argue that *p*-values are an ineffective measure of biological "reality" (Baker, 2016; Cohen, 2011), we use *p*-values here as a measurable quantity since they represent the common statistical measure from all tests applied to our data and because ultimately they are the values that allow for tests of Null Hypotheses. However, we are not advocating the use of *p*-values as a measure of effective strength of the statistical test in future research, since they are measurements of probability and not measurements of scale. We compare *p*-values here because we are holding sample size constant throughout each of the tests being run, making them comparable with each other. Despite this choice, the authors acknowledge the importance of testing for effect strength, instead of *p*-values, in studies that are not interested in comparing the efficacy of multiple tests to reject the Null Hypotheses being tested.

Figure 1 represents the *p*-values obtained for each joint complex using the statistical tests between the Early and Middle, Middle and Late, and Early and Late periods. Specifically, Figure 1a represents the results of the univariate statistical tests that do not use age-at-death as a covariate, while Figure 1b represents the results of ANCOVAs and Factorial ANOVAs that include age-at-death as a covariate or second independent variable. As can be seen, the tests produce similar *p*-values, regardless of the significant effects of age in several of the joint complexes. The Fisher's exact test is a more conservative test than the others (Levin & Fox, 2011; Sokal & Rohlf, 2011), so it consistently yielded higher *p*-values than the other tests. The ANCOVAs and Factorial ANOVAs return slightly smaller *p*-values than both the

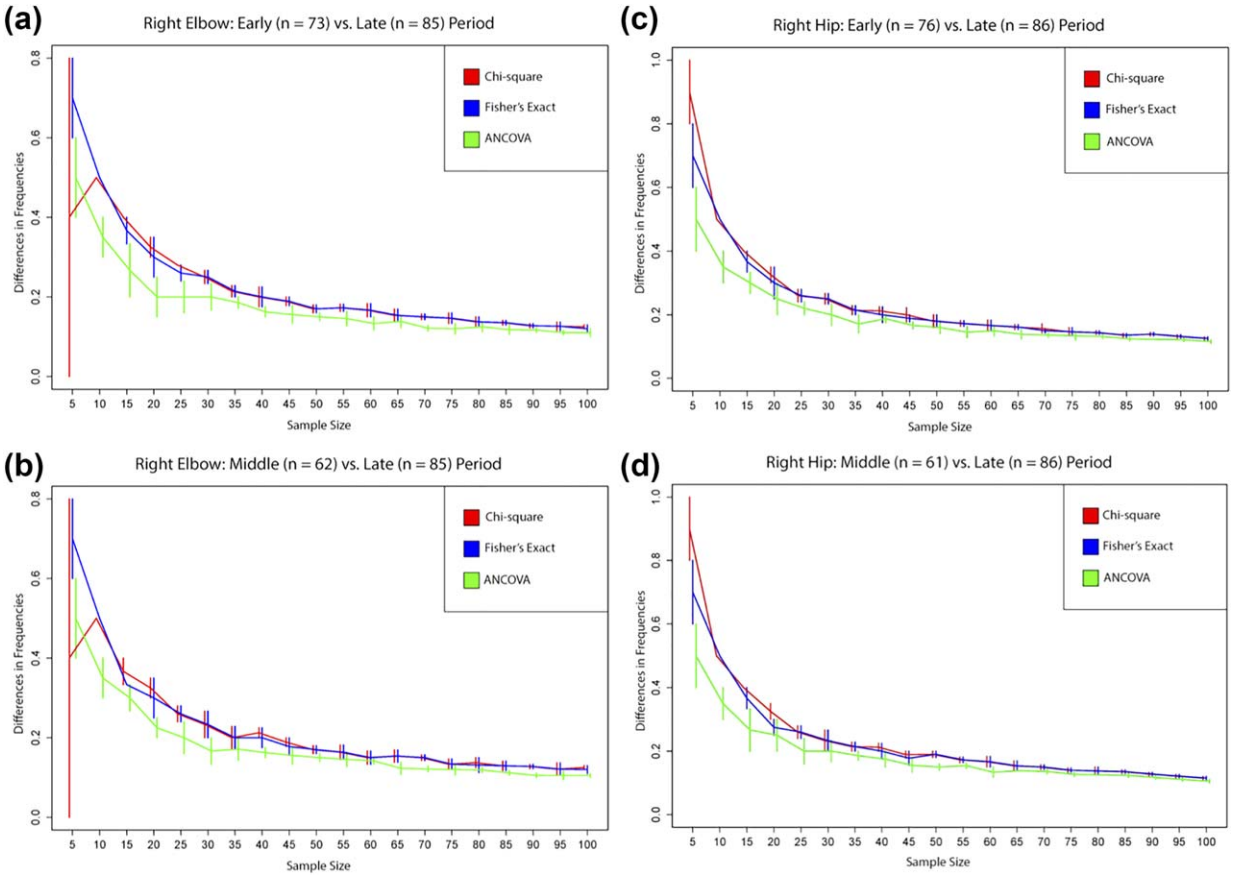

FIGURE 3 Probabilities of Type-I error for each of the statistical tests given increased sample size derived from 1,000 random subsets from the original data. The original data for A and C show no significance between periods, while the original data for B and D show significant differences at $\alpha = 0.05$. The vertical lines plotted at each sample size represent the window of uncertainty for that particular test at that sample size (*x*-axis) after 1,000 permutations, and runs from the minimum difference that was never significant to the maximum difference that was always significant

chi-square and odds ratio tests. However, despite the small differences observed, the tests yield similar conclusions in term of rejecting the Null Hypothesis. That is, they observe significant differences in the same joint for each pairwise comparison, so they have similar abilities to reject or fail to reject the Null Hypothesis in each case presented here.

3.2 | Bootstrap #1: fluctuating frequencies

The first bootstrap explores the effect of fluctuating frequencies on the observed *p*-values. Figure 2 shows the results of the first set of bootstraps between Early and Middle and between Middle and Late Periods for two sets of articulations, one where significant differences were observed in the original data and another where no statistically significant differences were found (Figure 1). The remaining 24 articulations show identical patterns, and therefore are not shown here. The four graphs illustrate the expected pattern of the tests to refute the Null Hypothesis using chi-square, Fisher's exact, odds ratios, and ANCOVA tests according to the difference in percentage of osteoarthritis prevalence obtained from multiple bootstraps of the data. As expected, the probability of Type I error diminishes exponentially when the magnitude of difference between the pairwise samples increases. The four tests perform similarly, though the Fisher's exact test yields higher *p*-values because of its conservative nature. Each of the 24 tests that use different combinations of joints and time periods yields significance beginning with differences of frequencies between 11% and 17%. As expected, sites that have smaller overall sample sizes (see below as well) required larger differences in frequencies between groups before the tests cross the alpha of 0.05. No differences were observed in this pattern when the results were based on articulations that were originally significantly different (Figure 2B,D) and those that were not (Figure 2A,C).

3.3 | Bootstrap #2: changes in sample size

The second test explores the effect of fluctuating sample sizes on the observed *p*-values. Figure 3 represents similar results to the 24 (data not shown) separate pairwise analyses evaluated in this study. The graphs compare the differences in osteoarthritis frequency between pairwise samples in the *y*-axis to the progressive sample sizes selected from the original data. The vertical lines represent the window of uncertainty in the rejection of the Null Hypothesis for that particular test at that sample size after 1,000 resamples of the data. Anything above the line represents differences that were always significant,

while anything below the line represents differences in frequencies that were never significant. In other words, the vertical line represents the interval between the minimum difference that is always significant and the maximum difference that is never significant. Differences between these limits were sometimes significant and sometimes not significant, depending on the overall frequency of presence and absence in the random sub-samples.

Besides the expected reduction of the magnitude of differences required to identify significance as sample size increases, the results of this analysis illustrate two interesting patterns. First, the window of uncertainty decreases as the sample size increases, especially for ANCOVAs. Second, ANCOVAs identify significant differences between groups with smaller magnitudes of difference between the two samples, since part of the overall variance in osteoarthritis prevalence is explained by the covariate (i.e., age-at-death in this case). In other words, by removing the confounding effect of age, the ANCOVAS are able to refute the Null Hypothesis with smaller overall differences. Therefore, ANCOVAs are more efficient at rejecting the Null Hypothesis in this context.

3.4 | Bootstrap #3: changes in age distributions (ANCOVAs)

The third set of bootstraps investigates how differences in age distributions between two samples affect the results of ANCOVAs. For these bootstraps, osteoarthritis prevalence and sample size are held constant, while age-at-death is permutated 1,000 times to generate a distribution of data that have different correlations with age. Figure 4 represents the results of analyses that compare the *p*-value obtained using ANCOVAs to the age-at-death correlation between the pairwise samples. From the permutations of the data, correlation coefficients vary from approximately −0.3 to 0.3. The graphs demonstrate parabolic trends between the strength of the correlation of data with age and the probability of Type I error), where smaller *p*-values are obtained with increased correlations between age-at-death distributions and higher *p*-values are observed when there is no correlation between age-at-death distributions. This association is observed in all 24 comparisons (data not shown). An interesting pattern to note is the different distributions observed between initial tests with significant frequencies between periods and initial tests with insignificant frequencies between periods. The bootstraps where significance initially occurred (Figure 4B) result in larger variances than the bootstraps where there were no significant differences (Figure 4A).

However, it must be remarked that these results are expected because there are no differences in age-at-death distributions in the original sample. In this case, the amount of the variance of the residuals that is explained by the covariate increases when the association with the covariate is larger, which will increase the F-values obtained for the factor being tested. Nonetheless, the differences in *p*-value were observed in the third or fourth decimal case in each of these analyses, which indicates that the impact of the covariate is not noticeable and therefore it has no implication for decisions made about the Null Hypothesis.

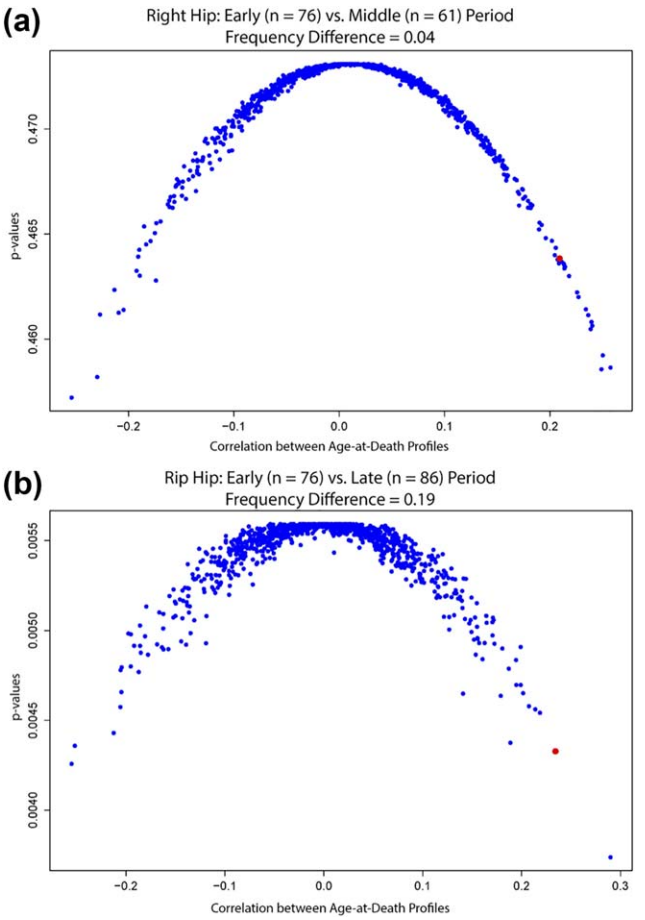

FIGURE 4 Scatterplots comparing the *p*-values obtained from ANCOVAs using age-at-death as a covariate, when the correlations with age are recalculated based on 1,000 bootstraps of the original age-at-death data (blue dots). The red dot represents the results for the original data. The original data for A shows no significant differences between periods, while the original data for B shows significant differences at $\alpha = 0.05$

4 | DISCUSSION

It is important to consider several types of analytical techniques when discussing how various studies can be compared, including choice of data collection protocols (Bridges, 1993; Waldron & Rogers, 1991) and statistical techniques (Bridges, 1993). These considerations help researchers decide which method will be most powerful to test hypotheses with the obtained data. This study investigates how choice of statistical techniques impacts the ability to refute the Null Hypothesis when interpreting the prevalence of age-related skeletal markers. Several important conclusions can be drawn from the results. First, all tests perform similarly with larger sample sizes, beginning at approximately 30 individuals in each group. While the Fisher's exact test is the most conservative test for smaller sample sizes, it achieves significance at similar magnitudes of difference when studies use larger sample sizes. ANCOVAs and Factorial ANOVAs also present very similar results to the other tests, despite including age as a covariate in the analysis.

Previous research indicates Fisher's exact tests should be used instead of chi-square tests when analyzing smaller samples (Levin & Fox, 2011; Sokal & Rohlf, 2011). The results of these analyses indicate ANCOVAs may be another viable option because they consistently identify significance with fewer overall differences between groups. They also partially control for differing age-at-death distributions between groups. Given the low power and high potential for Type-II error associated with smaller sample sizes, which affect many bioarchaeological studies, the inclusion of a covariate in the analyses may be beneficial for the identification of significant patterns when studying skeletal traits that progress with age.

The results obtained using the second set of bootstraps indicate that sample sizes under 35 individuals require larger frequencies to be significant. This effect clearly illustrates the impact that smaller sample sizes have in our ability to refute Null Hypotheses. These results also highlight the trend where sample sizes of 35 or more individuals perform relatively similar, supporting the assertion that bioarchaeological samples should aim to include at least 30–35 individuals in each subgroup for analyses. However, this relatively large sample size often cannot be met in bioarchaeological studies, which limits our ability to make conclusions based on differences in prevalence between groups.

Based on the first two sets of bootstraps, ANCOVAs may be useful in studies of binary skeletal indicators even though the use of binary data violates several assumptions such as homoscedasticity and normal distribution of the data. Despite these limitations, it should be considered a better alternative to handling the influence of age than breaking samples into age categories, which would further reduce sample sizes for analyses. Moreover, despite the fact that a categorical age variable was assumed in the ANCOVAs as a numerical covariate, which also violates the assumption of this test, the results of the ANCOVAs were undistinguishable from the ones obtained in a Full Factorial ANOVA model. The results from the third bootstrap indicate that the strength of age correlations affects *p*-values obtained between the main factor and the data. However, the correlation coefficients vary between -0.3 and $+0.3$ using data from the initial joints and time periods, and the differences in *p*-values are on the third or fourth decimal case. Therefore, the age correlations did not impact the ability to refute the Null Hypothesis in our examples.

In future studies, researchers should build on the simple general linear models presented here to collect data that will make the ANCOVAs more powerful. Specifically, ordinal data should be used instead of presence/absence data, when possible, because ANCOVAs will perform better using variables that reflect biological changes that occur on continuous or semi-continuous scales. In addition, studies that investigate age-related skeletal markers should use aging techniques that more reliably assess age-at-death in older individuals. Using point estimates from techniques such as Transition Analysis will allow the ANCOVA to incorporate age as a numerical independent variable, thus constructing age-at-death as an appropriate covariate in ANCOVA models. The true advantages of using ANCOVAs will be observed when using these higher-level variables; nonetheless, these results indicate that ANCOVAs can perform better than other tests when dichotomous variables are used, such as those commonly observed in bioarchaeological research. However, it is probable that the advantages of using ANCOVAs to study age-related skeletal markers are being underestimated in this study because of the limitations of the data adopted here. To properly construct their variables in the future, bioarchaeologists who wish to use ANCOVAs for their studies should rethink their scoring criteria and experimental design before data collection begins to ensure they have higher-level variables that will increase the utility of ANCOVAs in their research.

5 | CONCLUSIONS

Bioarchaeologists can choose to apply many different statistical tests to address age-related skeletal indicators to their research. This paper tested differences between chi-square, Fisher's exact test, and odds ratios to determine whether the results of studies that use these tests are comparable. It also tested whether ANCOVAs can be used in similar contexts, even though the use of binary data violates assumptions of linearity, homoscedasticity, and normal distribution of the data. The four tests performed similarly when frequency fluctuated between groups, suggesting the results can be compared irrespective of the test used.

However, ANCOVAs require fewer differences between groups to refute the Null Hypothesis with all sample sizes. They also have a smaller window of uncertainty than chi-square and Fisher's exact tests, meaning they are more efficient in refuting the Null Hypothesis than the other two tests. In addition, ANCOVAs are able to include a covariate to remove effects of age from variables that progress linearly with age. While changes in age distributions did not affect the results of ANCOVAs in this study, it will be interesting to see the results when differences between age distributions are exaggerated in future studies.

ANCOVAs identify significance with lower magnitudes of differences than the other tests. Therefore, they should be considered in studies that test the different prevalence of age-progressive osteological markers among past populations, especially given the low power of many bioarchaeological studies with small sample sizes. This conclusion applies to age-related markers such as osteoarthritis, as evidenced by the results. It is also valid for other markers, such as caries or occlusal dental wear, because ANCOVAs are able to include age as a covariate to reduce the uncertainty in interpreting age-related markers.

ACKNOWLEDGMENTS

The authors thank Dr. Tim White, Natasha Johnson, and Dr. Socorro Baez-Molgado for facilitating access to the research collections at the Phoebe Hearst Museum. We also thank Dr. Eric Bartelink for his support and advice during data collection and Kathleen Downey for support during the initial phases of this analysis. We are extremely thankful to two anonymous reviewers for providing constructive criticism that strengthened our final manuscript.

REFERENCES

Aspden, R. M. (2008). Osteoarthritis: A problem of growth not decay? *Rheumatology*, *47*, 1452–1460.

Baker, J., & Pearson, O. M. (2006). Statistical methods for bioarchaeology: applications of age-adjustment and logistic regression to comparisons of skeletal populations with differing age-structures. *Journal of Archaeological Science 33*(2):218–226.

Baker, M. (2016). Statisticians issue warning over misuse of *p* values. *Nature News*, *531*, 151.

Bartelink, E. J. (2001). *Elbow osteoarthritis in the prehistoric San Francisco Bay: A bioarchaeological interpretation of resource intensification and the sexual division of labor* (Published MA thesis). California State University, Chico, Salinas.

Bennyhoff, J. A. (1994). Central California Augustine: Implications for northern California archaeology. In J. A. Bennyhoff & D. A. Fredrickson (Eds.), *Toward a new taphonomic framework for central California archaeology*. Berkeley: Contributions of California Archaeological Research Facility.

Bitton, R. (2009). The economic burden of osteoarthritis. *The American Journal of Managed Care*, *15*, 230–235.

Boldsen, J. L., & Milner, G. R. (2012). An epidemiological approach to paleopathology. In A. L. Grauer (Ed.), *A companion to paleopathology* (pp. 114–132) Hoboken: Wiley-Blackwell.

Boldsen, J. L., Milner, G. R., Konigsberg, L. W., & Wood, J. W. (2002). Transition analysis: A new method for estimating age from skeletons. In Hoppa R. D & J. W. Vaupel (Eds.), *Paleodemography: Age distributions from skeletal samples* (pp. 73–106). Cambridge: Cambridge University Press.

Bouey, P. D. (1995). *Final report on the archaeological analysis of CA-SAC-43, cultural resource mitigation for the Sacramento Urban Area Levee Reconstruction Project*. Sacramento County, CA. Far Western Anthropological Research Group.

Bridges, P. S. (1991). Degenerative joint disease in hunter-gatherers and agriculturalists from the southeastern United States. *American Journal of Physical Anthropology*, *85*, 379–391.

Bridges, P. S. (1993). The effect of variation in methodology on the outcome of osteoarthritis studies. *International Journal of Osteoarchaeology*, *3*, 289–295.

Bridges, P. S. (1994). Vertebral arthritis and physical activities in the prehistoric southeastern United States. *American Journal of Physical Anthropology*, *93*, 83–93.

Buikstra, J. E., & Ubelaker, D. H. (1994). *Standards for data collection from human skeletal remains: Proceedings of a seminar at the Field Museum of Natural History (Arkansas Archaeological Report Research Series)*. Fayetteville: Arkansas Archaeological Survey Press.

Cheverko, C. M. (2013). *A bioarchaeological analysis of osteoarthritis patterns in prehistoric Central California* (Unpublished MA thesis). California State University, Chico.

Cohen, H. W. (2011). P-values: Use and misuse in medical literature. *American Journal of Hypertension*, *24*, 18–23.

Crubezy, E., Goulet, J., Bruzek, J., Jelinek, J., Rouge, D., & Ludes, B. (2002). Epidemiology of osteoarthritis and enthesopathies in a European population dating back 7700 years. *Joint Bone Spine*, *69*, 580–588.

Debono, L., Mafart, B., Jeusel, E., & Guipert, G. (2004). Is the incidence of elbow osteoarthritis underestimated? Insights from paleopathology. *Joint Bone Spine*, *71*, 397–400.

Fontanals-Coll, M., Subira, M. E., Diaz-Zorita Bonilla, M., & Gibaja, J. F. (2017). First insight into the Neolithic subsistence economy in the north-east Iberian Peninsula: Paleodietary reconstruction through stable isotopes. *American Journal of Physical Anthropology*, *162*, 36–50.

Hair, J. F., Black, W. C., Babin, B. J., Anderson R. E. 2009. Multivariate data analysis, seventh edition. Upper Saddle River, NJ: Prentice Hall.

Heizer, R. F. (1949). The archaeology of central California I: The early horizon. *University of California Anthropological Records*, *12*, 1–84.

Hemphill, B. E. (1999). Wear and tear: Osteoarthritis as an indicator of mobility among the Great Basin hunter-gatherers. In B. E. Hemphill & C. S. Larsen (Eds.), *Prehistoric lifeways in the Great Basin wetlands: Bioarchaeological reconstruction and interpretation*. Salt Lake City: University of Utah Press.

Hoffman, M. J. (1987). *The descriptive physical anthropology of the Cardinal Site, CA-SJO-154: A late Middle Horizon-Early Phase I site from Stockton, California*. Colorado College Publications in Anthropology, No. 12. Colorado Springs: Department of Anthropology, Colorado College.

Hollimon, S. E. (1988). Age and sex related incidence of degenerative joint disease in skeletal remains from Santa Cruz Island, California. In G. Richards (Ed.), *Human skeletal biology: Contributions to the understanding of California's prehistoric populations*. Archives of California's prehistory No. 24. Salinas, CA: Coyote Press.

Hollimon, S. E. (1992). Health consequences of sexual division of labor among Native Americans: The Chumash of California and the Arikara of the northern Plains. In C. Classen (Ed.), *Exploring gender through archaeology: Selected papers from the 1991 Boone conference*. Madison: Prehistory Press.

Johnson, E. M. (1937). *Notes on Hotchkiss Mound, near Knightsen, California. Manuscript 14A-2. Berkeley, California: Phoebe A. Hearst Museum of anthropology*, Berkeley: University of California.

Jurmain, R. (1990). Paleoepidemiology of a central California prehistoric population from CA-ALA-329: II. Degenerative disease. *American Journal of Physical Anthropology*, *83*, 83–94.

Jurmain, R. (1991). Degenerative changes in peripheral joints as indicators of mechanical stress: Opportunities and limitations. *International Journal of Osteoarchaeology*, *1*, 247–252.

Jurmain, R. (1999). *Stories from the skeleton: Behavioral reconstruction in human osteology*. Amsterdam: Gordon and Breach.

Jurmain, R., Cardosa, F. A., Henderson, C., & Villotte, S. (2012). Bioarchaeology's Holy Grail: The reconstruction of activity. In A. L. Grauer (Ed.), *A companion to paleopathology*. Oxford: Wiley Blackwell.

Jurmain, R. D. (1980). The pattern of involvement of appendicular degenerative joint disease. *American Journal of Physical Anthropology*, *53*, 143–150.

Kinaston, R. L., Roberts, G. L., Buckley, H. R., & Oxenham, M. (2016). A bioarchaeological analysis of oral and physiological health on the south coast of New Guinea. *American Journal of Physical Anthropology*, *160*, 414–426.

Klaus, H. D., Larsen, C. S., & Tam, M. E. (2009). Economic intensification and degenerative joint disease: Life and labor on the postcontact North coast of Peru. *American Journal of Physical Anthropology*, *139*, 204–221.

Knudson, K. J., & Stojanowski, C. M. (2008). New directions in bioarchaeology: Recent contributions to the study of human social identities. *Journal of Archaeological Research*, *16*, 397–432.

Levin, J., & Fox, J. (2011). *Elementary statistics in social research: The essentials*. Boston, MA: Allyn and Beacon.

Lieverse, A. R., Bazaliiskii, V. I., Goriunova, O. I., & Weber, A. W. (2013). Lower limb activity in the Cis-Baikal: Entheseal changes among middle Holocene Siberian foragers. *American Journal of Physical Anthropology*, *150*, 421–432.

Lieverse, A. R., Weber, A. W., Bazaliiskii, V. I., Goriunova, O. I., & Savel'ev, N. A. (2007). Osteoarthritis in Siberia's Cis-Baikal: Skeletal

indicators of hunter-gatherer adaptation and cultural change. *American Journal of Physical Anthropology*, *132*, 1–16.

Lillard, J. B., Heizer, R. F., & Fenenga, F. (1939). *An introduction to the archaeology of central California*. Sacramento, CA: Sacramento Junior College, Department of Anthropology, Bulletin 2.

Machicek, M. L., & Beach, J. B. (2013). Stresses of life: A preliminary study of degenerative joint disease and dental health among ancient populations of inner Asia. In P. Pechenkina & M Oxenham (Eds.), *The bioarchaeology of East Asia: Movement, contact, health*. Gainesville: University Press of Florida.

Marklein, K. E., Leahy, R. E., & Crews, D. E. (2016). In sickness and in death: Assessing frailty in human skeletal remains. *American Journal of Physical Anthropology*, *161*, 208–225.

Milner, G. R., & Boldsen, J. L. (2012). Estimating age and sex from the skeleton, a paleopathological perspective. In A. L. Grauer (Ed.), *A companion to paleopathology*. Chichester: Wiley-Blackwell.

Milner, G. R., Wood, J. W., & Boldsen, J. L. (2008). Advances in paleodemography. In M. A. Katzenberg & S. R. Saunders (Eds.), *Biological anthropology of the human skeleton* (2nd ed.). Chichester: Wiley-Blackwell.

Molnar, P., Ahlstrom, T. P., & Leden, I. (2011). Osteoarthritis and activity—An analysis of the relationship between eburnation, musculoskeletal stress markers (MSM) and age in two Neolithic hunter-gatherer populations from Gotland, Sweden. *International Journal of Osteoarchaeology*, *21*, 283–291.

Novak, M., & Slaus, M. (2011). Vertebral pathologies in two early modern period (16th–19th century) populations from Croatia. *American Journal of Physical Anthropology*, *145*, 270–281.

Ortner, D. J. (1968). Description and classification of degenerative bone changes in the distal joint surfaces of the humerus. *American Journal of Physical Anthropology*, *28*, 139–155.

Palmer, J. L. A., Hoogland, M. H. L., & Waters-Rist, A. L. (2014). Activity reconstruction of post-medieval Dutch rural villagers from upper limb osteoarthritis and entheseal changes. *International Journal of Osteoarchaeology*, *26*, 78–92.

Ragir, S. (1972). *The early horizon in central California prehistory. Contributions of the University of California Archaeological Research Facility 15*. Berkeley: University of California Press.

Rando, C., & Waldron, T. (2012). TMJ osteoarthritis: A new approach to diagnosis. *American Journal of Physical Anthropology*, *148*, 45–53.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Roach, H. I., & Tilley, S. (2009). The pathogenesis of osteoarthritis. In F. Bronner & M. C. Farach-Carson (Eds.), *Bone and osteoarthritis*. London: Springer.

Rogers, J., & Waldron, T. (1995). *A field guide to joint disease in archaeology*. Chichester: Wiley.

Rosenthal, J. S. (2009). The Blossom Mound, Brazil Mound, and Hotchkiss Mound. In F. P. McManaman (Ed.), *Archaeology in America* (Vol. *4*). Westport, CT: Greenwood Press.

Schrader, S. A. (2012). Activity patterns in New Kingdom Nubia: An examination of entheseal remodeling and osteoarthritis at Tombos. *American Journal of Physical Anthropology*, *149*, 60–70.

Scott, A. B., Choi, K. Y., Mookherjee, N., Hoppa, R. D., & Larcombe, L. A. (2016). The biochemical signatures of stress: A preliminary analysis of osteocalcin concentrations and macroscopic skeletal changes associated with stress in the 13th–17th centuries Black Friars population. *American Journal of Physical Anthropology*, *159*, 596–606.

Sharma, L. (2007). Epidemiology of osteoarthritis. In R. W. Moskowitz, R. D. Altman, M. C. Hochberg, J. A. Buckwalter, & V. M. Goldberg (Eds.), *Osteoarthritis: Diagnosis and medical/surgical management* (4 ed.). Philadelphia, PA: Lippincott Williams and Wilkins.

Smith, H. M. (2004). *Degenerative joint disease in the Windover population* (Unpublished MA thesis) Florida State University, Tallahassee.

Sokal, R. R., & Rohlf, F. J. (2011). *Biometry* (4th ed.). New York, NY: W. H. Freeman.

Ullinger, J. M., Sheridan, S. G., & Ortner, D. J. (2012). Daily activity and lower limb modification at Bab edh-Dhra', Jordan, in the Early Bronze Age. In M. A. Perry (Ed.), *Bioarchaeology and behavior: The people of the ancient near East*. Gainesville: University Press of Florida.

Vrezas, I., Elsner, G., Bolm-Audorff, U., Abolmaali, N., & Seidler, A. (2010). Case–control study of knee osteoarthritis and lifestyle factors considering their interaction with physical workload. *The International Archives of Occupational and Environmental Health*, *83*, 291–300.

Waldron, T. (1995). Changes in the distribution of osteoarthritis over historical time. *International Journal of Osteoarchaeology*, *5*, 385–389.

Waldron, T., & Rogers, J. (1991). Inter-observer variation in coding osteoarthritis in human skeletal remains. *International Journal of Osteoarchaeology*, *6*, 76–83.

Walker, P. L., & Hollimon, S. E. (1989). Changes in osteoarthritis associated with the development of a maritime economy among southern California Indians. *International Journal of Anthropology*, *4*, 171–183.

Weiss, E. (2006). Osteoarthritis and body mass. *Journal of Archaeological Sciences*, *33*, 690–695.

Weiss, E. (2007). Muscle markers revisited: Activity pattern reconstruction with controls in a central California Amerind population. *American Journal of Physical Anthropology*, *133*, 931–940.

Williamson, M. A. (2000). A comparison of degenerative joint disease between upland and coastal prehistoric agriculturalists from Georgia. In P. M. Lambert (Ed.), *Bioarchaeological studies of life in the age of agriculture: A view from the southeast*. Tuscaloosa: The University of Alabama Press.

Woo, E. J., & Pak, S. (2014). The relationship between the two types of vertebral degenerative joint disease in a Joseon Dynasty population, Korea. *International Journal of Osteoarchaeology*, *24*, 675–687.

Woo, E. J., & Sciulli, P. W. (2011). Degenerative joint disease and social status in the terminal Late Archaic Period (1000–500 B.C.) of Ohio. *International Journal of Osteoarchaeology*, *23*, 529–544.

Yaussy, S. L., DeWitte, S. N., & Redfern, R. C. (2016). Frailty and famine: Patterns of mortality and physiological stress among victims of famine in medieval London. *American Journal of Physical Anthropology*, *160*, 272–283.

Yonemoto, S. (2016). Differences in the effects of age on the development of entheseal changes among historical Japanese populations. *American Journal of Physical Anthropology*, *159*, 267–283.

Zampetti, S., Mariotti, V., Radi, N., & Belcastro, M. G. (2016). Variation of skeletal degenerative joint disease features in an identified Italian modern skeletal collection. *American Journal of Physical Anthropology*, *160*, 683–693.

How to cite this article: Cheverko CM, Hubbe M. Comparisons of statistical techniques to assess age-related skeletal markers in bioarchaeology. *Am J Phys Anthropol*. 2017;163:407–416. https://doi.org/10.1002/ajpa.23206