Received: 15 September 2021 | Revised: 12 April 2023 | Accepted: 24 May 2023

DOI: 10.1002/ajpa.24790

RESEARCH ARTICLE

AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY

WILEY

The effect of reference sample composition and size on dental age interval estimates

Valerie Sgheiza[1] | Helen M. Liversidge[2]

[1]Department of Anthropology, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

[2]Institute of Dentistry, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Correspondence
Valerie Sgheiza, Department of Anthropology, University of Illinois at Urbana-Champaign, Champaign, IL, USA.
Email: sgheiza2@illinois.edu

Abstract

Objectives: Validation studies in juvenile dental age estimation primarily focus on point estimates while interval performance for reference samples of different ancestry group compositions has received minimal attention. We tested the effect of reference sample size and composition by sex and ancestry group on age interval estimates.

Materials and Methods: The dataset consisted of Moorrees et al. dental scores from panoramic radiographs of 3334 London children of Bangladeshi and European ancestry and 2–23 years of age. Model stability was assessed using standard error of mean age-at-transition for univariate cumulative probit and sample size, group mixing (sex or ancestry), and staging system as factors. Age estimation performance was tested using molar reference samples of four sizes, stratified by year of age, sex, and ancestry. Age estimates were performed using Bayesian multivariate cumulative probit with 5-fold cross-validation.

Results: Standard error increased with decreasing sample size but showed no effect from mixing by sex or ancestry. Estimating ages using a reference and target sample of different sex reduced success rate significantly. The same test by ancestry groups had a lesser effect. Small sample size (n < 20/year of age) negatively affected most performance metrics.

Discussion: We found that reference sample size, followed by sex, primarily drove age estimation performance. Combining reference samples by ancestry produced equivalent or better estimates of age by all metrics than using a single-demographic reference of smaller size. We further proposed that population specificity is an alternative hypothesis of intergroup difference that has been erroneously treated as a null.

KEYWORDS
dental age estimation, method optimization, population specificity

1 | INTRODUCTION

It is often assumed that ideal skeletal methods will be constructed from the same population as the target individual, however in juvenile age estimation, the assumption that dental development differs greatly between ancestral or geographic groups has not been examined while extensively controlling for other factors that may impact estimates of age. Many methods of estimating dental age in

contemporary individuals appear to be population specific; error rates found in follow-up studies are larger than those originally reported (Chaillet et al., 2005) or better results are obtained by using a population-specific reference (Baghdadi & Pani, 2012; Jayaraman et al., 2018). Similarly, several studies report consistent bias in estimated age (Jayaraman et al., 2013; Kırzıoğlu & Ceyhan, 2012; Kumaresan et al., 2016; Maia et al., 2010; Mani et al., 2008; Phillips & van Wyk Kotze, 2009; Prasad & Kala, 2019; Tunc & Koyuturk, 2008).

One possible explanation for these discrepancies between methods and follow-up studies is that there are biological differences in dental development between groups. Such biological difference would imply that combining ancestry groups in reference samples or estimating age in a target sample from one group using a method developed with a reference sample from another will produce worse estimates of age than if the reference and target sample were both drawn from a single ancestral group. Such assumptions are complicated by the fact that many factors other than biological difference have the potential to affect estimates of age, with most of these factors related to method and not the people in the dataset (Corron et al., 2018). Some such factors include reference sample size, reference sample age distribution, and model type. Here we will address potential effects from reference sample size, as well as issues with how the population specificity assumption is framed during hypothesis construction and testing.

Reference sample size is a plausible explanation for discrepancies in error rates because small samples will appear different even when drawn from the same population, the phenomenon of small-sample effects, which has been documented in statistics for over a century (Lin, 2018; Student, 1908; Welch, 1958). We do not currently have a good understanding of what constitutes a small sample in dental age estimation. It is possible that total reference sample sizes in dental age estimation will need to be quite large (thousands rather than hundreds) in order to avoid these effects because partitioning continuous dental development into a large categorical scale will result in many partitions having very few individuals. Uneven age distributions can exacerbate the issue, producing bias and variable error rates across the age range (Buckberry, 2015). This is similar to sex bias in sex estimation methods with unequal sex ratios in the reference sample (Boldsen et al., 2015; Milner & Boldsen, 2012).

The assumption that population specific methods are necessary is of practical importance because requiring population-specific methods means requiring ancestry estimates from juvenile skeletal remains as well as dividing already sparse datasets of juvenile dental development data. If dental age estimation is truly population-specific, this will require serious efforts to improve ancestry estimation in children, and to obtain reference samples that are directly representative of specific target populations. If small sample effects are to blame, the entirely opposite solution of combining reference samples should be applied instead.

There are several problems with the population specificity assumption. Of the three that we will discuss here, two are methodological challenges to identifying population specificity in the first place and the third is related to human variation. The first problem is that results are not consistent across studies. Some show no differences in error rates between groups or biases in dental age estimates (Braga et al., 2005; Kiran et al., 2015; Liversidge, 2011; Thevissen et al., 2010). Follow-up studies using different individuals from the same geographic/demographic population as those included in the development of the original method are uncommon. Willems et al. (2001) provide an excellent example of this procedure and to some extent so do AlQahtani et al. (2010, 2014) although sources are added to the follow-up sample that were not included in the original. This dearth of follow-up studies on original groups means that reported error rates in original studies may be underestimates due to overfitting (error rate based on data used to fit the model) and model selection bias (error rate based on data used to select the best model).

The second challenge to identifying population specificity is that there have been no systematic studies of these effects on age interval estimates. Most existing research has focused on mean errors in point estimates of age (Jayaraman et al., 2013). This is concerning because existing best practices in forensic age estimation recommend estimating age intervals with associated error rates or point estimates with associated standard errors (Corron et al., 2018).

The third problem with the assumption is that since human variation is continuous and does not follow typological categorizations of race, it is not possible to construct a reference sample larger than monozygotic siblings that has truly homogeneous ancestry. There must come a point at which it is no longer practical to divide a sample further. Moorrees et al. (1963) found consistent differences in development between children from Ohio and Boston. Building city-specific reference samples is a nearly impossible task and would have limited applicability in forensic identification. It may be that developmental differences exist at a finer level than it is practical to capture in our methods. Not all detectable differences will be useful or meaningful. Further, these findings suggest that phenotypic plasticity in the dentition (rather than genetic ancestry) may be driving intergroup differences in development through environmental effects. This has subsequently been supported in additional contexts (Cardoso, 2007).

The mismatch between age estimation methods and the landscape of human variation is partly the result of how hypotheses in validation studies are framed. A null hypothesis is of "no difference" between populations, or of "no effect" from an experimental manipulation (Nickerson, 2000). Inter-group difference in age estimation is fundamentally an alternative hypothesis; the corresponding null hypothesis should be that there is "no difference." The population-specificity *assumption* frames intergroup difference as the null hypothesis. This is problematic both because it is incorrect from the standpoint of hypothesis construction and because treating intergroup difference as something that must be falsified reinforces erroneous typological ideas about human variation. We therefore reframe intergroup difference as an alternative hypothesis as follows:

> **Ho1.** Demographic differences in sex and ancestry in the reference and target sample **do not** cause sufficient differences in dental development to impact estimates of age.

Ha1. Demographic differences in sex and ancestry in the reference and target sample **do** cause sufficient differences in dental development to impact estimates of age.

Note that this set of hypotheses does not speak to the *effect* of experimental manipulation, which includes factors such as the size and age distributions of the reference and target samples. When "difference" and "effect" hypotheses are conflated or combined there is a very real risk that an effect of experimental manipulation will be interpreted as a difference between groups. Separating difference from effect requires controlling experimental conditions by holding them constant or varying them systematically. Systematic variation of experimental conditions necessitates a second null hypothesis of "no effect" along with a corresponding suite of appropriate alternatives. We are interested in the effect of reference sample size on both the stability of modeling parameters and on age estimation performance. We therefore use a separate set of hypotheses for each line of inquiry:

Ho2. Reference sample size has no effect on model stability (model mean-age-at-transition).

Ha2. Larger reference samples produce *more stable models* than smaller ones that are more homogeneous.

Ho3. Reference sample size has no effect on age estimation performance (estimated vs. true age).

Ha3. Larger reference samples produce *better estimates of age* than smaller ones that are more homogeneous.

In this approach, we reframe intergroup *similarity* as a null hypothesis of no difference by sex and ancestry, include a second and third null hypothesis of no effect by sample size (an experimental manipulation), and test appropriate alternatives to both. We also leverage Bayesian hypothesis testing, which unlike frequentist hypothesis testing, allows for evaluating the strength of evidence for null hypotheses (Krueger, 2001). We use sex as a positive control for inter-group differences, since dental developmental differences by sex are well-documented (Demirjian & Levesque, 1980; Garn et al., 1958; Moorrees et al., 1963). By testing these hypotheses, we will address the pressing questions of whether (A) it is necessary to estimate an individual's age from a reference sample of the same ancestry and (B) reference samples can be combined even when ancestry is heterogeneous between them.

2 | MATERIALS AND METHODS

2.1 | Dataset

The initial data consisted of Moorrees et al. (1963) scores of the left permanent mandibular dentition from panoramic radiographs of 3334 London children between 2 and 23 years of age, of known sex, and clear medical history other than dental caries and associated pathologies. Images were taken during the course of normal diagnosis and treatment at Institute of Dentistry, Barts, and The London School of Medicine and Dentistry, London. All images were scored by Dr. Helen Liversidge (intra-observer weighted kappa = 0.952, $n = 30$ individuals for eight teeth). For H_1 and H_3 we used a subsample ($N = 1120$) of molar scores (three teeth) with a uniform age distribution by year of age between 5 and 19 years stratified by sex and ancestry. For H_2 we used a subsample ($N = 2607$) of complete cases (eight teeth) between 2 and 23 years but did not control the age, sex, or ancestry distributions.

2.2 | Demographic correspondence of reference and target samples (H_1)

Since estimating age intervals requires calculating the full residual correlation matrix between dental developmental variables (after controlling for age), a time-consuming computation step, only the molars were included for this hypothesis. The subsample of molar scores was divided into four groups of $n = 280$ by sex and ancestry (European females, Bangladeshi males, etc.). Bayesian multivariate cumulative probit models with a uniform prior bounded from 2 to 23 years were fit with age on a log scale to each group after using stage collapsing and a Lagrange multiplier goodness-of-fit test with a cutoff p-value of 0.1 to ensure model fit (Bera et al., 1984; Johnson, 1996). See Konigsberg et al. (2016) for details of this procedure. Equation (1) gives Bayes theorem as it was applied to dental stages here with the posterior probability of age given stage on the left, the likelihood and prior in the numerator, and the normalizing constant in the denominator.

$$P(age|stage_1,\ldots stage_k) = \frac{\prod_{i=1}^{k} P(stage_i|age)P(age)}{\int_{age=2}^{23} \prod_{i=1}^{k} P(stage_i|age)P(age)dage} \quad (1)$$

Next, each group was used as a reference sample to estimate ages for every other group. Performance testing within the same group was not conducted. This would require cross-validation, which would produce metrics not directly comparable to those from the other reference-target combinations (which do not need to be cross-validated because the reference and target are different samples). Individuals were excluded from the target sample if they had tooth stages that were not represented in the reference sample and the number of exclusions was tracked. This is necessary in categorical data modeling because unlike with continuous data, it is not possible to extrapolate beyond the range of the data. Ages were calculated as both the 95% highest posterior density region (HPD) and the maximum likelihood estimate from the posterior (i.e., the mode of the posterior).

Following age estimation, we calculated residual error of maximum likelihood estimates, root-mean-square error (RMSE) of residuals, precision as the scaled age interval width (width of HPD divided by true age), and accuracy as the success rate of the HPD capturing true age (the rate at which the estimated age interval included the

true age of the individual). Success rate was calculated using two different values for the denominator: the total number of individuals in the target sample, and the number of individuals in the target sample *for whom age could be estimated*. If an individual in the target sample has tooth stages not present in the reference sample, their age cannot be estimated in a categorical model. We used both residual error and RMSE because residual error provides an evaluation of bias, while RMSE is more sensitive to large errors due to the squared term. Since the squared terms are summed before taking the root, larger errors are weighted more heavily in the final calculation. If two models have the same residual error but one has larger RMSE, this would indicate that large errors in the second model are less frequent but more severe when they occur.

Success rates were compared via Bayes factors calculated using integration of the binomial logistic distribution (Morey et al., 2015). A Bayes factor is the ratio of the probability of the alternative hypothesis to the probability of the null hypothesis, each of which can be expressed as an integral. Here, the alternative hypothesis is in the numerator and the null is in the denominator, so a Bayes factor greater than 1 supports the alternative hypothesis and a value less than 1 supports the null.

2.3 | Reference sample size and model fitting (H_2)

We fit cumulative probit models to each tooth in the dataset with log-scale age as the explanatory variable and sex and ancestry as covariates. We then performed variable selection on each model using backward stepwise Akaike Information Criterion (AIC). AIC penalizes both poor model fit and model complexity, so larger values indicate a worse model. In backwards stepwise AIC, a model is fit with all variables included and with each variable removed. AIC is calculated for each model and the model with the lowest AIC is chosen as the new model. This process is repeated until removing a variable no longer produces a lower AIC. The results of the stepwise AIC were used to test whether the positive control (sex) was appropriate in our dataset and determine if the subsequent factorial design was sensitive to effects from sex.

The effect of reference sample size on model parameters was assessed using a full factorial design. The factors were staging system (Demirjian et al., 1973; Moorrees et al., 1963), percent original sample size (60, 80, 100), and group mixing percentage (0, 10, 20, 30, 40, 50), for a total of 36 factor-level combinations in each experimental run. We ran two versions, one with sex as the mixing variable and one ancestry group as the mixing variable. Ten bootstrap runs were completed for each version. For each tooth in each run, we fit a Bayesian univariate cumulative probit model with age on a log scale and computed the standard error of mean age at transition between model stages. The final test metric was the average standard error for all stage transitions for a particular tooth and run of our full factorial design. Averaging the standard error across stages was a necessary generalization due to the number of stages for each tooth (up to 16) and number of factor-level combinations in the design.

In order to compare the effect of total sample size against the effect of within-stage sample size we employed two staging systems with very different numbers of stages: Moorrees et al. (1963) (15 stages) and Demirjian et al. (1973) (eight stages). For each system, we included an additional initial crypt stage for a total of 16 and nine stages respectively. The Moorrees et al. stages in the original data were collapsed into Demirjian stages according to Liversidge (2008a, 2008b).

Before beginning a single experimental run for either sex or ancestry the larger of the two groups was randomly trimmed to match the size of the smaller group. At each percentage level of original sample size, the groups were again randomly trimmed by the specified amount. Similarly, for each percentage level of group mixing the necessary number of individuals were randomly selected and switched to the opposite group. It is the randomness of this trimming and mixing that allowed for bootstrapping multiple runs.

2.4 | Reference sample size and age interval estimates (H_3)

The experimental design for H_3 used the same $N = 1120$ subsample from the first design and again considered only the molars. Each reference sample was constructed with a uniform age distribution of 10, 20, 40, or 80 individuals in each year of age with equal numbers by sex and ancestry group if applicable. Type 1 reference samples contained a single sex and ancestry group, type 2 a single ancestry group, type 3 a single sex, and type 4 both sexes and ancestry groups. Age estimates were performed using full Bayesian multivariate cumulative probit with 5-fold cross-validation.

For each sample type 10, 20, 40, or 80 individuals were included per year of age by drawing 5, 10, or 20 individuals from each stratum of ancestry, sex, and year of age. The sample was divided into reference and target and individuals were excluded from the target if they had tooth stages not present in the reference sample. We collapsed stages based on a Lagrange multiplier goodness-of-fit test with a cut-off p-value value of 0.1 as was done for H_1. We then fit a Bayesian multivariate cumulative probit model with a uniform prior bounded from 2 to 23 years and age on a log scale (see Equation 1). We then estimated MLEs and highest posterior density regions from the target sample. This was 5-fold cross-validated and repeated for every combination of reference sample type and size.

What made this a fractional factorial design was that not every sample size was available for every reference sample type. Twenty was the maximum within-year reference sample size for type 1 reference samples (20 from a single sex and ancestry) but the minimum within-year size for type 4 reference samples (five from each sex-ancestry combination). An additional consideration was that while unequal sample sizes were needed for fitting models, equal numbers of estimated ages were preferred for comparing performance metrics. Including every person in the sample in every type-size combination ensured that the only variables in our model comparisons were model characteristics, not the individuals in the dataset.

TABLE 1 Performance metrics for testing demographic correspondence of reference and target samples.

Type	Mean residual	RMSE	Int width	SR	BF	# Exclusions
same_all[a]	-	-	0.3455	0.9366	1.0132	5
same_anc	0.0880	1.2307	0.3455	0.9152	1.51E+04	7
same_sex	0.0824	1.2181	0.3457	0.9259	4.88E+01	10
diff_all	0.0874	1.2199	0.3448	0.9223	2.72E+02	7

Abbreviations: BF, Bayes factor; SR, success rate.
[a]Cross-validated type 1 model with 20 children per year of age.

TABLE 2 AIC values from stepwise AIC for teeth with a significant contribution from ancestry.

Tooth	All variables	Removing ancestry	Removing sex	Removing age
1st premolar	4509.7	4520.8	4589.5	10600.1
3rd molar	8515.6	8529.7	8516.3	12864.9

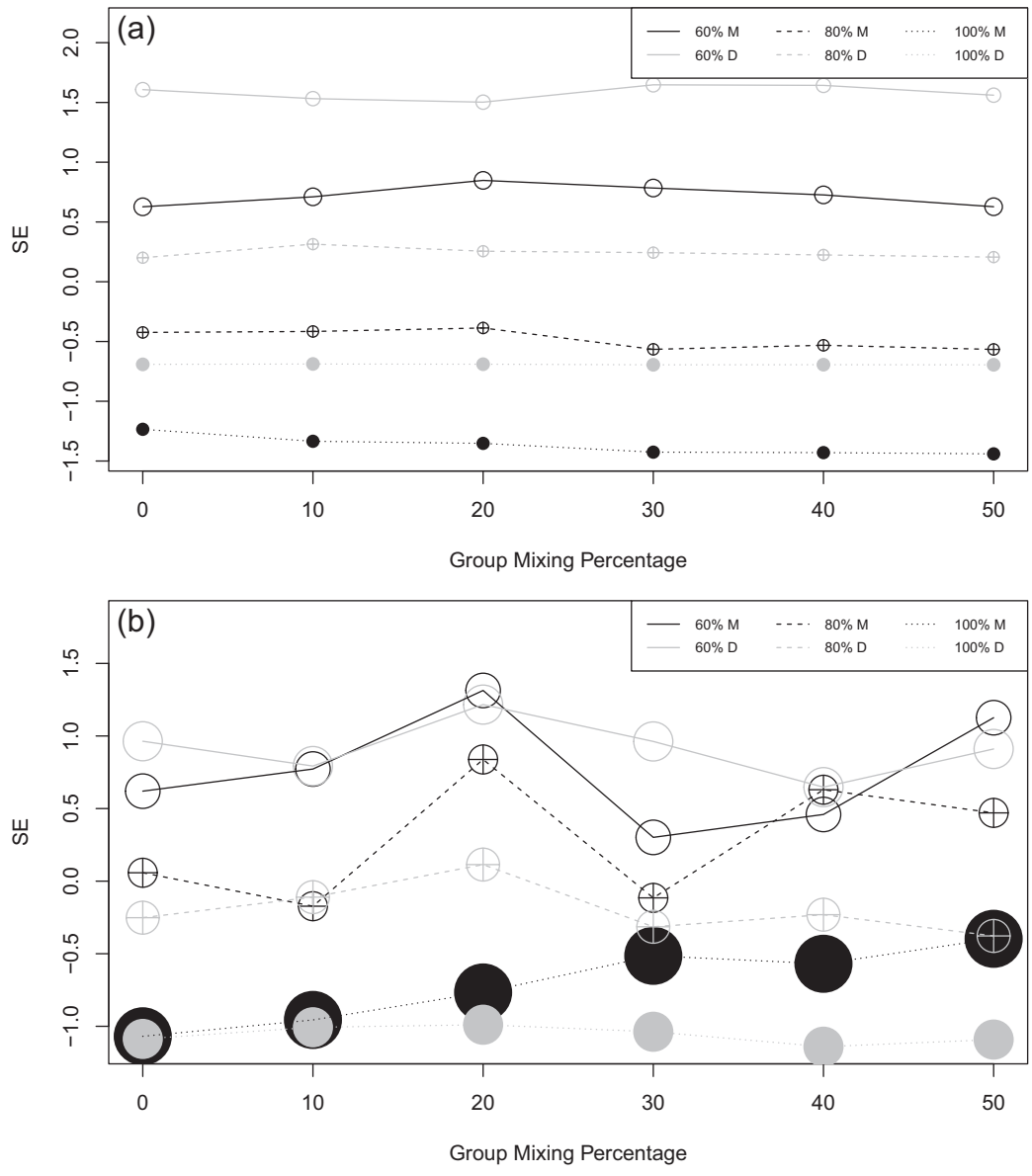

FIGURE 1 Standard error results for I1 sex (a) and P3 ancestry (b).

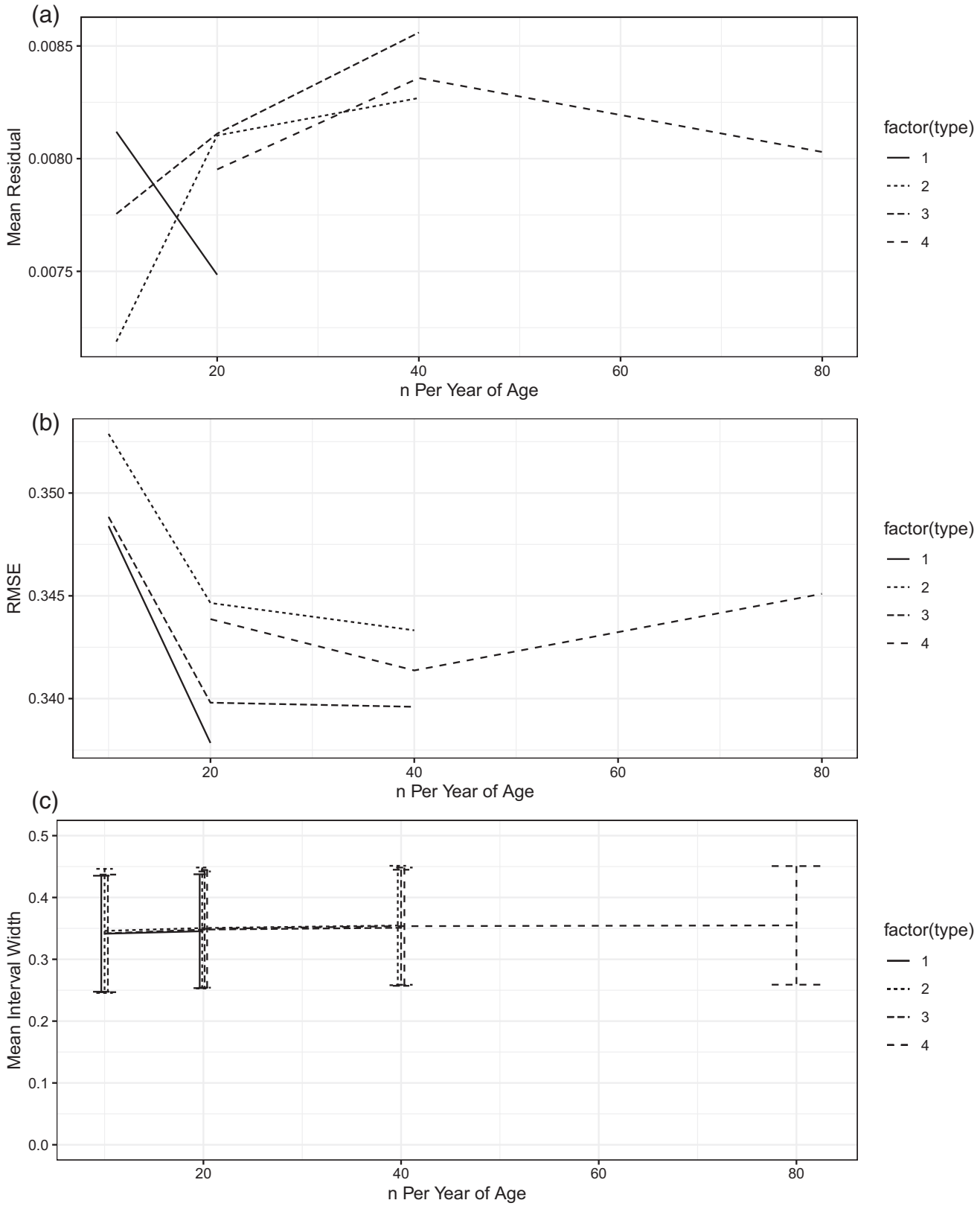

FIGURE 2 Performance metrics for all sample types: mean residuals (a), RMSE (b), and age interval width with standard deviations (c). Factor labels are reference sample types: (1) single sex, single ancestry group, (2) single ancestry group, combined sex, (3) single sex, combined ancestry, (4) combined sex, combined ancestry.

This was accomplished using partitioning. For smaller sample sizes, only ½ or ¼ of the available sample was used to cross-validate. The solution was to partition the sample into halves or quarters by giving every individual within each year-sex-ancestry stratum an index between one and 20 and repeat the cross-validation procedure on left-out partitions for smaller sample sizes. For example, in Bangladeshi boys (type 1) with 10 individuals per year of age we cross-validated on individuals indexed 1–10 in the first partition and on individuals indexed 11–20 in the second partition. There were 47 total type-size combinations when accounting for partitions. This made for a total of 235 model fits after cross-validation. The performance metrics were residual error, RMSE, relative age interval width, and success rate (both calculations as detailed above). Success rates were compared via Bayes factors, as described for H_1.

3 | RESULTS

3.1 | Demographic correspondence of reference and target samples (H_1)

The demographic correspondence of reference and target sample had no effect on age interval width. The largest mean residual and RMSE was found when the reference and target were the same, however, the difference in RMSE was small. This indicated that while there was more error overall, there was not a corresponding increase in large errors. Success rate was lowest when ancestry was the same. According to Bayes factors, there was moderate evidence for the alternative hypothesis of success rate different from 0.95 for same sex, different ancestry (Bayes factor between 30 and 100) and decisive evidence for the alternative for same ancestry, different sex and for different sex and ancestry (Bayes factor > 100). The number of individuals excluded from the target sample due to having tooth stages not represented in the reference sample was similar for all three reference-target combinations (Table 1).

3.2 | Reference sample size and model fitting (H_2)

Based on backward stepwise AIC of univariate probit models, sex was a significant covariate for all teeth. Ancestry was a significant covariate for the first premolar and third molar only (Table 2). This is consistent with the findings of Liversidge (2011) on a subset of our dataset, in which the first premolar and third molar were the only teeth with stage differences between ancestry groups.

Standard error of mean age at transition increased with decreasing reference sample size. Standard error was also consistently higher with Demirjian et al. staging than with Moorrees et al. staging. Standard error results for mixing by sex in the first incisor are similar to those found for the remaining teeth (Figure 1a). There was no consistent effect from group mixing for either sex or ancestry. The standard deviation of this metric remained consistent within a tooth type, group size, and staging system, also showing no trends by mixing percentage. The first premolar had somewhat erratic standard error results for both sex and ancestry (Figure 1b). The remaining 14 plots are available as Supporting Information.

3.3 | Reference sample size and age interval estimates (H_3)

All reference sample types and sizes showed positive mean residuals. Increasing sample size increased the mean residual but decreased RMSE for most combinations of type and size (Figure 2a,b). In addition, combining reference samples by sex (type 2) resulted in smaller mean residuals but larger RMSE than combining by ancestry (type 3). Sample size and type did not have a significant effect on age interval width (Figure 2c). When exclusions were not included in the success rate calculation, only the smallest three reference samples showed evidence of difference from 0.95. These three samples also had the largest numbers of exclusions. When exclusions were included in the success rate calculation, the 20-per-year reference samples for types 2, 3, and 4 all had Bayes factors between 10 and 30 (moderate to strong evidence of difference from null). All of the larger samples

TABLE 3 Success rates for all sample types with and without exclusions.

Type	*n*	SR	BF	# Exclusions	SR with exclusions	BF with exclusions
1	10	0.9175	3.24E+03	17	0.9036	4.88E+07
1	20	0.9408	0.3898	5	0.9366	1.0132
2	10	0.9187	1.70E+03	13	0.9080	1.76E+06
2	20	0.9309	6.0471	5	0.9268	32.7869
2	40	0.9462	0.1925	5	0.9420	0.3191
3	10	0.9210	4.76E+02	19	0.9054	1.25E+07
3	20	0.9336	2.4034	5	0.9295	10.7712
3	40	0.9418	0.3271	3	0.9393	0.5305
4	20	0.9344	1.8770	7	0.9286	15.4043
4	40	0.9382	0.6722	3	0.9357	1.2951
4	80	0.9400	0.4545	3	0.9375	0.8044

along with the 20-per-year sample for type 1 had Bayes factors below 1.3 (Table 3).

We found mixed support for H_{a1}. The highest success rate for models in part 1 was achieved when the reference and target samples had the same sex and ancestry, however, the success rate for estimates where the reference and target differed by both sex and ancestry was intermediate between estimates differing by ancestry and those differing by sex. There was no effect on age interval width. Residual error and RMSE were not compared due to cross-validation concerns. H_{a2} was strongly supported for total reference sample size. Reference sample size, and to a lesser extent, staging system, were the only factors that affected model stability. We cannot speak conclusively here about the effects of within-stage sample size. We found strong support for H_{a3}. Combining reference samples produced equal or better estimates of age up to a point of diminishing returns in success rates for increasing reference sample size.

4 | DISCUSSION

Existing literature shows conflicting results regarding population specificity in juvenile dental age estimation with some follow-up studies showing consistent bias (Jayaraman et al., 2013; Kırzıoğlu & Ceyhan, 2012; Kumaresan et al., 2016; Maia et al., 2010; Mani et al., 2008; Phillips & van Wyk Kotze, 2009; Prasad & Kala, 2019; Tunc & Koyuturk, 2008), while others find no consistent differences in dental development between groups (Braga et al., 2005; Liversidge, 2011; Thevissen et al., 2010). This raises the practical questions of whether reference and target samples must have the same ancestry and whether reference samples of different ancestries should be combined to produce larger sample sizes. Mixed support for H_{a1} suggests that it may be better to estimate age using a demographically similar reference sample, but matching sex is more important than matching ancestry, and this does not take into account sample size, which will be discussed below. Support for H_{2a} and H_{3a} indicates that combining reference samples across ancestry groups is a viable strategy for improving age estimation performance.

We demonstrated that there are important differences in model performance between reference and target samples that have non-corresponding demographics (H_1) and those that have equally mixed demographics (H_3). This was particularly true for sex. Success rates were poor with a large Bayes factor when the sex of the reference and target sample did not match (Table 1), but when both the reference and target had mixed sex, success rate was higher, reducing the Bayes factor by about three orders of magnitude (Table 3). Some caution is warranted for combining by sex because it may increase *large* errors relative to combining by ancestry (residual error vs. RMSE).

Ancestry displayed a similar pattern, although to a lesser degree. The Bayes factor decreased to about 25% of non-corresponding ancestry of reference and target sample for mixed ancestry of reference and target sample. However, the primary question we are asking is whether it is useful to combine reference samples across demographic variables to improve model performance. Here we first compare types 2 and 3 where $n = 10$ to type 4 where $n = 20$. These reference sample type/size combinations can be thought of as separating by either ancestry or sex (types 2 and 3) and then doubling the size of the reference sample by combining. There is a large gain in success rate going from $n = 10$ to $n = 20$ per year of age even when combining across both demographics. The success rate Bayes factor decreases by five to six orders of magnitude.

Similarly, we double the model reference sample size of type 1 where $n = 10$ by combining sex or ancestry (types 2 and 3 where $n = 20$) and then combining both (type 4 where $n = 40$). The Bayes factor of the success rate again decreases by five to six orders of magnitude when the reference sample size is increased from $n = 10$ to $n = 20$ and by one order of magnitude when the sample size increased from $n = 20$ to $n = 40$. This trend in success rate relative to sample size demonstrates two things. First, it is advantageous to combine reference samples across demographic variables in order to produce a larger sample. Second, there are diminishing returns in success rate with increasing reference sample size.

These diminishing returns are especially apparent for types 2 and 3 where $n = 40$ compared to type 4 where $n = 80$. Here, the success rate for type 4 is slightly lower and the Bayes factor slightly higher than the rates for the type 2 and 3 models. The rates are very similar, and the Bayes factors are all below 1, so all three model type and size combinations were equally successful at estimating ages to the target error rate. This indicates that combining reference samples to increase sample size from $n = 40$ to $n = 80$ did not improve model performance.

Part of the reason for this patterned shift in performance is the decrease in numbers of exclusions with increasing sample size. Reference samples with $n = 10$ do not capture sufficient variation in dental stages to reliably cover the variation expressed in the target sample, resulting in large numbers of exclusions. A reference sample of $n \geq 20$ was large enough to capture this variation. Additional performance improvements beyond $n = 20$ seem to mostly result from a larger sample size directly benefitting model parameter estimation, supported by a nearly identical pattern in RMSE (Figure 2b).

There are several potential reasons why consistently reported differences in age estimates by ancestry group did not manifest here. These include using age intervals instead of point estimates, a large sample size, a uniform age distribution for two out of three hypotheses, and a Bayesian modeling framework. Bias in point estimates will not necessarily translate to poor success rates if the bias is small relative to the uncertainty of the estimate. Note that we saw positive bias (mean residuals) for all models in H_3. Small sample effects can produce apparent differences between samples that are not representative of the population as a whole (Student, 1908; Welch, 1958), and sampling error due to small sample sizes can result in biases in metanalyses (Lin, 2018).

A uniform reference distribution and a Bayesian framework should reduce the potential for age mimicry of reference samples by target samples. Age mimicry is a more serious problem in adult age estimation where traits have lower correlations with chronological age, but age mimicry is possible in any instance where the correlation between trait and age is less than one (Boldsen et al., 2002; Konigsberg & Frankenberg, 1992). Age mimicry has been documented for age estimation from the developing dentition (Sgheiza & Liversidge, 2023). Using a uniform reference sample age distribution ensures that any effect from the reference sample age distribution is consistent across all comparisons (Konigsberg & Frankenberg, 1992). A Bayesian modeling framework regresses dental stage (dependent variable) on chronological age (independent variable) and solves for age rather than the reverse scenario of inverse calibration methods where the independent variable (age) is regressed on the dependent variable (stage). This former method is less sensitive to the effect of age mimicry but requires Bayes' theorem to estimate since true ages in the target sample are treated as unknown (Boldsen et al., 2002; Hoppa & Vaupel, 2002).

Examining the five studies identified previously that reported no significant differences in dental development by various measures of population affinity (ancestry, ethnic group, national origin, geographic location) four had overall sample sizes of at least 1000 individuals (Braga et al., 2005; Liversidge, 2011; Liversidge et al., 2017; Thevissen et al., 2010), two used approximately uniform age distributions (Liversidge, 2011; Liversidge et al., 2017), and three used a Bayesian modeling framework (Braga et al., 2005; Liversidge et al., 2017; Thevissen et al., 2010). By approximately uniform, we mean that an effort was made to have similar numbers of individuals by year of age, but the numbers were not identical.

When we compare the findings of two third molar studies, Liversidge (2008a, 2008b) and Liversidge et al. (2017), the potential impact of methodological characteristics on findings of population specificity is apparent. Liversidge (2008a, 2008b) had a total sample size of 3224, an approximately uniform age distribution, modeled dental development using logistic regression, and found significant differences in development between groups. Liversidge et al. (2017) had a total sample size of 4555, an approximately uniform age distribution, and modeled dental development using a Bayesian framework. In this instance, applying a Bayesian framework showed that while there were consistent differences in development between groups, these differences were small relative to the standard deviations of mean age at transition between tooth stages.

Our findings may not be generalizable between geographic areas. Everyone in our sample was from a single location (London). We do not account for population differences due to environmental factors. Groups of different socioeconomic status may show differences in dental development, regardless of ancestry, due to environmental effects (Cardoso, 2007). Individual-specific information on socio-economic status was unknown, so we cannot speak to its effects. What we do provide is evidence against population specificity in a single location and demonstrates the effect of sample size as a confounding factor in identifying developmental differences between groups. We also propose that our design for H_3 may be used as a framework for testing population differences in other samples. A cross-validated factorial design using uniform age distributions should reduce overfitting and age mimicry effects, both of which are possible explanations for observed differences in age estimation performance between original methods and follow-up studies.

5 | CONCLUSIONS

Through hypothesis testing we address two key questions: if it is necessary to estimate the age of an individual using a reference sample of the same ancestry and whether reference samples can be combined across ancestry groups. These questions are motivated by conflicting results in the literature that suggest that estimates of age may be population specific. Here we frame population specificity as a hypothesis of difference in contrast with reference sample size, which is a hypothesis of effect. Our goals are to deconstruct the population specificity assumption and to contribute to the process of identifying sampling and modeling characteristics that will facilitate stronger comparisons between studies.

There are several implications for age estimation that can be drawn from these results. First and foremost, model stability and age estimation performance are strongly driven by reference sample size. Here we found that the minimum effective reference sample size for a uniform age distribution was 20 to 40 individuals per year of age. There were diminishing returns in metrics of model performance beyond 40 individuals. Second, using a model with a reference sample that is heterogeneous by sex may be especially advantageous when sex is unknown. Lastly, combining reference samples by sex or by ancestry to produce a larger reference will produce better estimates of age *up to a point of diminishing returns from sample size*. To answer our initial research questions, reference samples can be combined when ancestry is heterogeneous between them. It is not necessarily better to use a reference sample with the same ancestry as the target individual. It is more important to use a reference sample from the same sex as the target individual, or of mixed sex if the sex of the target individual is unknown.

AUTHOR CONTRIBUTIONS

Valerie Sgheiza: Conceptualization (equal); formal analysis (lead); methodology (lead); software (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Helen Liversidge:** Conceptualization (equal); data curation (lead); investigation (lead); resources (lead); writing – review and editing (supporting).

ACKNOWLEDGMENTS

The authors are grateful to Dr. Lyle Konigsberg for his suggestions on early drafts of this manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the second author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Valerie Sgheiza https://orcid.org/0000-0002-9896-9252
Helen M. Liversidge https://orcid.org/0000-0002-7056-9337

REFERENCES

AlQahtani, S. J., Hector, M. P., & Liversidge, H. M. (2010). Brief communication: The London atlas of human tooth development and eruption. *American Journal of Physical Anthropology*, *142*(3), 481–490.

AlQahtani, S. J., Hector, M. P., & Liversidge, H. M. (2014). Accuracy of dental age estimation charts: Schour and Massler, Ubelaker and the London atlas. *American Journal of Physical Anthropology*, *154*(1), 70–78.

Baghdadi, Z. D., & Pani, S. C. (2012). Accuracy of population-specific Demirjian curves in the estimation of dental age of Saudi children. *International Journal of Paediatric Dentistry*, *22*(2), 125–131.

Bera, A. K., Jarque, C. M., & Lee, L.-F. (1984). Testing the normality assumption in limited dependent variable models. *International Economic Review*, *25*, 563–578.

Boldsen, J. L., Milner, G. R., & Boldsen, S. K. (2015). Sex estimation from modern American Humeri and femora, accounting for sample variance structure. *American Journal of Physical Anthropology*, *158*(4), 745–750.

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

Braga, J., Heuze, Y., Chabadel, O., Sonan, N. K., & Gueramy, A. (2005). Non-adult dental age assessment: Correspondence analysis and linear regression versus Bayesian predictions. *International Journal of Legal Medicine*, *119*(5), 260–274.

Buckberry, J. (2015). The (Mis) use of adult age estimates in osteology. *Annals of Human Biology*, *42*(4), 323–331.

Cardoso, H. F. V. (2007). Environmental effects on skeletal versus dental development: Using a documented subadult skeletal sample to test a basic assumption in human osteological research. *American Journal of Physical Anthropology*, *132*(2), 223–233.

Chaillet, N., Nyström, M., & Demirjian, A. (2005). Comparison of dental maturity in children of different ethnic origins: International maturity curves for clinicians. *Journal of Forensic Science*, *50*(5), JFS2005020-11.

Corron, L., Marchal, F., Condemi, S., & Adalian, P. (2018). A critical review of sub-adult age estimation in biological anthropology: Do methods comply with published recommendations? *Forensic Science International*, *288*, 328.e1–328.e9.

Demirjian, A., & Levesque, G.-Y. (1980). Sexual differences in dental development and prediction of emergence. *Journal of Dental Research*, *59*(7), 1110–1122.

Demirjian, A., Goldstein, H., & Tanner, J. M. (1973). A new system of dental age assessment. *Human Biology*, *45*, 211–227.

Garn, S. M., Lewis, A. B., Koski, K., & Polacheck, D. L. (1958). The sex difference in tooth calcification. *Journal of Dental Research*, *37*(3), 561–567.

Hoppa, R. D., & Vaupel, J. W. (2002). The Rostock manifesto for paleodemography: The way from stage to age. In *Paleodemography: Age distributions from skeletal samples* (pp. 1–8). Cambridge University Press.

Jayaraman, J., Roberts, G. J., Wong, H. M., & King, N. M. (2018). Dental age estimation in southern Chinese population using panoramic radiographs: Validation of three population specific reference datasets. *BMC Medical Imaging*, *18*(1), 1–8.

Jayaraman, J., Wong, H. M., King, N. M., & Roberts, G. J. (2013). The French–Canadian data set of Demirjian for dental age estimation: a systematic review and meta-analysis. *Journal of Forensic and Legal Medicine*, *20*(5), 373–381.

Johnson, P. A. (1996). A test of the normality assumption in the ordered Probit model. *Metron*, *54*, 213–221.

Kiran, C. H. S., Sudhakara Reddy, R., Ramesh, T., Sai Madhavi, N., & Ramya, K. (2015). Radiographic evaluation of dental age using Demirjian's eight-teeth method and its comparison with Indian formulas in south Indian population. *Journal of Forensic Dental Sciences*, *7*(1), 44–48.

Kırzıoğlu, Z., & Ceyhan, D. (2012). Accuracy of different dental age estimation methods on Turkish children. *Forensic Science International*, *216*(1–3), 61–67.

Konigsberg, L. W., & Frankenberg, S. R. (1992). Estimation of age structure in anthropological demography. *American Journal of Physical Anthropology*, *89*(2), 235–256.

Konigsberg, L. W., Frankenberg, S. R., & Liversidge, H. M. (2016). Optimal trait scoring for age estimation. *American Journal of Physical Anthropology*, *159*(4), 557–576.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16–26.

Kumaresan, R., Cugati, N., Chandrasekaran, B., & Karthikeyan, P. (2016). Reliability and validity of five radiographic dental-age estimation methods in a population of Malaysian children. *Journal of Investigative and Clinical Dentistry*, *7*(1), 102–109.

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One*, *13*(9), e0204056.

Liversidge, H. M. (2008a). 10 dental age revisited. *Technique and Application in Dental Anthropology*, *53*, 234.

Liversidge, H. M. (2008b). Timing of human mandibular third molar formation. *Annals of Human Biology*, *35*(3), 294–321.

Liversidge, H. M. (2011). Similarity in dental maturation in two ethnic groups of London children. *Annals of Human Biology*, *38*(6), 702–715.

Liversidge, H. M., Kalaiarasu Pearlasamy, M. P. D., Ngom, P. I., Shimada, Y., Kuroe, K., Tvete, I. F., & Kvaal, S. I. (2017). A radiographic study of the mandibular third molar root development in different ethnic groups. *The Journal of Forensic Odonto-Stomatology*, *35*(2), 97–108.

Maia, M. C., Martins, M. D., Germano, F. A., Neto, J. B., & Da Silva, C. A. (2010). Demirjian's system for estimating the dental age of northeastern Brazilian children. *Forensic Science International*, *200*(1–3), 177.e1–177.e4.

Mani, S. A., Naing, L. I. N., John, J., & Samsudin, A. R. (2008). Comparison of two methods of dental age estimation in 7–15-year-old Malays. *International Journal of Paediatric Dentistry*, *18*(5), 380–388.

Milner, G. R., & Boldsen, J. L. (2012). Humeral and femoral head diameters in recent white American skeletons. *Journal of Forensic Sciences*, *57*(1), 35–40.

Moorrees, C. F. A., Fanning, E. A., & Hunt Jr, E. E. (1963). Age variation of formation stages for ten permanent teeth. *Journal of Dental Research*, *42*(6), 1490–1502.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'Bayesfactor'. Retrieved Accessed 1006 15 from Http://Cran/r-Projectorg/Web/Packages/BayesFactor/BayesFactor.Pdf

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Phillips, V. M., & van Wyk Kotze, T. J. (2009). Testing standard methods of dental age estimation by Moorrees, Fanning and Hunt and Demirjian, Goldstein and Tanner on three south African children samples. *The Journal of Forensic Odonto-Stomatology*, *27*(2), 20–28.

Prasad, H., & Kala, N. (2019). Accuracy of two dental age estimation methods in the Indian population—A meta-analysis of published studies. *The Journal of Forensic Odonto-Stomatology*, *37*(3), 2–11.

Sgheiza, V., & Liversidge, H. (2023). Reference and target sample age distribution impacts between model types in dental developmental age estimation. *International Journal of Legal Medicine*, *137*(2), 383–393.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Thevissen, P. W., Ali Alqerban, J., Asaumi, F. K., Jaswinder Kaur, Y. K., Kim, P., Van Pittayapat, M., Vlierberghe, Y. Z., & Fieuws, S.

(2010). Human dental age estimation using third molar developmental stages: Accuracy of age predictions not using country specific information. *Forensic Science International*, *201*(1–3), 106–111.

Tunc, E. S., & Koyuturk, A. E. (2008). Dental age assessment using Demirjian's method on northern Turkish children. *Forensic Science International*, *175*(1), 23–26.

Welch, B. L. (1958). Student'and small sample theory. *Journal of the American Statistical Association*, *53*(284), 777–788.

Willems, G., Van Olmen, A., Spiessens, B., & Carels, C. (2001). Dental age estimation in Belgian children: Demirjian's technique revisited. *Journal of Forensic Science*, *46*(4), 893–895.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sgheiza, V., & Liversidge, H. M. (2023). The effect of reference sample composition and size on dental age interval estimates. *American Journal of Biological Anthropology*, 1–11. https://doi.org/10.1002/ajpa.24790