



# Advancing estimation of chronological age by utilizing available evidence based on two radiographical methods

Øyvind Bleka<sup>1</sup> · Torbjørn Wisløff<sup>2,3</sup> · Pål Skage Dahlberg<sup>1</sup> · Veslemøy Rolseth<sup>1</sup> · Thore Egeland<sup>1,4</sup>

Received: 7 November 2017 / Accepted: 9 April 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

## Abstract

This paper describes a strategy for estimating chronological age of individuals based on age indicators of X-ray of the hand and the third molar tooth. The great majority of studies in the field provide group-wise data of different formats, which makes them difficult to compare and utilize in a model. In this paper, we have provided a framework to utilize different types of data formats to build a common model for estimating chronological age. We used transition analysis to describe the relationship between the age indicators and chronological age. Further, likelihood ratio weight of evidence and posterior distribution of chronological age were used to model the distribution of chronological age given the observed age indicators. Being able to utilize such a large amount of data, with different data formats, from different studies, as presented in this paper improves previous age estimation methods.

**Keywords** Age estimation · Bayesian inference · Likelihood ratio · Demirjian's · Greulich and Pyle

## Introduction

The estimation of chronological age<sup>1</sup> is an important issue in several forensic fields. In forensic anthropology, the remains of bones and teeth have been used to assess chronological age. Estimating the chronological age of dead bodies or trace samples is important forensic applications, while the estimation of age in living individuals is the largest field in terms of publications. In the latter category, age assessment of adolescent asylum seekers is most important in the western world. In these cases, age estimation is commonly based on development of the

skeleton or teeth [1]. Traditionally, an X-ray of the hand and wrist is the most widely applied method to assess skeletal development. When it comes to dental age assessment, a panoramic X-ray of all teeth is commonly performed. In the late teens, the third molar is usually the only tooth still in development. This paper focuses on the two grading systems that are arguably the most used in the field: the Greulich and Pyle (GP) atlas for X-ray of the hand and wrist [2] and Demirjian's grading of the third molar tooth [3, 4]. Based on these techniques, our goal was to build a model to estimate the chronological age for an individual with an observed age indicator (separately for each systems, or combined). Gelbrich et al. [5] studied the development of teeth and skeleton of the same individuals, showing that the two variables appear to be independent given chronological age. Other studies also support this finding [6, 7]. This simplifies the combination of results from the two grading systems and thereby obtaining more accurate estimates. Dahlberg et al. [8] and Rolseth et al. [9] recently performed comprehensive systematic reviews of age estimation based on the GP atlas for X-ray of the hand and Demirjian's grading of the third molar, respectively. We utilize many of their identified studies in the present work.

Many studies estimate chronological age using the conditional age distribution of given age indicators directly. However, the estimation of chronological age with this representation points in the direction of (or mimics) the chronological ages of the individuals included in the study [10]. The solution offered by Boldsen et al. [11] was to calculate the probabilities of observing

<sup>1</sup> The time between date of birth and the date of assessment

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00414-018-1848-y>) contains supplementary material, which is available to authorized users.

✉ Øyvind Bleka  
oyvble@hotmail.com

<sup>1</sup> Department of Forensic Sciences, Oslo University Hospital, Oslo, Norway

<sup>2</sup> Department of Infectious Disease Epidemiology and Modelling, Norwegian Institute of Public Health, Oslo, Norway

<sup>3</sup> Department of Health Management and Health Economics, University of Oslo, Oslo, Norway

<sup>4</sup> Norwegian University of Life Sciences, Ås, Norway

different age indicators given chronological age. To build such a model, individual-based information is typically needed (that is both the age indicator and chronological age for every individual). However, many studies only report group-wise data. In this paper, we provide a framework for utilizing such group-wise information in the model.

There are two typical frameworks for chronological age information: (1) estimating the conditional distribution of chronological age given the observed age indicators and (2) estimating an evidential value of the observed age indicators as a likelihood ratio [12, 13] value where two hypotheses involving chronological age are compared. The former framework (1) is also referred to as “the two-step approach” by Konigsberg et al. [14]. In this approach, the likelihood of chronological age for a given age indicator is estimated, and Bayes’ theorem is applied to return the posterior distribution of chronological age given age indicator. This strategy requires that a prior distribution of chronological age is defined. The second framework (2) does not require the definition of such a prior distribution, but a construction of two hypotheses,  $H_1$  and  $H_2$ . The likelihood ratio indicates whether the data support hypothesis  $H_1$  over  $H_2$ , or vice versa. In this paper, these hypotheses distinguish whether an individual is above or below a threshold of  $T$  years of age. This is of particular importance in relation to asylum seekers where the authorities typically use a threshold  $T=18$  years to define adulthood.

The method part of this paper is comprehensive: First, we provide a framework to impute individual-based information used for further modeling. Second, we suggest different candidate models for estimating the probability of age indicators given chronological age, for the hand and tooth methods. Third, we present the evidential and predictive values for different age indicators for the two methods, and the methods combined.

## Material and data

Dahlberg et al. [8] performed a systematic review considering the GP atlas grading system based on hand skeleton while Rolseth et al. [9] addressed studies considering Demirjian’s stage system based on third molar teeth. Data from studies included in these two comprehensive reviews were extracted. In addition, authors were contacted and asked to provide individual level data. An overview of all the considered datasets is given in Table 1.

For our analyses, we only included individuals with skeletal age at least 10 (hand) and the lower left third molar tooth. The age indicators for the GP system (the hand method) are presented as in the GP atlas [2], given as skeletal ages (SAs) (10, 11, 11.5, 12.5, 13, 13.5, 14, 15, 15.5, 16, 17, 18, and 19) for males and (10, 11, 12, 13, 13.5, 14, 15, 16, 17 and 18) for females. For Demirjian’s stage system (the tooth formation classification system), the eight letters (A, B, C, D, E, F, G, and H) are given for both sexes. The age indicators for both

systems are ordered (for a given individual), meaning that the hand or teeth development cannot reverse as an individual gets older.

A challenge with many of the datasets in Table 1 is that we only have access to individual-level data for about half of the studies (i.e., the type 1 format). The other studies instead present summary statistics based on the individual-based information only (type 2 and type 3 format; see [Supplementary material](#) section “A Data material” for detailed examples). For these studies, we impute the individual-based information using summary information from the publications. The number of samples in total is 4082/4513 (males/females) for the tooth method and 1724/1154 samples for the hand method.

## Methods

We assume the data from each study to be independent and follow the same underlying model (i.e., all studies are assumed to have the same parameters). This causes the final model to smooth out the information between the studies with respect to their sample size (giving “study-averaged” inference). To investigate the contribution from each study, we made comparison plots which show the non-parametric estimates of the age indicator probabilities (for given chronological age) per study and over all studies (see [Supplementary material](#) section “B.2 Stage probabilities as a function of age per study” for results). We return to this topic in the discussion.

### Imputing individual-based data information

The studies with type 2 format present data as frequency tables, giving the number of individuals between integer age ranges (for instance between 12.00 and 12.99 years) corresponding to a specific age indicator. We assume that chronological age (for these individuals) is uniformly distributed within the specified age ranges. The study with type 3 (Chaumoitre et al. [19]) presents data as summary statistics in terms of number of individuals and mean and standard deviation of the chronological ages for a given age indicator. This study investigated the distribution of chronological age for given specific SAs (the age indicator for hand) and suggested that these were approximately normally distributed.

*The model for imputation of individual data may be formulated more concisely in mathematical language:*

- *Type 2: chronological age of individuals in age range  $[a, a + 1)$  at age indicator  $S = s$  is distributed as  $Uniform(a, a + 1)$ .*
- *Type 3: chronological age of individuals at age indicator  $S = s$  is distributed  $Normal(mean_s, sd_s)$ , where  $mean_s$  and  $sd_s$  are the mean and standard deviation from the tables.*

**Table 1** The number of individuals from different studies in the literature. The column “Format” is the format of how data is presented: type 1 is individual-based (private communication); type 2 gives the number of individuals within an integer age range having a specific age

indicator; and type 3 data presents the mean and standard deviation of chronological age for a group of individuals with a specific age indicator. The column “Age span” gives the lower and upper chronological age (in filled years) observed in the dataset

Method	Format	Study	Males	Females	Country	Age span
Tooth	Type 1	Malta dataset	553	650	Malta	8–26
Tooth	Type 1	South China dataset	682	617	China	8–26
Tooth	Type 2	Lee (2009) [15]	786	964	South Korea	7–25
Tooth	Type 2	Johan (2012) [16]	540	539	Malaysia	14–26
Tooth	Type 2	Duangto (2017) [17]	872	983	Thailand	8–23
Tooth	Type 2	Li (2012) [18]	649	760	China	7–24
Hand	Type 1	Santos (2011)	136	94	Portugal	12–20
Hand	Type 1	VanRijn (2001)	178	197	Netherlands	9–20
Hand	Type 1	Zafar (2010)	165	64	Pakistan	7–18
Hand	Type 1	Tise (2011) [20]	359	126	Italy	11–19
Hand	Type 3	Chaumoitre (2016) [19]	886	673	France	7–20

A complete dataset is created by generating chronological age for individuals representing the type 2 and type 3 format data, and adding these to the type 1 individual data.

### Models for stage probabilities

We aim to estimate the stage probabilities  $Pr(S=s|Age=a)$  as a continuous function of chronological age  $a$  based on the study data. One simple non-parametric estimate for  $Pr(S=s|Age=a)$  would be to consider the proportion of individuals in the age segment  $[a-b, a+b]$  which are in stage  $s$  for a given chronological age  $a$  and specified bandwidth  $b$ :  $\frac{x_s(a)}{\sum_{s'} x_{s'}(a)}$  where  $x_s$  is the number of individuals having stage  $s$  in the age segment. Such a model could lead to overfitting if not carefully selected by applying cross-validation techniques. We only use this model type to provide a graphical examination of model validity for other parametric model suggestions (this has also been suggested previously [11, 14]).

Since the age indicators based on the hand or teeth methods are non-reversible as chronological age increases, it would be reasonable to assume that the age indicators are ordinal discrete variables. Also important, the individuals will “pile up” at the last indicator (the end stage) as chronological age increases. A parametric model family which considers such variables is called “transition analysis” [11]; it models the passage of individuals from a given development stage to the next higher stage in an ordered sequence [14]. We abbreviated this to “transition model.” In the next section, we present different candidate models of this type.

### Candidate models and model selection

In this paper, we define variants of the models as presented by Boldsen et al. [11]. We consider both a proportional-odds

cumulative model and a continuation-ratio model, in which the probability of observing one of the ordinal stages given chronological age is defined differently (see “Appendix A.1 Candidate models” for details) [21]. For these two model types, we use either a *logit* link or a *probit* link. Konigsberg et al. [14] suggested that the chronological age variable is log-transformed before being applied in order to model asymmetry. Motivated by this, we allow the data to decide the transformation of the chronological age variable by raising it to the power of an extra parameter (see “Appendix A.1 Candidate models”). We define the linear predictor to consist of an intercept parameter for each transition between two stages and a common slope parameter as part of the age variable. To infer the candidate models, we apply the function *vglm* from the R-package VGAM (version 1.0–3) [22, 23].

The considered candidate models have the same number of parameters, such that the final chosen is the one which fits the observed data best, i.e., the one with largest maximum likelihood value. The model which fits best based on 100 generated complete datasets is chosen. This procedure is carried out for each method (hand and tooth) and sex (males and females) separately. Hence, the outcome is four selected models.

### Model validation

The purpose of model validation is to investigate whether the fitted model is adequate for the observed data. The model describes the stage probabilities as a function of chronological age, and we are interested in whether these probabilities are representative for given ages. We first performed an exploratory goodness-of-fit analysis, where we performed a likelihood ratio test for each stage and each integer of chronological age (plus/minus a half year). The  $p$  values for these tests were calculated exactly (see “Appendix A.2 Details of the likelihood

ratio test statistics and the calculation of the  $p$  value” for details). After, we performed likelihood ratio tests for each integer of chronological age (over all stages). The  $p$  values from these tests were estimated with bootstrap using 100,000 iterations. Details of these analyses are given in the [Supplementary material](#) section “C Exploratory goodness of fit.” Last, we calculated goodness-of-fit statistics for the whole transition model based on the Lipsitz test where individuals were categorized into ten almost equally sized groups [24, 25] (see “Appendix A.3 Details of the Lipsitz test statistics and the calculation of the  $p$  value” for details). We present the 5, 50, and 95% quantiles of the  $p$  values (from this test) over 1000 simulated datasets (to take into account that some data types are generated). The Lipsitz test is suitable to use for ordinal regression models when only continuous covariates are considered [24, 26].

### Combining age indicators for independent methods

Because of the great biological variability affecting the hand and tooth methods, it is desirable to combine these in order to obtain a more precise estimate of chronological age. With the assumption that the age indicators for the hand and tooth are conditionally independent, given the information about the chronological age, we can present the joint distribution of observing a skeletal age for hand, and a development stage for tooth as a product of the stage probabilities for each tooth and hand separately:  $Pr(\text{skeletal age} = s, \text{tooth stage} = t | \text{Age} = a) = Pr(\text{skeletal age} = s | \text{Age} = a) * Pr(\text{tooth stage} = t | \text{Age} = a)$ . This requires the X-ray pictures of the hand and tooth to be taken approximately at the same time (for the same individual). The formula implies that we can use the fitted models for the two methods separately to model the combined information.

### The predictive/evidential value of an observed age indicator

Our aim is to describe the predictive/evidential value of chronological age for a new individual based on observed age indicator(s) such as skeletal age or/and tooth stage. In this section, we present two different approaches: first the evidential value in terms of likelihood ratio, and then the predictive value in terms of posterior distribution of chronological age.

### Likelihood ratio as weight of evidence

In forensic science, the recommended format to present results is in terms of weight of evidence where two hypotheses are compared [27, 28]. The common statistical measure for this weight is called the likelihood ratio (LR). We compare two hypotheses related to chronological age:

- $H_1$ : The chronological age is more than  $T$  years.
- $H_2$ : The chronological age is less than or equal to  $T$  years.

We calculate the likelihood ratio as  $LR = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_2)}$ . Hence, the LR value itself depends on a specified probabilistic model for the data (i.e., the observed stage indicator(s)). For an observed stage  $s$ ,

$$LR = \frac{\max_a Pr(S = s | a > 18y)}{\max_a Pr(S = s | a \leq 18y)}$$

where the probabilities under each of the hypotheses are maximized under the corresponding assumptions. If LR is greater than 1, this supports that the hypothesis  $H_1$  is true, whereas if LR is less than 1, this supports that the hypothesis  $H_2$  is true. If the LR is close to 1, the evidence is neutral. We consider a verbal scale [28] for the LRs as shown in Table 2. This makes it easier to understand the importance of the LR numbers. More theory about likelihood ratios can be found in [12].

### The conditional distribution of chronological age given age indicator

In this section, we apply a Bayesian approach to yield the conditional distribution of chronological age given an observed age indicator. This approach requires a prior distribution of chronological age to be specified (see section “[The choice of the prior distribution of chronological age](#)”). The prior is combined with the fitted stage probabilities given chronological age so that a posterior distribution of chronological age is obtained:

$$Pr(\text{Age} = a | \text{Stage} = s) = Pr(\text{Stage} = s | \text{Age} = a) \times Pr(\text{Age} = a) \times c$$

where  $c$  is a normalization constant such that the area under the distribution  $Pr(\text{Age} = a | \text{Stage} = s)$  sums to 1. Based on the

**Table 2** The verbal equivalents to likelihood ratio values (LRs), where the LRs within the intervals mean that the evidence give “verbal scale” support to  $H_1$  (over  $H_2$ )

Likelihood ratio value	1–10	10–100	100–1000	1000–10,000	10,000–1 mill.
Verbal scale	Weak	Strong	Moderately strong	Strong	Very strong

posterior distribution, we are interested in calculating the following statistics:

1. The  $(1 - \alpha) \times 100\%$  prediction interval (i.e., credibility interval) of chronological age given the observed age indicator, with the interval  $I = [L, U]$  defined as the  $\frac{\alpha}{2}$  and  $1 - \frac{\alpha}{2}$  percentiles:

$$Pr(\text{Age} \leq L | \text{Stage} = s) = \frac{\alpha}{2}$$

$$Pr(\text{Age} \leq U | \text{Stage} = s) = 1 - \frac{\alpha}{2}$$

2. The (posterior) probability that chronological age is below  $T$  years, i.e.

$$P = \int_0^T Pr(\text{Age} = a | \text{Stage} = s) da$$

For the first statistic, we consider both  $\alpha = 0.05$  and  $\alpha = 0.25$  for comparison, giving a 95% and a 75% prediction interval.<sup>2</sup> For the second statistic, we consider the age thresholds  $T = 16$  and  $T = 18$  years, since these age thresholds are particularly important for asylum seekers. However, we emphasize that other percentages and age thresholds can be selected, depending on the application.

### The choice of the prior distribution of chronological age

As mentioned, the Bayesian approach requires that a prior distribution of chronological age is specified. Such a prior describes the uncertainty of chronological age before obtaining skeletal age or tooth stage information, and hence, it is subjective. Choosing an appropriate prior is an important part of the Bayesian framework, where relevant data (possibly based on alternative sources) should be considered [29], e.g., expert opinions for each individual, or the frequency of occurrence in a population. However, such information is often not available. Choosing an evenly distributed prior (uniform) is sometimes referred to as a flat prior.<sup>3</sup> Different authors have considered other types of priors, such as a distribution around the age threshold of interest [30] or a distribution around the stated age [5] of the individual to be predicted.

The purpose of this paper is to provide an “up-to-date” unbiased description of the distribution of chronological age for given stages. As an example, we define the prior to be uniformly distributed ranging from age 7 up to 21 years. The upper age limit of 21 years is set to be the same for all methods and both sexes to make the results comparable. Importantly, the upper age limit is essential for the resulting

<sup>2</sup> The use of “probability overweight” 51% in practical legal use has led us to something between 51 and 95%.

<sup>3</sup> We prefer “flat prior” to the commonly used “non-informative” as there in our view is no such thing as a non-informative prior.

effects since a higher defined age limit would provide higher prediction intervals and higher probabilities of being above a certain age threshold. These effects are also investigated in terms of a sensitivity analysis.

### Choosing statistics based on the generated datasets

For each of the statistics in “Likelihood Ratio as weight-of-evidence” and “The conditional distribution of chronological age given age indicator” sections, there are  $M$  values based on  $M$ -generated complete datasets. Based on these  $M$  values, we use the 5 or the 95% quantile to get only one value. For the prediction interval statistic  $I = [L, U]$ , the 5% quantiles of the  $ML$ -values are chosen as the lower limit, while the 95% quantiles of the  $M U$ -values are chosen as the upper limit. This gives the interval. For the probability of being below  $T$  years, the resulting statistic is estimated as the 95% quantile of the  $M P$ -values. For the LR weight-of-evidence statistics, the 95% quantile of the  $M LR$ -values is used if the 5% quantile is at least 1, whereas the 5% quantile of the  $M LR$ -values is used if the 95% quantile of LR is less than 1; otherwise, the LR is set to 1. Missing individual data will widen the prediction interval, as is reasonable, or increase the probability of being below  $T$  years, or pointing the LR towards 1 (leading to a more neutral evidentiary weight).

## Results

### Model selection and model validation

#### Selected model

Different candidate models for describing the stage probabilities as a function of chronological age (as described in “Appendix A.1 Candidate models”) were fitted using maximum likelihood estimation. For each candidate, the sum of the 100 maximum likelihood values (one per generated dataset) were calculated, and the candidate model with highest sum was selected.

Table 3 gives an overview of the models selected to describe the stage probabilities as a function of chronological age, for the different methods (hand or tooth) and sexes (males or females). We found that the proportional-odds cumulative model (with *logit* link function) with no transformation of age fitted the data best for the hand method (both sexes). Hence, the skeletal age stage probabilities are estimated to be symmetrical as a function of chronological age (except for the first and last stages). For tooth, a continuation-ratio model with a square root transformation of age fitted the observations best. Here, the link functions *logit* and *probit* fitted best for males and females respectively. The fitted models differ in that the females have wider upper tails for chronological age, indicating that there is more variation of the ages in which females transition to the next stage (particularly from stage D), compared to males.

**Table 3** The best fitting candidate models for each method and sex

Method	Sex	Age-transformation	Model type	Link function
Hand	Male	Age <sup>1</sup>	Proportional-odds cumulative	<i>logit</i>
Hand	Female	Age <sup>1</sup>	Proportional-odds cumulative	<i>logit</i>
Tooth	Male	Age <sup>0.5</sup>	Continuation-ratio	<i>logit</i>
Tooth	Female	Age <sup>0.5</sup>	Continuation-ratio	<i>probit</i>

### Model validation and goodness of fit

The first part of the model validation was carried out as an exploratory analysis where we investigated the calculated  $p$ -values from the goodness-of-fit statistics for each integer of chronological age (from 7 to 27 years<sup>4</sup>) for each stage (see [Supplementary material](#) section “C Exploratory goodness of fit” for more details). From this analysis, we found that the fitted transition models (for each methods and sexes) were adequate for the observations for most stages and ages. We highlight the most important situations where this was not the case:

– *Hand—males:*

*SA = 11.5: fewer individuals than expected at age 11*  
*SA = 13: two to three individuals observed at age 17 (high age). More individuals than expected at age 14*  
*SA = 13.5: more individuals than expected at age 11*  
*SA = 16: more individuals than expected at age 13*  
*SA = 17: more individuals than expected at age 17*  
*SA = 18: one individual observed at age 12 (very low age)*

– *Hand—females:*

*SA = 12: fewer individuals than expected at age 11*  
*SA = 14: one individual observed at age 9*

– *Tooth—males:*

*Stage E: one individual observed at age 26 (very high age)*  
*Stage G/H: more/fewer individuals than expected at age 23–24*

– *Tooth—females:*

*Stage C: one individual observed at age 23 (high age)*  
*Stage D: zero to three individuals observed at age 9 (low age)*  
*Stage G: one individual observed at age 13 (low age)*  
*Stage H: one individual observed at age 13 (low age)*

There are also other less important situations (because of limited amount of samples) where the models tended not to be adequate. A full overview of the situations, together with graphical representations of the differences, can be found in [Supplementary material](#) section C.

In the second part, we calculated the  $p$  values for the Lipsitz test. The (5, 50, and 95%) quantiles of the  $p$  values were given as follows: hand/males: [3e-7, 4e-4, 0.04]; hand/females: [3e-8, 6e-5, 0.01]; tooth/males: [2e-05, 0.006, 0.1]; and tooth/females: [2e-06, 0.001, 0.07]. Hence, the  $p$  values from the Lipsitz tests indicate that the models (as a whole) are not adequate for the observations (considering rejection level 5% for the median  $p$  values). By comparing the  $p$  values, we find that the models for tooth fit better than the models for hand.

### Evidential/predictive value of observed stages

To provide the statistics described in “[Choosing statistics based on the generated datasets](#),” we generated  $M = 1000$  complete (simulated) datasets and stored the  $M = 1000$  stage probabilities (for all stages) for the chronological ages  $a = 7.00, 7.01, \dots, 20.99, 21.00$  (i.e., age grid-size 0.01).

### Results based on likelihood ratio evidential values

The results from the LR methods (Table 4) show moderately strong and strong support for the hypothesis that an individual is above 16 years old when tooth stage H is observed, for males and females respectively. For the combined method, this is also the case for the 18/H and 19/G combinations, whereas the 19/H combination gives very strong support. The skeletal stage 19 and the combinations 17/H, 19/F, and 18/G give moderate support for the hypothesis that an individual is above 16 years old. Only the last tooth stage H and the combination 19/H (and 18/H for female) give moderate support for the hypothesis that an individual is above 18 years old. Importantly, the LR value increases when the last stage for the two methods is observed (combined), compared to one of the methods alone (both sexes). When for instance observing only tooth stage H for a male, the LR is 17.5. If a skeletal age of 19 is also observed (by itself, it only gives LR = 2.0), the LR increases to 35.2. If on the other hand a skeletal age of 18 is observed, the LR decreases to 1.2.

<sup>4</sup> The validation method takes care of having no samples for particular ages.

**Table 4** The table consists of three smaller tables showing the likelihood ratio (LR) values for each method (hand, tooth and combined) and sex for given observed stages. “T” is the year thresholdconsidered in the hypothesis. For the combined method only the top 10 ranked (with respect to LR for male  $T=16$ ) stage combinations are considered

Hand method with skeletal ages				Tooth method with stages						
Gender	17	18	19	Gender	F	G	H			
Male ( $T=16$ )	1.2	3.2	24.8	Male ( $T=16$ )	1.3	6.7	545			
Female ( $T=16$ )	1.3	8.1		Female ( $T=16$ )	1.6	9.8	1128			
Male ( $T=18$ )	0.4	0.9	2.0	Male ( $T=18$ )	0.8	1.2	17.5			
Female ( $T=18$ )	0.5	1.3		Female ( $T=18$ )	1.0	1.6	27.8			
Combined method with skeletal ages and tooth stages										
Gender	17/G	16/H	18/F	19/E	17/H	19/F	18/G	18/H	19/G	19/H
Male ( $T=16$ )	3.6	3.8	3.8	4.6	15.0	14.0	15.6	112	152	13,917
Female ( $T=16$ )	4.0	4.1	10.6		22.0		80.0	9488		
Male ( $T=18$ )	0.8	1.0	0.8	1.0	1.0	1.1	1.0	1.2	2.3	35.2
Female ( $T=18$ )	0.9	1.0	1.1		1.0		2.1	35.7		

### Results based on posterior distribution of chronological age

The definition of the upper age limit may have effect on the resulting statistics based on the posterior distribution of age. In particular, the largest effect is for the last age indicator as this stage “never stops.” A sensitivity analysis was carried out to explore these effects (i.e., changes in the results) by changing the assumed upper age limit from age 18 and up to 27 years. Figure D1 in [Supplementary material](#) section “D The effect of the prior age distribution” shows examples considering skeletal age 18 and tooth stage G (males only). Here, we found that the effects were largest for the 87.5 and 97.5% percentile results since these are closest to the upper age limits. In general, all stages having non-zero stage probabilities for ages above the upper age limit are affected. In the following results, only the upper age limit 21 years is considered.

The results for the posterior prediction interval statistics for methods hand and tooth are given in Fig. 1, whereas the posterior probabilities are given in Fig. 2. The prediction intervals and probabilities for the two methods combined are presented in [Supplementary material](#) section “E Additional results (the combinations).”

### The posterior prediction interval results

Consider Fig. 1. If skeletal age 16 is observed, the upper limit for the 95% prediction interval is 18 years both for males and females. The lower limit is 13.5 years for males and slightly lower for females. The estimated 95% prediction interval of a male with observed skeletal age 18 is [15.2, 19.9]. If the same individual additionally has tooth stage F (giving a

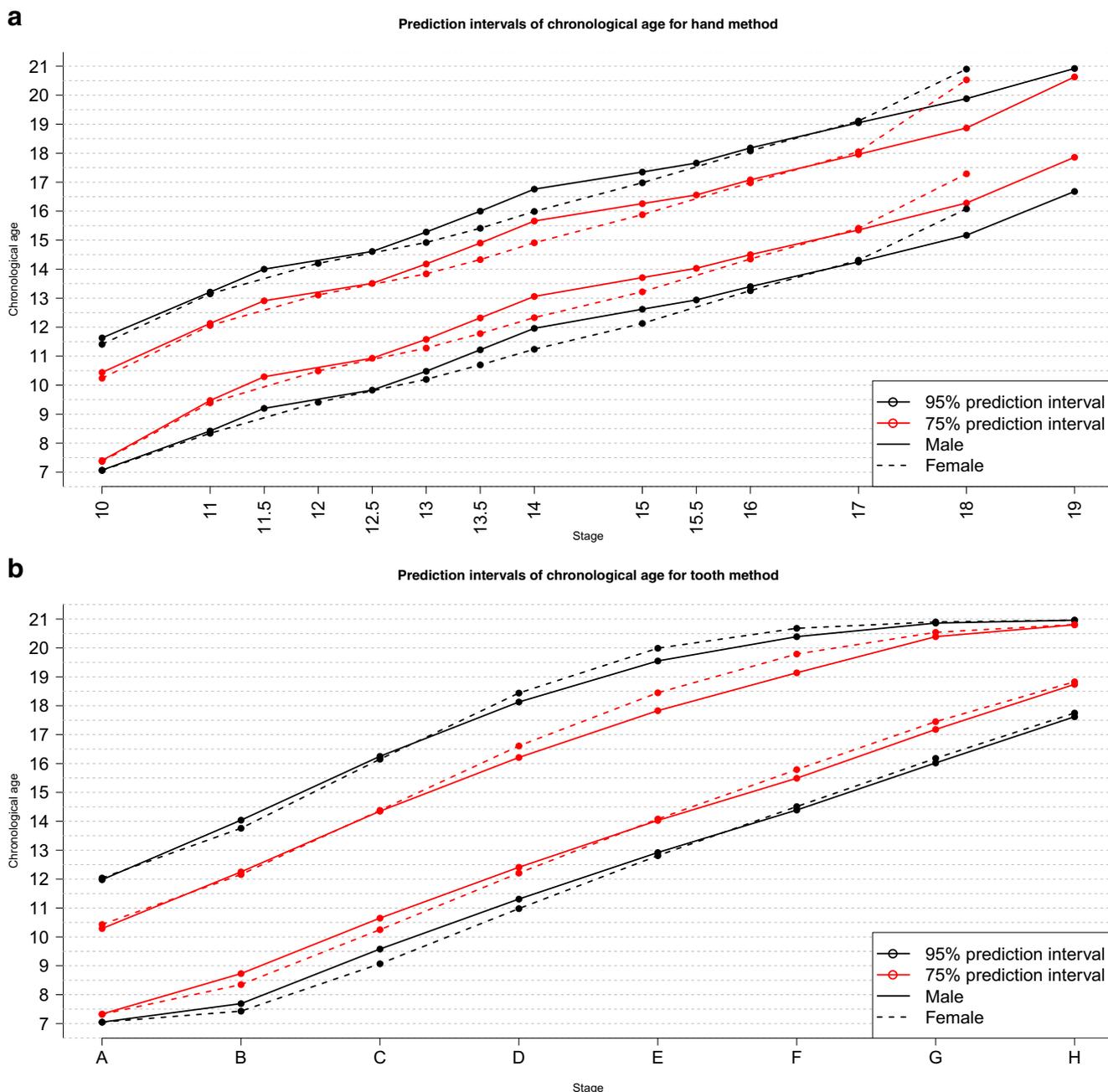
95% prediction interval [14.4, 20.4]), the combination 18/F would give a 95% prediction interval of [15.6, 19.3] (see Fig. E1 in the [Supplementary material](#) section “E Additional results (the combinations)”). Hence, the combination of the two methods for this case only causes slightly narrower prediction intervals over the hand method alone.

For males, the average width of the 95% prediction intervals was 4.7 years for hand and 5.7 years for tooth, whereas for females, this was 4.8 and 5.9 years, respectively. Correspondingly, the average width of the 75% prediction intervals for males was 2.6 years for hand and 3.3 years for tooth, whereas for females, this was 2.7 and 3.6 years, respectively. Overall, the intervals are slightly narrower for males than for females. For the combined methods, the average width of the 95% prediction intervals was 4.5 years for males and 4.6 years for females, whereas for the 75% prediction intervals, the average widths were both 2.7 years for both sexes.

It is also noteworthy how prediction intervals “jump” for skeletal age 18 for females. This is a consequence of the included ages. The jump would be lower with prior age limit 20 years for instance. Last, we see from Fig. 1 that the prediction intervals for chronological age of the two methods increase quite linearly as a function of development stages.

### The probability of being-below-a-certain-age-threshold results

For the hand method, the end stages 19 years for males and 18 years for females gave a probability of being below 18 years of 0.15 and 0.25 respectively (Fig. 2). In comparison, the likelihood ratio values (Table 4) were correspondingly 2.0



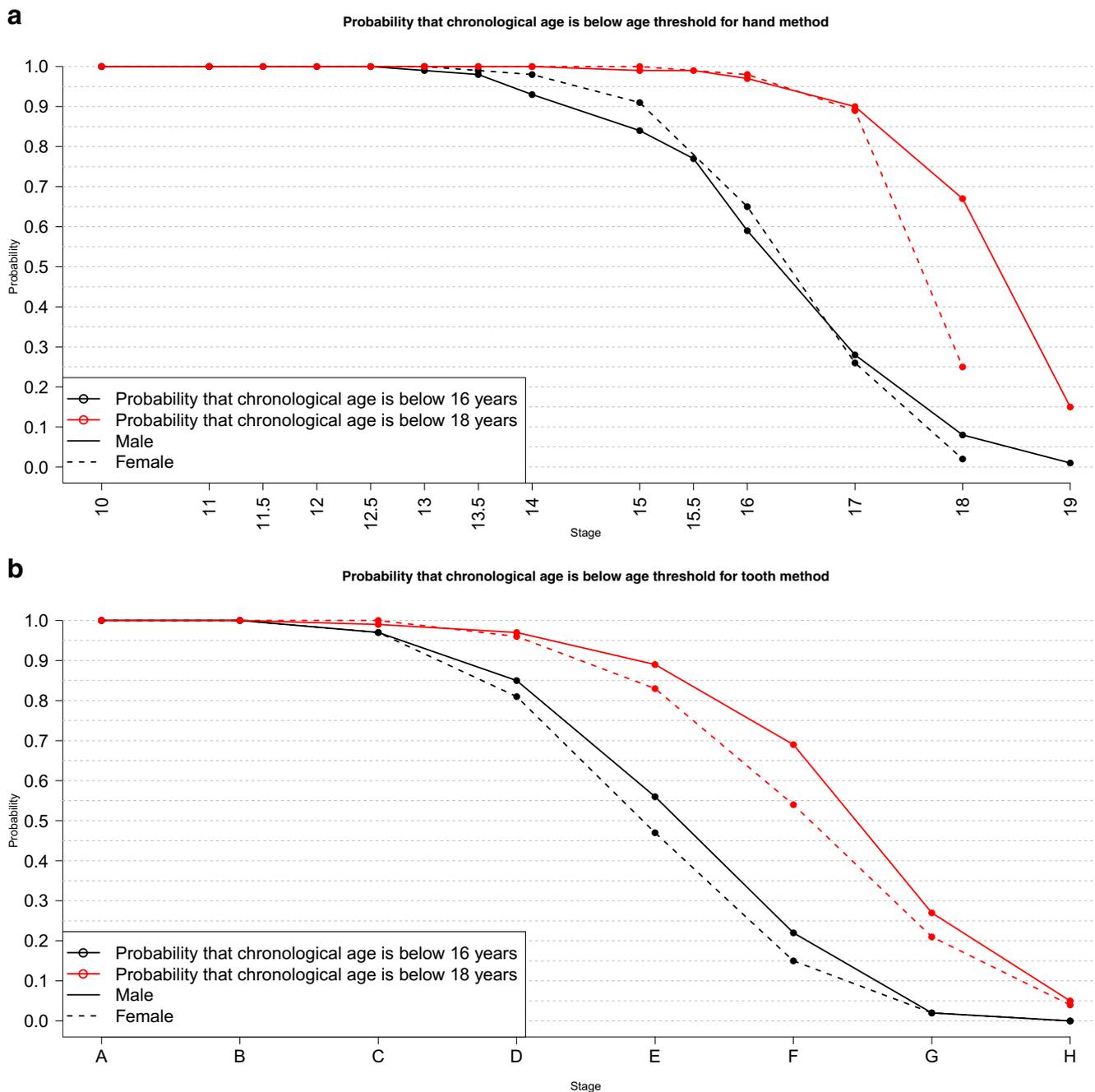
**Fig. 1** The 95% (black) and 75% (red) posterior prediction intervals of chronological age for given stages for each sex (hand in **a** and tooth in **b**). The solid and dashed lines represent the males and the females, respectively

and 1.3, giving only weak support for the hypothesis that the individual was above 18 years.

For the tooth method, the stage G gives a probability of being below 18 years of about 0.28 for males and 0.21 for female, whereas the end stage H gives a probability of about 0.05 for both sexes. In comparison, the corresponding likelihood ratio values (Table 4) were 17.5 and 27.8 for males and females, which indicate a moderate support for the hypothesis that the individual was above 18 years old.

Females need to be 18/H to have probability of being below 18 years less than 0.05, whereas males need to be 19/H for the same. The combinations 18/G for females and 19/G for males give a probability of being below 18 years around 0.1. In comparison, the corresponding likelihood ratio values were just above 2, giving only weak support of being above 18 years old.

From Fig. 2, we see that for tooth, the probabilities of being below a certain age threshold (16 or 18 years) are slightly smaller for females than for males, indicating



**Fig. 2** The posterior probabilities that chronological age is below a certain age threshold for given stages for each sex (hand in **a** and tooth in **b**). The black and red lines give the corresponding age thresholds 16 and 18, where the solid and dashed lines represent the males and the females, respectively

that the female development of the lower left third molar tooth is slower than for males.

### Discussion and conclusions

The aim of this paper is to provide updated evidential and predictive values regarding the chronological age of individuals based on age indicators from the grading systems (I)

Demirjian’s stages of the lower left third molar tooth and (II) GP skeletal age of the hand. The evidential value using likelihood ratio (LR) requires definition of two comparing hypotheses, which we considered “the chronological age is more than 16/18 years” versus “the chronological age is less than or equal to 16/18 years.” From our results, we found that the last stage H of third molar development and the combination 19/H (and 18/H for female) give moderate support (LR between 10 and 100) for the hypothesis that an individual is above 18 years

old. These are important findings, since in many countries, the age threshold of 18 years is critical to decide if an asylum seeker is treated as a child or as an adult. The LR could have been used further in order to define costs for making wrong decisions; however, this was not the scope of this study.

As an alternative to likelihood ratio we also considered a Bayesian approach in order to get the predictive information about chronological age for given observed age indicators (as suggested by Boldsen et al. [11] and Konigsberg et al. [14]). We did this by estimating prediction intervals which indicate expected ranges of chronological age corresponding to various combinations of development stages. The reason for considering such intervals rather than point estimates is that the latter could easily be misused as being very precise (when in fact they are not). Since the biological variation is considerable for these methods, the 95% prediction intervals give a wide age span. For practical legal use, we therefore chose to calculate the 75% prediction intervals (of chronological age) for comparison. Of course, the percentage is arbitrary, but the choice in real casework must reflect a user's tolerance (in terms of error)—this is not a decision that a scientist can make. It is worth mentioning that the Bayesian approach is a controversial topic because the choice of the prior is a subjective matter. However, if there is an obvious prior or a consensus prior, the Bayesian approach is unproblematic and enables results which typically are easier for laymen to interpret than the likelihood ratio.

An important issue is whether our fitted model is representative for predicting the chronological age of individuals. In this paper, we have assumed that the different studies follow the same development distributions (i.e., the stage probabilities) giving study-averaged inference. By inspection ([Supplementary material](#) “B.2 Stage probabilities as a function of age per study”), we observed that some of the stage probabilities differed between studies. Therefore, large sample studies that behave differently from other studies have great impact on the inference. An extension of the model where a random study intercept is added to the linear predictor would take care of such situations; however, it would shrink the age distribution towards the global mean of chronological age [31]. Recognizing common effects across several studies as in a meta-analysis is not the aim of this paper (see [32] for more discussions about the difference between fixed and mixed effects models). Because the fitted model is based on several studies combined, it is important to note the underlying study material. In our study material, the Chaumoitre 2016 study (for hand) represents over half of the reference data, with around 70 individuals per “whole-age” intervals. In contrast, other studies typically only have around 20 (except for Tise 2011 which has considerably more for some specific ages for males). Hence, the Chaumoitre 2016 study has a lot of impact

on the hand model. The tooth studies contribute about equally in terms of total sample size, however some more/less for specific ages than others ([Supplementary material](#) section “B.1 Number of individuals as a function of age per study”). Hence, the different studies provide about equal impact on the tooth model.

When combining the two methods, we assumed that the development of hand skeleton and the third molar is conditionally independent (given chronological age). Even though some publications support this assumption [5–7], more research is needed to explore whether this assumption actually holds true for different chronological ages and circumstances. The prediction intervals of chronological age were substantially wider for the tooth method than for the hand method. Because of this, the width of the prediction interval based on a combination of the two methods was only slightly smaller than for the hand method alone. However, combining the two methods is of great importance for testing the below/above 18-year threshold. Observing the last stage for both methods strengthens the evidence for the hypothesis that an individual is above 18 years (compared to observing the last stage for only one of the methods).

This paper also provides a framework to utilize several studies where information about the chronological age is given group-wise. This was done by recreating the chronological age for each individual based on additional model assumptions. For instance, the study Chaumoitre 2016 [19] illustrates that the distribution of chronological age (for given age indicators) was approximately normal for most situations. A clear weakness of adapting this assumption in our analysis is that we do not know how much the true individual values deviate from our recreated ones. Obviously, there is a cost in not having individual-based information as uncertainty then increases. Hence, results would be more precise if all datasets were openly available. For the common purpose of scientific progress, we therefore encourage research professionals in the field to willingly share such data in the future.

From our goodness-of-fit analyses, we found that the fitted model for the stage probabilities coincided with the observations for most situations (see [Supplementary material](#) section “C Exploratory goodness of fit”). However, we discovered some combinations of stages and chronological ages where the observations were “further off” than expected (if the observations were assumed to originate from the fitted model). None of the fitted transition analysis models as a whole were adequate for all observations, partially due to the highlighted situations found in the explorative goodness-of-fit analysis. However, we argue that we have pin-pointed the situations causing this and that the deviations are so small that it would not have any impact on the end results.

In conclusion, we have circumvented the problem of providing results which are biased in the direction of the ages of individuals included in the study (a problem also known as

“age mimicry”) [11]. This has been dealt with by building a statistical model for stage/age indicator probabilities as a function of age. Also importantly, we have been able to include several studies where the information about the chronological age is given group-wise. However, notice that there is a clear limitation of how representative the model is for practical usage since most hand studies are from Europe whereas most tooth studies are from East Asia.

The source code for running all the modeling and simulations and providing the results can be obtained by contacting the first author.

**Acknowledgments** We thank Jayakumar Jayaraman, Simon Camilleri, Rick R. van Rijn, Marco Tisè, Eugénia Cunha, and Abdul Mueed Zafar for providing data and Peter Gill for proofreading. We also want to thank the reviewers for their useful comments.

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Appendix**

**Candidate models for describing stage probabilities as a function of chronological age**

Let the ordinal stage variables be given as  $j = 1, \dots, J$ , with  $x$  as the chronological age variable and model parameters  $\theta = (\alpha_1, \dots, \alpha_{J-1}, \beta, \lambda)$ . Here, we define  $\lambda$  to be one of the discrete values  $\{0.1, 0.2, \dots, 0.9, 1\}$ . By letting  $Y$  be the discrete variable of the stage outcomes  $1, \dots, J$  which an individual  $l$  with age  $x$  may be in, the candidate models (1–4) can be written as

1. Proportional-odds cumulative with logit link
  - a.  $\text{logit}(\text{Pr}(Y \leq j | \theta, x)) = \alpha_j + \beta \times x^\lambda$
2. Proportional-odds cumulative with probit link
  - a.  $\text{probit}(\text{Pr}(Y \leq j | \theta, x)) = \alpha_j + \beta \times x^\lambda$
3. Continuation-ratio with logit link
  - a.  $\text{logit}(\text{Pr}(Y = j | Y \geq j, \theta, x)) = \alpha_j + \beta \times x^\lambda$
4. Continuation-ratio with probit link
  - a.  $\text{probit}(\text{Pr}(Y = j | Y \geq j, \theta, x)) = \alpha_j + \beta \times x^\lambda$

For  $j = J$  (last stage), we have  $\text{Pr}(Y \leq J | \theta, x) = 1$  and  $\text{Pr}(Y = J | Y \geq J, \theta, x) = 1$ . The link function  $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ , while the link function  $\text{probit}$  is the cumulative standard normal distribution.

**Details of each individual transition model**

The candidate models are with the form  $f(\text{Pr}(Y \leq j | \theta, x))$  and  $f(\text{Pr}(Y = j | Y \geq j, \theta, x))$  where the link function  $f$  is either *logit* or *probit*. To calculate the likelihood function, we need an expression for the stage probabilities  $\text{Pr}(Y = j | \theta, x)$ . We now describe this mathematically for each type of model:

*Proportional-odds cumulative model*

$$\begin{aligned} \text{Pr}(Y = 1 | \theta, x) &= f^{-1}(\alpha_1 + \beta \times x^\lambda) \\ \text{Pr}(Y = j | \theta, x) &= f^{-1}(\alpha_j + \beta \times x^\lambda) - f^{-1}(\alpha_{j-1} + \beta \times x^\lambda) \\ &\quad \text{for } j = 2, \dots, J-1 \\ \text{Pr}(Y = J | \theta, x) &= 1 - f^{-1}(\alpha_{J-1} + \beta \times x^\lambda) = 1 - \sum_{j=1}^{J-1} \\ \text{Pr}(Y = 1 | \theta, x) \end{aligned}$$

*Continuation-ratio model*

$$\begin{aligned} \text{Pr}(Y = 1 | \theta, x) &= f^{-1}(\alpha_1 + \beta \times x^\lambda) \\ \text{Pr}(Y = j | \theta, x) &= f^{-1}(\alpha_j + \beta \times x^\lambda) \prod_{l=1}^{j-1} [1 - f^{-1}(\alpha_{l-1} + \beta \times x^\lambda)] \\ &\quad \text{for } j = 2, \dots, J-1 \\ \text{Pr}(Y = J | \theta, x) &= 1 - \sum_{j=1}^{J-1} P(Y = 1 | \theta, x) \end{aligned}$$

**Model fitting using maximum likelihood estimation**

In order to fit the candidate models (1–4) to the data, a maximum likelihood estimate is chosen for the unknown parameters  $\theta$ . In other words, the parameters are chosen such that the data are those most likely to be observed. These are found by maximizing the likelihood function (on the log scale)

$$L(\theta) = \prod_{i=1}^I \prod_{j=1}^I \prod_{l=1}^{n_{i,j}} \text{Pr}(Y_{i,l} = j | \theta, x_{i,l})$$

where  $n_{i,j}$  is the number of individuals in stage  $j$  for study  $i$  (out of  $I$  studies) and  $x_{i,l}$  is the chronological age for individual  $l$  in study  $i$  observed with stage  $Y_{i,l} = j$ . Here, the unknown parameters are equal/common across the studies  $i = 1, \dots, I$ .

## Details of the likelihood ratio test statistics and the calculation of the $p$ value

For a given age  $a$ , say  $a=16$  years, we define the total number of observations between 15.5 and 16.5 years to be  $n$ . Of these,  $x_j$  are observed to have stage  $j$  (for the  $j = 1, \dots, J$  stages). The corresponding stage probabilities based on a fitted model (based on one of the candidate models) for these stages are

$$\hat{p}_j = \Pr(S = j | \text{Age} = a). \text{ We also have } n = \sum_{j=1}^J x_j \text{ as the total}$$

number of individuals. We now define a likelihood ratio test statistics which is based on comparing the fitted model (i.e., the model under the null hypothesis) and an extended model (i.e., the number of observations at stage  $j$  is binomial distributed as the alternative hypothesis).

$$LRT_j = \frac{\text{bin}\left(x_j | p_j = \frac{x_j}{n}\right)}{\text{bin}\left(x_j | p_j = \hat{p}_j\right)}$$

where  $\text{bin}$  is the binomial density function taking values  $x_j = 0, 1, \dots, n$ .

ratio test statistics is given as

$$Q_j = x_j \times \log\left(\frac{x_j}{n \times \hat{p}_j}\right) + (n - x_j) \times \log\left(\frac{n - x_j}{n \times (1 - \hat{p}_j)}\right)$$

Here, if  $x_j = 0$  the first term is set to 0, whereas if  $n - x_j = 0$ , the last term is set to 0.

We are interested in calculating the  $p$  value for an observation  $Q_j$  for any of the stages  $j$ , i.e.,  $\Pr(Q \geq Q_j | \text{Fitted model} = \text{model under null hypothesis})$ , where  $Q$  is the random variable of likelihood ratio statistics under the fitted model. This measure is used to investigate whether the fitted model is adequate for the observed dataset for a particular stage and a chronological age. We calculate this  $p$  value exactly by our own R-script (Where we utilize that  $\log \text{bin}(x | n, p) = \log \Gamma(n + 1) - \log \Gamma(n - x + 1) - \log \Gamma(x + 1) + x \log p + (n - x) \log(1 - p)$ ). This means we can vectorize the probability of the outcomes very easily with utilizing the  $\text{lgamma}$  function in R. The procedure to calculate the  $p$  value was as follows:

Let  $Q(x)$  be the obtained test value for an observation  $x$ , where  $x$  is binomial distributed following the fitted model (the model under the null hypothesis). The  $p$  value of the test can then be calculated exactly as

$$p \text{ value} = \sum_{x=0}^n \text{Ind}(Q(x) \geq Q_j) \times \text{bin}(x | p = \hat{p}_j)$$

where  $\text{Ind}$  is the indicator function taking value 1 if true, or 0 if false.

## Details of the Lipsitz test statistics and the calculation of the $p$ value

There are several steps involved in order to carry out the Lipsitz test statistics [24, 25]: Individuals were categorized into ten ( $G = 10$ ) almost equally sized groups. This categorization was based on sorting the individuals based on an assigned score. Group one contained the individuals with lowest scores and group ten contained the individuals with highest scores (etc.). We used that the (ordinal) score for an individual with age  $a$  was given as the weighted sum of fitted stage probabilities,  $\sum_j j \times \Pr(S = s_j | \text{Age} = a)$ , where  $s_j$  is the  $j$ th ordinal stage. Define the indicator variable  $I_{i, g}$  to be value 1 if individual  $i$  ( $i = 1, \dots, n$ ) was categorized to be in group  $g$  ( $g = 1, \dots, G$ ), and 0 otherwise. Then, the linear predictor of the “transition model” (see section A.1 in the “Appendix”).

$$\alpha_j + \beta \times x^\lambda \quad (\text{Model0})$$

is extended to

$$\alpha_j + \beta \times x^\lambda + \sum_{g=0}^{G-1} \gamma_g \times I_{i, g} \quad (\text{Model1})$$

where  $\gamma_1, \dots, \gamma_{G-1}$  are unknown parameters. By letting  $L_0$  being the maximum likelihood value for Model 0 and  $L_1$  being the maximum likelihood value for Model 1, the Lipsitz test statistics  $-2(L_1 - L_0)$  is chi-square distributed with degrees of freedoms  $G - 1 = 9$  under the null hypothesis;  $\gamma_1 = \dots = \gamma_{G-1} = 0$ , i.e., that the fitted transition model (Model 0) is adequate. We used the R-package VGAM (version 1.0–3) to fit the models.

## References

1. European Asylum Support Office (2013) EASO Age assessment practice in Europe. Luxembourg
2. Greulich W, Pyle S (1959) Radiograph Atlas of Skeletal Development of the Hand and Wrist, 2nd edn. Stanford
3. Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. Hum Biol 45(2):211–227
4. Mincer HH, Harris EF, Berryman HE (1993) The A.B.F.O. study of third molar development and its use as an estimator of chronological age. J Forensic Sci 38(2):379–390
5. Gelbrich B, Frerking C, Weiss S, Schwerdt S, Stellzig-Eisenhauer A, Tausche E et al (2015) Combining wrist age and third molars in forensic age estimation: how to calculate the joint age estimate and its error rate in age diagnostics. Ann Hum Biol 42(4):389–396
6. Varkkola O, Ranta H, Metsäniitty M, Sajantila A (2011) Age assessment by the Greulich and Pyle method compared to other skeletal X-ray and dental methods in data from Finnish child victims of the Southeast Asian Tsunami. Forensic Sci Med Pathol 7(4):311–316
7. Kanbur NO, Kanlı A, Derman O, Eifan A, Ataç A (2006) The Relationships between Dental Age, Chronological Age and Bone Age in Turkish Adolescents with Constitutional Delay of Growth. J Pediatr Endocrinol Metab 19(8):979–985

8. Dahlberg P, Mosdøl A, Ding KY, Bleka Ø, Rolseth V, Straumann GH et al (2017) Agreement between chronological age and bone age based on the Greulich & Pyle-atlas for age estimation: a systematic review. *Folkehelseinstituttet, Oslo*
9. Rolseth V, Mosdøl A, Dahlberg PS, Ding KY, Bleka Ø, Skjerven-Martinsen M et al (2017) Demirjian's development stages on wisdom teeth for estimation of chronological age: a systematic review. *Folkehelseinstituttet, Oslo*
10. Bocquet-Appel J-P, Masset C (1982) Farewell to paleodemography. *J Hum Evol* 11(4):321–333
11. Boldsen JL, Milner GR, Konigsberg LW, Wood JW (2002) Transition analysis: a new method for estimating age from skeletons. *Cambridge Studies in Biological and Evolutionary Anthropology*. 73–106
12. Royall RM (1997) *Statistical Evidence: A Likelihood Paradigm*. Hall Ca, editor. London
13. Blume JD (2002) Likelihood methods for measuring statistical evidence. *Stat Med* 21(17):2563–2599
14. Konigsberg LW, Herrmann NP, Wescott DJ, Kimmerle EH (2008) Estimation and evidence in forensic anthropology: age-at-death. *J Forensic Sci* 53(3):541–557
15. Lee SH, Lee JY, Park HK, Kim YK (2009) Development of third molars in Korean juveniles and adolescents. *Forensic Sci Int* 188(1):107–111
16. Johan NA, Khamis MF, Abdul Jamal NS, Ahmad B, Mahanani ES (2012) The variability of lower third molar development in Northeast Malaysian population with application to age estimation. *J Forensic Odonto-Stomatol* 30(1):45–54
17. Duangto P, Iamaroon A, Prasitwattanaseree S, Mahakkanukrauh P, Janhom A (2017) New models for age estimation and assessment of their accuracy using developing mandibular third molar teeth in a Thai population. *Int J Legal Med* 131(2):559–568
18. Li G, Ren J, Zhao S, Liu Y, Li N, Wu W et al (2012) Dental age estimation from the developmental stage of the third molars in western Chinese population. *Forensic Sci Int* 219(1):158–164
19. Tise M, Mazzarini L, Fabrizio G, Ferrante L, Giorgetti R, Tagliabracci A (2011) Applicability of Greulich and Pyle method for age assessment in forensic practice on an Italian sample. *Int J Legal Med* 125(3):411–416
20. Chaumoitre K, Saliba-Serre B, Adalian P, Signoli M, Leonetti G, Panuel M (2017) Forensic use of the Greulich and Pyle atlas: prediction intervals and relevance. *Eur Radiol* 27(3):1032–1043
21. Agresti A (2003) *Logit models for multinomial responses. Categorical data analysis*. John Wiley & Sons, Inc. p 267–313
22. Yee TW (2010) The VGAM Package for Categorical Data Analysis. *J Stat Softw* 1(10)
23. Yee TW (2017) VGAM: Vector Generalized Linear and Additive Models. Available from: <https://CRAN.R-project.org/package=VGAM>
24. Fagerland MW, Hosmer DW (2016) Tests for goodness of fit in ordinal logistic regression models. *J Stat Comput Simul* 86(17): 3398–3418
25. Lipsitz SR, Fitzmaurice GM, Molenberghs G (1996) Goodness-of-Fit Tests for Ordinal Response Regression Models. *Appl Stat* 45: 175–190
26. Pulkstenis E, Robinson TJ (2004) Goodness-of-fit tests for ordinal response regression models. *Stat Med* 23(6):999–1014
27. Aitken CGG, Taroni F (2005) *The evaluation of evidence. Statistics and the evaluation of evidence for forensic scientists*. John Wiley & Sons, Ltd. p 69–118
28. Association of Forensic Science P (2009) Standards for the formulation of evaluative forensic science expert opinion. *Sci Justice* 49(3):161–164
29. Spiegelhalter DJ (2004) Incorporating Bayesian Ideas into Health-Care Evaluation. *Stat Sci* 19(1):156–174
30. Sironi E, Gallidabino M, Weyermann C, Taroni F (2016) Probabilistic graphical models to deal with age estimation of living persons. *Int J Legal Med* 130(2):475–488
31. Agresti A (2003) *Random effects: generalized linear mixed models for categorical responses. Categorical data analysis*. John Wiley & Sons, Inc. p 491–537
32. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) *Fixed-effect versus random-effects models. Introduction to meta-analysis*. John Wiley & Sons, Ltd. p 77–86