Published in final edited form as:

Ann Hum Biol. 2013 ; 40(6): 505–514. doi:10.3109/03014460.2013.806591.

An update of the statistical methods underlying the FELS method of skeletal maturity assessment

Ramzi W. Nahhas¹, Richard J. Sherwood^{1,2,3}, Wm. Cameron Chumlea^{1,2,4}, and Dana L. Duren^{1,5}

¹Division of Morphological Sciences and Biostatistics, Lifespan Health Research Center, Department of Community Health

²Department of Pediatrics, Boonshoft School of Medicine, Wright State University, Dayton, OH 45435, USA

³Department of Orthodontics, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA

⁴Division of Epidemiology, Lifespan Health Research Center, Department of Community Health

⁵Department of Orthopaedic Surgery, Sports Medicine and Rehabilitation, Boonshoft School of Medicine, Wright State University, Dayton, OH 45435, USA

Abstract

Background: Evaluation of skeletal maturity provides clinicians and researchers a window into the developmental progress of the skeleton. The FELS method for maturity assessment provides a point estimate and standard error based on 98 skeletal indicators.

Aim: This paper outlines the statistical methodology used by the original FELS method and evaluates improvements that address the following: serial correlation in the calibration sample is now considered, a Bayesian estimation method is now employed to improve estimation near ages 0 and 18 years and uncertainty in the calibration due to sampling is now accounted for when computing confidence limits.

Subjects and methods: The original FELS method was calibrated using 677 Fels Longitudinal Study participants. In the improved method, serial correlation is accounted for using GEE, a Bayesian analysis with a prior centred on chronological age is used and the bootstrap is used to account for all sources of variation.

Results: Accounting for serial correlation resulted in larger slopes for ordinal indicators. The Bayesian paradigm led to narrower confidence limits and a natural interpretation of skeletal age. Sampling variability in the calibration parameters was negligible.

© 2013 Informa UK Ltd

Correspondence: Ramzi W. Nahhas, Department of Community Health, Wright State University, 3171 Research Boulevard, Dayton, OH 45420, USA. Tel: 937-775-1421. ramzi.nahhas@wright.edu.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Conclusion: Improvements to the statistical basis of the FELS method provide a more effective method of estimating skeletal maturity.

Keywords

Bayesian; bone age; development; longitudinal; skeletal age

Introduction

Children mature at different rates such that, at a given chronological age, there is a large amount of population variation in maturity status. Skeletal age, a metric of skeletal maturation, can be used to diagnose developmental delay (Pludowski et al., 2005), predict adult stature (Roche et al., 1975a, 1988; Tanner et al., 1975b) and inform treatment decisions that depend on how close a child is to their pubertal growth spurt (Spadoni & Cianfarani, 2010). Also, in both clinical and epidemiological studies, skeletal age explains a portion of the between-child variation observed in body composition, physical status and performance (Beunen et al., 1992; Claessens et al., 2006; Malina et al., 1999).

The hand and wrist have been the focus of skeletal maturity assessment due to their easy accessibility and positioning and their large number of bones located in a very small area (helpful in limiting radiation exposure) (Cox, 1997; Roche, 1980). There are three primary methods of assessing skeletal age of the hand–wrist: the Greulich–Pyle (GP) atlas method (Greulich & Pyle, 1950, 1959), Tanner–Whitehouse (TW3) (Tanner & Whitehouse, 1959; Tanner et al., 1962, 1975a, 2001) and the FELS method (Chumlea et al., 1989; Roche et al., 1988). Both TW3 and the FELS method were developed to overcome the significant shortcomings of the GP atlas method and, while each has its own benefits and drawbacks, the FELS method remains the only skeletal maturity assessment procedure that provides a standard error, allowing computation of confidence limits for the calculated skeletal age (Chumlea et al., 1989).

For researchers wishing to use the FELS method, it is, at present, difficult to find a detailed description of how the method arrives at its skeletal age estimate (Roche et al., 1975b). The purpose of this article is to document the statistical basis of the FELS method and to outline three improvements: (a) accounting for serial correlation when calibrating the method with longitudinal data, (b) a skeletal age point estimate and confidence region based on a Bayesian reformulation of the estimation procedure and (c) accounting for uncertainty in the calibration parameters due to sampling variability when providing confidence limits for a skeletal age estimate.

Methods

The improvements to the FELS method are best understood after first describing the statistical methods used in the original version of the method.

The original FELS method of skeletal maturity assessment

The FELS method provides an estimated skeletal age (SA) and its associated standard error for a child based on an assessment of the maturity of the bones of the hand–wrist. This is

accomplished by examining the distribution of chronological ages (CA) among children in the reference sample with skeletal maturity similar to the child being assessed. SA is, therefore, defined as a relative measure of skeletal maturity, relative to a particular reference group. For the FELS method, the reference group is the participants in the Fels Longitudinal Study (Roche, 1992). At the time the method was developed, this group consisted of 677 Caucasian children (355 boys, 322 girls) measured (via a hand–wrist radiograph) on a total of 13 823 occasions from 1932–1977 at ages from 1 month to 22 years. These children were born between 1928–1974 and lived in southwestern Ohio, US (Roche et al., 1988). Thus, any skeletal age estimate produced by the existing FELS method must be interpreted relative to that sample. For example, the statement “this girl has a FELS method skeletal age of 10 years” should be interpreted as “this girl’s skeletal maturity is similar to that of 10 year old girls in the Fels Longitudinal Study sample used to calibrate the FELS method”.

In the FELS method, 98 skeletal maturity indicators are measured (58 binary indicators, 27 ordinal indicators and 13 continuous measures) from a hand–wrist radiograph. At any given age, only 20–66 of the indicators are actually assessed. Which indicators are assessed depends on which were found to vary in the reference sample at that age. The set of indicators assessed at a given age remains unchanged in the revised method presented herein. Each indicator progresses from less mature to more mature as a child develops and indicators differ in their average timing of maturation and in the between-child variability of that timing. The FELS method summarizes the information from these individual indicators using a maximum likelihood (ML) procedure. SA can be interpreted as a measure of the centre of the distribution of CA in the calibration sample given an individual child’s observed indicator values. If all the indicators were categorical, a non-parametric estimate of this conditional distribution could be obtained by noting the distribution of ages within every possible combination of indicator values. With a large number of indicators, however, this would require an enormous reference group to obtain any kind of reliable estimate.

The statistical procedure used in the FELS method is based on Samejima (1969), in which the value of a continuous latent trait is estimated based on binary and ordinal indicators of that trait. For example, in educational testing, the latent trait is a student’s ability and the indicators are the test questions. In SA estimation, the latent trait is skeletal maturity and the indicators are the aforementioned 98 traits, with Samejima’s model extended to include continuous indicators. The application of Samejima’s method to skeletal maturity was first done in the context of the knee joint (Roche et al., 1975b). Both the original FELS method and the improvements to the FELS method to be described herein make use of this latent trait model.

The FELS method requires calibration before it can be used for estimation. The calibration of the FELS method consists of estimating the parameters of the probability distribution (for binary and ordinal indicators) or density function (for continuous indicators) of each indicator conditional on chronological age. In the estimation step, these functions are combined to form a joint likelihood function which is maximized, providing an estimate of the age in the reference group at which children are most similar in skeletal maturity to the child being assessed. The following documents the statistical procedures that were used for calibration and estimation in the original FELS method (Roche et al., 1988).

Calibration in the original method

A binary maturity indicator can take on levels 1 (immature) or 2 (mature), while an ordinal indicator with J levels can take on levels 1 (immature), 2, ..., $J - 1$ or J (mature). For example, the onset of ossification of the pisiform (indicator “P-1”) is a binary indicator with grade 1 assigned when the centre of ossification is not present and grade 2 assigned when the centre is present. As with P-1, the absence/presence of the ossification centre of the lunate is also graded, but the indicator “L-1” also includes information regarding the shape of the lunate. Grades 1 and 2 again indicate absence or presence of the ossification centre and additional levels 3, 4 and 5 occur as the shape of the lunate changes. Each continuous indicator is a ratio of the epiphyseal to metaphyseal width for a particular bone (such as the first proximal phalanx; indicator PP I-1) and generally ranges in value from 0.2–1.4. Overall, the indicators were chosen to create a set that jointly changes with age in a manner that is fairly consistent across children but with enough variability to provide salient information regarding maturational status. Specific reasons for the choice of the particular 98 indicators included in the FELS method are described elsewhere (Roche et al., 1988).

Calibrating the FELS method to the reference group involves estimating the parameters of the conditional probability functions for the 98 individual skeletal maturity indicators. The probability distribution function for a binary indicator X with possible levels 1 (immature) and 2 (mature), conditional on CA , is modelled with the logistic model

$$P(X=2|CA) = \{1 + \exp[-d(CA - \tau)]\}^{-1}$$

where τ represents the median age at maturity and d is a rate parameter. A large d represents a steep slope on the logit scale and corresponds to low variability in the calibration sample in terms of when maturity occurs; smaller d indicates more variability. For an ordinal indicator X with J levels 1, ..., J , the probability distribution function is modelled with the ordinal logistic model

$$P(X \geq j|CA) = \{1 + \exp[-d(CA - \tau_{j-1})]\}^{-1}$$

for $j = 2, \dots, J$, with $P(X = 1|CA) = 1$ and $P(X = J + 1|CA) = 0$. Here, τ_{j-1} represents median age of achieving maturity level j . Again, d is a rate parameter reflecting between-individual variability in the timing of maturity and is assumed to be the same for all j . The probability of being at maturity level j is $P(X = j|CA) = P(X \geq j|CA) - P(X \geq j + 1|CA)$. A continuous indicator X is modelled using the linear regression model $Z = \mu + \beta CA + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, and X is transformed as $Z = X^3$ in order to better meet the linearity, normality and homoscedasticity assumptions of the linear regression model. Thus, the probability density function at $Z = z$ is

$$\begin{aligned} p(z|CA) &= \phi\left(\frac{z - \mu - \beta CA}{\sigma}\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - \mu - \beta CA}{\sigma}\right)^2\right] \end{aligned}$$

where $z = x^3$ and $\phi()$ is the probability density function of the standard normal distribution. Calibration of the FELS method is carried out separately for boys and girls.

Estimation in the original method

Let θ be the quantity to be estimated. With respect to the reference group, θ is CA. With respect to the individual being assessed, θ is SA. That is, the FELS method defines *skeletal* age for an individual to be the most likely *chronological* age in the sub-group of the calibration sample which has similar skeletal maturity. The likelihood function for an observed set of indicators x is the joint distribution function of those indicators given age. Calibration of such a high-dimensional distribution would be very difficult and the FELS method, following Samejima (1969), makes the simplifying “local independence” assumption that the indicators are independent given CA. This is *not* the same as assuming that the indicators are totally independent, which would be a much stronger and less reasonable assumption. The local independence assumption implies that any dependence in the traits is a consequence of the fact that they all change from less to more mature with chronological age. Local independence is still a fairly strong assumption, but is necessary in order to obtain a tractable likelihood function. Given local independence, the likelihood function is the product of the indicator-specific probability functions

$$\begin{aligned} L(\theta|x, d, \tau, \mu, \beta, \sigma) &= \prod_{X_i \text{ binary}} [1 - P(X_i=2|\theta, d_i, \tau_i)]^{I(x_i=1)} P(X_i=2|\theta, d_i, \tau_i)^{I(x_i=2)} \\ &\times \prod_{X_i \text{ ordinal}} \prod_{j=1}^{J_i} [P(X_i \geq j|\theta, d_i, \tau_{i,j-1}) - P(X_i \geq j+1|\theta, d_i, \tau_{i,j})]^{I(x_i=j)} \\ &\times \prod_{X_i \text{ continuous}} \phi\left(\frac{x_i^3 - \mu_i - \beta_i \theta}{\sigma_i}\right) \end{aligned}$$

where the index i is over the 98 indicators, the function $I(x=j)$ is 1 if $x=j$ and 0 otherwise and all parameters other than θ are assumed to be known (with values supplied by the calibration step). Each term in the likelihood function corresponds to the probability of the value that was observed (or, for continuous indicators, the probability density). The FELS method SA estimate is the value of θ that maximizes $L(\theta|x, d, \tau, \mu, \beta, \sigma)$.

At times, maximization of the likelihood results in $SA < 0$ years or $SA > 18$ years and the FELS method then assigns SA to be 0 or 18. Also, in cases where all the assessed indicators are immature or all are mature, and no continuous indicators have been assessed, the likelihood will be monotonic and, thus, no global maximum exists. Again, in this case, SA is set to 0 or 18. Thus, a set of indicators that result in $SA = 0$ or 18 means that individuals with that level of skeletal maturity in the calibration sample were generally age 0 or age 18, respectively. Skeletal ages equal to 0 or 18 do not necessarily imply immaturity or maturity in all indicators, although, given the existing FELS method calibration, immaturity in all indicators does result in $SA = 0$ and full maturation in all the indicators does result in $SA = 18$.

In most statistical analyses, the sample consists of multiple individuals and the likelihood function is a product over those individuals. In SA estimation, however, there is only a single individual and the “sample” is the set of 98 indicators. The number of individuals in the *calibration* sample does, however, come into play in that it affects the precision of estimation of d , τ , μ , β and σ . In the estimation step, the FELS method assumes that these values are known when in fact they were estimated. The impact of this failure to account for all sources of variation will be assessed below and turns out to be negligible due to the large calibration sample.

As an illustration of this estimation procedure, consider the following simplified example in which SA is estimated based on two binary indicators: presence of a radiopaque line or zone within the lateral margin of the triquetral (TRI-4) and presence of a radiopaque line or zone within the distal margin of the trapezium (TPM-3). Figure 1 illustrates the probabilities of maturity for TRI-4 and immaturity for TPM-3 as functions of age (left panel) and the product of these probabilities which is the likelihood function (right panel). The likelihood function is greatest at age 10.6 years, which is interpreted as meaning that in the reference group used to calibrate the FELS method, the *chronological* age at which you are most likely to observe (mature, immature) for (TRI-4, TPM-3) is 10.6 years. Thus, the *skeletal* age of a child with this level of maturity is estimated to be 10.6 years. Incorporating additional indicators into the SA estimation procedure is simply an extension of this two-indicator example, where, instead of simply multiplying two probability functions, one multiplies as many functions as there are indicators assessed at that CA.

SA standard error estimation in the original method

The FELS method computes the variance of the SA estimate as the inverse of the observed information at the maximum likelihood estimate (MLE) $\hat{\theta}$, or

$$\left(-\frac{\partial^2}{\partial \theta^2} \log L(\theta | x, d, \tau, \mu, \beta, \sigma) | \theta = \hat{\theta} \right)^{-1}$$
 (the inverse of the second derivative of the log-likelihood). The standard error (SE) of the MLE is the square root of this quantity. For the example shown in Figure 1, the SE is 1.4 years, which is quite large as only two indicators were assessed. Typical SE values in practice are ~0.3 years. Using a normal distribution as an approximation to the sampling distribution of the MLE, a $100 \times (1 - \alpha)\%$ confidence interval (CI) can be constructed as $\hat{\theta} \pm z_{1-\alpha/2} SE$ where z_p is the p^{th} percentile of a standard normal distribution. The bounds of the CI are truncated to be within $[0, 18]$ *post hoc*. When maximization of the likelihood occurs at $SA < 0$ or $SA > 18$, the SE is often extremely large. Also, when the likelihood is monotonic, the SE is not defined. The resulting very wide or undefined CI is misleading; one should, in fact, have a lot of confidence that SA is either 0 or 18. This issue is resolved using the Bayesian method to be described below.

In most settings, the MLE is asymptotically normal and a normal approximation is reasonable for a large sample size. As mentioned previously, in SA estimation the “sample size” is the number of skeletal maturity indicators which can be at most 98 (in fact, it is always fewer, as only a sub-set of indicators need be assessed at any given age). In the

Bayesian method described below, there is no need to use a normal approximation for the sampling distribution of the MLE in order to obtain confidence limits.

Improvement to the FELS method #1: Accounting for serial correlation

An assumption implicit in the calibration step is that the units in the calibration sample are independent; the estimation methods for the parameters of the logistic, ordinal and linear regression models assume independent observations. In fact, since the reference group subjects were participants in a longitudinal study and many participants are related, there is serial correlation as well as dependence due to family relatedness. Regarding dependence due to family relatedness, Roche et al. (1988) re-did the calibration after randomly selecting only one child of each sex among first-degree relatives and found no significant differences.

An improvement is to account for serial correlation through the use of generalized estimating equations (GEE) (Heagerty & Zeger, 1996; Liang & Zeger, 1986; Zeger & Liang, 1986) when estimating the parameters of the logistic, ordinal and linear regression models. Marginal models fit with GEE are more appropriate in this setting than random effects models since, in the calibration step, the target of inference is the population rather than the individual (Fitzmaurice & Verbeke, 2008). The parameters of the models for the indicators describe their population distributions given CA. For all indicators, GEE with an exchangeable correlation structure is used, unless the estimation algorithm fails to converge, in which case independence is assumed.

To evaluate the impact of this improvement, the calibration was conducted with and without accounting for serial correlation using the original FELS method calibration sample. Calibration was carried out in R 2.10 (R Development Core Team, 2009) with the functions (packages) `geeglm` (geepack) for GEE for binary and continuous variables (Halekoh et al., 2006; Liang & Zeger, 1986; Prentice & Zhao, 1991); `ordgee` (geepack) for GEE for ordinal variables (Heagerty & Zeger, 1996; Halekoh et al., 2006); and `polr` (MASS) (Venables & Ripley, 2002) for ordinal logistic regression assuming independence for cases where `ordgee` failed to converge even with the independence assumption. Of the 27 ordinal traits, eight were modelled using GEE with an exchangeable correlation structure, about half (14 for boys, 13 for girls) used GEE assuming independence and the remaining used ordinal logistic regression.

Improvement to the FELS method #2: Bayesian reformulation of the estimation step

The likelihood based estimation method encounters difficulties when SA is near 0 or 18. CIs based on the MLE's asymptotic normal distribution can contain values less than 0 or greater than 18 (requiring *post hoc* truncation). Also, as discussed above, the SE can be extremely large or undefined, resulting in a CI which is not very helpful or is undefined. Additionally, it is natural (but incorrect) to interpret SA as a summary of the distribution of CA in a particular sub-set of the reference group, for example as the average CA among those in the reference group with similar skeletal maturity. The ML estimation method does not allow that interpretation. The method finds the CA that maximizes the likelihood function, but it is incorrect to interpret this as a summary of the probability distribution of ages in the reference group. Such an interpretation, however, does follow from a Bayesian analysis in

which, rather than maximizing a likelihood function, one directly estimates the distribution of ages in the reference group given the assessed values of the skeletal maturity indicators for a particular child.

Using Bayes' rule, $f(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)f(\theta)/f(\mathbf{x})$ where $f(\theta|\mathbf{x})$ represents the “posterior distribution” of CA in the sub-set of the reference group with indicator values $\mathbf{x}$. The term $f(\theta)$ is the “prior distribution” of CA which represents one's beliefs about the distribution of skeletal age prior to assessing any indicators. Thus, the prior and posterior distributions represents estimates of the distribution of skeletal age before and after collecting data, respectively. The term $f(\mathbf{x})$ is the marginal distribution of the indicators and can be left unspecified as it is constant with respect to θ . The term $f(\mathbf{x}|\theta)$ is the joint distribution of the indicators given age and is equivalent to the likelihood function.

Given local independence, and the fact that $f(\mathbf{x})$ is a constant with respect to θ ,

$$f(\theta|\mathbf{x}) \propto \left[\prod_{i=1}^{98} f(x_i|\theta) \right] \times f(\theta) = L(\theta|\mathbf{x}, \mathbf{d}, \tau, \mu, \beta, \sigma) \times f(\theta).$$

The existing FELS method maximizes the product $\prod f(x_i|\theta) = L(\theta|\mathbf{x}, \mathbf{d}, \tau, \mu, \beta, \sigma)$ with respect to θ . This implicitly assumes that there is a non-informative prior distribution on θ , uniform over the real line (i.e. $f(\theta) = 1$). Such a non-informative prior implies that, prior to knowledge of the skeletal maturity indicators $\mathbf{x}$, any value of SA is just as likely as any other.

Thus, maximizing the FELS method likelihood is equivalent to a Bayesian analysis with a non-informative prior for SA. Using the Bayesian interpretation we see that, in the FELS method, the reported SA is an estimate of the mode of the distribution of chronological ages in the reference group among those with indicator grades $\mathbf{x}$, and the reported SE is an estimate of the standard deviation of that distribution.

A non-informative prior for θ , however, can lead to an improper posterior distribution, in particular when the likelihood is monotonic. Also, a non-informative prior is too conservative as there is some information regarding SA even before any indicators are assessed. Given a child's chronological age, one can be fairly confident that their SA is not too far from their CA. As an extreme example, the proportion of 15 year old children in the Fels Longitudinal Study that have SA = 7 years is 0. In the entire Fels Longitudinal Study collection of radiographs of children aged 0–22 years, SA (estimated by maximization of the likelihood) is at most 4.0 years different from CA. Thus, it makes sense to place a prior on θ that has almost all of its support over the age range within $\delta = 4$ years of the CA of the child being assessed (but not all of its support, allowing the possibility of a more extreme SA). A reasonable prior is a normal distribution centred at CA where $p_\delta = 0.99$ of the prior density is in $[CA - \delta, CA + \delta]$. This results in a proper posterior distribution which can be written as

$$f(\theta|\mathbf{x}) = C \times \left[\prod_{i=1}^{98} f(x_i|\theta) \right] \times f(\theta) \\ = L(\theta|\mathbf{x}, \mathbf{d}, \tau, \mu, \beta, \sigma) \times \phi \left[z_{1-\frac{1-P\delta}{2}} (\theta - CA) / \delta \right]$$

where C is a normalizing constant. SA is defined to be the mode of the posterior distribution. Unless SA is near 0 or 18, this estimate will be very similar to the MLE. All posterior values of θ 0 or 18 are interpreted as total immaturity ($SA = 0$) or full maturity ($SA = 18$), respectively. Thus, as with the MLE, SA is effectively truncated to be within [0, 18]. However, in the Bayesian paradigm, this truncation has a meaningful interpretation in terms of point masses at 0 and 18. Let $p_0 = P(\theta = 0|\mathbf{x})$ and $p_{18} = P(\theta = 18|\mathbf{x})$ be posterior probabilities that the $SA = 0$ and 18, respectively. If the MLE = 18, say, one can only state that those in the reference group most like the individual being assessed are “generally” 18 years or older. Using the Bayesian formulation, one can actually estimate the proportion of the reference group with skeletal maturity indicators $\mathbf{x}$ that are age 18 years or older. Note that, to compute posterior probabilities, the product of the likelihood and prior is normalized so that it integrates to one, resulting in a proper posterior distribution.

One could, at this point, compute the standard deviation of the posterior distribution. This will be almost exactly the same as the SE of the MLE described above when the posterior is symmetric. However, reporting SE can be misleading when SA is close to 0 or 18; a symmetric CI centred on a SA close to 0 or 18 computed as the posterior mode $\pm$ some multiple of the posterior standard deviation can contain values outside of [0, 18]. Using the posterior distribution, a more sensible method is to compute a posterior credible interval as the highest posterior density (HPD) region (Berger, 1985). The $100 \times (1 - \alpha)\%$ HPD region is the interval of ages with the highest posterior density with total probability $1 - \alpha$. As with the truncated ML-based CI, the HPD region is always within [0, 18] but, unlike the CI, this follows directly from collapsing all of the posterior probability outside of [0, 18] into point masses at 0 and 18.

The impact of this improvement is illustrated via comparison of the ML and Bayesian point and interval estimates for cases where the MLE is (a) 0, (b) near 0, (c) far from 0 and 18, (d) near 18 and (e) 18, as well as a comparison of the ML-based CI and HPD region widths for the entire Fels Longitudinal Study sample of 19 487 hand–wrist radiographs.

Improvement to the FELS method #3: Accounting for all sources of variation

The likelihood function is a product of probability functions, all of which are conditioned on the parameters of the distributions of the indicators given age. As discussed previously, the FELS method implicitly assumes that $\omega = (\mathbf{d}, \tau, \mu, \beta, \sigma)$ are known. Since the elements of ω were, in fact, *estimated* from the calibration sample, the method ignores a portion of the total variation when reporting the variation (SE, ML-based CI or HPD region) for SA . The likelihood (and, therefore, also the posterior distribution) only accounts for the variability of CA in the calibration sample given sex and the observed levels of skeletal maturity traits. It does not include the uncertainty in the calibration process (the estimation of ω) which results from sampling variability (i.e. the calibration sample is not the entire population of

children who will be assessed by the FELS method). The actual total variation should include both the variation within the calibration sample and sampling variability.

The total variation is difficult to quantify analytically, but can be approximated using the bootstrap (Efron & Tibshirani, 1993; Ogden & Tarpey, 2006). Let ω and $\hat{\omega}$ represent the calibrations based on the population and calibration sample, respectively. Subjects from the calibration sample were sampled with replacement $B = 500$ times to produce B bootstrap calibration datasets, each of which were used to estimate ω , resulting in the calibrations $\hat{\omega}_1, \dots, \hat{\omega}_B$. The posterior distribution that was used above should actually be written as $f(\theta|x, \hat{\omega})$, making explicit the dependence on the calibration sample. Using the bootstrap calibrations and noting that ω (the unknown population calibration) and x (the indicator data for a new individual) are effectively independent, the actual posterior distribution of θ given the skeletal maturity indicators can be approximated as

$$\begin{aligned} f(\theta|x) &= \int f(\theta, \omega|x) d\omega = \int f(\theta|x, \omega) f(\omega|x) d\omega \\ &= \int f(\theta|x, \omega) f(\omega) d\omega \approx \frac{1}{B} \sum_{b=1}^B f(\theta|x, \hat{\omega}_b) \end{aligned}$$

Thus, incorporating the error in the estimation of ω is straightforward. The average of the B posterior distributions is an estimate of the distribution of ages in the sub-group of the population from which the reference group was drawn which has indicators x , taking into account both population variability in skeletal maturity and variability in drawing the calibration sample from the population.

The impact of ignoring sampling variability was evaluated by estimating HPD regions for Fels Longitudinal Study participants with and without accounting for sampling variability.

Results

Improvement #1: Impact of accounting for serial correlation

Figure 2 illustrates the impact on the calibration of accounting for serial correlation. The impact of the adjustment for serial correlation was minimal for all parameters except the ordinal regression rate parameters (d) which were, in general, larger after accounting for serial correlation. Thus, this improved calibration method will result, in general, in narrower HPD regions than if serial correlation were ignored, since a larger rate parameter indicates a steeper logistic regression slope resulting in a more peaked distribution when the probabilities are combined to form the posterior.

Improvement #2: Impact of Bayesian reformulation of the estimation step

Figure 3 illustrates the likelihood (dashed curve), posterior density (solid curve), ML-based CI after truncation (vertical dashed lines) and the HPD region (vertical solid lines) for cases where the MLE is (a) 0, (b) near 0, (c) far from 0 and 18, (d) near 18 and (e) 18. Table 1 contains the MLE, ML-based 95% CI, posterior mode, 95% HPD region, $P(SA = 0)$ and $P(SA = 18)$ for these five scenarios. In cases (a), (c) and (d), the HPD region is narrower than the CI; in (b) the two are equal; and in (e) the ML-based CI does not exist since the

likelihood is monotonic, yet the Bayesian analysis still provides a HPD region. Also, the Bayesian analysis provides explicit estimates of $P(SA = 0)$ and $P(SA = 18)$ for each scenario.

The performance of the Bayesian method relative to ML on the Fels Longitudinal Study sample is illustrated in Figure 4 and Table 2. Figure 4 plots the difference in the widths of the truncated 95% ML-based CI and the 95% HPD region (values >0 indicate that the CI is wider than the HPD region) vs chronological age, with points in black corresponding to the HPD region being narrower and points in gray corresponding to the MLE CI being narrower. Since many points overlap, it is difficult to visually judge the relative proportion of gray and black points; in fact, as shown in Table 2, there are relatively few points for which the MLE CI is narrower. Within four age groups, and by sex, Table 2 presents the number of cases for which the CI is undefined (due to a monotonic likelihood; and including cases where the SE is extremely large due an almost flat likelihood at the maximum), the HPD region is narrower (better), the CI is narrower and the HPD region and CI are approximately equal in length ($|\text{difference in widths}| < 0.05$ years), as well as the proportion of cases for which the HPD region is at least as good as the CI (CI undefined, HPD narrower or equal width). Overall, the HPD region is at least as good 96.2% of the time.

A feature of the FELS method is that it provides a SA estimate regardless of how few indicators were assessed, with the SE increasing as fewer indicators are assessed. The one very large outlier for boys in Figure 4 (difference in widths = 3.3 years) is from a 3.5-year old where only three indicators were assessed (perhaps due to poor quality of the radiograph). In this case, the prior distribution has a large impact on the confidence limits due to the relatively small amount of information in the data.

A sensitivity analyses was conducted to check how sensitive the Bayesian SA estimate and the width of the HPD region are to the magnitude of the prior standard deviation which is a function of δ and p_δ . δ was varied from 1–10 by 0.25 and p_δ was varied from 0.09–0.99 by 0.10. Table 3 shows the resulting ranges of SA and HPD region widths for the five examples in Figure 3, both over the ranges of δ and p_δ values resulting in a more informative (less conservative) prior and over the ranges of values resulting in a less informative (more conservative) prior. A more informative prior draws the estimate closer to the prior mean (CA) while priors less informative than the analysis prior (with $\delta = 4$ and $p_\delta = 0.99$) all result in approximately the same SA estimate. Similarly, for the width of the HPD region, a less informative prior results in almost the same interval width as the analysis prior, with one exception. The width of the HPD region for Figure 3(e) becomes very small when the prior is more diffuse. This is because in this case most of the area under the likelihood is well above 18 years, so a more diffuse prior results in an HPD region with a lower bound approaching 18 years and a width approaching zero.

In summary, the Bayesian framework provides a more effective and intuitive method of estimating SA. In particular, (1) truncation at 0 and 18 are interpreted as summarizing the posterior distribution below 0 and above 18 as point masses p_0 and p_{18} corresponding to $P(SA = 0)$ and $P(SA = 18)$, respectively; (2) the posterior mode always exists, as does the HPD region, even when the likelihood is monotonic or the ML-based SE is extremely large;

(3) the HPD region is almost always narrower than the ML-based CI; and (4) a $100 \times (1 - \alpha)\%$ HPD region can be validly interpreted in the manner that is most intuitive, namely as an estimate of the age range in the reference group containing $100 \times (1 - \alpha)\%$ of the individuals with similar skeletal maturity as the individual being assessed.

Improvement #3: Impact of accounting for all sources of variation

Within the Fels Longitudinal Study sample, ignoring sampling variability understates the width of the HPD by at most less than 1 month and, in almost half of the cases, by less than 1 week. In other words, the reference group is so large that most of the variation in SA estimation comes from the variation in CA within the calibration sample and very little from uncertainty in the calibrated parameters. Arguably, the extra computation time required to carry out the bootstrap procedure outweighs the negligible gain. If one views the calibration sample as the reference population, then treating ω as known is appropriate and the width of the HPD region is not under-stated. If one wishes to interpret SA relative to some larger population from which the calibration sample was drawn, then ignoring sampling variability results in an HPD region for which the width is slightly under-stated.

Discussion

Assessment of skeletal maturation is an important clinical tool in the management of children with growth disorders, genetic syndromes and those receiving hormonal, orthopaedic or orthodontic treatments (Baccetti et al., 2006; Cox, 1997; Gilbert et al., 2004; Kaufman & Sy, 1999; Sanders et al., 2007; Sidlauskas et al., 2005). It is also an informative research measure that can be used to compare children's biological development across populations.

This article documents the statistical basis of the FELS method of skeletal age estimation and outlines three improvements. First, accounting for serial correlation in the calibration results in steeper rate parameters for ordinal indicators and, therefore, narrower confidence limits for the SA estimate. Second, inclusion of only a small amount of prior information improves the estimation procedure, in particular for cases with SA near 0 and 18 years, and results in more appropriate (and almost always narrower) confidence limits. Additionally, the Bayesian paradigm allows a more natural interpretation of a child's SA as the centre of the distribution of chronological ages in a child's skeletal maturity peer group within the reference sample and the reported confidence region as the chronological age range that contains 95% of individuals in the reference group with skeletal maturity similar to that child. Third, accounting for the fact that the reference group is a sample, not a population, results in a more appropriate estimate of the confidence limits for a SA estimate, although the previously unaccounted for variation is quite small for a reference sample the size of the Fels Longitudinal Study.

The authors are in the process of updating the FELS method calibration using a modern and racially diverse reference sample. Given current trends for earlier puberty (Chumlea et al., 2003; Sun et al., 2005) and the fact that the timing of puberty is more closely related to skeletal maturation than it is to chronological age (Simmons & Greulich, 1943), revisiting methods for assessing skeletal maturation is a particularly timely endeavour. Using the

resulting calibration (or calibrations, if there are significant differences between races), adjusted for serial correlation as described herein, an updated FELS method will be available in the future. The updated method will estimate SA and the HPD region using the Bayesian method as described in this paper. Users will have the option of varying the prior distribution, in particular by increasing the standard deviation of the prior in order to remove all influence of CA on the estimate. Such an age-independent estimate is useful, for example, when a pathology is suspected such that the assumption that CA provides prior information about SA is not valid. Users will also have the option of either using the bootstrap to obtain the best variation estimate or of minimizing computing time by ignoring the small amount of sampling variability (which would be useful, for example, when processing large batches of SA estimates).

Conclusion

The improvements to the statistical basis of the FELS method described in this article provide a more effective method of estimating skeletal maturity and a more natural interpretation of the resulting point and interval estimates. The FELS method is unique among skeletal maturity estimation procedures in providing confidence limits for its estimate of skeletal age. This feature is critical, providing vastly more information than a point estimate alone, information that is instrumental for clinicians to better evaluate the discrepancy between chronologic age and skeletal age and, thus, arrive at a more informed treatment decision. The revised FELS method is not yet publicly available; however, readers may contact the corresponding author for information regarding its future availability.

Acknowledgments

The authors thank the participants of the Fels Longitudinal Study and the data collection and management staff of the Lifespan Health Research Center. The authors also thank Thaddeus Tarpey for helpful comments that led to improvements in the manuscript.

This work was supported by National Institutes of Health grants R01AR055927 and R01HD012252.

References

- Baccetti T, Franchi L, De Toffol L, Ghiozzi B, Cozza P. The diagnostic performance of chronologic age in the assessment of skeletal maturity. *Prog Orthod*. 2006; 7:176–188. [PubMed: 17143345]
- Berger, JO. Statistical decision theory and Bayesian analysis. Springer; New York: 1985.
- Beunen GP, Malina RM, Renson R, Simons J, Ostyn M, Lefevre J. Physical activity and growth, maturation and performance: a longitudinal study. *Med Sci Sport Exer*. 1992; 24:576–585.
- Chumlea WC, Roche AF, Thissen D. The FELS method of assessing the skeletal maturity of the hand-wrist. *Am J Hum Biol*. 1989; 1:175–183.
- Chumlea WC, Schubert CM, Roche AF, Kulin HE, Lee PA, Himes JH, Sun SS. Age at menarche and racial comparisons in US girls. *Pediatrics*. 2003; 111:110–113. [PubMed: 12509562]
- Claessens AL, Lefevre J, Beunen GP, Malina RM. Maturity-associated variation in the body size and proportions of elite female gymnasts 14–17 years of age. *Eur J Pediatr*. 2006; 165:186–192. [PubMed: 16344993]
- Cox LA. The biology of bone maturation and ageing. *Acta Paediatr Suppl*. 1997; 423:107–108. [PubMed: 9401555]
- Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall/CRC; New York: 1993.

- Fitzmaurice, G.; Verbeke, G. Parametric modeling of longitudinal data: introduction and overview. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. *Longitudinal data analysis*. Chapman & Hall/CRC; Boca Raton, FL: 2008. p. 31-41.
- Gilbert SR, Gilbert AC, Henderson RC. Skeletal maturation in children with quadriplegic cerebral palsy. *J Pediatr Orthop*. 2004; 24:292–297. [PubMed: 15105725]
- Greulich, W.; Pyle, SI. Radiographic atlas of skeletal development of the hand and wrist. Stanford University Press; Stanford, CA: 1950.
- Greulich, W.; Pyle, SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd. Stanford University Press; Stanford, CA: 1959.
- Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw*. 2006; 15:1–11.
- Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. *J Amer Statist Assoc*. 1996; 91:1024–1036.
- Kaufman FR, Sy JP. Regular monitoring of bone age is useful in children treated with growth hormone. *Pediatrics*. 1999; 104:1039–1042. [PubMed: 10506261]
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Malina RM, Katzmarzyk PT, Beunen GP. Relation between birth weight at term and growth rate, skeletal age, and cortical bone at 6–11 years. *Am J Hum Biol*. 1999; 11:505–511. [PubMed: 11533970]
- Ogden RT, Tarpey T. Estimation in regression models with externally estimated parameters. *Biostatistics*. 2006; 7:115–129. [PubMed: 16020616]
- Pludowski P, Lebiedowski M, Lorenc RS. Evaluation of practical use of bone age assessments based on DXA-derived hand scans in diagnosis of skeletal status in healthy and diseased children. *J Clin Densitom*. 2005; 8:48–56. [PubMed: 15722587]
- Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*. 1991; 47:825–839. [PubMed: 1742441]
- R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2009.
- Roche, A. The measurement of skeletal maturation. In: Johnston, F.; Roche, A.; Susanne, C., editors. *Human physical growth and maturation: methodologies and factors*. Plenum; New York: 1980.
- Roche, AF. Growth, maturation and body composition: The Fels Longitudinal Study 1929-1991. Cambridge University Press; Cambridge: 1992.
- Roche, AF.; Chumlea, WC.; Thissen, D. Assessing the skeletal maturity of the hand-wrist: FELS Method. Charles C. Thomas; Springfield, IL: 1988.
- Roche AF, Wainer H, Thissen D. The RWT Method for the prediction of adult stature. *Pediatrics*. 1975a; 56:1026–1033.
- Roche, AF.; Wainer, H.; Thissen, D. Skeletal maturity: the knee joint as a biological indicator. Plenum; New York: 1975b.
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr*. 1969; 17:1–100.
- Sanders JO, Browne RH, McConnell SJ, Margraf SA, Cooney TE, Finegold DN. Maturity assessment and curve progression in girls with idiopathic scoliosis. *J Bone Joint Surg Am*. 2007; 89:64–73. [PubMed: 17200312]
- Sidlauskas A, Zilinskaite L, Svalkauskiene V. Mandibular pubertal growth spurt prediction. Part one: method based on the hand-wrist radiographs. *Stomatologija*. 2005; 7:16–20. [PubMed: 16254472]
- Simmons K, Greulich WW. Menarcheal age and the height, weight, and skeletal age of girls age 7 to 17 years. *J Pediatr*. 1943; 22:518–548.
- Spadoni GL, Cianfarani S. Bone age assessment in the workup of children with endocrine disorders. *Horm Res Paediatr*. 2010; 73:2–5. [PubMed: 20190534]
- Sun SS, Schubert CM, Liang R, Roche AF, Kulin HE, Lee PA, Himes JH, Chumlea WC. Is sexual maturity occurring earlier among U.S. children? *J Adolesc Health*. 2005; 37:345–355. [PubMed: 16227118]

- Tanner, JM.; Healy, MJR.; Goldstein, H.; Cameron, N. Assessment of skeletal maturity and prediction of adult height (TW3 Method). W. B. Saunders; London: 2001.
- Tanner, JM.; Whitehouse, RH. Part I. Paris: International Children's Centre. 1959. Standards for skeletal maturity.
- Tanner, JM.; Whitehouse, RH.; Cameron, N.; Marshall, WA.; Healy, MJR.; Goldstein, H. Academic Press; London: 1975a. Assessment of skeletal maturity and the prediction of adult height (TW2 Method).
- Tanner JM, Whitehouse RH, Healy MJR. A new system for estimating skeletal maturity from the hand wrist, with standards derived from a study of 2,600 healthy British children. Part II. Paris: International Children's Centre. 1962
- Tanner JM, Whitehouse RH, Marshall WA, Carter BS. Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. Arch Dis Child. 1975b; 50:14–26. [PubMed: 164838]
- Venables, WN.; Ripley, BD. Modern applied statistics with S. 4th. Springer; New York: 2002.
- Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–130. [PubMed: 3719049]

Figure 1.

Example of probability functions for two indicators (left panel) and their joint probability, the likelihood function (right panel), which is maximized at age 10.6 years (at the vertical dashed line).

Figure 2.

Comparison of calibration assuming independence vs calibration adjusting for serial correlation. The diagonal dashed line is the 45° line. Units for each parameter are given in parentheses (y = years).

Figure 3.

ML vs Bayesian estimation of SA for examples where the MLE is (a) 0, (b) near 0, (c) far from 0 and 18, (d) near 18 and (e) 18. The dashed curves are the normalized likelihoods and the vertical dashed lines represent the end-points of the truncated ML-based 95% CI. The solid curves are the posterior distributions and the vertical solid lines represent the boundaries of the 95% HPD region. If a CI end-point is identical to an HPD region end-point, then the vertical dashed line is hidden. See Table 1 for more information.

Figure 4.

Difference in the width of the confidence regions (truncated ML-based 95% CI vs 95% HPD region) vs chronological age for assessments of 19 487 radiographs from the Fels Longitudinal Study. Points in grey are those where the CI is narrower than the HPD region. See Table 2 for more information.

Table 1

ML versus Bayesian estimates of skeletal age for the five examples in Figure 3.

Figure	CA	Maximum likelihood		Bayesian			
		MLE	95% CI	Mode	95% HPD region	P(SA = 0)	P(SA = 18)
3a	0.76	0.00	(0.00, 0.43)	0.01	(0.00, 0.36)	0.64	0.00
3b	0.50	0.08	(0.00, 0.45)	0.08	(0.00, 0.45)	0.49	0.00
3c	9.51	8.89	(8.37, 9.41)	8.91	(8.40, 9.40)	0.00	0.00
3d	16.00	17.85	(16.28, 18.00)	17.50	(16.31, 18.00)	0.00	0.27
3e	17.00	18.00	–	18.00	(17.05, 18.00)	0.00	0.80

Table 2

Comparison of widths of the truncated 95% ML-based CI and the 95% HPD region for SA assessed on Fels Longitudinal Study participants using the existing FELS Method calibration (width = CI width – HPD region width, rounded to nearest 0.1 year).

Sex	Age range, years	<i>n</i>	CI undefined	width > 0 (HPD better)	width < 0 (CI better)	width = 0 (HPD ≈ CI)	% width 0 or CI undefined (HPD at least as good as CI)
Boys	0–4	2 784	228	2 088	330	138	88.1%
	>4–10	3 352	0	3 352	0	0	100.0%
	>10–14	2 063	0	2 063	0	0	100.0%
	>14–22	2 042	451	1 485	75	31	96.3%
	Total	10 241	679	8 988	405	169	96.0%
Girls	0–4	2 478	151	1 866	298	163	88.0%
	>4–10	3 236	0	3 236	0	0	100.0%
	>10–14	1 911	1	1 907	1	2	99.9%
	>14–22	1 621	581	980	46	14	97.2%
	Total	9 246	733	7 989	345	179	96.3%
Total		19 487	1412	16 977	750	348	96.2%

Table 3

Sensitivity to prior standard deviation for the five examples in Figure 3. The Bayesian SA and 95% HPD region were based on a prior with $\delta = 4$ and $p\delta = 0.99$. The ranges of results below are the minimum and maximum values after varying δ from 1–10 by 0.25 and $p\delta$ from 0.09–0.99 by 0.10.

Figure	CA (years)	Bayesian SA (years)	Range for more informative priors	Range for less informative priors	Width of 95% HPD region	Range for more informative priors	Range for less informative priors
3a	0.76	0.01	0.01–0.14	0.00–0.01	0.36	0.36–0.44	0.36–0.37
3b	0.50	0.08	0.08–0.15	0.07–0.08	0.45	0.45–0.47	0.45–0.45
3c	9.51	8.91	8.91–9.09	8.90–8.91	1.00	0.84–1.00	1.00–1.01
3d	16.00	17.50	16.46–17.48	17.50–17.80	1.69	1.26–1.93	1.59–1.70
3e	17.00	18.00	17.35–18.00	18.00–18.00	0.95	0.95–1.29	0.02–0.94