

# Automated determination of bone age from hand X-rays at the end of puberty and its applicability for age estimation

Hans Henrik Thodberg<sup>1</sup> · Rick R. van Rijn<sup>2</sup> · Oskar G. Jenni<sup>3</sup> · David. D. Martin<sup>4</sup>

Received: 31 March 2016 / Accepted: 6 October 2016 / Published online: 18 October 2016  
© Springer-Verlag Berlin Heidelberg 2016

**Abstract** The BoneXpert method for automated determination of bone age from hand X-rays was introduced in 2009, covering the Greulich–Pyle bone age ranges up to 17 years for boys and 15 years for girls. This paper presents an extension of the method up to bone age 19 years for boys and 18 years for girls. The extension was developed based on images from the First Zurich Longitudinal Study of 231 healthy children born in 1954–1956 and followed with annual X-rays of both hands until adulthood. The method was validated on two cross-sectional studies of healthy children from Rotterdam and Los Angeles. We found root mean square deviations from manual rating of 0.69 and 0.45 years in these two studies for boys in the bone age range 17–19 years. For girls, the deviations were 0.75 and 0.59 years, respectively, in the bone age range 15–18 years. It is shown how the automated bone age method can be applied to infer the age probability distribution for healthy Caucasian European males. Considering a population with age 15.0–21.0 years, the method can be used to decide whether the subject is above 18 years with a false positive rate (children classified as adults) of 10 % (95% confidence interval = 7–13%) and a false negative rate of 30 % (adults classified as children). To apply this method in other ethnicities will require a study of the average of “bone age –

age” at the end of puberty, i.e. how much this population is shifted relative to the Greulich–Pyle standard.

**Keywords** Bone age · Age assessment · Bayes · Hand radiographs · Asylum seekers

## Introduction

Bone age rating from hand X-rays is associated with a considerable inter- and intrarater variability that limits its usefulness. To remove this, a fully automated method for bone age determination was introduced in 2009 [1, 2] for the bone age ranges 2.5–17 years for boys and 2–15 years for girls. This is widely used in clinical practice in Northern Europe, mainly in relation to paediatric endocrinology. To date, it has not been used in forensic medicine.

This paper presents an extension of the automated method up to bone age 19 years for boys and 18 years for girls. The method was developed on longitudinal data and validated on two cross-sectional studies against manual rating according to the Greulich–Pyle (GP) method [3].

In the discussion, we show how this can be used to derive the probability distribution of age for a given observed bone age. Finally, we present performance measures (sensitivity, specificity, etc.) for the use of this method to determine whether a male is above 18 years old.

Readers who are interested only in age assessment may wonder why the authors do not develop a method that predicts age directly from the image, without going via bone age. There are, however, many reasons for using bone age as a stepping stone.

- Firstly, this is the workflow that has been used for age assessment by manual raters for decades.

✉ Hans Henrik Thodberg  
thodberg@visiana.com

<sup>1</sup> Visiana, Søllerødvej 57C, 2840 Holte, Denmark

<sup>2</sup> Academic Medical Center, Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

<sup>3</sup> Child Development Center, University Children’s Hospital Zürich, Steinwiesstrasse 75, 8032 Zürich, Switzerland

<sup>4</sup> Tübingen University Children’s Hospital, Hoppe-Seyler-Strasse 1, 72076 Tübingen, and Filderklinik, Im Haberschlag 7, 70794 Filderstadt, Germany

- Secondly, because bone age rating from hand X-rays is so common, there is an excellent working knowledge for doing this, and this can generate accurate “target values” for a machine learning method, which can thereby learn from the experts.
- Thirdly, using bone age rather than age as target for a regression from the image means that the residual error is smaller, around 0.6 years, whereas the residual for age prediction is about 1.0 years. Thus, the teacher is much “sharper”, and this helps the machine learning method to extract the salient features. This is the most important reason.
- Fourthly, using bone age as target allows pooling of studies from populations reaching full maturity at different ages.
- Finally, bone age assessment at the end of puberty has its own clinical value, e.g. in paediatric orthopaedics.

## Material and methods

### Material

This study was based on hand X-rays from four data sets, which had all been used in previous publications, but here, we exploited the higher end of the bone age more extensively.

1. The First Zurich Longitudinal Study (1ZLS) of 231 healthy children with both left- and right-hand X-rays taken at every anniversary<sup>1</sup> until the age of 20—for late maturers, even one or more years longer. These 119 boys and 112 girls were born in 1954–1956, and the X-rays were rated according to the GP method at the time of the study. These data were used previously to establish a model for adult height prediction [4, 5]. Two thousand eight images were used from this study.
2. The Björk longitudinal study of 162 healthy children enrolled for orthodontic treatment at the Royal Dental College in Copenhagen [6]. The subjects were born on average in 1952, and X-rays of the non-dominant hand were taken annually, typically from age 7 to 21 years. For each subject, the first author identified the first image reaching bone age 18 years for boys and bone age 16 years for girls. Five hundred forty-seven images from this study were used.
3. The Erasmus cross-sectional study of 542 normal children from Rotterdam imaged in 1997. Each image was rated by one of two paediatric radiologists [7].

4. The Los Angeles (LA) cross-sectional study of 1103 normal children of four ethnicities. Each image was rated independently by two paediatric radiologists [8].

The first two studies were used for development and the last two for validation.

### Method

The BoneXpert method for automated determination of bone age is a CE-marked medical device for routine clinical use [9], intended to be able to replace the manual rating altogether. The method locates 13 bones in the hand: radius, ulna and the 11 short bones in rays 1, 3 and 5—the so-called RUS bones. It determines the GP bone age in each bone, and the final bone age is formed as the average.

The BoneXpert method was originally developed using the principle (due to Tanner) [10] that the bone ages of the 13 RUS bones should agree amongst each other, and this was used in the development of the method based on cross-sectional data [1] by training each bone to predict the average bone age of all the other bones. In this work, we were interested in the bone age ranges above 17 years for boys and 15 years for girls, the so-called *end-of-puberty range*, and here, this principle can no longer be applied because the maturation of the short bones has finished; only the ulna and radius display signs of continued maturation. Tanner therefore refrained from assessing bone age in the end-of-puberty range, so the Tanner–Whitehouse 3 bone age scale, which agrees, on average, well with the GP scale before the end of puberty, stops at bone ages 16.5 years for boys and 15 years for girls. The GP bone age scale, however, continues up to bone ages 19 years for boys and 18 years for girls.

To extend BoneXpert, we employed the 1ZLS and Björk study in the following manner: For each boy, we selected the visit, where GP 18 was reached, as the *anchor* of that subject. For girls, the anchor was chosen as the visit where GP 16 was reached, which has the same appearance as GP 18 in boys. This stage of maturity is described as “fusion of the epiphysis has begun”—a relatively clear maturity indicator because fusing of the epiphysis occurs over a short span of time, so it is relatively easy for a rater to identify the visit that reaches this stage. We pooled males and females by treating the females as males with GP 18 at the anchor visit.

Males and females mature by going through the same sequence of visual appearances, but shifted in age by, on average, about 2 years, and this average shift is constant over a rather wide age range, at least from the age of 11 years of boys. The GP method ends at 19 for males, so one would expect the female scale to end at 17; however, it ends at 18 for females, so it would not be surprising if the progression from 17 to 18 in females is determined less reliably. However, rating females up to bone age 18 years

<sup>1</sup> Ninety-four per cent were taken within 14 days of the anniversary; 99% within a month.

has become a convention, so the new automated system adheres to this. The automated system yields a smoothly progressing bone age, and for numerical reasons, it is designed to end not strictly at 19 for males and 18 for females, but just below 19.3 and 18.3, respectively.

This idea of pooling males and females is originally due to Tanner [10], whose scheme for manual rating employs nine stages of bone appearance common to males and females. Human raters thus have to learn only one rating system, not one for each gender. Likewise, our machine learning system only had to learn one system, namely the bone age of the image assuming that it is a male. This has the benefit of doubling the amount of training data for the method. A relatively simple formula is then used for the females to transform the computed male bone age to the female scale.

We picked the images taken 1 and 2 years before the anchor and assigned them nominal bone ages 17 and 16 years. Likewise, the visits 1 and 2 years after the anchor were defined as having nominal bone ages 19 and 20 years. Figure 1 shows examples of these training data for two subjects.

The computation of bone age from the images was implemented using random forests [11], suitable for interpreting data with many input variables. We used a technique similar to [12], where features are formed from average image intensities in rectangles placed in arbitrary locations across the bone image.

The random forest has 160 decision trees (used as regression trees), and each tree was trained on a subset of the *subjects*, each subject having a left- and, for the 1ZLS, also a right-hand series of five images centred at the anchor visit. This process of introducing randomness, called bagging, allows for an elegant way to use the training data also for validation through out-of-bag cross-validation [11]. To exploit this, we trained 500 trees, and for each subject, we found 160 trees, which did not use this subject for training, and these were used to form a random forest to cross-validate the model on this subject. This was done for each subject in turn. We exploited this in the discussion section to form distributions of bone ages observed at a given age.

The new method requires that at least 2 cm of radius is included in the image. Figure 2 is an example where there is just about enough included.

A critical element of the new method is the localisation of the radius and ulna. These bones are more difficult to delineate than the short bones because their shapes and pose vary more. In particular, the ulna can be rotated around its axis; the ideal rotation presents the tip of ulna on the left side, but it can also be in the middle (as in Fig. 2). Also, the amount of profusion of the tip of ulna varies. Finally, there can be some overlap of the radius and ulna—preferably this should be avoided.

The new method automatically determines whether the bones have been located with sufficient reliability for a bone age assessment. If the radius is not found reliably, and the

bone age assessment is above 17 years, it is reported as “unreliable”. The ulna, however, is allowed to be missing.

In the bone age range 2.5–15 years for boys, the bone age is determined as a simple average over the 13 bones, i.e. the bones have equal weight. When bone age becomes larger than 15 years, the radius and ulna are assigned a progressively larger weight, and at the end of the bone age scale, they have all the weight. The relative importance of the radius and ulna is initially 2:1, but as bone age progresses above 18, the ulna loses its contribution, so from 18 to 19, the bone age is almost exclusively determined by the radius. This is similar to what a manual rater does. For girls, the same rules apply, but shifted by 2 years.

## Results

We validated the new method on two studies, which had been used previously to validate the method in the bone age ranges up to 17 years for boys and 15 years for girls.

For images at the end of puberty, the radius was found reliably in 97 % of the images in the Erasmus study and in 87 % in the LA data. Most of the rejections were due to an insufficient amount of radius included in the exposed image, namely 2 % of the Erasmus images and 11 % of the LA images.

The comparison of the manual and automated ratings for the Erasmus study is shown in Fig. 3. Figure 4 compares the automated and manual ratings for the LA study. As “manual rating”, we used the average of the two manual ratings.

As usual in such comparisons, it is instructive to inspect the images with the highest deviation; in this case, we chose to define such images (arbitrarily) as those with a deviation more than 1.6 years, as indicated by the horizontal lines. We have also drawn vertical lines to delimit the end-of-puberty range.

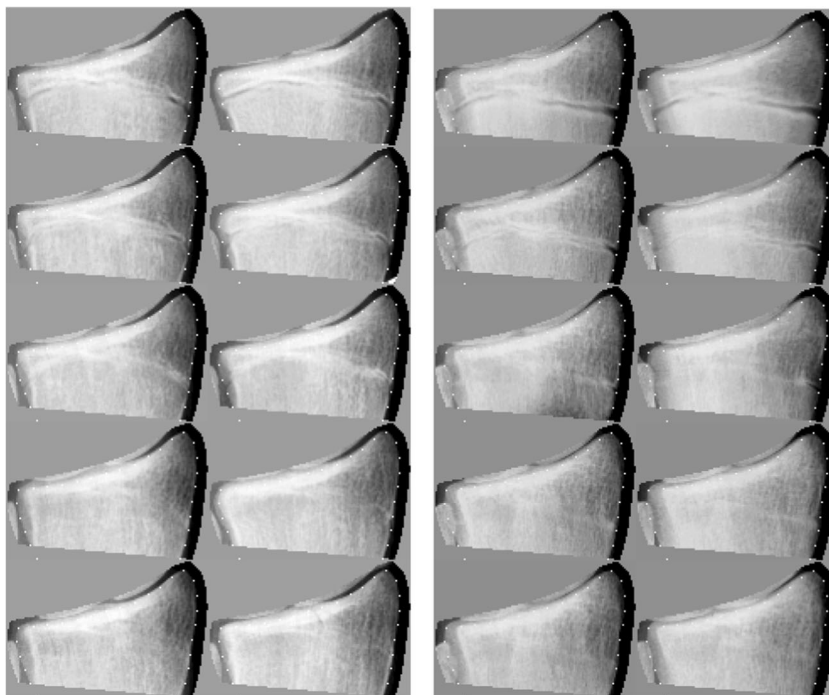
The cases at the end of puberty, where the deviation was more than 1.6 years, are encircled. The authors rerated these images to decide whether the original manual rating or the automated rating was most correct, and this showed that two had wrong automated rating, whilst three had an error in the manual rating, i.e. no method was significantly better than the other in these disputed cases, but interestingly, they were all females.

The root mean square deviations between the bone age determinations in the bone age ranges above 17 years for boys and 15 years for girls are summarised in Table 1, which also includes the root mean square deviations between the two manual LA raters in this bone age range; the latter was remarkably large for males.

For the LA study, there were 96 males in the end-of-puberty range, and the 95 % confidence interval for the root mean square error is 0.39–0.52 years.

The conclusion of this validation is that the automated method performed as well at the end of puberty as in the rest of the bone age range, except for females in the bone age range

**Fig. 1** The *left half* shows images of the distal radius from the same boy, his left hand to the left and his right hand (mirrored) to the right. The five rows correspond to five subsequent anniversaries selected such that the manual GP bone age is 18 in the *third row*. Thus, the rows correspond roughly to GP bone ages 16, 17, 18, 19 and “20” years. The *right half* shows images from another boy. The images have been warped to the average shape of the radius

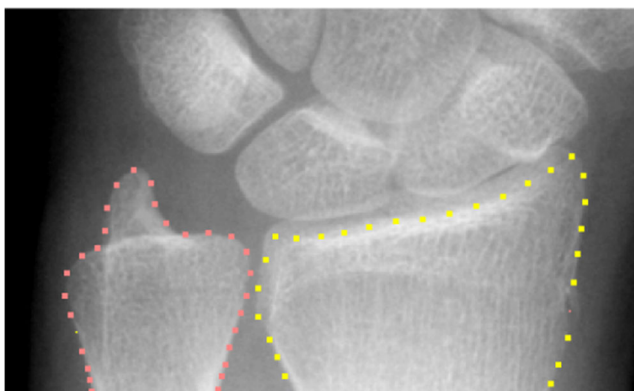


17–18 years. The accuracy of the automated method for males in the LA study is significantly better than the agreement between the two manual raters.

## Discussion

### Bone age assessment

In the LA study, 11 % of the images were rejected because the amount of radius included in the exposed image was insufficient. At least 2 cm is required, so this can be avoided in the future by specifying a protocol for bone age exams that includes at least 2 cm of radius, or 3 cm to be safe. Then the rejection rate will be 2 % or less for normal subjects, but it could be higher for clinical patients, which can show



**Fig. 2** In this example, just about enough of radius is included for the analysis to succeed

deformities in the wrist, e.g. Madelung deformities. It can be considered acceptable that such images are rejected, as they should be reviewed by a radiologist, because such deformations are likely to interfere with bone maturation.

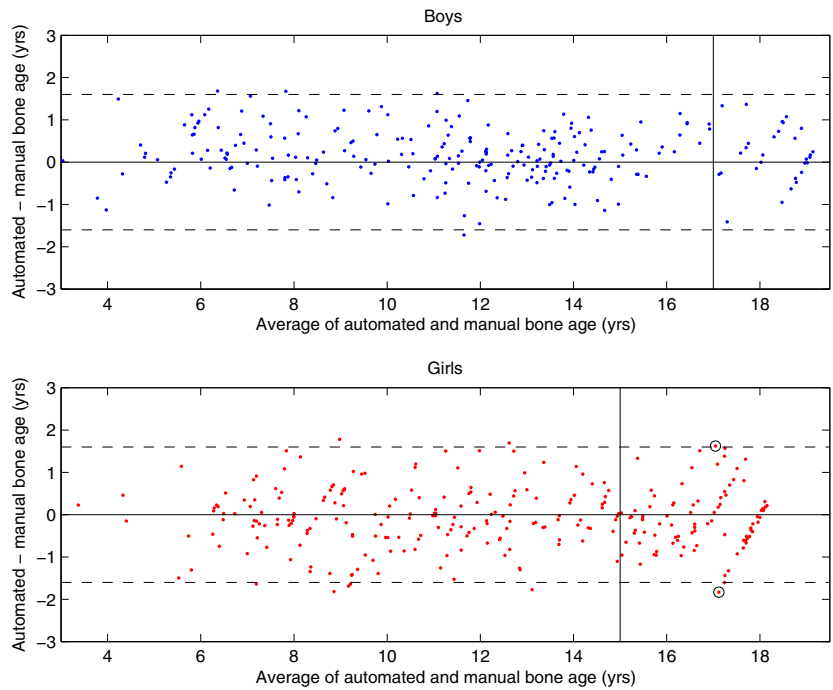
The agreement with manual rating was better in males than in females, judging from Figs. 3 and 4. For females, the agreement was particularly poor above 17 years of bone age. We interpret this as support for the view that the GP bone age scale, which ends at 19 years for males, should perhaps have ended at 17 years for females, rather than at 18, as suggested from the 2-year offset of maturation in males and females.

### Challenges in age assessment

In forensic [13] and sports medicine [14], bone age is used as input to an assessment of chronological age. A particularly important application is to determine whether a male is above 18 years. To discuss this in detail, we introduce the abbreviations BA for bone age and CA for chronological age. There are three challenges with this usage.

1. Manual BA assessment is associated with considerable rater variability.
2. The median of BAs observed for subjects of a given CA is only equal to the CA in the population originally used to set up the GP scale; other Caucasian populations have typically been found to have a median BA lower than CA. For modern European Caucasians, the median BA is typically 0.2–0.4 years below the CA, whilst for other ethnicities this population bias can be larger, and one

**Fig. 3** Bland–Altman plot showing the agreement between the automated and manual ratings in the Erasmus study. The vertical lines delimit the end-of-puberty range and the circles indicate the disputed cases in this range

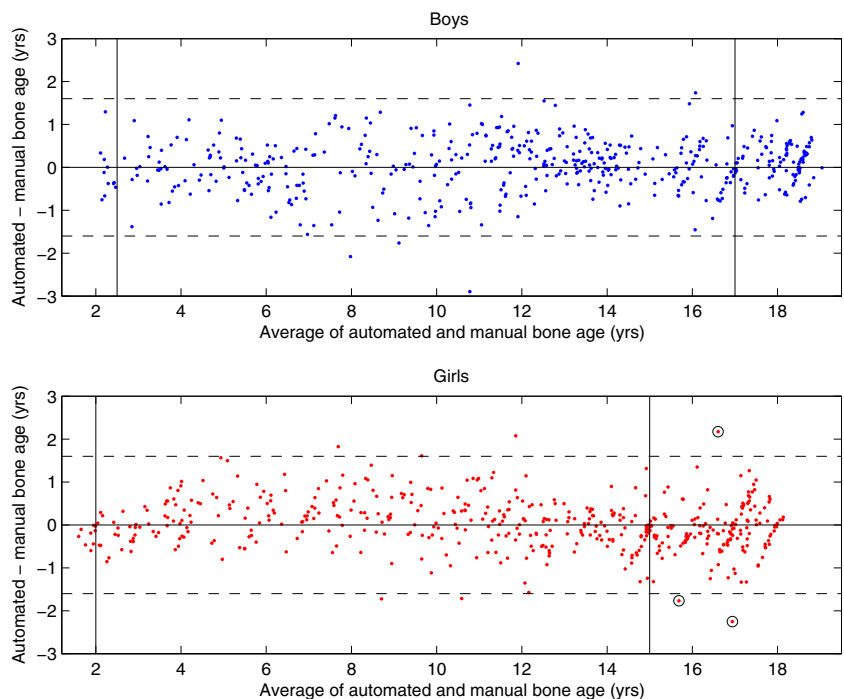


should always take this into account in age assessment based on BA.

3. The distribution of bone ages observed for subjects from a given population and with a given age has a SD of typically 1 year. This implies that when BA is used as a “clock”, it is not perfect—the errors in its timing across a population exhibits a SD of typically 1 year.

Can automated BA determination mitigate these challenges? As for the first challenge, automated rating eliminates the rater variation completely. The only variability left is a precision error associated with the physical measurement, defined as the SD of BAs obtained when repeated X-ray images are made and analysed with the software. The SD of this error is 0.18 years in the BA

**Fig. 4** Bland–Altman plot showing the agreement between the automated and manual bone age ratings in the Los Angeles study. The manual rating is the average of the two manual ratings



**Table 1** Root mean square deviations between bone age determinations (in years), each computed from  $N$  subjects

	Males		Females		Both sexes RMS error
	$N$	RMS error	$N$	RMS error	
Erasmus: manual – automated	29	0.69	89	0.75	0.73
LA: AverageManual – automated	96	0.45	157	0.59	0.54
LA: Manual1 – Manual2	96	0.78	157	0.52	0.62

range 2.5–17 years for boys [15], much smaller than the manual rater variability of typically 0.58 years [16].

The second challenge can also be addressed efficiently with the automated method because reference curves for average BA – CA versus CA are being collected across the world. So far, curves for six populations have been presented [7, 8, 17]. In this discussion, we will assume BA – CA = –0.3 years, as found in 1ZLS. The 1ZLS was followed up by the Zurich Generation Study of children with one parent in the 1ZLS. This study showed no secular trend in BA – CA from 1ZLS; in other words, although the 1ZLS is a rather old study, these children are compatible with modern children. In the Erasmus study, we also observed an average BA – CA of approximately –0.3 years, so it appears that BA – CA = –0.3 years is our best assumption for the BA offset relative to the GP standard in present-day Caucasian children in Europe.

The third challenge is *not* alleviated by an automated method because it relates to the “imperfection” of bone maturation in an individual when used as a clock. It is a *biological* limitation, so any age assessment method based on bone maturity will have a SD error contribution from this cause of approximately 1 year. So whilst the error of BA determination has been reduced to SD 0.18 years by the automated method, age assessment (above age 7) through bone age can never obtain a SD lower than about 1 year.

There are two additional challenges in age assessment, related to how the results are presented.

4. The age assessment is conventionally communicated as a centre value age and a “confidence interval” with poor rational justification, and not easily understood by the authorities.
5. Performance measures are not well defined and standardised.

In the next section, we provide a more satisfactory solution to the last two challenges.

### Inferring the age distribution from a bone age determination

What we can observe in studies is the distribution of BAs at a given age. In age assessment, we want to turn this around and

obtain the probability distribution of age corresponding to a given observed BA [18]. This turning-around is just another day at the office for a statistician because it is an application of Bayes’ theorem. But for people not trained in statistics, this can be difficult to grasp, so in the following we will perform this inference in a graphical and intuitive manner so that also non-statisticians can appreciate its validity.

To start with, the observed automated bone ages at five anniversaries are shown for the 1ZLS in Fig. 5—these are out-of-bag cross-validated results. It is practical to parameterise these distributions as Gaussians, and we have made this possible by transforming the observed BA into a modified BA\*, which stretches the upper end of the BA scale, so that BA\* extends to 20 years, whereas BA extends only to 19.3 years. BA\* is defined as

$$\begin{aligned} \text{BA}^* &= \text{BA} && \text{for BA} < 18.7 \\ \text{BA}^* &= 18.7 + 5/3 * (\text{BA} - 18.7) && \text{for BA } 18.7\text{--}19.0, \\ \text{BA}^* &= 19.2 + 8/3 * (\text{BA} - 19.0) && \text{for BA} > 19.0. \end{aligned}$$

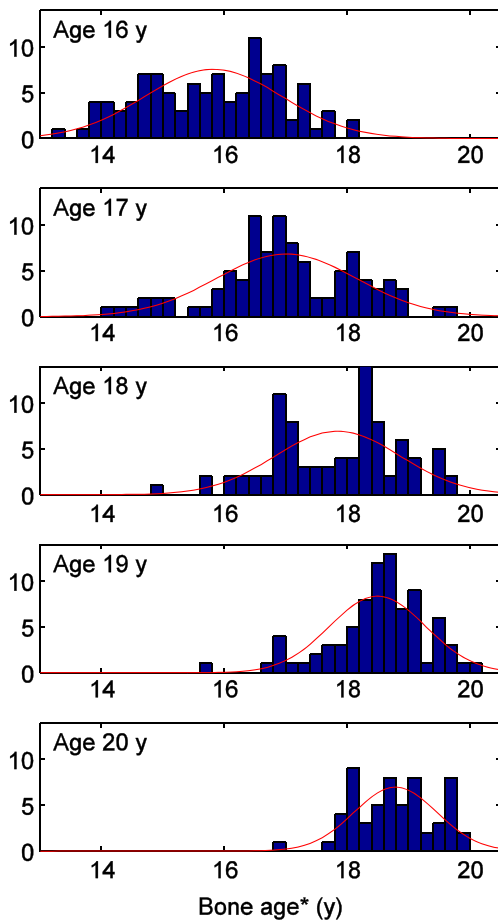
Figure 5 shows BA\*. This trick eliminates the piling up of values near 19. The observed means and SDs of BA\* are given in Table 2.

The next step in the Bayesian approach is to select a *prior distribution of ages*, and here it is customary to use a flat distribution within reasonable bounds. This describes our knowledge of age before the test, and the flat distribution is chosen in order not to bias the inference—we want the data (i.e. the X-ray image) to speak for themselves.

In Fig. 6, we then generate a population of 20,000 subjects uniformly in the selected age range, and for each subject, we sample a BA randomly from a Gaussian with mean and SD pertaining to that age, obtained by interpolation in Table 2.<sup>2</sup> We do not have the BA distribution at age 21, so for ages 20–21 years we use extrapolation, which seems justified.

If we now observe a BA for a new subject, we use the simulated population to generate the corresponding age distribution by sampling the density of points along a horizontal line at that BA. We normalise this density to sum to 1 so that it

<sup>2</sup> We actually sample BA\* values, which we then transform to BA values in Fig. 6.



**Fig. 5** The observed bone ages at five different anniversaries for the males in the First Zurich Longitudinal Study. As described in the text, the bone age scale has been stretched above 18.7 years to render the distributions compatible with Gaussians, shown *superimposed*

becomes a probability distribution that we can interpret as our *belief* in various ages after the measurement. Figure 7 shows these so-called *posterior* age distributions for four different bone ages.<sup>3</sup>

For an observed BA of 17 years, we see from Fig. 7 that the age distribution falls to zero at either end and is well described as a Gaussian and the distribution has mean 17.3 years and SD 1.02, so in this case we can report the age assessment by the mean and SD of the posterior distribution. We can even understand these values intuitively: the mean age is 0.3 years higher than BA because this population is shifted 0.3 years relative to the GP scale, and the 1.02 years is similar to the SDs in Table 2.

In general, it is good practice to provide the result of the age assessment as the posterior age distribution. This indicates directly what we can know about the age of this person, and the graphical representation as a bar plot aids the user to assess the weights of probabilities.

<sup>3</sup> To generate these, we actually sampled 2 million subjects and measured the density in a band covering ±0.1 years around BA.

**Table 2** For each age, this table gives the means and SDs of the bone age distributions in Fig. 5 (in years)

Age	Mean BA*	SD BA*
15	14.69	0.99
16	15.82	1.12
17	17.01	1.15
18	17.85	1.02
19	18.50	0.78
20	18.79	0.67

BA\* is the “stretched” bone age; see text

We saw that the bounds of the prior distribution do not matter for inferring age at BAs 17 and 17.5 years. But at BA 18 years and even more at 18.5, the posterior age distribution is truncated abruptly at the upper bound. Although the user will understand that the distribution could be extended beyond age 21, the choice of bounds matters for the normalisation to a sum of 1. So if one wants to compute the probability that the age is larger than 18 years, the bounds matter at BA 18 years and above whilst they are almost irrelevant at BAs 17 and 17.5 years. To compute such probabilities, one *must* adopt a standard for the prior age intervals—this is unavoidable.

It is therefore relevant to argue in more detail for our choice of an interval ±3 years around the age being tested for. The argument is that those being tested will in practice have an age not too far from 18 years, and it is reasonable to assume that when we get down to 15 years, only about half of the subjects would be sent for such a test. Likewise, around age 21, it would start to be clear from the physical appearance that this person is above 18. So a prior distribution over effectively 6 years seems reasonable, and the choice of a flat distribution that falls off abruptly at the bounds is appealing by its simplicity—the flatness ensures that the shape of the posterior age distribution is not affected by the prior.

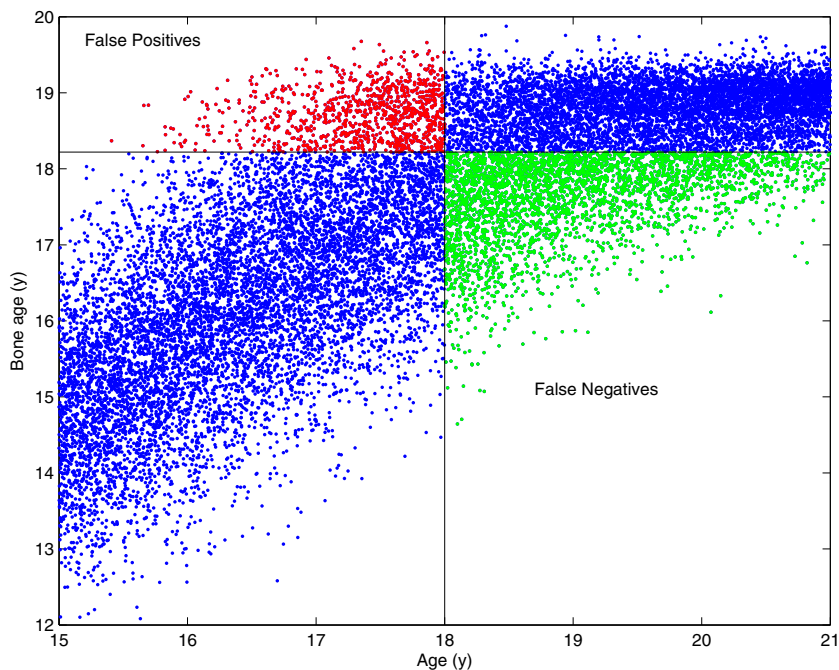
We conclude this graphical tour of Bayesian age estimation from BA by relating it to Bayes’ theorem.

$$P(CA|BA) = P(BA|CA) P(CA) / P(BA)$$

The components are:

- $P(BA|CA)$  is the BA distribution at a given CA; examples are shown in Fig. 5.
- $P(CA)$  is the prior distribution of age, a uniform age distribution with reasonable upper and lower bounds.
- $P(CA|BA)$  is the posterior distribution of age for a given BA; examples are shown in Fig. 7.
- $P(BA)$  is a mere normalisation factor, ensuring that  $P(CA|BA)$  summed over all CAs yields 1.

**Fig. 6** Monte Carlo simulation of 20,000 males with age uniformly distributed in the range 15–21 years. The bone ages are generated according to the curves in Fig. 5. The horizontal line represents a threshold at bone age 18.2 years. When this is used to classify the subjects into children and adults, there are two types of errors: false positives, i.e. children classified as adults, and false negatives, i.e. adults classified as children



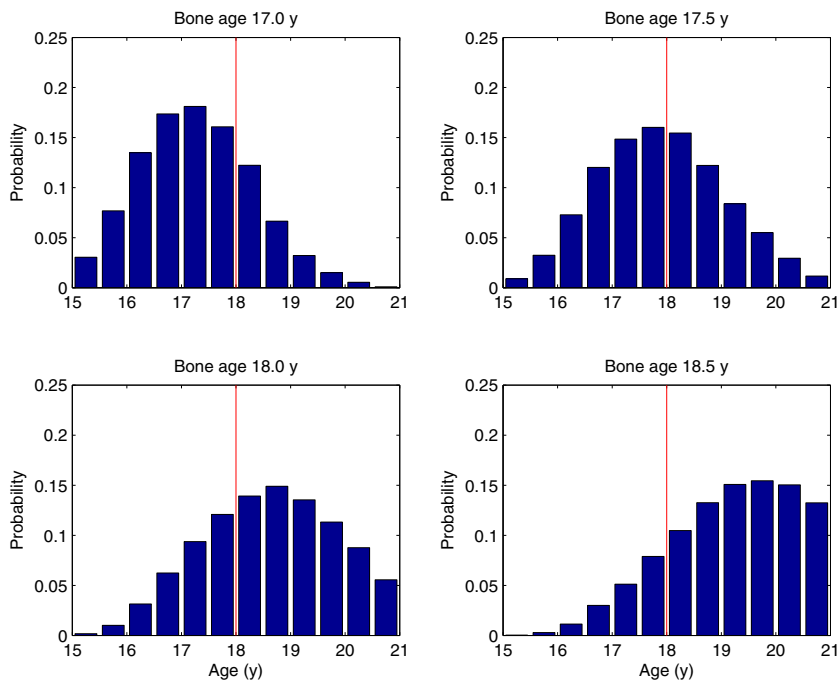
**Performance measures**

Age assessment in forensics or sports medicine is used to make a decision whether a person is above a certain age; typically whether a person is an adult, i.e. above 18 years. This means that one must settle on a certain BA threshold, above which the subject is classified as an adult, and when doing this, there is a trade-off between two types of errors illustrated in Fig. 6.

- False positives: children classified as adults
- False negatives: adults classified as children

If one wants to minimise the total number of misclassifications, the BA threshold should be set at 17.8 years. This yields 17 % false positives and 17 % false negatives, meaning that 17 % of the children are falsely classified as adults and 17 % of adults are classified as children. The total error rate is also 17 %.

**Fig. 7** Posterior probability distributions of age corresponding to observed bone ages 17, 17.5, 18.0 and 18.5 years for males. This is our recommended format for reporting the result of an age assessment based on bone age





Often there are different costs associated with making the two types of errors. As an example, we can decide that false positives are three times more expensive than false negatives, and the optimal BA threshold is then 18.2 years. Then there are 10 % false positives and 30 % false negatives, and the total error rate is now 20 %.

Using a BA threshold of 18.5 reduces the false positives to 7 %, and with a threshold of 19.0, they drop to 2 %. It is not possible to find a threshold that yields 0 % false positives, except by deeming all subjects to be children.

When computing these error rates, the prior distribution of age is essential. For instance, lowering the lower bound from 15 to 9 years would reduce the total error rate by a factor of 2, a clearly unreasonable way to “improve” the performance of the classifier. We believe that our choice of the prior age interval 15–21 years is a reasonable representation of the group of persons subjected to this age test.

These performance measures are based on the 119 male subjects in the 1ZLS, and as such, they are associated with an uncertainty from the limited size of the data set. We have estimated the 95 % confidence interval on the false positive rate to be 7–13 % for the situation where we keep the false negative rate fixed at 30 %. The computation was done using resampling methods [19], where the analysis is repeated by sampling 119 males with replacement from the 1ZLS data.

The performance of this method for detecting adults is summarised here.

- False positive rate: 10 %—the fraction of children classified as adults
- False negative rate: 30 %—the fraction of adults classified as children
- Sensitivity: 70 %—the fraction of adults classified as adults
- Specificity: 90 %—the fraction of children classified as children
- Positive predictive value: 87 %—the fraction of those classified as adults, which are indeed adults
- Negative predictive value: 75 %—the fraction of those classified as children, which are indeed children
- Accuracy: 80 %—the percentage of correctly classified subjects

Finally, we have computed the performance for detecting whether the age is above 15 years. Optimising for best accuracy, we find an accuracy of 86 % for females and 88 % for males. Again, we assume a prior age interval of width 6 years centred at the age in question.

To apply this method of age assessment to populations other than the European Caucasians, one will need to perform a study of automated BAs of healthy subjects from that population in order to derive the average BA – CA at the end of puberty. This was found to be –0.3 years for the Europeans,

and if this is found to be, for instance, –0.5 years in the new population, the inferred age distributions in Fig. 7 should be shifted 0.2 years *upwards*. One will also need to shift the age axis in Fig. 6, reconsider the BA threshold and evaluate the percentages of errors committed by the method.

## Conclusion

We have presented an extension of the automated determination of bone age to the end of puberty. The validation of the method in two studies showed good agreement with manual rating, with root mean square errors of 0.73 years in the Erasmus study and 0.54 years in the LA study for the two sexes combined. The smaller deviations in the latter can be understood as due to a more precise manual rating, defined as the average of two independent ratings.

The Bland–Altman plots showed no particular increase of deviations at 17–19 years for males and at 15–17 years for females compared to the deviations at lower bone ages. But for females in the bone age range 17–18 years, the deviations tended to be larger, which suggests that the GP scale extends a year too far for the females.

The method was able to analyse 98 % of images with at least 3 cm of radius included.

The coverage of ages up to 20 years gives a reliable foundation for a Bayesian inference of the age probability distribution corresponding to a given observed bone age, and the result of the age assessment is presented as the entire probability distribution.

For a population of European Caucasian males uniformly distributed in the age interval 15–21 years, the automated bone age method can decide whether a subject is above 18 years with an overall error rate of 20 %, and with 10 % (95% CI = 7–13%) of the children falsely classified as adults when operating at the point where 30 % of the adults are falsely classified as children.

## Compliance with ethical standards

**Conflict of interest** HHT is the owner of Visiana, which develops and markets the BoneXpert medical device for bone age assessment. The other authors have nothing to declare.

## References

1. Thodberg HH, Kreiborg S, Juul A, Pedersen KD (2009) The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging* 28:52–66
2. Thodberg HH (2009) Clinical review: an automated method for determination of bone age. *J Clin Endocrinol Metab* 94:2239–2244

3. Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of the hand and wrist, 2nd edn. Stanford University Press, Stanford
4. Thodberg HH, Jenni OG, Caflisch J, Ranke MB, Martin DD (2009) Prediction of adult height based on automated determination of bone age. *J Clin Endocrinol Metab* 94:4868–4874
5. Thodberg HH, Neuhofer J, Ranke M, Jenni OG, Martin DD (2010) Validation of bone age methods through their ability to predict adult height. *Horm Res Paediatr* 74:15–22
6. Björk A (1968) The use of metallic implants in the study of facial growth in children: method and application. *Am J Phys Anthropol* 29:243–254
7. van Rijn RR, Lequin MH, Thodberg HH (2009) Automatic determination of Greulich and Pyle bone age in healthy Dutch children. *Pediatr Radiol* 39:591–597
8. Thodberg HH, Säwendahl L (2010) Validation and reference values of automated bone age determination for four ethnicities. *Acad Radiol* 17:1425–1432
9. Martin DD, Deusch D, Schweizer R, Binder G, Thodberg HH, Ranke MB (2009) Clinical application of automated Greulich-Pyle bone age in children with short stature. *Pediatr Radiol* 39:598–607
10. Tanner JM, Healy MJR, Goldstein H, Cameron N (2001) Assessment of skeletal maturity and prediction of adult Height (TW3 Method). WB Saunders, London
11. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
12. Urschler M, Grassegger S, Stern D (2015) What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. *Ann Hum Biol* 42:356–365
13. Schmeling A, Grundmann C, Fuhrmann A et al (2008) Criteria for age estimation in living individuals. *Int J Legal Med* 122:457–460
14. Malina RM, Peña Reyes ME, Figueiredo AJ, Coelho E, Silva MJ, Horta L, Miller R, Chamorro M, Serratos L, Morate F (2010) Skeletal age in youth soccer players: implication for age verification. *Clin J Sport Med* 20:469–474
15. Martin DD, Neuhofer J, Jenni OG, Ranke MB, Thodberg HH (2010) Automatic determination of left- and right-hand bone age in the First Zurich Longitudinal Study. *Horm Res Paediatr* 74:50–55
16. Kaplowitz P, Srinivasan S, He J, McCarter R, Hayeri MR, Sze R (2010) Comparison of bone age readings by pediatric endocrinologists and pediatric radiologists using two bone age atlases. *Pediatr Radiol* 41:690–693
17. Zhang S-Y, Liu G, Ma C-G, Han Y-S, Shen X-Z, Xu R-L, Thodberg HH (2013) Automated determination of bone age in a modern Chinese population. *ISRN Radiol* 2013:1–8
18. Thevissen PW, Fieuws S, Willems G (2010) Human dental age estimation using third molar developmental stages: does a Bayesian approach outperform regression models to discriminate between juveniles and adults? *Int J Legal Med* 124:35–42
19. Armitage P, Berry G, Matthews JNS, Corporation E (1994) Statistical methods in medical research. Wiley Online Library