Pediatr Radiol (2009) 39:591–597
DOI 10.1007/s00247-008-1090-8

ORIGINAL ARTICLE

Automatic determination of Greulich and Pyle bone age in healthy Dutch children

Rick R. van Rijn · Maarten H. Lequin · Hans Henrik Thodberg

Received: 25 July 2008 / Revised: 15 September 2008 / Accepted: 21 September 2008 / Published online: 6 January 2009

Abstract
Background Bone age (BA) assessment is a routine procedure in paediatric radiology, for which the Greulich and Pyle (GP) atlas is mostly used. There is rater variability, but the advent of automatic BA determination eliminates this.
Objective To validate the BoneXpert method for automatic determination of skeletal maturity of healthy children against manual GP BA ratings.
Materials and methods Two observers determined GP BA with knowledge of the chronological age (CA). A total of 226 boys with a BA of 3–17 years and 179 girls with a BA of 3–15 years were included in the study. BoneXpert's estimate of GP BA was calibrated to agree on average with the manual ratings based on several studies, including the present study.
Results Seven subjects showed a deviation between manual and automatic BA in excess of 1.9 years. They were re-rated blindly by two raters. After correcting these seven ratings, the root mean square error between manual and automatic rating in the 405 subjects was 0.71 years (range 0.66–0.76 years, 95% CI). BoneXpert's GP BA is on average 0.28 and 0.20 years behind the CA for boys and girls, respectively.
Conclusion BoneXpert is a robust method for automatic determination of BA.

Keywords Bone age · Skeleton · Radiography · Automated recognition

R. R. van Rijn (✉)
Department of Radiology, Academic Medical Center Amsterdam,
Meibergdreef 9,
1105 AZ Amsterdam Zuid-Oost, The Netherlands
e-mail: r.r.vanrijn@amc.uva.nl

M. H. Lequin
Radiology Department, Erasmus MC,
Rotterdam, The Netherlands

H. H. Thodberg
Visiana,
Holte, Denmark

Introduction

The assessment of skeletal maturity or bone age (BA) is a routine procedure in paediatric radiology, for which the Greulich and Pyle (GP) method is by far the most commonly employed technique. The second edition of the GP atlas contains high-quality reproductions of hand radiographs [1]. Greulich and Pyle derived their atlas from Todd's large study of well-off children from Ohio examined between 1931 and 1942 [2]. For each chronological age (CA) they selected an image close to the median maturity to represent the standard for that CA. The children from Todd's study came from upper middle class homes, i.e. they had what Todd would call a better-than-average constitution, and this might explain why the BAs in the GP atlas have been found to be advanced relative to almost all the normal populations studied since then. However, the tendency of modern children to mature faster implies that children are slowly catching up with the GP standard [3].

It is well known that different populations have a different tempo of maturation, so it is not to be expected that the average GP BA of children of different populations agrees with their CA [4, 5]. Instead, BA assessment should be regarded as a quantification of the aspects of bone morphology that are related to maturation. This measure is conveniently expressed in years by reference to the GP

atlas, but this value must be viewed as an arbitrary scale of maturation [6].

In a clinical context dealing with patients from a particular region, the clinician should ideally establish a local BA reference for healthy children. If the population consists of several clearly distinguishable segments, e.g. different ethnicities, there should be one reference for each segment. As a minimum, one should determine the average BA deviation of the local population relative to the GP scale in a relevant age interval for boys and girls. Thus if boys are known to be on average 0.6 years behind the GP BA scale, an observed BA of 9 for a 10-year old boy means that his maturity is 0.4 years behind expectation.

We report here the performance of BoneXpert's GP BA in the Erasmus study. In addition we report the average differences between GP BA and CA in this population.

Materials and methods

The Erasmus study was performed in 1997 in children from the Erasmus Gymnasium in Rotterdam by researchers at the Erasmus Medical Center Rotterdam (EMCR) [7]. The younger subjects were children of employees (and their relations) at the EMC institutions. For the initial study, IRB approval was given to obtain radiographs of the left hand in all children, and subsequent use of these data was permitted by the IRB. For all children younger than 12 years of age informed consent was obtained from the parents or guardians; for children aged 12 years and older informed consent was obtained from the parents or guardians and from the child. This is in keeping with Dutch guidelines on clinical studies in children. A total of 255 boys (median age 12.4 years, range 3.8–20.1 years) and 276 girls (median age 12.6 years, range 3.8–20.0 years), all Caucasian, were included, yielding in total 531 healthy children.

Radiographs of the left hand were recorded on mammography film (Diagnost H; Philips, Eindhoven, The Netherlands) or GTU film (Imation, Oakdale, MN), and Alfa-II Trimax intensifying screens (3M, Maplewood, MN). Radiographic parameters were: small 0.6-mm focus, film–focus distance 1.5 m, 45 kVp, 16 mAs. The images were digitized to 300 dpi with 12 bits per pixel using a Vidar Diagnostic Pro Advantage scanner (Vidar, Hemdon, VA) using TWAIN v5.2.

The films were bone-age rated by two paediatric radiologists who each rated approximately half of the images. The radiologists had knowledge of the CA, which reflects the daily practice of most paediatric radiologists. Intraobserver coefficient of variation of duplicate assessment of skeletal age for investigator 1 was 2.4% and for investigator 2 was 1.5%. We found no significant systematic differences between the two observers regarding variability and levels of measurement, and the interobserver agreement was good [8].

The new computerized approach for BA assessment (BoneXpert, v1.0; Visiana, Holte, Denmark, www.BoneXpert.com) consists of three computational layers [9]:

1. The first layer reconstructs the borders of 15 bones – the five metacarpals, the phalanges of fingers 1, 3 and 5, and the radius and ulna, as shown in Fig. 1. The bone reconstruction algorithm is based on a so-called generative model of image analysis. This enables the method to determine to what extent the bone appears normal. Abnormal bones, as well as wrongly posed normal bones, are automatically rejected.
2. The second computational layer determines bone maturity values, called intrinsic bone ages, for 13 of these 15 bones based on the appearance of the bone. If a BA value deviates more than 2.4 years from the average of all the bones, it is deemed unacceptable. If fewer than eight bones are accepted the image is rejected and no BA values are reported.
3. The third layer transforms the computed intrinsic bone ages to agree on average with GP BA based on a training set of images with manual ratings. (The method can also determine the Tanner and Whitehouse BA, but this was outside the scope of this study.)

The first two layers, which are by far the most complex, were developed from the radiographs of Danish and

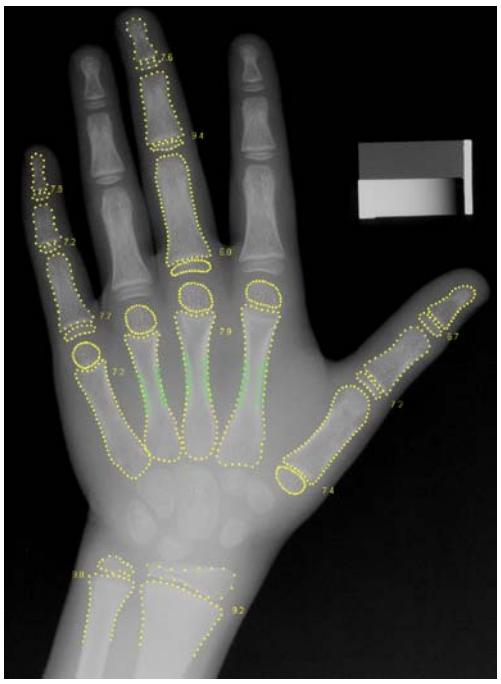

Fig. 1 BoneXpert reconstructs the borders of 15 bones (metacarpals 2 and 4 are reconstructed in order to determine bone mass by radiogrammetry) and estimates the GP BA for 13 bones as indicated

Belgian children (age range 7–17 years), supplemented by radiographs from various sources to extend the age range to 2.5–19 years for boys and 2–18 years for girls; in total 1,678 images [9]. BoneXpert's accuracy is recognized to be poor for boys above 17 years and girls above 15 years, so the intended age range for the clinical use of BoneXpert v1.0 is GP BA 2.5–17 years for boys and 2–15 years for girls, and the performance tested in this work was therefore restricted to these age ranges.

The adjustment of BoneXpert v1.0 to GP BA (the third layer) was made by pooling three datasets in order to average over several manual raters – this study (Erasmus study), a study performed in Tübingen [10], and the GP atlas. Based on these data, a nonlinear transformation of the intrinsic BA into the BoneXpert v1.0 GP BA was constructed. The fact that the Erasmus data were used (together with other data) to develop layer C of BoneXpert v1.0 and also to validate the accuracy of this version in this study requires a careful explanation, because such a strategy could potentially weaken the study, and this is addressed in the appendix.

In the main analysis only the 226 boys and 179 girls with average BA of the manual and automatic methods younger than 17 years or 15 years, respectively, were used.

BoneXpert was marketed as a medical device in Europe in April 2008.

Results

Quality of films

As described in the previous section, BoneXpert automatically rejects images with poor image quality or abnormal bone structure, but no images were rejected by BoneXpert in the Erasmus study. All radiographs were of good quality – the hands were correctly positioned, the images contained all hand bones and film noise was low.

Analysis of deviations

The agreement between BoneXpert v1.0 and the manual rating is shown in Fig. 2. Of the total number of observations, 89% lay within the band ±1.9 year deviation on the plot of manual BA versus automated BA. The observations with larger deviations are particularly interesting because they could represent gross errors in the BoneXpert method. In order to clarify this issue, these cases were subjected to a new blinded rating with two independent raters (R.R. and H.H.T.). In this rating only the sex was known, i.e. neither the CA, nor the previous manual rating, nor BoneXpert's rating were revealed. The radiograph shown in Fig. 1 is among these seven; it is from a 10.1-year-old boy who was initially rated as having a BA of 10.0 years. BoneXpert rated it to 7.8 years, and the manual re-rating yielded 8.3 years. The original and the new ratings for the seven cases versus CA are shown in Fig. 3. The new ratings correlated strongly with the BoneXpert ratings (root mean square, rms, deviation 0.38 years), while they were at odds with the original ratings. The new blind rating showed a BA that deviated considerably from the CA, and the arrows suggest how the initial ratings were biased towards the CA. For the seven re-rated observations, the original ratings were replaced by the new ratings in the subsequent analysis.

Accuracy

The detailed agreement between manual and automatic BA determination was studied by way of a Bland-Altman plot (Fig. 4), where the difference between two measurements is plotted versus their average [11].

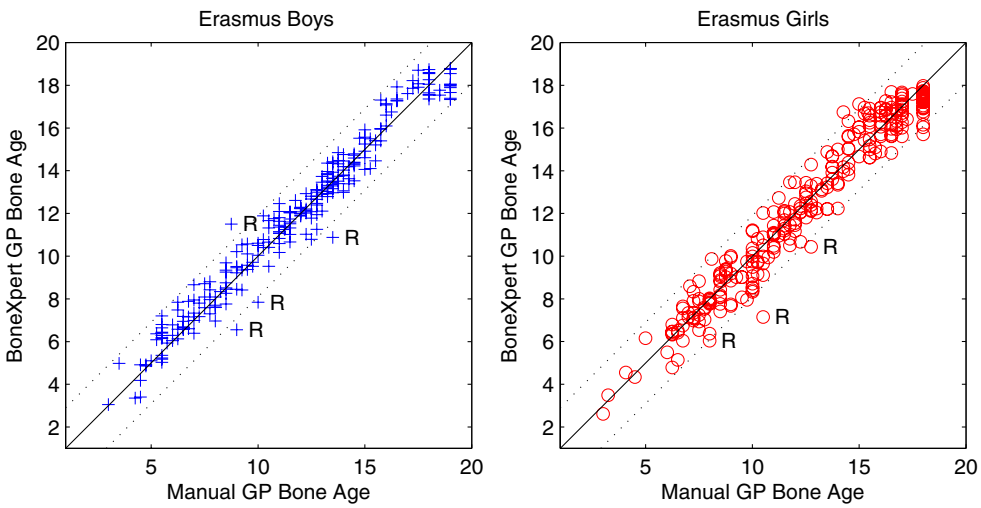

Fig. 2 Comparison of automatic and manual BA rating for all 538 children, using version 1.0 of BoneXpert, in which Layer C was designed based on the Erasmus and Tübingen data. The cases marked *R* deviate by more than 1.9 years and were subjected to a blind re-rating

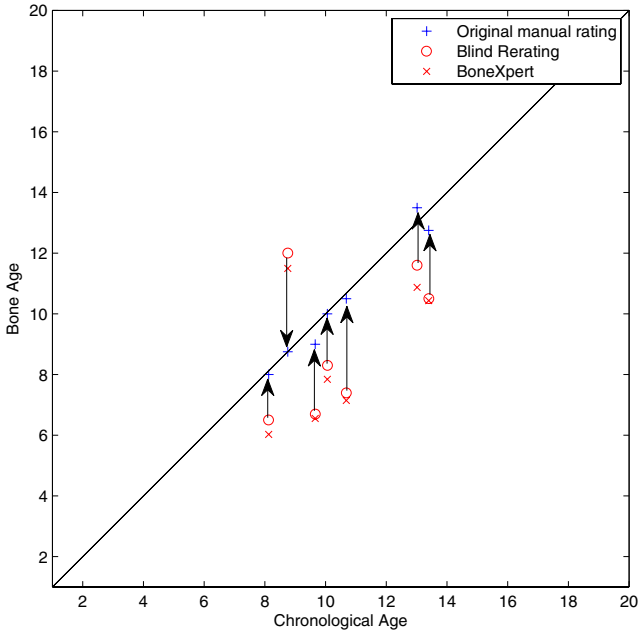

Fig. 3 Analyses of the seven cases with >1.9 years deviation between the original manual rating and the BoneXpert rating. The blind re-rating and BoneXpert's rating agreed very well. The arrows connect the re-rating values (considered to be the most correct) with the original values, and they represent the bias observed in the original rating due to knowledge of CA

The agreement of the individual observations was quantified with the rms error, rather than with the standard deviation, because the latter hides an overall bias. The agreement was poorer for boys with an average BA above 17 years and for girls with an average BA above 15 years. These data were, in accordance with the intended use of BoneXpert, excluded from the results. The agreement between the BoneXpert and manual GP BA ratings for boys was 0.65 years and for girls was 0.76 years (rms errors).

Retardation relative to GP BA

Since the Erasmus study is representative of ethnic Dutch middle and upper class children today, one can use this study to estimate a reference for GP BA across CA for this population. This is illustrated in Fig. 5, which depicts the age difference BA−CA versus CA. The smooth line is the running average over a 3-year interval around the CA. The average deviation between BoneXpert BA and CA averaged over all ages up to 17 years and 15 years for boys and girls, respectively, is −0.28 years and −0.20 years. Similar results for manual GP BA were reported for the Erasmus study [8].

Standard deviation between BA and CA

The standard deviation between CA and BA is shown in Table 1 for three types of ratings: the manual rating, the

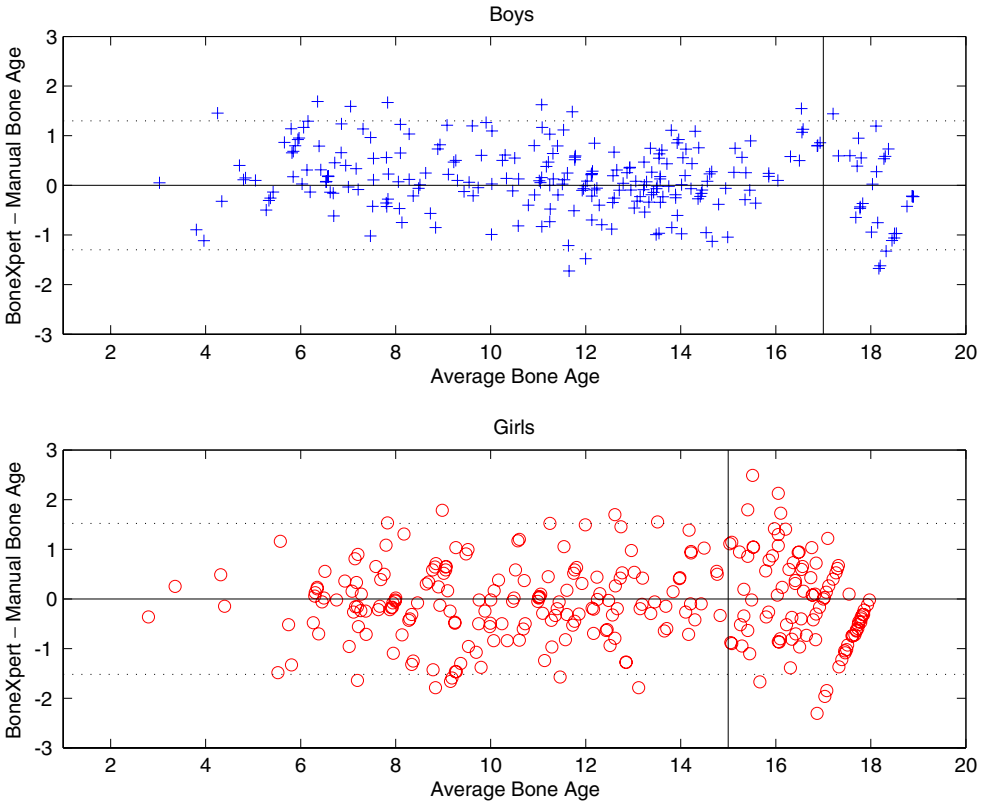

Fig. 4 Bland-Altman plot of the manual and automatic BA rating. The average of the two methods of rating is shown along the horizontal axis, and their difference along the vertical axis. BoneXpert is not intended to be used for boys older than 17 years and girls older than 15 years, and it is seen that the deviations are indeed larger here. For boys BoneXpert overshoots the manual BA slightly at 6–9 years. Elsewhere the two methods agree well on average (dashed lines are drawn at two times the rms deviations of Table 1)

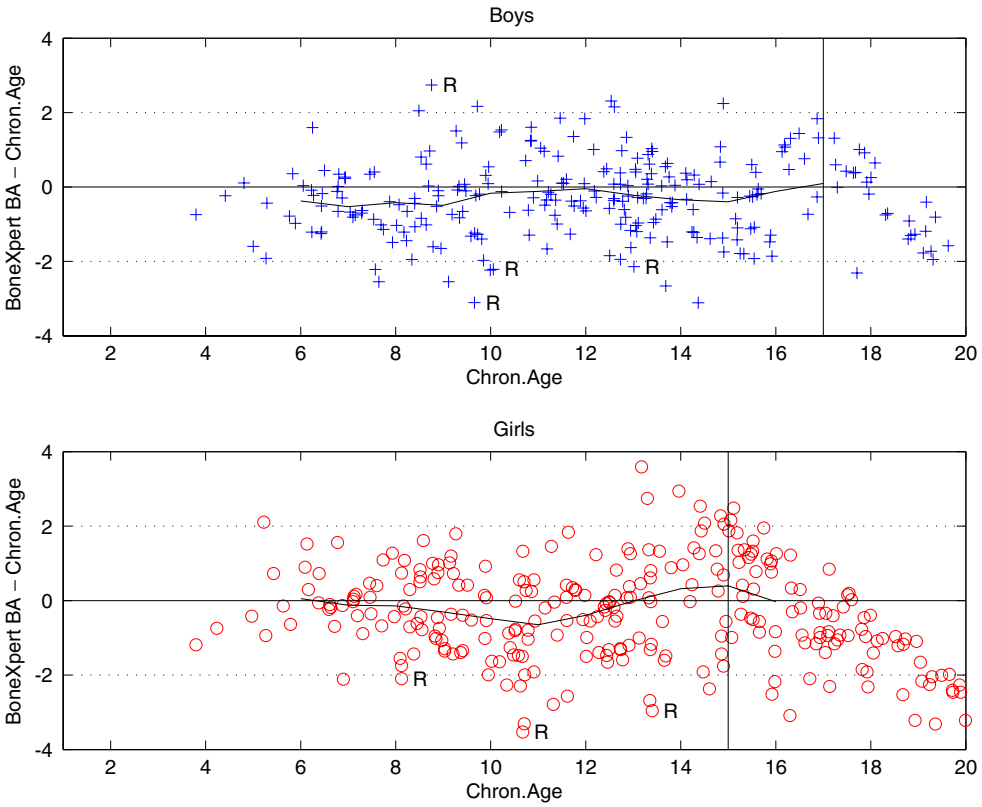

Fig. 5 BoneXpert's BA minus CA versus CA. The smooth curve indicates the running average. The average BA is 0–0.6 years behind the CA, except for girls older than 13 years. The seven cases marked *R* are those that were re-rated (six of them are extremely retarded in BA)

manual rating after correction of the seven cases, and for BoneXpert. The standard deviation was remarkably small for the original manual rating, again evidence of a bias in the manual rating due to knowledge of the CA.

Discussion

The purpose of the re-rating was to generate BA values close to the "true" values, so a discussion of this concept is relevant. There is no objective reference for BA rating. We define the true BA of a radiograph statistically as the average of the ratings by many qualified raters. Thus we regard the intra- and interrater variability as random effects that can be eliminated by taking the average of a large number of ratings. We consider this "aggregate reading" a better procedure than a consensus reading, which is applied in many clinical studies where it improves sturdiness of the dataset. Aggregate reading also underlies the design of BoneXpert where we used the ratings of five different raters to pinpoint the transformation in Layer C of BoneXpert v1.0.

Table 1 The SD between BA and CA for various BA rating methods (computed for CA <17 for boys and CA <15 for girls).

Rating method	SD between BA and CA (years)	
	Boys	Girls
Original manual rating	0.82	0.78
Original manual rating after correcting seven cases	0.87	0.84
BoneXpert rating	1.05	1.23

It is well known that there is interobserver variation in BA rating – as well as in many other radiological procedures – and that the rating can be biased by various expectations [12]. This is inherent in human nature and as such tends to be accepted in the community as unavoidable. However, with the advent of automatic BA determination, rater variability is eliminated. Analysis of the seven radiographs in which manual and automatic rating differed by more than 1.9 years was striking. In all cases the error was on the human side, underlining the robustness of BoneXpert.

We hypothesize that the origin of the large deviations between BoneXpert and the original manual rating was an interpretation bias due to knowledge of the CA. Figure 3 supports this hypothesis. Such an effect was also reported by Berst et al. [13], but our study displayed a more dramatic effect. This finding is a problem in a PACS-based environment, where in daily clinical routine it is virtually impossible to blind radiologists to the CA. The best remedy seems to be to inform the radiologists about the severity of this bias and to encourage them not to look at the CA while rating. The computerized method receives only the image

and the gender as input so by design there is no bias from any other factors.

The susceptibility of raters to bias could be particularly large in BA rating because the result is a continuous value which can easily slide. We have been able to study only the bias from knowing the CA, but there could be other biases. For instance, in a clinical trial where excessive advancement of BA is an undesired effect, the rater could be biased to underestimate the BA. There could also be a bias from looking at the radiograph taken 1 year earlier; if that radiograph was overrated by 1 year there would be a tendency for the new examination also to be overrated. There could be bias from knowledge of the sexual development or height of the child. These biases are undesired because BA rating should be a procedure defined strictly as an isolated interpretation of the hand radiograph without knowing anything other than the sex.

BoneXpert has been calibrated by reference to five different human raters, and therefore embodies a well-supported standardization of GP BA rating. The agreement with the two raters of the Erasmus study in Fig. 4 attests to the extent to which these raters were consistent with the new BoneXpert standard. In general there was good agreement. The bias for boys at age 6–9 years (where BoneXpert overshot the manual rating) is counterbalanced by an opposite bias in the other studies used for the calibration; for instance, BoneXpert underestimates the nominal BA of the GP atlas at these ages. These biases are, therefore, interpreted as rater idiosyncrasies, which the calibration method diluted through the use of many raters. These bias effects are considerably smaller than the observed rms errors, so it is concluded that the participating GP raters and the GP atlas were fairly consistent with each other, and this consistent rating is reflected in BoneXpert's standardized rating.

The fact that BoneXpert has been designed to agree on average with manual GP ratings is of great practical importance because this allows clinicians to adopt the new method while still being able to relate the results to previous manual ratings. BoneXpert's standardized ratings are particularly useful in multicentre studies where geographic location does not have to be a limitation per se in the study set-up, and could serve as a replacement for "central reading". Currently BoneXpert is a Windows-based application, serving as a PACS node to which PACS stations can send DICOM files, which can then be analysed in the Windows program. Full integration into a PACS environment is currently work in progress.

The validation of the BoneXpert method in the Erasmus study showed that BoneXpert was able to analyse all images, and blind re-rating of the seven subjects with the largest deviations from the manual rating showed that in these subjects the manual rating was wrong. These subjects had a very retarded or advanced BA, revealing the radiologist's bias as they were aware of the CA. After correcting the rating of the seven outliers, and omitting boys older than 17 years and girls older than 15 years, the rms deviation between manual and automatic BA was 0.65 years and 0.76 years for boys and girls, respectively. The deviation for both sexes combined was 0.71 years (range 0.66–0.76 years, 95% CI).

Studies in healthy children are rare in recent times because they are difficult to justify ethically. The Erasmus study is therefore of unique value for establishing a reference for GP BA for modern Western European children. The study showed that boys and girls in this population are expected, on average, to have a BoneXpert GP BA 0.28 years and 0.20 years below the CA, respectively.

The study has the limitation that the Erasmus data were used both to adjust the overall BA scale and to validate a range of other aspects of the same system. As discussed in the appendix, this does not, in our opinion, significantly reduce the strength of the study. However, it does make the

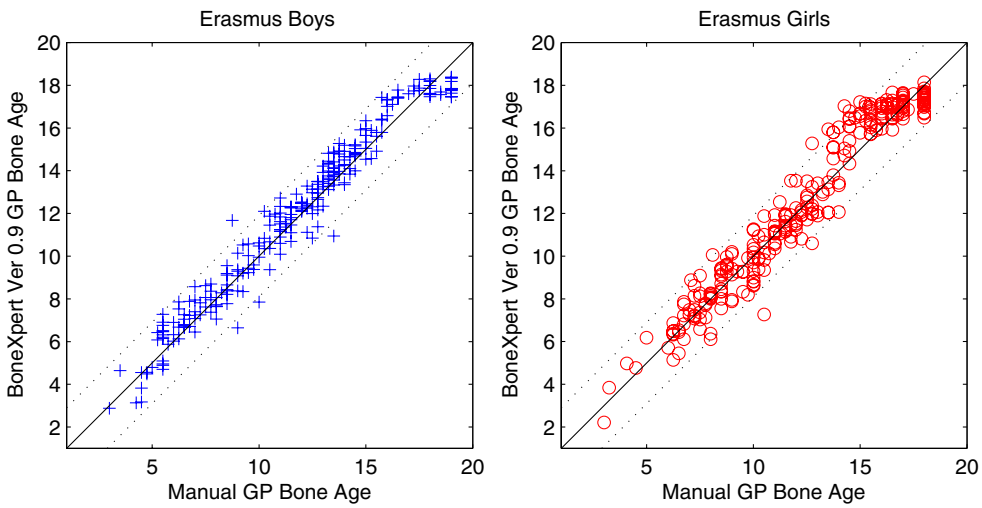

Fig. 6 Agreement between manual ratings and BoneXpert v0.9 for all 538 children (v0.9 was developed without any reference to the Erasmus data). The outliers represent the same patients as identified in Fig. 2

presentation of the study more complicated. A more serious limitation is that this study included only healthy Caucasian children recorded on high-quality radiographs. It is, therefore, appropriate to mention that the study is complemented by the Tübingen study [10], where the children had various endocrine disorders and where the image quality was more typical.

Acknowledgement Novo Nordisk is acknowledged for providing access to the film scanner.

Disclosure The BoneXpert technology is proprietary to Visiana, a company owned by H. H. Thodberg. BoneXpert has been marketed as a medical device. A patent application on BoneXpert has been filed.

Appendix

The purpose if this appendix is to demonstrate the lack of concern regarding the fact that the Erasmus data are used both for the adjustment of "layer C" of BoneXpert v1.0 and also for validation of the accuracy of BoneXpert 1.0.

The previous version of BoneXpert (v0.9) has been described [9]. This version did not use the Erasmus and Tübingen data to adjust layer C, but only the GP atlas. The difference between v0.9 and v1.0 is only in layer C, which transforms the intrinsic BA ($BA_{intrinsic}$) into the predicted GP BA by means of a so-called shift function:

$$GP = BA_{intrinsic} + Shift(BA_{intrinsic}, sex)$$

The purpose of the shift function is to adjust the bias of the BA determination to agree on average with a training set, which for v1.0 consisted of the Tübingen data, the Erasmus data and the GP atlas. All images of the same sex and $BA_{intrinsic}$ are adjusted by the same amount, so this adjustment cannot change the *ordering* of the images on the BA scale.

Figure 6 presents the agreement between BoneXpert v0.9 and the manual rating, and Fig. 2 shows the same for v1.0. It is seen that the *width* of the band of data points is approximately the same in the two versions, but the band follows the diagonal closer in v1.0. The squared correlations were R^2=0.962 and 0.965 for boys in v0.9 and v1.0, respectively, and 0.946 and 0.950 for girls, respectively (including all data in the figures). The average absolute difference between predictions of v0.9 and v1.0 is 0.26 years. The largest difference between v0.9 and v1.0 is in girls of BA 13–14 years.

Figures 2 and 6 demonstrate that the adjustment of layer C only affects the average BA; it does not affect the deviations between the automatic and the manual ratings. The strength of a computerized BA system lies not in its ability to produce correct answers on average, but in its ability to order the cases correctly according to their maturity. There is, therefore, no concern about the use of the Erasmus data both for adjusting layer C of BoneXpert v1.0 and also for validating its accuracy.

The reason that all the data were used for calibration (rather than putting, for example, one-third apart for validation) is that reliable calibration requires a large sample size. Other studies are in preparation in which BoneXpert will be validated using independent data.

References

1. Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of hand and wrist, 2nd edn. Stanford University Press, Stanford
2. Todd TW (1937) Atlas of skeletal maturity. Part I. Hand. Kimpton, London
3. Tanner JM (1981) A history of the study of human growth. Cambridge University Press, Cambridge
4. Loder RT, Estle DT, Morrison K et al (1993) Applicability of the Greulich and Pyle skeletal age standards to black and white children of today. Am J Dis Child 147:1329–1333
5. Ontell FK, Ivanovic M, Ablin DS et al (1996) Bone age in children of diverse ethnicity. AJR 167:1395–1398
6. Tanner JM, Whitehouse RH, Cameron N et al (1975) Assessment of skeletal maturity and prediction of adult height (TW2 method), 2nd edn. Academic Press, London
7. Lequin MH, van Rijn RR, Robben SG et al (2000) Normal values for quantitative tibial ultrasonometry in a Caucasian pediatric population (aged 6 to 19 years). Calcif Tissue Int 67:101–105
8. Van Rijn RR, Lequin MH, Robben SG et al (2001) Is the Greulich and Pyle atlas still valid for Dutch Caucasian children today? Pediatr Radiol 31:748–752
9. Thodberg HH, Kreiborg S, Juul A et al (2008) The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging (in press). doi:10.1109/TMI.2008.926067
10. Martin D, Deusch D, Schweizer R et al (2007) Validation of automatic Greulich and Pyle bone age on GHD, UTS, SGA and Silver-Russell syndrome children. Horm Res 68 [Suppl 1]:69
11. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
12. Brealey S, Westwood M (2007) Are you reading what we are reading? The effect of who interprets medical images on estimates of diagnostic test accuracy in systematic reviews. Br J Radiol 80:674–677
13. Berst MJ, Dolan L, Bogdanowicz MM et al (2001) Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR 176:507–510