Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods

R K Bull, P D Edwards, P M Kemp, S Fry, I A Hughes

Abstract

***Purpose*—Comparison of bone age assessed using either the "atlas matching" method of Greulich and Pyle or the "point scoring system" of Tanner and Whitehouse (TW2).**

***Materials and methods*—362 consecutive "bone age" radiographs of the left hand and distal radius performed in a large provincial teaching hospital. Data were analysed using the "method comparison" statistical technique. Ten per cent of the radiographs were re-analysed to assess intra-observer variation.**

***Results*—The 95% confidence interval for the difference between the two methods was 2.28 to −1.52 years. Intra-observer variation was greater for the Greulich and Pyle method than for the TW2 method (95% confidence limit, −2.46 to 2.18 *v* −1.41 to 1.43).**

***Conclusion*—The two methods of bone age assessment as used in clinical practice do not give equivalent estimates of bone age and we suggest that one method only (preferably the TW2) should be used when performing serial measurements on an individual patient.**

(*Arch Dis Child* 1999;**81**:172–173)

Keywords: bone age; Greulich and Pyle; Tanner and Whitehouse

Department of Radiology, Addenbrooke's Hospital, Hills Road, University of Cambridge, Cambridge CB2 2QQ, UK
R K Bull
P D Edwards

Department of Nuclear Medicine, Addenbrooke's Hospital
P M Kemp

University of Cambridge, Department of Paediatrics, Box 116, Addenbrooke's Hospital
S Fry
I A Hughes

Correspondence to:
Professor Hughes.

Accepted 30 March 1999

Bone age is commonly assessed by one of two main methods: either the Greulich and Pyle method[1] or the Tanner and Whitehouse (TW2) method.[2] The Greulich and Pyle method as originally described involves a complex comparison of all of the bones in the hand and wrist against reference "normal" radiographs of different ages. In most institutions a "rapid" modified version of this technique is used, whereby the overall appearance of a given radiograph is compared with the reference radiographs and the nearest match is selected. Although this modified approach is considerably faster than the original it may be less accurate. The TW2 method relies on the systematic evaluation of the maturity of all the bones in the hand and wrist. Several small studies have compared the two methods,[3 4] and have suggested that there is close agreement between them. However, the data from these studies were analysed by regression analysis, which is inappropriate for this type of comparison.

Our study compares the rapid Greulich and Pyle method, as used commonly in clinical practice, with the TW2 method in a large group of subjects. Data are analysed using the more appropriate "method comparison" technique.[5]

Materials and methods

All bone age radiographs of the left hand, including the wrist and distal radius, performed in our institution between 1992 and 1996 for assessment of bone age were analysed (362 radiographs). The children were aged between 2 and 18 years and came from the general population of a large provincial teaching hospital (children aged < 2 years were excluded because bone age assessment from radiographs of the wrist in this age group is unreliable).

Over the four year period of study, the radiographs were assessed by a succession of radiology trainees (total 12) according to the method of Greulich and Pyle.[1] The same radiographs were also assessed by the TW2 method[2] by one of two nurse auxologists: specialist nurses who had received specific training in the use of the TW2 method.

Thirty nine of the radiographs (~ 10%) were then reassessed by both methods by the same observers to assess intra-observer variation for each method.

Results

Statistical analysis involved comparison of bone age assessed by these two methods. Results are shown on a scatter graph (fig 1)

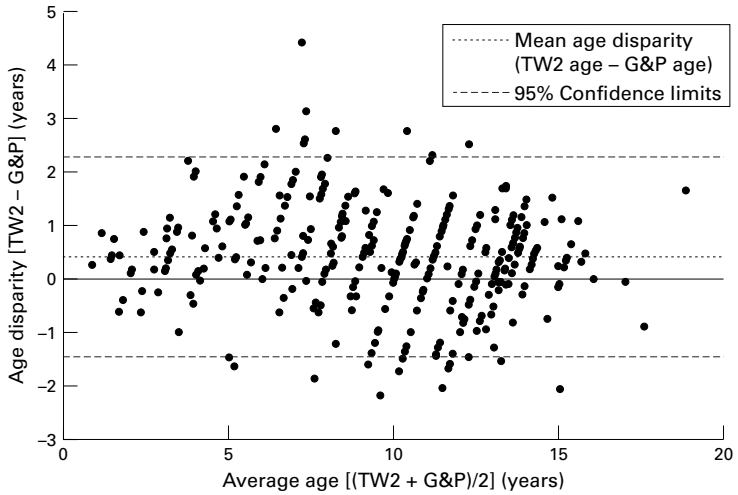

Figure 1 Age disparity versus average age. G & P, Gruelich and Pyle; TW2, Tanner and Whitehouse.

Table 1 Intra-observer variation and variation between methods

	Mean age disparity (years)	*SD of disparity (years)*	*95% CL*
Intra-observer variation			
Greulich and Pyle method	0.14	1.16	−2.46 to 2.18
Tanner and Whitehouse method	0.01	0.71	−1.41 to 1.43
Variation between methods	0.38	0.95	2.28 to −1.52

For intra-observer variation, mean age disparity is reading 2 − reading 1
For variation between methods, mean age disparity is TW2 age − G & P age.
CL, confidence limits.

plotting mean age as calculated by the two methods against the age disparity between the two methods.

Similar statistical techniques were then used to assess the 39 repeated studies to measure intra-observer variation. This variation can be compared with the variation between the two methods (table 1).

Discussion

This large study using the method comparison technique confirms the finding that bone ages assessed with the TW2 method are slightly greater than those measured with the Greulich and Pyle method. Our mean age disparity of 0.38 years (fig 1; table 1) is similar to that of previous studies,[3 4] and this difference is significantly different from zero ($p < 0.01$).

Our study is the first of this type to use method comparison scatter plots instead of regression analysis. These "Bland-Altman" scatter plots[5] of differences against mean bone age are readily interpretable by the reader. The range of differences is easily discernible, which is important in clinical usage. The 95% confidence interval for the difference between the two methods is 2.28 to −1.52 years. In clinical practice an error of this size is unacceptably large. This disagrees with previous studies. If our data are re-analysed using the inappropriate method of regression analysis (as used in previous studies) the *r* value obtained is 0.96, which initially appears highly impressive. However, it does not convey the relevant information about absolute and maximum differences between the results of the two techniques. The correlation coefficient measures the strength of an association between two variables, not the agreement between them; furthermore, the wider the range of values being compared (in this case from 2 to 18 years), the greater the correlation.[5]

Our measured intra-observer variation (table 1) is greater for the Greulich and Pyle method than for the TW2 method (95% confidence limits, −2.46 to 2.18 *v* −1.41 to 1.43). This magnitude of intra-observer variation seen for the Greulich and Pyle method probably accounts for much of the discrepancy between the two methods. The subjects on whom the two bone age methods were originally based came from very different social backgrounds. Greulich and Pyle studied American children of high socioeconomic status in the 1940s, whereas Tanner and Whitehouse studied Scottish children of low socioeconomic status in the 1950s. It must also be noted that the two different methods of bone age assessment were performed by different observers with differing levels of experience (radiology registrars (Greulich and Pyle) and nurse auxologists (TW2)). All of the above factors probably contributed to the higher intra-observer variation seen with the Greulich and Pyle method. However, the greatest potential source of error probably comes from the fact that we compared the overall appearance of the radiographs with the standard reference radiographs to obtain the best match. Although this is the approach commonly used, this is not the method originally suggested by Greulich and Pyle. If this more time consuming approach had been used in our study, it is possible that both intra-observer variation and variation between methods would have been reduced.

Conclusion

We conclude that the Greulich and Pyle and TW2 methods produce different values for bone age, which are significant in clinical practice. This disagrees with previous smaller studies, all of which were somewhat flawed by the use of regression analysis, which is an inappropriate statistical technique for this type of study. In addition, we have shown that the TW2 method is more reproducible than the Greulich and Pyle method. We hypothesise that the rapid Greulich and Pyle method, as used in common clinical practice, is potentially less accurate than the more rigorous time consuming approach originally suggested by these authors. Therefore, we suggest that only one method of bone age assessment (preferably the TW2 method) should be used when performing serial measurements on an individual patient.

We gratefully acknowledge the work undertaken by Ms C Hawley and Ms P Moreland. The study was supported by the Child Growth Foundation and Pharmacia and Upjohn.

1 Greulich WW, Pyle SI, Waterhouse AM. *A radiographic standard of reference for the growing hand and wrist.* Chicago: Case Western Reserve University, 1971.
2 Tanner JM, Whitehouse RH, Cameron N, Marshall WA, Healy MJR, Goldstein H. *Assessment of skeletal maturity and prediction of adult height*, 2nd ed. London: Academic Press, 1983.
3 Milner GR, Levick RK, Kay R. Assessment of bone age: a comparison of the Greulich and Pyle and the Tanner and Whitehouse methods. *Clin Radiol* 1986;**37**:119–21.
4 King DG, Steventon DM, O'Sullivan MP, *et al.* Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. *Br J Radiol* 1994;**67**:848–5.
5 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.