A Comparative Evaluation of Bone Age Estimation Models

Dr Daniel Brzozowski, Mrs Caroline Parker and Dr Phillip Wardle

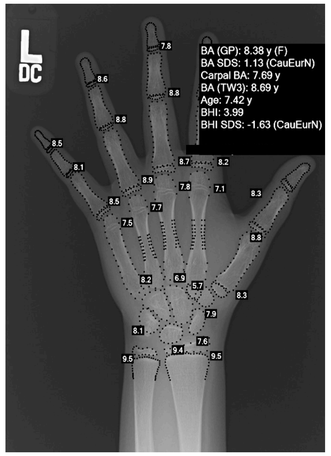
Academi Ddelweddu Genedlaethol Cymru
National Imaging Academy Wales

PURPOSE

To evaluate performance of different AI models used to estimate bone age against the gold standard of experienced Consultant Radiologist reports.

METHODS

50 hand radiographs were analysed retrospectively, and bone age estimation data was calculated. Original consultant reports utilising the Greulich and Pyle model were collected, and second blind reports were produced by a Radiology Trainee and Consultant Radiologist using the same method. Four different AI models were used to estimate bone age: 1. BoneXpert using the Greulich and Pyle method; 2. BoneXpert using the Tanner-Whitehouse 2 method; 3. BoneXpert using the Tanner-Whitehouse 3 method; and 4. Gleamer using the Greulich and Pyle method.





RESULTS

The average bone age estimation of Radiology Trainee and Consultant Radiologists reports was 10 years and 7 months which was taken as the gold standard for the purpose of AI model comparison.

The average deviation from gold standard for the AI models were +1.9 months (SD=0.42), +8.3 months (SD=0.51), -2 months (SD=0.68) and +1.2 months (SD=0.67) for models 1-4 respectively. The maximum difference was: 1.1, 1.5, 1.7 and 2.3 years respectively.

CONCLUSION

The Bone Age estimation AI models that use the Greulich and Pyle method have performed similarly, and both were close to the Radiologist gold standard reports. There was significant variation between the estimations of both BoneXpert models using the Tanner-Whitehouse methods 2 and 3. This highlights the difficulties of evaluating the AI models which utilise different calculation methods to the perceived gold standard.

Despite model 4 performing closest to the mean, the standard deviation of its results was significantly higher than that of other models and the highest discrepancy was more than twice that of the first model. This is important, as the reliability and reproducibility of the data is crucial to inform confidence of users of the AI models, both Radiologists and referrers. This further evidences the need for ongoing close scrutiny and monitoring of AI tools when implementing and using Radiology AI models in clinical practice.

