Validation of an AI system for the automatic calculation of bone age in hand x-rays

Martina Aineseder¹ Meng Candelaria Mosquera¹ Andres Brandan¹ Yulian Battellini¹ Federico Lipsich¹ Patricia Manganaro¹ Sonia Benitez¹ Cristian Demeco¹ Tamara Kreindel¹

¹Hospital Italiano de Buenos Aires, Almagro, Argentina

URL: https://tghncollections.pubpub.org/pub/ohmp7pr8
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
Objective: To evaluate the performance of automated bone age prediction compared to measurement by imaging professionals.

Methods: An artificial intelligence model was developed that calculates bone age (BA) from a hand radiograph using a convolutional neural network.

For validation, hand radiographs acquired at HIBA were included with BA data in the radiology report.

Diagnostic performance on 200 radiographs was compared between the model, two imaging subspecialists, a resident and the original radiograph reports based on Greulich & Pyle and Oxford Altas.

Results: To measure agreement in OE in months using the Kappa coefficient. It was observed that the highest agreement is between the two experts (0.9), and that the resident presents an average agreement of 0.835 with the experts, while Carpian presents an average agreement of 0.805.

To evaluate the EO calculation performance, the average EO reported by both experts was used as the reference standard. Quadratic error was calculated. The resident presented the best results, with a quadratic error of 0.07, followed by Carpian with 0.11 and finally the original report with 0.13.

Confusion matrices were calculated for the detection of abnormal BA. We observed that in 13 cases Carpian reported an OE that would have indicated an abnormality when the OE according to experts was in accordance with chronological age (false positives). In 14 cases the opposite was true: the experts reported an OE more than two years different than chronological, whereas Carpian would have reported an EO according to chronological (false negatives). The resident had 15 false positives and 10 false negatives, showing comparable performance to Carpian.

Conclusion: Carpian's performance is comparable to that of a resident trained in BA analysis. The difference with the report is not significant, suggesting that Carpian could complement a practitioner's report. This would improve diagnostic performance in centers without pediatric radiology specialists, improving patient care in an equitable and universal way.