Abstract of talk at ESPE 2019 in Vienna, to be published in Hormone Research in Paediatrics

Validation of a new version of BoneXpert bone age in children with congenital adrenal hyperplasia (CAH), precocious puberty (PP), growth hormone deficiency (GHD), Turner syndrome (TS), and other short stature diagnoses

Hans Henrik Thodberg[1], David D Martin[2,3]

[1]Visiana, Hørsholm, Denmark. [2]University Children's Hospital , Tübingen, Germany. [3]Witten/Herdecke University, Herdecke, Germany

Abstract

Background: The BoneXpert method for automated determination of bone age from hand X-rays is based on machine learning, so it lends itself naturally to be improved by adding more training data and using better learning algorithms. Currently, version 2 is running in 145 hospitals across Europe, and a new version 3 is rolled out in 2019.

Objective and hypotheses: The aim was to validate version 3 against manual ratings in retrospective studies, for which the performance of the previous version of BoneXpert has already been published.

Method: The training set included 14036 public images from the 2017 RSNA Bone Age Challenge, 1642 images of normal Dutch and Californian children, and three studies from Tubingen collected 1976-2006: 6743 images of short stature (GHD, TS, Silver-Russell Syndrome, idiopathic short stature etc), 775 images of CAH, and 732 images of PP. The learning algorithm included more accurate and robust localisation of the bones, an extension of the bone age range down to new-borns, and adding of carpals and finger 2 and 4. We report the results as the cross-validated root mean square errors (RMSE) of the method relative to the original manual rating.

Results: The RMSE in short stature improved from 0.74 years for the current version to 0.64 years for version 3. For CAH, the RMSE improved from 0.67 to 0.57 years and for PP from 0.68 to 0.60 years. The overall improvement was from 0.72 to 0.63 years.

Conclusion: The accuracy of automated bone age rating is now so good, that the observed error relative to a single manual rating is dominated by the uncertainty of the manual rating. The standard deviation of manual ratings, when repeated by different observers, is 0.52-0.64 years in clinical routine, and the Tubingen raters are believed to lie at the lower end of this interval.

The observed RMSE thus tells more about the rater variability of the particular raters than uncertainty of the method. Therefore, we recommended that future validation studies compare the AI method to the average of three or more raters. This will reduce the error of the “reference”, and at the same time allow an estimate of the interrater variability. In the RSNA Challenge, a test set of 200 images was rated by six raters, and the new version obtains an RMSE of 0.45 years against the average. Automated bone age rating is now clearly better than a single manual rating.