Radiology



# Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs[1]

David B. Larson, MD, MBA
Matthew C. Chen, MS
Matthew P. Lungren, MD, MPH
Safwan S. Halabi, MD
Nicholas V. Stence, MD
Curtis P. Langlotz, MD, PhD

**Purpose:** To compare the performance of a deep-learning bone age assessment model based on hand radiographs with that of expert radiologists and that of existing automated models.

**Materials and Methods:** The institutional review board approved the study. A total of 14036 clinical hand radiographs and corresponding reports were obtained from two children's hospitals to train and validate the model. For the first test set, composed of 200 examinations, the mean of bone age estimates from the clinical report and three additional human reviewers was used as the reference standard. Overall model performance was assessed by comparing the root mean square (RMS) and mean absolute difference (MAD) between the model estimates and the reference standard bone ages. Ninety-five percent limits of agreement were calculated in a pairwise fashion for all reviewers and the model. The RMS of a second test set composed of 913 examinations from the publicly available Digital Hand Atlas was compared with published reports of an existing automated model.

**Results:** The mean difference between bone age estimates of the model and of the reviewers was 0 years, with a mean RMS and MAD of 0.63 and 0.50 years, respectively. The estimates of the model, the clinical report, and the three reviewers were within the 95% limits of agreement. RMS for the Digital Hand Atlas data set was 0.73 years, compared with 0.61 years of a previously reported model.

**Conclusion:** A deep-learning convolutional neural network model can estimate skeletal maturity with accuracy similar to that of an expert radiologist and to that of existing automated models.



An earlier incorrect version of this article appeared online. This article was corrected on January 19, 2018.

[1] From the Departments of Radiology (D.B.L., M.P.L., S.S.H., C.P.L.), Computer Science (M.C.C.), and Biomedical Informatics (C.P.L.), Stanford University School of Medicine, 300 Pasteur Dr, Stanford, CA 94305-5105; and Department of Radiology, Children's Hospital Colorado, Aurora, Colo (N.V.S.). Received February 1, 2017; revision requested April 3; revision received July 18; accepted August 1; final version accepted August 24. **Address correspondence to** D.B.L. (e-mail: *david.larson@stanford.edu*).

Deep learning is a form of machine learning that uses multiple levels of representations to enable automated classification of items in a data set (1). Specifically, deep-learning models contain layers of nodes, representing a hierarchy of features of increasing complexity, that are mathematically related to one another in networks (2,3). Through multiple layers, complex input data can be related to an output classifier to determine specific properties of the input data. Recent advances in computing power and machine learning techniques prompted the rise of a specific type of deep learning—convolutional neural networks—to be applied to image recognition tasks (4). Such tasks include facial recognition, object detection, and image classification. Deep-learning applications have important implications for diagnostic imaging (5–8).

Automated assessment of hand radiographs for determining skeletal maturity, or bone age, is a logical early application for deep learning because a manual technique of comparing a single radiographic image to a reference standard has been used for several decades, resulting in a single quantitative output: an estimated bone age (9). Most deep-learning models that were developed for nonmedical applications are designed to classify a single planar image, which enables rapid reuse and refinement of that foundational work for bone age classification. Furthermore, manual assessment of skeletal age is frequently criticized as tedious, time-consuming, and limited by considerable interrater and intrarater variability (10–12). In fact, bone age rating was one of the first radiologic procedures considered for automation when quantitative image analysis techniques became available (13).

Through use of specifically programmed techniques for feature extraction, automated algorithms that assess hand radiographs for determination of skeletal age have been developed; they are currently used in clinical practice and have an accuracy similar to that of raters (14).

The purpose of this study was to compare the performance of a deep-learning model for bone age assessment based on hand radiographs with that of expert radiologists and that of existing automated models.

### Advances in Knowledge

- Deep learning can be used to create an automated bone age assessment model.
- The average mean absolute difference (MAD) between the bone age estimates from each human reviewer and the mean estimate of the other human reviewers' estimates was 0.61 years, whereas the average MAD of the estimates from the model and the mean estimate from the same reviewers was 0.52 years ($P < .05$ for two of the four reviewers).
- The root mean square of the difference between the model's bone age estimates and that of a publicly reported data set was 0.73 years, compared with 0.61 years for a previously reported model based on feature extraction (statistical significance cannot be assessed).
- The clinical classification of the bone age as advanced, normal, or delayed was not significantly different for any of the reviewers or the model ($P = .59$).
- The model increased in accuracy with increasing training set size, with root mean square of the interobserver difference for the Digital Hand Atlas test set decreasing from 1.08 to 0.91 to 0.78 to 0.73 years for training set sizes of 1558, 3141, 6295, and 12611 images, respectively.

### Implication for Patient Care

- A deep learning–based automated software application with accuracy similar to that of a radiologist could be made available for clinical use.

## Materials and Methods

### Data Acquisition

A data set composed of 14036 clinical radiographs of the left hand from two institutions, Lucile Packard Children's Hospital at Stanford University (Palo Alto, Calif; $n = 2983$) and Children's Hospital Colorado (Aurora, Colo; $n = 11053$), obtained for bone age assessment, was used to create the model (Table 1). These images had been interpreted by pediatric radiologists, who documented skeletal age in the radiology report on the basis of a visual comparison to Greulich and Pyle's *Radiographic Atlas of Skeletal Development of the Hand and Wrist* (15). The images were drawn from the picture archive and communication systems of the respective hospitals. The institutional review boards of both institutions approved the study. Informed consent was waived. Bone age designations were extracted automatically from the radiology reports with a Python script (Python Software Foundation, Beaverton, Ore) and were used as ground truth for training the model. The number of examinations in which the reports described a skeletal dysplasia was also documented.

**https://doi.org/10.1148/radiol.2017170236**

**Content code:** PD

**Radiology 2018;** 287:313–322

**Abbreviations:**
MAD = mean absolute difference
RMS = root mean square

**Author contributions:**
Guarantors of integrity of entire study, D.B.L., M.C.C., C.P.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, D.B.L., M.C.C., M.P.L., S.S.H., C.P.L.; clinical studies, D.B.L.; experimental studies, M.P.L.; statistical analysis, D.B.L., M.C.C., M.P.L., C.P.L.; and manuscript editing, all authors

Conflicts of interest are listed at the end of this article.

See also the editorial by Summers in this issue.

**Table 1**

**Summary Information for the Training, Validation, and Test Image Data Sets**

| Variable | No. of Images in Males | No. of Images in Females | Total No. of Images | Mean Chronologic Age (y) | Mean Estimated Bone Age (y) |
|---|---|---|---|---|---|
| Training set | | | | | |
| Stanford | 1485 | 1200 | 2685 | 10.7 ± 4.1 | 10.9 ± 3.9 |
| Colorado | 5348 | 4578 | 9926 | 10.8 ± 3.4 | 10.5 ± 3.3 |
| Total | 6833 | 5778 | 12611 | 10.8 ± 3.5 | 10.6 ± 3.4 |
| Validation set | | | | | |
| Stanford | 174 | 124 | 298 | 10.8 ± 4.2 | 10.9 ± 4.0 |
| Colorado | 599 | 528 | 1127 | 10.8 ± 3.3 | 10.5 ± 3.3 |
| Total | 773 | 652 | 1425 | 10.8 ± 3.5 | 10.6 ± 3.5 |
| Test sets | | | | | |
| Stanford test set | 100 | 100 | 200 | 11.3 ± 3.8 | 11.0 ± 3.6 |
| Digital Hand Atlas test set | 434 | 479 | 913 | 8.8 ± 3.6 | 8.8 ± 3.8 |

Note.—Data are mean ± standard deviation or numbers of images. Stanford refers to Lucile Packard Children's Hospital at Stanford University and Colorado refers to Children's Hospital Colorado.

This image set was divided into two separate subsets: a training set, which was used to optimize the model parameters, and a validation set, which was used to tune the model hyper-parameters, such as learning rate and training duration. Of the images, 90% ($n$ = 12611) were randomly selected for the training set and 9% ($n$ = 1425) were used as the validation set.

Two data sets were used as test sets to evaluate the performance of the model relative to human reviewers and existing automated bone age software.

We used the first test set, a sex-stratified set of 200 hand radiographs (100 from male patients and 100 from female patients), to evaluate the performance of the model relative to that of human reviewers. These radiographs were randomly selected from the clinical picture archive and communication systems at Stanford but represented different studies than those of the training and validation sets used in the model development. The images in this test set were reviewed by three fellowship-trained pediatric radiologists, with 9 years (S.S.H.), 8 years (D.B.L.), and 2 years (M.P.L.) of post-fellowship experience, respectively. The bone age estimate stated in the original clinical report was used as an additional assessment, for a total of four human reviewers. Reviewers provided their bone age estimates to the nearest month. For examinations in which the clinical report provided a bone age estimate that was between two ages, the mean of the ages was used. The three study reviewers were blinded to the patient's chronologic age, the clinical report, and the assessments of the other reviewers. Of the 14036 clinical examinations from the original training set, these radiologists had reviewed 16 (S.S.H.), 81 (D.B.L.), and four (M.P.L.) examinations, respectively, for a total of 101 images, or 0.72% of the original training set.

The second test set was developed and used to evaluate the performance of the model relative to that of existing automated software (12). A total of 913 deidentified images were obtained from the publicly available Digital Hand Atlas developed by the University of Southern California Image Processing and Informatics Laboratory (16,17). Each image in this publicly available data set had been assigned a bone age rating by two independent radiologists. The mean of the bone age assessments of the two raters was considered the reference standard.

To assess the clinical significance of rating differences, we compared individual reviewers and the model according to the rates at which their bone age estimates would have resulted in a change in diagnosis compared with the reference standard. Normal was defined as a bone age within 2 standard deviations of the chronologic age, as defined by Greulich and Pyle, after adjustment according to the Brush Foundation data (15). Advanced or delayed bone age was defined as a bone age estimate higher than or lower than the limits of normal, respectively. With use of this definition, each rating was classified as advanced, normal, or delayed, according to the assessments by the clinical report, each of the study reviewers, and the model.

To evaluate the relationship between the size of the training set and the precision of the model, four separate models were created on the basis of randomly selected training subset sizes of 1558, 3141, 6295, and 12611 images. By use of the Digital Hand Atlas test set, a root mean square (RMS) was calculated for each subset size.

### Data Preprocessing

Images were converted from Digital Imaging and Communications in Medicine to Portable Network Graphics images by using Python (version 2.7) and the pydicom library (Python Software Foundation; version 0.9.9) and downsized to a resolution of 256 × 256 pixels by using Python Imaging Library (version 1.1.7; Python Software Foundation). Images were enhanced with contrast-limited adaptive histogram equalization (18), implemented in Python OpenCV (version 2.4.9) with three different threshold parameters used to create a false red, green, blue color channel (18,19).

**Figure 1**

**Figure 1:** Examples of preprocessed training images from four different patients. All preprocessing transformation types other than mean subtraction are shown, including application of false color channel by using contrast-limited adaptive histogram equalization and random contrast adjustments, flips, and crops.

Images were further cropped to a resolution of 224 × 224 to accommodate transformations on the augmented images, which included random flips, crops, and contrast adjustments of the image (4). Examples of preprocessed images are shown in Figure 1. The full data pipeline is shown in Figure 2.

### Model Implementation

The model used an architecture known as a deep residual network (20,21). The deep residual network architecture with 50 layers, containing $3.8 \times 10^9$ floating point operations (known as FLOPs), was used for this project because it achieved results similar to those observed with deeper architectures with shorter training times. The original 1000-class output layer was replaced with a final fully connected layer to output a probability score for each month and sex combination. The output of the classifier was a probability distribution over bone ages from 0 to 19 years, in increments of 1 month.

We implemented the model by using an open-source machine learning library (TensorFlow version 0.9.0; Google, Mountain View, Calif). The model took 6–8 hours to train on a single K80 Graphics Processing Unit (Nvidia, Santa Clara, Calif). Adam (22), a method used for stochastic optimization that is implemented within TensorFlow, was used as the optimization algorithm. We initialized our model parameters to pretrained weights optimized for the ImageNet data set and converted from Caffe Zoo, an online repository for pretrained model data (23–25).

We supplemented the quantitative evaluation of our model with a qualitative assessment of the areas in each image to which the model was the most sensitive. We used a viewing method called saliency maps to show sensitive regions of the image and displayed the results as a heat map (26). The map displays the absolute magnitude of the partial derivative of the loss function

with a given ground-truth label with respect to each input pixel. In the resulting graph, the magnitude at a pixel represents how important that pixel is in the overall choice of the model output.

### Statistical Analysis

To compare the performance of the human reviewers to paired interobserver differences reported in the literature, a mean paired interobserver difference was calculated for each reviewer pair.

The overall performance of the model was assessed by comparing the RMS and the mean absolute difference (MAD) between the model estimates and the reference standard bone ages from the 200 radiographs from the Stanford test set. MAD was included because it is less sensitive to outlier data than is RMS. Because an objective reference standard bone age cannot be determined, the mean estimate of all four human reviewers (including the clinical report) was used as the reference standard for this calculation.

The RMS was calculated as the square root of the sum of the squares of the paired differences. The MAD was calculated as the mean of the absolute values of the difference between the reviewer and model estimates and those of the reference standard bone age.

Both interrater and intrarater variability are known limitations of bone age assessment (27); this variability increases the difficulty of comparing model performance with human performance. To assess agreement between reviewers and agreement of the model with each reviewer, 95% limits of agreement were calculated in a pairwise fashion. A Bland-Altman plot was created to show the difference between the model estimates and the reference standard estimates over the range of the mean of the two estimates. To compare the performance of the model directly with that of each human reviewer, the paired interobserver difference between each human reviewer and the mean of the remaining three human reviewers was compared with the paired interobserver difference between the model and the same mean. Relative performance was evaluated by comparing the mean, RMS, and MAD of each human reviewer with those of the model. Statistical significance was determined by using paired *t* tests for comparing means (mean and MAD) and *F* tests for comparing variances (ie, RMS).

A $\chi^2$ test was used to assess for differences between the clinical ratings of the reviewers and the model. The test was based on a $3 \times 5$ contingency table, with rows and columns consisting of the diagnosis (advanced, normal, and delayed) and the source of the estimate (clinical report, reviewers 1–3, and model), respectively.

Limits of agreement were calculated by using Stata (Stata Corp LP, College Station, Tex) release 14.2 and R version 3.3.1 with version 0.7.2 of the equivalence package (R Foundation for Statistical Computing, Vienna, Austria). All other statistical calculations were performed with Excel 2016 (Microsoft, Redmond, Wash).

**Figure 2**

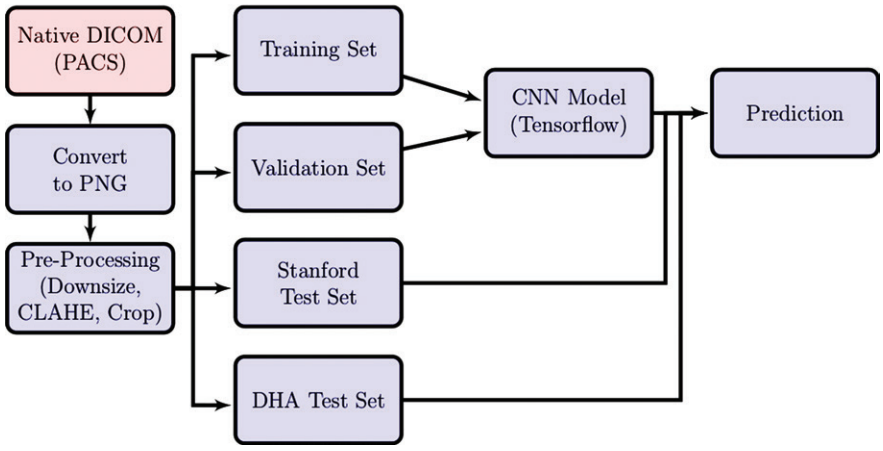



**Figure 2:** Data flow diagram from raw Digital Imaging and Communications in Medicine *(DICOM)* images to bone age assessment predictions made by the model on a test set, including creation of the convolutional neural network *(CNN)* model. *CLAHE* = contrast-limited adaptive histogram equalization; *DHA* = Digital Hand Atlas; *PACS* = picture archive and communication systems; *PNG* = portable network graphic file format.

## Results

The male-to-female ratio of the training and validation sets was 7606:6430 (54%:46%) (Table 1). The mean chronologic age of the patients was 10.8 years ± 3.4 (standard deviation; range, 0–21.4 years), and the mean bone age of the patients was estimated by the pediatric radiologists in the clinical report to be 10.6 years ± 3.5 (range, 0–19.0 years). The distribution of the estimated bone age of the training and validation sets is shown in Figure 3. Of the 14036 examinations in the training set, skeletal dysplasia was described in 30 examinations (0.21%).

The male-to-female ratio of the 200-examination Stanford test set was 100:100 (50%:50%) (Table 1). The mean chronologic age of the patients was 11.3 years ± 3.8 (range, 1.4–18.2 years), and the mean bone age of the patients was estimated by the human raters to be 11.0 years ± 3.6 (range, 0.9–17.9 years). The distribution of the estimated bone age of the 200-examination test set is shown in Figure 3.

In evaluating the human reviewer performance on the 200-examination Stanford test set, when reviewers were compared directly with each other in pairwise fashion, the RMS of the paired interobserver difference ranged from 0.93 to 1.17 years, with a mean of 1.00 years. The mean paired interobserver difference between the estimates of each human reviewer and the mean of the other reviewers' estimates ranged from −0.07 years to 0.08 years,

with a mean of 0 years (Table 2). The reviewers' RMS of the paired interobserver difference ranged from 0.73 to 0.95 years, with a mean of 0.82 years. The reviewers' MAD of the paired interobserver difference ranged from 0.53 to 0.69 years, with a mean of 0.61 years.

When we compared the performance of the model with the mean of the reviewer estimates, the mean difference in bone age estimates was 0 years. The mean RMS and MAD were 0.63 years and 0.50 years, respectively.

The limits of agreement between bone age estimates of the clinical report, reviewers, and the model are shown in Figure 4. All assessments were within the 95% limits of agreement of each other. The Bland-Altman plot illustrating the difference between the model estimates and the reference standard estimates over the range of the mean of the two estimates is shown in Figure 5.

Summary performance statistics for comparisons of the estimates of each reviewer and model to the mean of the other reviewers' estimates are shown in Table 2. The mean bone age estimated by the model was not significantly different from that estimated by any of the reviewers. The model's RMS and MAD were significantly lower when compared with the clinical report and reviewer 3, but not for reviewers 1 and 2. In other words, by these measures, the model outperformed the clinical report and reviewer 3 and performed at a level that was not statistically significantly different from that of reviewers 1 and 2.

The percentages of the 200 examinations that would be reclassified to a different diagnosis (advanced, normal, or delayed bone age) for the clinical report, the three reviewers, and the model, respectively, compared with the reference standard estimate, were as follows: 18.5%, 16.5%, 14.0%, 16.5%, and 15.5%. The $\chi^2$ analysis revealed no significant difference between clinical ratings ($P = .59$) (Table 3).

When applied to the Digital Hand Atlas data set, our model demonstrated an RMS of 0.73 years. This was after removal of images outside the range of 2.5–17 years for boys and 2–15 years for girls so that the analysis would be comparable to previous work (12).

Sample images from the test set with corresponding superimposed saliency maps are shown in Figure 6, and they highlight the general regions in the image to which the model output is most sensitive. The most sensitive regions corresponded to the proximal interphalangeal joints, the metacarpal-phalangeal joints, and the carpal bones, which correspond to maturity indicators as specified by the standards of Greulich and Pyle (15), although with different emphases for the different sample images. Specifically, the sensitivity to the carpal bones in many of the images resembles the approach of the Tanner-Whitehouse method of bone age assessment, which treats the carpal bones as one of two major portions of the assessment (28).

**Figure 3**

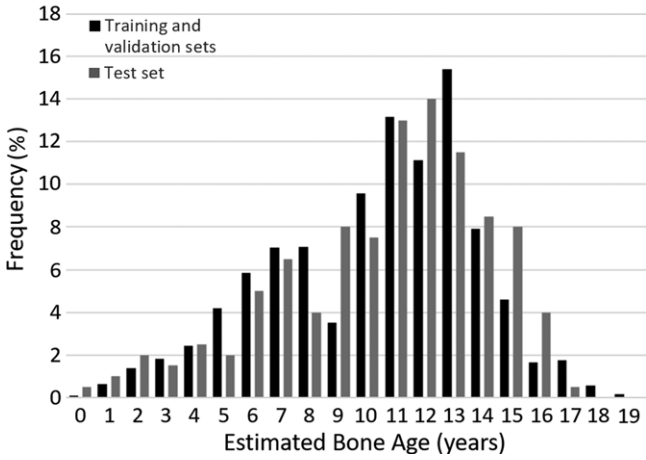



**Figure 3:** Distribution of estimated bone ages of the 14 036 training and validation examinations (combined) and the 200 examinations in the Stanford test set.

**Table 2**

**Summary Statistics of Paired Interobserver Difference between Bone Age Estimate of Each Reviewer and Mean of the Other Three Human Reviewers' Estimates, Compared with That of Model**

| Variable | Clinical Report | Reviewer 1 | Reviewer 2 | Reviewer 3 | Mean |
|---|---|---|---|---|---|
| Mean | | | | | |
| Reviewer | 0.08 | −0.07 | −0.07 | 0.06 | 0.00 |
| Model | 0.02 | −0.01 | −0.01 | 0.02 | 0.00 |
| *P* value (paired *t* test) | .41 | .34 | .36 | .57 | |
| RMS | | | | | |
| Reviewer | 0.87 | 0.73 | 0.73 | 0.95 | 0.82 |
| Model | 0.65 | 0.67 | 0.67 | 0.68 | 0.67 |
| *P* value (F test, comparing ratio of variances) | <.01 | .26 | .23 | <.01 | |
| MAD | | | | | |
| Reviewer | 0.65 | 0.55 | 0.53 | 0.69 | 0.61 |
| Model | 0.51 | 0.53 | 0.53 | 0.53 | 0.52 |
| *P* value (paired *t* test) | <.01 | .50 | .99 | <.01 | |

Note.—Unless otherwise noted, data are expressed as years. The authors of the clinical report were treated collectively as a single reviewer.

For the four models based on training set sizes of 1558, 3141, 6295, and 12611 images, the RMSs of the interobserver difference for the Digital Hand Atlas test set were 1.08, 0.91, 0.78, and 0.73 years, respectively.

## Discussion

We sought to develop a deep-learning model for estimation of bone age, based on comparison with the Greulich and Pyle atlas, with performance comparable to that of a human reviewer and to that of existing automated bone age software applications. Our model was developed on the basis of a training set size of 12611 clinical hand radiographs and the corresponding clinical radiology reports.

A key challenge in validating such a model is the determination of a reference standard for true bone age because variation in human reviewer estimates of bone age is a well-recognized phenomenon (10,13,27). Several studies (13,14,29) addressed this issue, reporting a standard error of the mean for interobserver observations falling in the range between 0.45 and 0.83 years; this corresponds to a standard deviation of 0.64–1.17 years. We found a standard deviation of paired interobserver differences of 1.00 years, which falls within that range, indicating that our reviewers performed similarly to other manual raters in clinical practice.

The model was within 95% limits of agreement of all the human reviewers. In head-to-head comparison with each of the reviewers, the performance of the model was similar to that of two of the reviewers and better than two of the reviewers. Therefore, we conclude that our deep learning–based model performs at a level similar to that of a trained human reviewer.

Automating the bone age assessment task by using hand radiographs is not new; the literature contains more than 15 documented attempts to perform this task, and the progression of this work mirrored the advancements in machine learning through the decades. The HANDX and CASAS systems were among the first works reported on the topic (30–32). These programs were based on feature extraction, with the automated assessment calculated after 13 individual bones were highlighted manually by the user, aided by templates presented on the screen. These programs, particularly CASAS, yielded a more consistent assessment of bone age compared with a manual rating in research settings (32,33), but they were not widely implemented because they were generally more time-consuming than manual assessment of bone age (34,35).

**Figure 4**

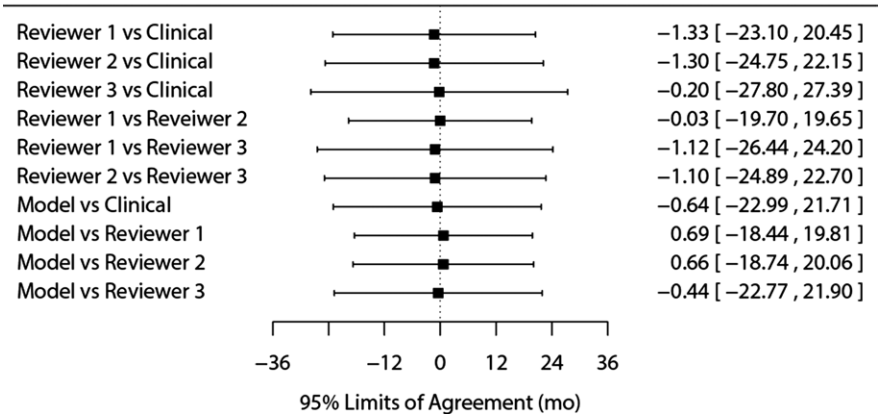



**Figure 4:** Ninety-five percent limits of agreement between bone age estimates of the clinical report, reviewers, and the model.

**Figure 5**

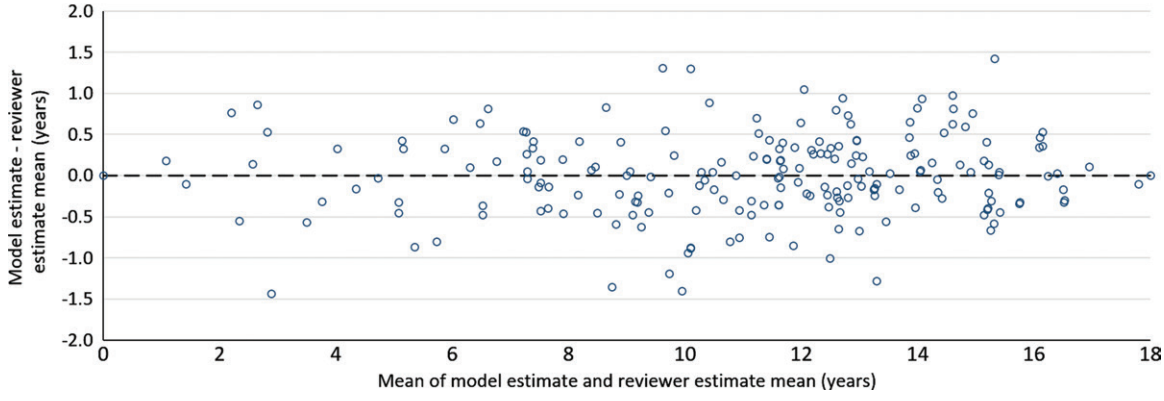



**Figure 5:** Bland-Altman plot showing the difference between the model estimate and the reviewer estimate mean over the range of the mean of these estimates.

A fully automated commercial system for bone age assessment (BoneXpert; Visiana Aps, Holte, Denmark, *http://www.bonexpert.com*) is now available. This system is based on a feature-extraction technique that reconstructs the borders of the bones (13). The application has been cleared for clinical use in Europe. Van Rijn and Thodberg (13) evaluated the accuracy of the model, as reported in six other peer-review publications. They reported that the mean standard deviation in the differences between the BoneXpert model and manual assessments ranged from 0.55 to 0.76 years, with a weighted average of approximately 0.68 years. This compares to our model's RMS of 0.63 years, calculated by using the mean of the reviewers' estimates as the reference standard.

More direct comparison of the two models can be performed on the basis of the Digital Hand Atlas data set. Thodberg et al (12) reported an RMS error of the BoneXpert model of 0.61 years, compared with 0.73 years for our model. However, they performed a post hoc reclassification of some of the examinations that produced the greatest error for their model, which may have resulted in a relatively optimistic assessment of the model accuracy. Regardless, the statistical significance of differences between the two models cannot be assessed.

On the basis of these assessments, we conclude that an automated model for assessment of bone age based on a convolutional neural network can have an accuracy similar to that of current state-of-the-art automated models by using feature-extraction techniques.

Our results suggest potential broad applicability of deep-learning models for a variety of diagnostic imaging tasks without requiring specialized subject matter knowledge or image-specific software engineering. Specifically, machine learning models developed for other vision tasks, such as the ImageNet Challenge (4), may also be generalized to tasks in the medical domain. By initializing weights optimized for ImageNet, which contains more than 1 million images and 1000 unique classes, we trained a model with fewer training examples that did not overfit the data. Such an approach may significantly lower the development costs of and increase the practical uses for automated image assessment. However, certain elements of assessing bone age make this task particularly suitable for a deep-learning solution, including the ready availability of classified images, the discrete and quantitative nature of the task, and the fact that a single planar image can be used for the task.

We found that the accuracy of the model improved as the size of the training image set increased. Specifically, the incremental improvement in the RMS of the paired interobserver difference decreased as the number of images in the training set was successively doubled. This suggests that a

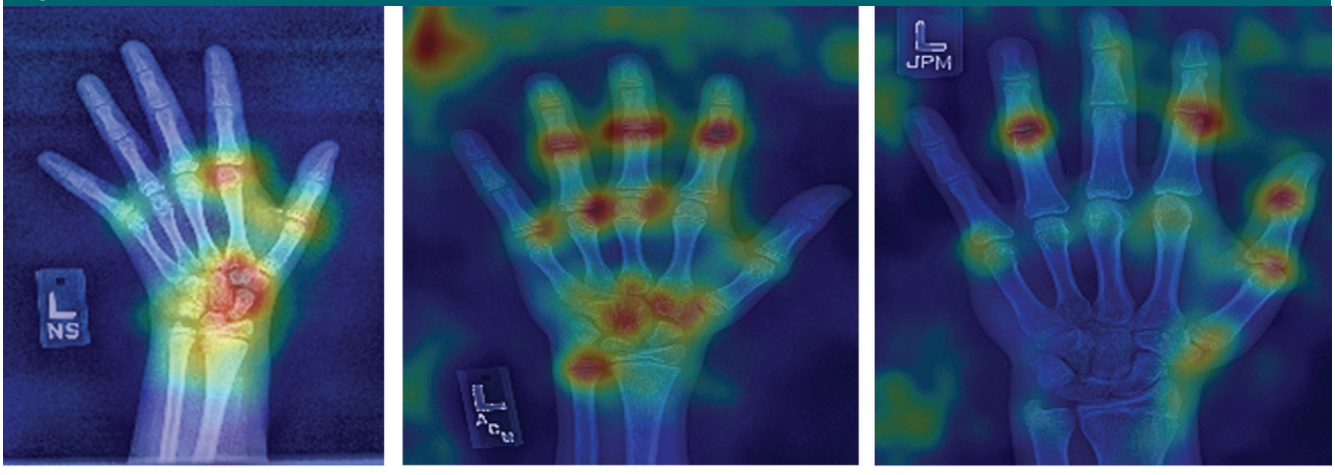
**Table 3**

**Clinical Ratings of Bone Age Assessments of the 200 Examinations in Stanford Test Set According to Clinical Report, Reviewers, and Model, Used as Basis for $\chi^2$ Analysis**

| Rating | Clinical Report | Reviewer 1 | Reviewer 2 | Reviewer 3 | Model |
|---|---|---|---|---|---|
| Advanced | 26 | 28 | 31 | 27 | 23 |
| Normal | 119 | 103 | 98 | 113 | 111 |
| Delayed | 55 | 69 | 71 | 60 | 66 |

Note.—Data are the number of examinations.

**Figure 6**




a. b. c.

**Figure 6:** Original image with superimposed saliency map for sample hand radiographic images in three male patients age 4 years (**a**), 15 years (**b**), and 17 years (**c**).

much larger data set would not necessarily produce substantially greater accuracy. Although we can conclude that the size of medical imaging data sets plays an important role in the development of deep-learning algorithms, we suspect that the specific data set size needed for model training needs likely varies according to the characteristics of the task.

While developing the model, we formulated a reference standard that enabled us to draw conclusions about the relative accuracy of both the model and radiologists. In the future, it is conceivable that a deep-learning model could be used as the reference for human reviewer performance rather than the reverse.

Our study had several limitations related to the specific application of bone image assessment as well as general aspects of deep learning. First, because of the lack of a reference standard and inherent variation of human assessment, it is difficult to determine with certainty how the performance of the model compares with that of human reviewers. Second, similar to other machine learning applications, in its current state, the model would not detect certain disorders that a human expert reviewer might detect from the image, such as hypochondroplasia, rickets, and congenital syndromes (13). Third, our model may be biased given that it was based on the comparison with the clinical assessments of radiologists from just two clinical sites by using the Greulich and Pyle atlas (15). However, the purpose of this study was to evaluate the feasibility of the use of a deep-learning model to automatically determine bone age; the model could probably be trained by using different bone age estimates from different sites or by using techniques other than the Greulich and Pyle atlas. In fact, such models most likely could be derived directly from normal hand radiographs rather than by relying on an atlas. Fourth, the model was not effective in predicting the bone age of patients younger than 2 years. This may have been related to the relatively low number of training set examinations for this age group and the fact that pediatric radiologists recognize the Greulich and Pyle method as less useful in this age group (9).

Finally, we emphasize that automated assessment of bone age is likely among the easiest applications for deep learning in medical imaging; more complex applications may be less accurate, less clinically relevant, or less feasible to develop. Although our results are encouraging for application of deep learning in medical images, they do not necessarily indicate how successful such applications will be when applied to more complex and nuanced imaging tasks.

In conclusion, we find that a deep-learning convolutional neural network model can estimate skeletal maturity with accuracy similar to that of an expert radiologist and to that of current state-of-the-art feature extraction–based automated models for assessment of bone age.

**Disclosures of Conflicts of Interest: D.B.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed a fee from Bayer Healthcare. Other relationships: disclosed no relevant relationships. **M.C.C.** disclosed no relevant relationships. **M.P.L.** disclosed no relevant relationships. **S.S.H.** disclosed no relevant relationships. **N.V.S.** disclosed no relevant relationships. **C.P.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is a founder, shareholder, and board member of Montage Healthcare Solutions, and author received consulting fees and travel reimbursement from Montage Healthcare Solutions. Other relationships: disclosed no relevant relationships.

## References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
2. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR Am J Roentgenol 2017;208(4):754–760.
3. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. RadioGraphics 2017;37(2):505–515.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, December 3–6, 2012. Red Hook, NY: Curran Associates, 2012; 1097–1105.
5. Petersen K, Nielsen M, Diao P, Karssemeijer N, Lillholm M. Breast tissue segmentation and mammographic risk scoring using deep learning. In: Fujita H, Hara T, Muramatsu C, eds. Breast imaging. IWDM 2014. Lecture Notes in Computer Science, vol 8539. Cham, Switzerland: Springer, 2014; 88–94.
6. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35(5):1153–1159.
7. Obermeyer Z, Emanuel EJ. Predicting the Future – Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016;375 (13):1216–1219.
8. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115–118.
9. Breen MA, Tsai A, Stamm A, Kleinman PK. Bone age assessment practices in infants and older children among Society for Pediatric Radiology members. Pediatr Radiol 2016; 46(9):1269–1274.
10. Roche AF, Rohmann CG, French NY, Dávila GH. Effect of training on replicability of assessments of skeletal maturity (Greulich-Pyle). Am J Roentgenol Radium Ther Nucl Med 1970;108(3):511–515.
11. Johnson GF, Dorst JP, Kuhn JP, Roche AF, Dávila GH. Reliability of skeletal age assessments. Am J Roentgenol Radium Ther Nucl Med 1973;118(2):320–327.
12. Thodberg HH, Neuhof J, Ranke MB, Jenni OG, Martin DD. Validation of bone age methods by their ability to predict adult height. Horm Res Paediatr 2010;74(1):15–22.
13. Van Rijn RR, Thodberg HH. Bone age assessment: automated techniques coming of age? Acta Radiol 2013;54(9):1024–1029.
14. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009;28(1):52–66.
15. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd ed. Stanford, Calif: Stanford University Press, 1971.
16. Gertych A, Zhang A, Sayre J, Pospiech-Kurkowska S, Huang HK. Bone age assessment of children using a digital hand atlas. Comput Med Imaging Graph 2007;31(4-5): 322–331.
17. Zhang A, Sayre JW, Vachon L, Liu BJ, Huang HK. Racial differences in growth pat-

terns of children assessed on the basis of bone age. Radiology 2009;250(1):228–235.

18. Teare PA, Morrison A, Elnekave EE. Malignancy risk assessment and localization on mammography using false color enhancement via contrast limited adaptive histogram equalization. Presented at the Society for Imaging Informatics in Medicine Conference on Machine Intelligence in Medical Imaging, Alexandria, Va, September 12, 2016.
19. Zuiderveld K. Contrast limited adaptive histogram equalization. Graphics gems IV. San Diego, Calif: Academic Press Professional, 1994.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv preprint December 10, 2015. arXiv:1512.03385. https://arxiv.org/abs/1512.03385. Published December 10, 2015. Accessed October 12, 2017.
21. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–252.
22. Kingma D, Ba J. Adam: A method for stochastic optimization. Presented at the International Conference on Representational Learning, Banff, Canada, April 14–16, 2014. arXiv preprint July 23, 2015: arXiv:1412.6980v8. https://arxiv.org/abs/1412.6980. Published December 22, 2014. Accessed October 12, 2017.
23. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. arXiv submission March 23, 2014: arXiv:1403.6382. https://arxiv.org/abs/1403.6382. Published March 23, 2014. Accessed October 12, 2017.
24. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, Fla, November 3–7, 2014: 675–678.
25. Kaiming H, Zhang X, Ren S. Sun. Deep residual networks. GitHub, Microsoft Research Web site. https://github.com/KaimingHe/deep-residual-networks. Published February 1, 2016. Accessed January 21, 2017.
26. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: isualizing image classification models and saliency maps. arXiv preprint April 19, 2014: arXiv:1312.6034v2. https://arxiv.org/abs/1312.6034. Published December 20, 2013. Accessed October 12, 2017.
27. Groell R, Lindbichler F, Riepl T, Gherra L, Roposch A, Fotter R. The reliability of bone age determination in central European children using the Greulich and Pyle method. Br J Radiol 1999;72(857):461–464.
28. Tanner JM, Healy MJR, Goldstein H, Cameron N. Assessment of skeletal maturity and prediction of adult height (TW3 method). 3rd ed. London, England: Saunders, 2001.
29. Bull RK, Edwards PD, Kemp PM, Fry S, Hughes IA. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. Arch Dis Child 1999;81(2):172–173.
30. Michael DJ, Nelson AC. HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs. IEEE Trans Med Imaging 1989;8(1):64–69.
31. Pietka E, McNitt-Gray MF, Kuo ML, Huang HK. Computer-assisted phalangeal analysis in skeletal age assessment. IEEE Trans Med Imaging 1991;10(4):616–620.
32. Tanner JM, Gibbons RD. A computerized image analysis system for estimating Tanner-Whitehouse 2 bone age. Horm Res 1994;42(6):282–287.
33. Tanner JM, Gibbons RD. Automatic bone age measurement using computerized image analysis. J Pediatr Endocrinol Metab 1994; 7(2):141–145.
34. Frisch H, Riedl S, Waldhör T. Computer-aided estimation of skeletal age and comparison with bone age evaluations by the method of Greulich-Pyle and Tanner-Whitehouse. Pediatr Radiol 1996;26(3):226–231.
35. Pietka E, Gertych A, Pospiech S, Cao F, Huang HK, Gilsanz V. Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction. IEEE Trans Med Imaging 2001;20(8):715–729.