Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Comparison of Commercial AI Software Performance for Radiograph Lung Nodule Detection and Bone Age Prediction

Kicky G. van Leeuwen, MSc • Steven Schalekamp, MD, PhD • Matthieu J. C. M. Rutten, MD, PhDs • Merel Huisman, MD, PhD • Cornelia M. Schaefer-Prokop, MD, PhD • Maarten de Rooij, MD, PhD • Bram van Ginneken, PhD • Bas Maresch, MD • Bram H. J. Geurts, MD • Cornelius F. van Dijke, MD, PhD • Emmeline Laupman-Koedam, MD • Enzo V. Hulleman, MD • Eric L. Verhoeff, MD • Evelyne M. J. Meys, MD, PhD • Firdaus A. A. Mohamed Hoesein, MD, PhD • Floor M. ter Brugge, MD • Francois van Hoorn, MD • Frank van der Wel, Ad • Inge A. H. van den Berk, MD • Jacqueline M. Luyendijk, MD • James Meakin, PhD • Jesse Habets, MD, PhD • Jonathan I. M. L. Verbeke, MD • Joost Nederend, MD, PhD • Karlijn M. E. Meys, MD • Laura N. Deden, MSc • Lucianne C. M. Langezaal, MD • Mahtab Nasrollah, MD • Marleen Meij, MD • Martijn F. Boomsma, MD, PhD • Matthijs Vermeulen, MD • Myrthe M. Vestering, MD • Onno Vijlbrief, MD • Paul Algra, MD • Selma Algra, MD, PhD • Stijn M. Bollen, MD • Tijs Samson, PDEng • Yntor H. G. von Brucken Fock, MD • for the Project AIR Working Group[1]

From the Department of Medical Imaging, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, the Netherlands (K.G.v.L., S.S., M.J.C.M.R., M.H., C.M.S.P., M.d.R., B.v.G., B.H.J.G., J.M.); Department of Radiology (M.J.C.M.R.) and Department of MICT and Imaging Techniques (T.S.), Jeroen Bosch Hospital, 's-Hertogenbosch, the Netherlands; Department of Radiology, Meander Medical Centre, Amersfoort, the Netherlands (C.M.S.P., M.V.); Department of Radiology, Hospital Gelderse Vallei, Ede, the Netherlands (B.M., M.M.V.); Department of Radiology, Noordwest Ziekenhuisgroep, Alkmaar, the Netherlands (C.F.v.D., P.A.); Department of Radiology & Nuclear Medicine, Máxima Medical Center, Eindhoven, the Netherlands (E.L.K., F.v.d.W.); Department of Radiology, Ziekenhuisgroep Twente, Almelo, the Netherlands (E.V.H., F.M.t.B., M.M., O.V., Y.H.G.v.B.F.); Center for Radiology and Nuclear Medicine, Deventer Hospital, Deventer, the Netherlands (E.L.V., J.M.L., M.N.); Department of Radiology, Catharina Hospital, Eindhoven, the Netherlands (E.M.J.M., J.N., K.M.E.M.); Department of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands (F.A.A.M.H.); Department of Radiology, Zaans Medisch Centrum, Zaandam, the Netherlands (F.v.H.); Department of Radiology and Nuclear Medicine, Amsterdam UMC–Location University of Amsterdam, Amsterdam, the Netherlands (I.A.H.v.d.B.); Department of Radiology & Nuclear Medicine, Haaglanden Medical Center, The Hague, the Netherlands (J.H.); Department of Radiology, Amsterdam University Medical Center, Amsterdam, the Netherlands (J.I.M.L.V.); Department of Radiology and Nuclear Medicine, Rijnstate, Arnhem, the Netherlands (L.N.D.); Department of Radiology, St Antonius Hospital, Nieuwegein, the Netherlands (L.C.M.L., S.A.); Department of Radiology, Isala Hospital, Zwolle, the Netherlands (M.F.B.); and Department of Radiology, Groene Hart Hospital, Gouda, the Netherlands (S.M.B.). Received April 26, 2023; revision requested July 6; final revision received November 22; accepted November 27. **Address correspondence to** K.G.v.L. (email: *kicky.vanleeuwen@radboudumc.nl*).

[1] Members of the Project AIR Working Group are listed in Table S1.

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

See also the editorial by Omoumi and Richiardi in this issue.

Radiology 2024; 310(1):e230981 • https://doi.org/10.1148/radiol.230981 • Content codes: CH MK AI

Background: Multiple commercial artificial intelligence (AI) products exist for assessing radiographs; however, comparable performance data for these algorithms are limited.

Purpose: To perform an independent, stand-alone validation of commercially available AI products for bone age prediction based on hand radiographs and lung nodule detection on chest radiographs.

Materials and Methods: This retrospective study was carried out as part of Project AIR. Nine of 17 eligible AI products were validated on data from seven Dutch hospitals. For bone age prediction, the root mean square error (RMSE) and Pearson correlation coefficient were computed. The reference standard was set by three to five expert readers. For lung nodule detection, the area under the receiver operating characteristic curve (AUC) was computed. The reference standard was set by a chest radiologist based on CT. Randomized subsets of hand (n = 95) and chest (n = 140) radiographs were read by 14 and 17 human readers, respectively, with varying experience.

Results: Two bone age prediction algorithms were tested on hand radiographs (from January 2017 to January 2022) in 326 patients (mean age, 10 years ± 4 [SD]; 173 female patients) and correlated strongly with the reference standard (r = 0.99; P < .001 for both). No difference in RMSE was observed between algorithms (0.63 years [95% CI: 0.58, 0.69] and 0.57 years [95% CI: 0.52, 0.61]) and readers (0.68 years [95% CI: 0.64, 0.73]). Seven lung nodule detection algorithms were validated on chest radiographs (from January 2012 to May 2022) in 386 patients (mean age, 64 years ± 11; 223 male patients). Compared with readers (mean AUC, 0.81 [95% CI: 0.77, 0.85]), four algorithms performed better (AUC range, 0.86–0.93; P value range, <.001 to .04).

Conclusion: Compared with human readers, four AI algorithms for detecting lung nodules on chest radiographs showed improved performance, whereas the remaining algorithms tested showed no evidence of a difference in performance.

Supplemental material is available for this article.

The market for artificial intelligence (AI) software in radiology is rapidly expanding, with over 200 products currently available in the European Union. Multiple vendors now offer similar solutions. For example, there are already 12 products to detect breast malignancy on mammographs and over 30 for brain region or lesion quantification (1).

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, RMSE = root mean square error

Summary

In independent validation, nine artificial intelligence products for detecting lung nodules on chest radiographs or predicting bone age on hand radiographs showed improved or comparable performance to human readers.

Key Results

- In this retrospective study validating commercial artificial intelligence products, two algorithms for predicting bone age tested on 326 hand radiographs showed no observable difference in root mean square error (0.63 and 0.57 years) compared with human readers (0.68 years).
- Of seven algorithms for detecting lung nodules tested on 386 chest radiographs, four performed better (area under the receiver operating characteristic curve [AUC] range, 0.86–0.93) than human readers (mean AUC, 0.81; *P* value range, <.001 to .04).

This abundance of options can make it challenging for users to determine which software best suits their needs (2). While various factors, such as workflow integration and time efficiency, should be considered when making a purchasing decision, adequate diagnostic performance is a crucial prerequisite for any AI tool to add value in clinical practice.

Transparency around the performance data of commercial AI products is often unsatisfactory (3–6). A recent review found that no scientific evidence on performance measures was available for two-thirds of Conformité Européenne–marked AI products (7) (Conformité Européenne, or CE, is the European Union's mandatory conformity marking). This review established that even when evidence was available, studies were often conducted and/or funded by the vendors themselves, making it difficult to assess the validity of the results (7). The lack of consistency of study protocols and data sets makes it challenging to compare the performance of different AI products (8,9).

Some studies have attempted to directly compare similar AI products for tasks such as detecting breast cancer, classifying breast density, or scoring tuberculosis (10–13). However, these are single snapshot validations, and as AI algorithms are continuously updated and improved, such studies need to be conducted repeatedly to ensure that the results are up-to-date and relevant. Frequently, validation data are shared or made public, which makes them nonreusable thereafter for independent validation, as they may have been used for, for example, retraining. An alternative approach to obtaining performance data is to conduct in-house validation studies of AI products before making a purchase (8,14,15). However, clinical centers rarely have the necessary resources and personnel to evaluate and compare multiple products prior to purchase.

The aim of the Project AIR initiative was to fill this gap in information and provide an independent measure of performance across different algorithms. A methodology was created for conducting validation studies to assess the standalone performance of commercially available Conformité Européenne–marked AI-based software for radiology. The products are tested on representative data sets, and the results are compared with radiologists' performance on the same sets. The data sets remain confidential, which allows the process to be repeated when new products and updated algorithms are brought to market. Test results are made publicly available on the globally used medical image challenge platform Grand Challenge (16). A summary of the Project AIR method is provided in Appendix S1 and Figure S1. It is more extensively described in the Project AIR general study protocol on Zenodo (17). The aim of this study was to determine the feasibility of the Project AIR methodology by performing an independent validation of commercially available AI products for bone age prediction based on hand radiographs and lung nodule detection on chest radiographs.

Materials and Methods

Study Design

Project AIR is an ongoing cohort study in which commercial AI products are externally validated on retrospective data sets from multiple medical centers. A subset of the data used originates from previously published work (18–20). Details can be found in Appendix S1. The study was reviewed by the research ethics committee of Radboud University Medical Center, and the requirement for informed consent was waived because the study data were collected retrospectively. All clinical data were anonymized before being used for analysis. The Grand Challenge platform maintains a list of Conformité Européenne–marked AI software for radiology *(https://grand-challenge.org/aiforradiology/)*. All vendors that had a product on this list between June and November 2022 addressing one or both use cases were invited to join Project AIR (bone age prediction, three vendors; lung nodule detection, 14 vendors). In total, nine products from eight vendors were validated between June 2022 and January 2023. The vendors had no control of the data and information submitted for publication.

To ensure a fair process and to enable reiteration, the validation data were not shared with the vendors. Instead, the vendors made their algorithms temporarily available locally or as an Amazon Machine Image (Amazon Web Services), enabling the execution of the algorithms in an isolated environment. A subset of the data, hereafter referred to as the public subset, was made publicly available under a CC BY license, allowing vendors to test the setup and retrospectively verify the processing. The public subset was not included in the analysis.

Study Sample

A consecutive set of conventional radiographs of the left hand of children (age range, 0–18 years) acquired between January 2017 and January 2022 was collected from seven medical centers (one academic hospital, five teaching hospitals, and one general hospital) in the Netherlands. Images showing extreme deformations of the hand were excluded. Images were obtained with radiographic equipment from multiple manufacturers, including Siemens, Philips, and Canon (Table 1), using local protocols, a voltage of 40–50 kV, and an exposure time of 3–17 milliseconds.

Table 1: Patient Characteristics and Data Inclusion for the Two Use Cases: Bone Age Prediction and Lung Nodule Detection

Characteristic	Hand Radiographs for Bone Age Prediction	Chest Radiographs for Lung Nodule Detection
No. of patients	326	386 (144 nodule cases, 242 controls)
No. of centers	7 (one academic hospital, five teaching hospitals, one general hospital)	7 (three academic hospitals, three teaching hospitals, one general hospital)
Sex		
Male	153 (47)	223 (58)
Female	173 (53)	163 (42)
Age (y)		
Mean ± SD	10 ± 4	64 ± 11
Range	0–18	26–89
Equipment manufacturer		
Siemens	146 (45)	65 (17)
Philips	90 (28)	187 (48)
Canon	88 (27)	14 (4)
Agfa	0	81 (21)
Hologic	0	31 (8)
Other	2 (1)	8 (2)

Note.—Except where noted, data are numbers of patients, with percentages in parentheses.

Chest radiographs, acquired between January 2012 and May 2022, for which matching chest CT scans were available that were obtained within 3 months of the radiograph were collected from seven medical centers (three academic hospitals, three teaching hospitals, and one general hospital) in the Netherlands. Images were obtained with radiographic equipment from multiple manufacturers, including Siemens, Philips, Canon, Agfa, and Hologic (Table 1), using local standardized protocols with automated exposure control and a voltage of 115–150 kV. CT was performed either with or without contrast material, with a section thickness of 3 mm or less. A convenience sample was selected at the discretion of the medical center based on report or registry search. Each patient was classified as either a control or nodule case. Controls were defined as patients with a normal-appearing chest radiograph as confirmed with CT. These control radiographs could contain signs of chronic obstructive pulmonary disease and/or emphysema, or minor bronchopathic changes. According to the Fleischner Society guidelines, nodule cases had to have one or more solid or part-solid nodules with a solid component with diameter measured at CT between 5 and 30 mm, based on the longest axis in the axial, coronal, or sagittal plane. Only one radiograph pair (frontal and lateral) per patient was included. Bedside radiographs, non–diagnostic quality images (eg, motion artifacts, incompletely imaged chest), and radiographs with diffuse or extensive pathology that could obscure nodules (eg, diffuse metastases, interstitial lung disease, large masses or consolidations) were not eligible for this data set. All radiographs and CT scans were reread by a specialized radiologist (S.S., with 8 years of experience) to ensure these eligibility criteria were met.

Table 2 provides a summary of the study design for both use cases. The study protocols for bone age prediction and lung nodule detection are available on Zenodo (21,22).

Reader Methods

For the bone age reference standard, three expert readers (M.J.C.M.R., J.I.M.L.V., and M.V.) who were pediatric or musculoskeletal radiologists (with 26, 23, and 11 years of experience, respectively) rated each image independently according to the Greulich and Pyle method (23). This method uses a bone age atlas according to sex to which readers compare an image to estimate the bone age. Images for which the difference between two of the three readers was greater than 2 years (n = 15) were reread by two other expert readers (B.M. and C.F.v.D., with 24 and 22 years of experience, respectively). The reference standard was then defined as the mean bone age of the five reads. Greulich and Pyle–based bone age predictions from the two algorithms and 14 readers were compared with the reference standard.

All chest radiographs and corresponding CT scans were read by a specialized radiologist (S.S., with 8 years of experience) to determine the reference standard (0, no nodule present; 100, one or more nodules present), measure the most prominent nodule on the CT scan (in millimeters), and provide a conspicuity classification on the basis of visual inspection (well visible, moderately visible, subtle, or very subtle). Lung nodule detection by the seven algorithms and 17 readers was compared with the reference standard.

Radiographs were read by a set of radiologists and radiology residents with varying experience to serve as the comparison for algorithm performance. Tables S2 and S3 provide the individual reader characteristics. Readers were recruited through the AI Network, a database of radiologists interested in AI established in 2019 by our center in collaboration with the Radiological Society of the Netherlands. Participating readers were requested to read a random subset of the images (95 hand radiographs or 140 chest radiographs) using the infrastructure offered by the Grand Challenge platform. This platform allows readers to access imaging data from anywhere and efficiently

Table 2: Summary of Study Design for the Two Use Cases: Bone Age Prediction and Lung Nodule Detection

Characteristic	Hand Radiographs for Bone Age Prediction	Chest Radiographs for Lung Nodule Detection
No. of invited vendors	3	14
No. of participating vendors	2	7
Input	Left hand radiograph	PA and lateral (available for 383 patients) chest radiographs
Outcome	Greulich and Pyle bone age, continuous scale	Probability of any nodule present, score 0–100 for each patient
Reference standard	Bone age based on mean of three expert reads, with two additional expert reads when the deviation was >2 y	Nodule presence (0 or 100) based on expert read of CT scan acquired within 3 mo of the radiograph
No. of readers	14 (8 radiologists, 6 residents)	17 (13 radiologists, 4 residents)
Mean ± SD experience of readers (y)	9.3 ± 9.6	11.5 ± 8.5
No. of radiographs read per reader	95	140
Metrics	Root mean square error, mean error	AUC, sensitivity (optional), specificity (optional)
Statistics	Pearson correlation coefficient	Multireader multicase analysis

Note.—AUC = area under the receiver operating characteristic curve, PA = posteroanterior.

go through images and provide predefined scores. Readers were blinded to the algorithm predictions and clinical data except for patient age and sex. The first five scans were used for practice and were excluded from the analysis.

For bone age prediction, human readers (n = 14) provided an age expressed in years and months, which for analysis was transformed to years in decimal. Algorithms provided a decimal prediction in years.

For lung nodule detection, algorithms and human readers (n = 17) provided a probability score between 0 and 100 for each patient of the likelihood that the patient was a nodule case. When an AI algorithm identified multiple findings, the nodule with the highest score was used. For algorithms that provided a score between 0 and 1, scores were scaled to range from 0 to 100. Algorithms that processed both frontal and lateral images had access to both, as did the human readers.

Statistical Analysis

For bone age prediction, the root mean square error (RMSE) (accuracy) and mean error (bias) were computed. Bland-Altman plots were constructed to provide a visual assessment of agreement between the algorithms and readers. The Pearson correlation coefficient *(r)* was calculated to determine the correlation between the ground truth and the algorithm and reader predictions. Statistical analysis was performed by one author (K.G.v.L.) using Python version 3.7.6 (Python Software Foundation) and Python libraries SciPy 1.7.3 and sklearn 0.23.2. Subanalyses were performed according to sex and age group.

For lung nodule detection, the area under the receiver operating characteristic curve (AUC) was computed. The AUC for the mean of reader performance was computed using the diagonal average. Multireader multicase analysis using iMRMC software (version 4.0.3, 2019; U.S. Food and Drug Administration) based on U-statistics was performed to determine the equivalence of the performance of individual algorithms to that of the human readers (24,25). A two-tailed P value of .05 was considered to indicate a statistically significant difference. When CIs could not be computed using U-statistics because of a negative estimate of variance, the nonparametric maximum likelihood estimate, also included in iMRMC, was used instead.

It was optional for vendors to provide a probability threshold between 0 and 100 above which the image was classified as containing a nodule. The threshold was used to compute sensitivity and specificity as secondary end points and was set at 50 for all readers. The 95% CIs for sensitivity and specificity were computed through bootstrapping.

Subanalyses were performed for different conspicuity levels and nodule size. All control patients were included in addition to the nodule cases in each class.

Results

Study Sample

For bone age prediction, a total of 1050 hand radiographs were collected. Only one radiograph per child was included. Fifty radiographs per center were randomly sampled to form a set of 350 images. Right hand radiographs (n = 4) and radiographs showing extreme deformations of the hand (n = 2) were excluded. A set of 18 random radiographs, stratified over the centers, were set aside as the public subset. The final set used for validation had a total of 326 radiographs (mean patient age, 10 years ± 4 [SD]; 173 girls) (Fig 1).

For lung nodule detection, a total of 561 chest radiographs were collected. After ensuring that the eligibility criteria were met, 404 chest radiographs remained (Fig 1). A set of 18 random radiographs, stratified over the centers, were set aside as the public subset. The final set used for validation consisted of radiographs from 386 patients (mean age, 64 years ± 11; 223

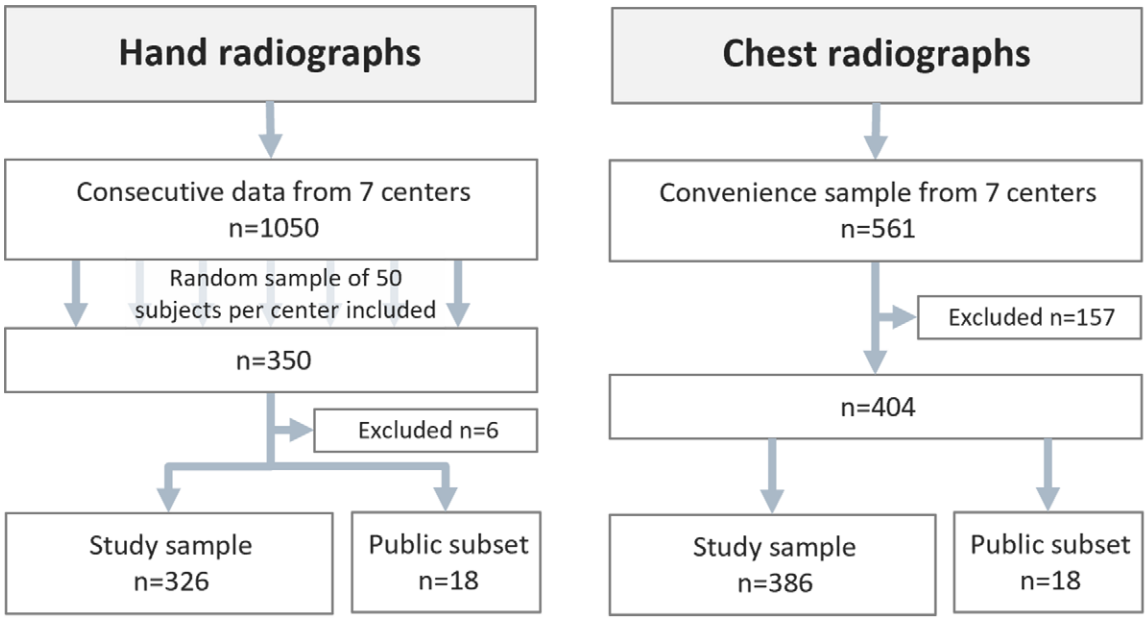

Figure 1: Flow diagrams of the data collection process for bone age prediction (hand radiographs) and lung nodule detection (chest radiographs).

Table 3: Characteristics of Eligible Artificial Intelligence Products for Bone Age Prediction and Lung Nodule Detection

Task and Vendor	Product Name	Response to Invitation	Radiograph Input	Version	Probability Threshold*
Bone age prediction on hand radiographs					
Visiana	BoneXpert	Participated	Left hand	v3.1.4	NA
VUNO	Med-BoneAge	Participated	Left hand	v1.1	NA
ImageBiopsy Lab	PANDA	Declined	NA	NA	NA
Lung nodule detection on chest radiographs					
Annalise.ai	Annalise Enterprise CXR	Participated	PA and lateral	v3.1	Not available
Infervision	InferRead DR Chest	Participated	PA	v1.0.0.1	47
Lunit	INSIGHT CXR	Participated	PA	v3.1.4.4	15
Milvue	Milvue Suite–SmartUrgences	Participated	PA	v1.24	25
Oxipit	ChestEye	Participated	PA	v2.6	Not available
Siemens Healthineers	AI-Rad Companion Chest X-ray	Participated	PA	v9	50
VUNO	Med-Chest X-ray	Participated	PA	v1.1.×	1
Gleamer	ChestView	No response	NA	NA	NA
JLK	JLD-02K	Declined	NA	NA	NA
Quibim	Chest X-Ray Classifier	Declined	NA	NA	NA
Qure.ai Technologies	qXR	Declined	NA	NA	NA
Rayscape	Rayscape CXR	Declined	NA	NA	NA
Riverain Technologies	ClearRead Xray Detect	No response	NA	NA	NA
Samsung Electronics	Auto Lung Nodule Detection	No response	NA	NA	NA

Note.—NA = not applicable, PA = posteroanterior.

* For the lung nodule detection task, it was optional for vendors to provide a probability threshold between 0 and 100 above which the image was classified as containing a nodule.

male patients), of whom 144 had at least one nodule according to the reference standard and were therefore considered nodule cases, and 242 were considered controls. Lateral radiographs were available for 383 patients. Table 1 shows the demographic characteristics of the study sample.

In total, nine commercially available AI products were validated. For the bone age prediction task, two of the three (67%) invited vendors participated. For the lung nodule detection task, seven of the 14 (50%) invited vendors participated. Table 3 shows the invited and participating vendors with their respective

Table 4: Bone Age Prediction Performance on Hand Radiographs by Two Commercial Artificial Intelligence Algorithms and Human Readers

	Visiana BoneXpert		VUNO Med-BoneAge		Mean Prediction of Human Readers (n = 14)	
Sample	RMSE	Mean Error	RMSE	Mean Error	RMSE	Mean Error
Total (n = 326)	0.63 (0.58, 0.69)	0.07 (0.00, 0.14)	0.57 (0.52, 0.61)	0.01 (−0.05, 0.07)	0.68 (0.64, 0.73)	−0.11 (−0.19, −0.04)
Sex						
Girls (n = 173)	0.60 (0.53, 0.67)	−0.06 (−0.15, 0.03)	0.57 (0.51, 0.63)	−0.13 (−0.21, −0.04)	0.69 (0.63, 0.74)	−0.19 (−0.26, −0.12)
Boys (n = 153)	0.67 (0.58, 0.76)	0.22 (0.12, 0.32)	0.56 (0.49, 0.63)	0.16 (0.08, 0.25)	0.67 (0.62, 0.73)	−0.03 (−0.13, 0.07)
Chronologic age (y)						
0–2 (n = 5)	0.21 (0.06, 0.34)	0.06 (−0.08, 0.25)	0.35 (0.18, 0.46)	0.09 (−0.22, 0.34)	0.27 (0.18, 0.37)	−0.12 (−0.25, 0.00)
3–6 (n = 53)	0.61 (0.42, 0.80)	0.26 (0.12, 0.41)	0.53 (0.39, 0.68)	0.25 (0.13, 0.38)	0.59 (0.51, 0.68)	0.03 (−0.10, 0.15)
7–10 (n = 114)	0.74 (0.64, 0.82)	0.13 (−0.00, 0.26)	0.67 (0.60, 0.75)	0.06 (−0.06, 0.19)	0.71 (0.65, 0.78)	−0.19 (−0.29, −0.10)
11–14 (n = 102)	0.56 (0.48, 0.65)	0.02 (−0.09, 0.13)	0.52 (0.45, 0.59)	−0.05 (−0.15, 0.05)	0.73 (0.66, 0.81)	−0.13 (−0.24, −0.02)
15–18 (n = 52)	0.58 (0.47, 0.68)	−0.15 (−0.30, 0.01)	0.44 (0.35, 0.54)	−0.24 (−0.35, −0.14)	0.55 (0.48, 0.62)	−0.02 (−0.09, 0.05)

Note.—Data are means (in years), with 95% CIs in parentheses. RMSE = root mean square error.

products and versions. Herein the individual products are referred to by their vendor's name. All algorithms were able to provide a prediction for each patient.

Bone Age Prediction: Algorithm and Reader Performance

Pearson correlation coefficients *(r)* between the reference standard and Visiana and VUNO were 0.987 (P < .001) and 0.989 (P < .001), respectively. The mean r value for the correlation between the human readers and the reference standard was 0.985 (95% CI: 0.98, 0.99).

RMSE (in years) was similar for Visiana (0.63 [95% CI: 0.58, 0.69]), VUNO (0.57 [95% CI: 0.52, 0.61]), and the human readers (mean, 0.68 [95% CI: 0.64, 0.73]). RMSE was largest for the chronologic age group of 7–10 years, at 0.74 years (95% CI: 0.64, 0.82) for Visiana and 0.67 years (95% CI: 0.60, 0.75) for VUNO. No evidence of a difference in RMSE (in years) between boys and girls was found for the algorithms (Visiana: 0.60 [95% CI: 0.53, 0.67] in girls and 0.67 [95% CI: 0.58, 0.76] in boys; VUNO: 0.57 [95% CI: 0.51, 0.63] in girls and 0.56 [95% CI: 0.49, 0.63] in boys) or the human readers (0.69 [95% CI: 0.63, 0.74] in girls and 0.67 [95% CI: 0.62, 0.73] in boys), as demonstrated in Table 4.

The average mean error (bias, in years) of the algorithms and human readers, as shown in Figure 2, was close to zero and clinically insignificant (Visiana, 0.07 [95% CI: 0.00, 0.14]; VUNO, 0.01 [95% CI: −0.05, 0.07]; human readers, −0.11 [95% CI: −0.19, −0.04]). Both algorithms predicted a more advanced age for boys (mean error [in years]: Visiana, 0.22 [95% CI: 0.12, 0.32]; VUNO, 0.16 [95% CI: 0.08, 0.25]) and a delayed age for girls (Visiana, −0.06 [95% CI: −0.15, 0.03]; VUNO, −0.13 [95% CI: −0.21, −0.04]). The same was observed for the human readers' predictions for girls (mean error [in years], −0.19 [95% CI: −0.26, −0.12]) but not for boys (−0.03 [95% CI: −0.13, 0.07]).

The positive mean error (advanced age prediction, in years) was largest for the chronologic age group of 3–6 years for both Visiana (0.26 [95% CI: 0.12, 0.41]) and VUNO (0.25 [95% CI: 0.13, 0.38]). The negative mean error (delayed age prediction, in years) was largest for the chronologic age group of 15–18 years for both Visiana (−0.15 [95% CI: −0.30, 0.01]) and VUNO (−0.24 [95% CI: −0.35, −0.14]), but not for the readers (−0.02 [95% CI: −0.09, 0.05]). Figure 3 provides several example images from the public subset, with reference values and predicted values from readers and algorithms.

Lung Nodule Detection: Algorithm and Reader Performance

The algorithms and human readers showed a wide performance spread regarding the AUC, as shown in Figure 4. The mean AUC for the readers (n = 17) was 0.81 (95% CI: 0.77, 0.85) (Table 5). Compared with human readers, multireader multicase analysis demonstrated superior performance for Annalise.ai (AUC, 0.90 [95% CI: 0.87, 0.94]; P < .001), Lunit (AUC, 0.93 [95% CI: 0.91, 0.96]; P < .001), Milvue (AUC, 0.86 [95% CI: 0.82, 0.90]; P = .04), and Oxipit (AUC, 0.88 [95% CI: 0.85, 0.92]; P = .005). No evidence of a difference was found between the human readers and the algorithms from Infervision (AUC, 0.79 [95% CI: 0.74, 0.84]; P = .33), Siemens Healthineers (AUC, 0.80 [95% CI: 0.75, 0.85]; P = .60), and VUNO (AUC, 0.84 [95% CI: 0.80, 0.88]; P = .26).

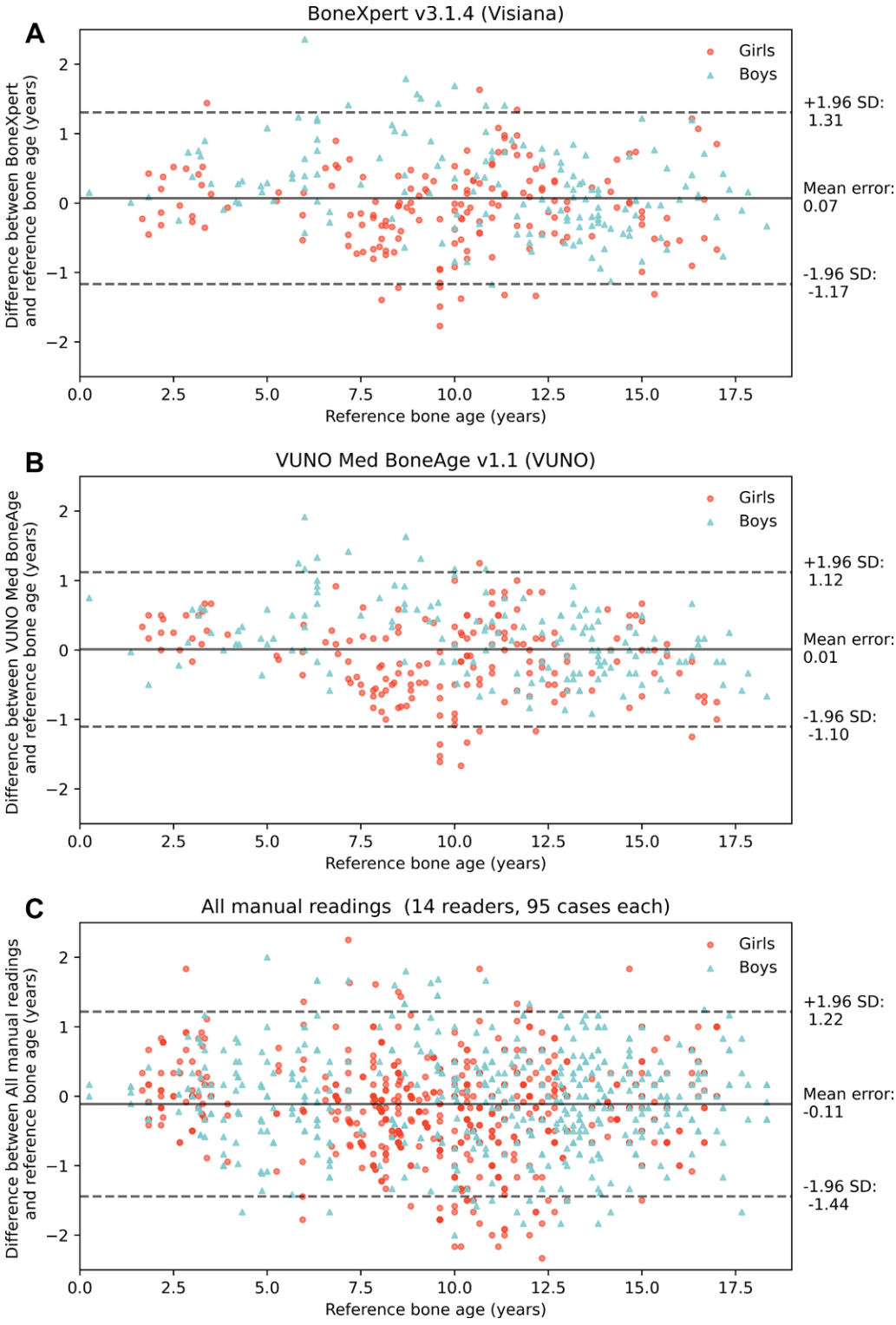

Figure 2: Bland-Altman plots showing bone age prediction by artificial intelligence algorithms and human readers minus the reference bone age as a function of the reference bone age for girls and boys. **(A)** Plot shows the results for Visiana BoneXpert v3.1.4. The average mean error (bias) of Visiana was less than 1 month (0.07 years [95% CI: 0.00, 0.14]) and was considered clinically insignificant. Visiana predicted a more advanced age for boys (mean error, 0.22 years [95% CI: 0.12, 0.32]) and a slightly delayed age for girls (mean error, −0.06 years [95% CI: −0.15, 0.03]). **(B)** Plot shows the results for VUNO Med-BoneAge v1.1. The average mean error (bias) of VUNO was less than 1 month (0.01 years [95% CI: −0.05, 0.07]) and was considered clinically insignificant. VUNO predicted a more advanced age for boys (mean error, 0.16 years [95% CI: 0.08, 0.25]) and a delayed age for girls (mean error, −0.13 years [95% CI: −0.21, −0.04]). **(C)** Plot shows the results for 14 human readers who each read 95 random radiographs from the set. The average mean error (bias) of the readers was −1.3 months (−0.11 years [95% CI: −0.19, −0.04]). The readers predicted a more delayed age for girls (mean error, −0.19 years [95% CI: −0.26, −0.12]). For boys, the bias was close to zero (mean error, −0.03 years [95% CI: −0.13, 0.07]).

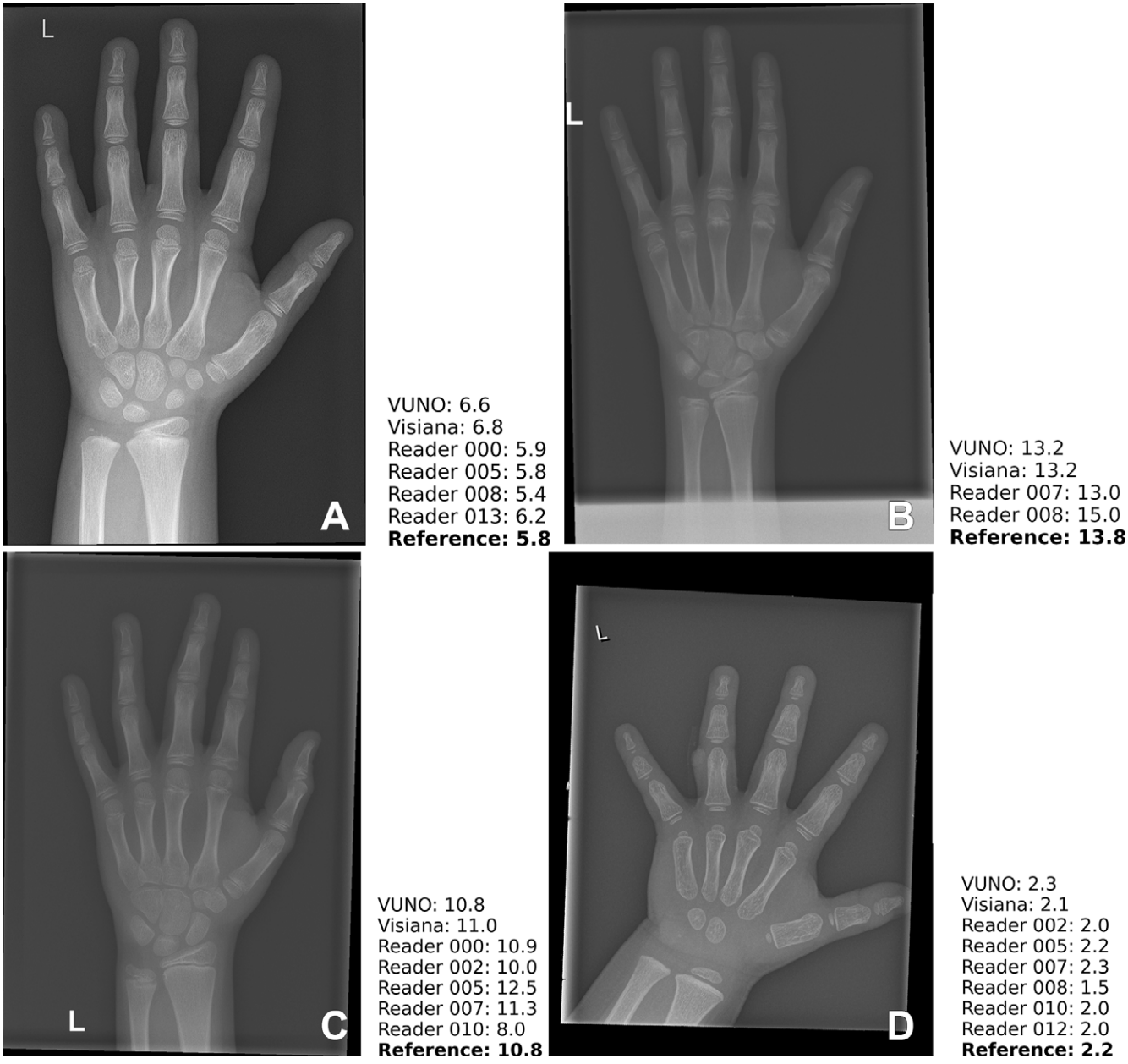

Figure 3: Example left hand radiographs from the public test set illustrate similarities and discrepancies in bone age predictions by artificial intelligence (AI) algorithms and the reference standard. Bone age predictions by human readers are provided for comparison. **(A)** Radiograph in a girl (chronologic age, 5 years) shows a large mean difference between the predictions of the AI products and the reference standard. **(B)** Radiograph in a boy (chronologic age, 17 years) shows a large mean difference between the predictions of AI products and the reference standard. **(C)** Radiograph in a boy (chronologic age, 12 years) shows a small mean difference between the predictions of the AI products and the reference standard. **(D)** Radiograph in a girl (chronologic age, 3 years) shows a small mean difference between the predictions of the AI products and the reference standard. The images shown in this figure were part of a public subset and not part of the set on which metrics are reported, which remains confidential for reevaluation in the future.

The readers showed a mean sensitivity of 71% (102 of 144 patients; 95% CI: 66%, 75%) and a mean specificity of 80% (194 of 242 patients; 95% CI: 73%, 85%) (Table 5). Sensitivity and specificity were optional metrics for the vendors. Infervision, Milvue, Siemens Healthineers, and VUNO showed a higher specificity (83% [95% CI: 79%, 88%], 99% [95% CI: 97%, 100%], 87% [95% CI: 83%, 91%], and 88% [95% CI: 83%, 92%], respectively) than sensitivity (64% [95% CI: 56%, 72%], 50% [95% CI: 42%, 58%], 66% [95% CI: 58%, 74%], and 75% [95% CI: 68%, 82%]), and Lunit showed the opposite (sensitivity, 89% [95% CI: 84%, 94%]; specificity, 80% [95% CI: 75%, 85%]). The full results for the lung nodule detection task are presented in Table 5.

All lung nodule detection algorithms as well as the reader mean showed a performance decline with decreasing nodule conspicuity class. Nodule size showed limited correlation with AUC for most algorithms and the reader mean. Figure 5 provides several example images from the public subset, with reference scores and scores from readers and algorithms. The algorithm scores shown in Figure 5 are raw, uncalibrated scores and cannot

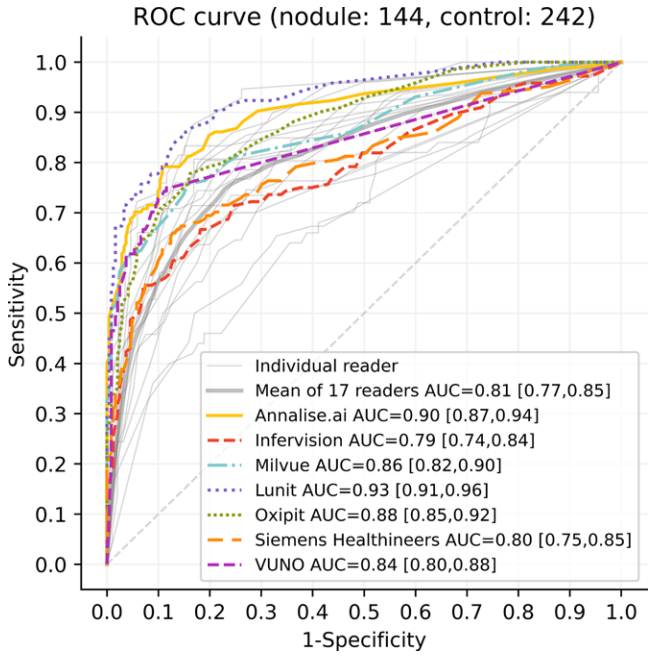

Figure 4: Receiver operating characteristic (ROC) curves for lung nodule detection by seven commercial artificial intelligence products and 17 readers in a data set of 386 chest radiographs from seven centers. The vendor name, area under the receiver operating characteristic curve (AUC), and 95% CI are listed. Compared with human readers, multireader multicase analysis demonstrated a superior performance for Annalise.ai, Lunit, Milvue, and Oxipit. No evidence of a difference in performance was found between the human readers and the algorithms from Infervision, Siemens Healthineers, and VUNO. Product names: Annalise.ai, Annalise Enterprise CXR v3.1; Infervision, InferRead DR Chest v1.0.0.1; Lunit, INSIGHT CXR v3.1.4.4; Milvue, Milvue Suite–SmartUrgences v1.24; Oxipit, ChestEye v2.6; Siemens Healthineers, AI-Rad Companion Chest X-ray v9; VUNO, Med-Chest X-ray v1.1.×. Each reader read a random subset of 140 images (or image pairs where lateral image was also available) of the total set. Receiver operating characteristic curves were generated using iMRMC. The receiver operating characteristic curve for the mean of reader performance was computed using the diagonal average.

be directly compared to each other; they are provided for indicative purposes only.

Individual reader characteristics and performance can be found in Table S2 (bone age prediction) and Table S3 (lung nodule detection). Results of subanalyses according to radiograph acquisition device manufacturer are available in Table S4 (bone age prediction) and Table S5 (lung nodule detection). The main results are shared as leaderboards online on the Grand Challenge platform (bone age prediction, *https://pairboneage22.grand-challenge.org/*; lung nodule detection, *https://pairlungnodulexr22.grand-challenge.org/*), which can be updated with the results of new or updated products.

Discussion

Performance data and fair comparison of commercially available artificial intelligence (AI) products for radiology are often lacking (4,6,7,9). The Project AIR validation process makes it possible to compare the performance of multiple AI products in a head-to-head manner and put the performance of AI products in the context of radiologist performance. The design allows for reiteration of the evaluation as new products or new versions of products are released. The aim of this study was to validate the stand-alone performance of commercially available AI software for bone age prediction and lung nodule detection on multicenter data sets from the Netherlands. For bone age prediction based on left hand radiographs, two of three eligible vendors participated, and the predictions showed excellent correlation with the reference standard (Visiana, $r = 0.987$ [$P < .001$]; VUNO, $r = 0.989$ [$P < .001$]). For lung nodule detection on chest radiographs, seven of 14 eligible vendors participated. The areas under the receiver operating characteristic curve (AUCs) varied substantially for both the algorithms (range, 0.79–0.93) and human readers (range, 0.69–0.91). Four of the AI products for lung nodule detection performed better (AUC range, 0.86–0.93) than readers (mean AUC, 0.81 [95% CI: 0.77, 0.85]; *P* value range, <.001 to .04).

Our results for the bone age prediction algorithms are similar to those presented in the latest studies from Visiana and VUNO. A study on the same product version by Visiana showed an RMSE of 0.61 years in a single manually rated set and 0.45 years in an independent test of 200 radiographs with six readers determining the ground truth (26). A 2017 study from VUNO showed an RMSE of 0.60 years and *r* of 0.992 for an independent test set of 200 images (27). These values are in the same range as our study, which showed RMSE values of 0.63 (95% CI: 0.58, 0.69) and 0.57 (95% CI: 0.52, 0.61) for Visiana and VUNO, respectively.

Regarding lung nodule detection products, there were no comparable studies available for Milvue and Infervision. A study by Seah et al (28) evaluated the performance of Annalise.ai (v1.2.0). Subanalyses showed an AUC of 0.95 for multiple masses or nodules and 0.88 for solitary lung nodules, comparable to the AUC of 0.90 (95% CI: 0.87, 0.94) in our study. A previous study (29) on Lunit (v3.1.2) showed a lower AUC for nodule detection (0.86) than was found in our study (0.93 [95% CI: 0.91, 0.96]). However, a similar gap is seen between the two studies for the mean reader AUC (0.75 vs 0.81), which may indicate a difference in the data distribution. The Oxipit product is intended to autonomously report on normal chest radiographs in cases where it is highly certain of the results. In the last study by the vendor (30), sensitivity was 99.8% at a specificity level of 28%. As other abnormalities were included as well, results are not directly comparable, but the receiver operating characteristic curve from our study confirms that this algorithm is optimized for high sensitivity. Previous results from Siemens Healthineers on its product vary. A study from 2021 (on an unspecified version) that tested the algorithm on 100 images, of which 50 contained a nodule, demonstrated a sensitivity and specificity of 64% and 92%, respectively (31), similar to the findings of our study (sensitivity, 66% [95% CI: 58%, 74%]; specificity, 87% [95% CI: 83%, 91%]). However, another study (on version VA23A) reported 83% for both sensitivity and specificity for the detection of lung lesions (masses and nodules) (32). The difference may be explained by the included sample (consecutive vs convenience), the inclusion of masses, the different threshold applied, and the different version of the product used. A study by Park et al (33) on VUNO reported a higher AUC (0.97) than our study (0.84 [95% CI: 0.80, 0.88]). In the study by Park et al, masses were included,

Table 5: Lung Nodule Detection Performance on Chest Radiographs by Seven Commercial AI Algorithms and Human Readers

Measure and Sample	Annalise.ai	Infervision	Lunit	Milvue	Oxipit	Siemens Healthineers	VUNO	Mean Score of Human Readers (n = 17)
AUC*								
Total (n = 386)	0.90 (0.87, 0.94)	0.79 (0.74, 0.84)	0.93 (0.91, 0.96)	0.86 (0.82, 0.90)	0.88 (0.85, 0.92)	0.80 (0.75, 0.85)	0.84 (0.80, 0.88)	0.81 (0.77, 0.85)
P value of difference[†]	<.001 (0.04, 0.13)	.33 (−0.09, 0.03)	<.001 (0.08, 0.16)	.04 (0.00, 0.10)	.005 (0.02, 0.11)	.60 (−0.07, 0.04)	.26 (−0.02, 0.08)	
Nodule conspicuity[‡]								
Well visible (n = 42)	0.99 (0.98, 1.00)	0.88 (0.81, 0.95)	0.99 (0.97, 1.00)	0.97 (0.94, 0.99)	0.97 (0.95, 1.00)	0.94 (0.90, 0.98)	0.97 (0.95, 1.00)	0.92 (0.88, 0.96)
Moderately visible (n = 43)	0.97 (0.95, 0.99)	0.88 (0.82, 0.94)	0.97 (0.95, 1.00)	0.94 (0.91, 0.98)	0.94 (0.90, 0.98)	0.81 (0.72, 0.90)	0.91 (0.85, 0.97)	0.86 (0.80, 0.92)
Subtle (n = 34)	0.85 (0.77, 0.93)	0.75 (0.65, 0.85)	0.94 (0.90, 0.97)	0.84 (0.76, 0.92)	0.87 (0.81, 0.92)	0.78 (0.70, 0.87)	0.76 (0.67, 0.85)	0.74 (0.70, 0.79)[§]
Very subtle (n = 25)	0.70 (0.59, 0.82)	0.51 (0.38, 0.65)	0.76 (0.66, 0.85)	0.58 (0.46, 0.71)	0.64 (0.53, 0.74)	0.57 (0.42, 0.72)	0.62 (0.51, 0.72)	0.65 (0.56, 0.73)
Nodule diameter (mm)[‡]								
25–30 (n = 14)	0.90 (0.79, 1.00)	0.77 (0.63, 0.91)	0.97 (0.93, 1.00)	0.83 (0.70, 0.96)	0.95 (0.90, 0.99)	0.81 (0.69, 0.93)	0.87 (0.73, 1.00)	0.79 (0.69, 0.90)
20–24 (n = 27)	0.93 (0.88, 0.97)	0.73 (0.60, 0.86)	0.96 (0.91, 1.00)	0.85 (0.75, 0.95)	0.90 (0.84, 0.96)	0.73 (0.59, 0.87)	0.84 (0.74, 0.93)	0.81 (0.76, 0.85)
15–19 (n = 51)	0.90 (0.84, 0.96)	0.79 (0.72, 0.87)	0.92 (0.87, 0.96)	0.88 (0.83, 0.94)	0.86 (0.80, 0.91)	0.79 (0.72, 0.87)	0.85 (0.78, 0.91)	0.80 (0.74, 0.87)
10–14 (n = 41)	0.89 (0.82, 0.96)	0.82 (0.73, 0.91)	0.92 (0.87, 0.97)	0.90 (0.83, 0.97)	0.88 (0.82, 0.95)	0.83 (0.73, 0.92)	0.86 (0.79, 0.94)	0.83 (0.77, 0.89)
5–9 (n = 11)	0.87 (0.70, 1.00)	0.78 (0.57, 1.00)	0.93 (0.86, 0.99)	0.73 (0.51, 0.95)	0.83 (0.70, 0.96)	0.89 (0.77, 1.00)	0.72 (0.54, 0.90)	0.86 (0.73, 0.98)
Sensitivity (%)	NA	64 (56, 72)	89 (84, 94)	50 (42, 58)	NA	66 (58, 74)	75 (68, 82)	71 (66, 75)
Specificity (%)	NA	83 (79, 88)	80 (75, 85)	99 (97, 100)	NA	87 (83, 91)	88 (83, 92)	80 (73, 85)

Note.—Values in parentheses are 95% CIs. The 95% CIs for sensitivity and specificity were computed through bootstrapping. AI products are as follows: Annalise.ai, Annalise Enterprise CXR v3.1; Infervision, InferRead DR Chest v1.0.0.1; Lunit, INSIGHT CXR v3.1.4.4; Milvue, Milvue Suite–SmartUrgences v1.24; Oxipit, ChestEye v2.6; Siemens Healthineers, AI-Rad Companion Chest X-ray v9; VUNO, Med-Chest X-ray v1.1.×. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, NA = not available.

* All AUC-related metrics were computed using iMRMC software based on U-statistics.

[†] P value is for the difference in AUC between the AI algorithm and the human readers. Values in parentheses are the 95% CIs of the difference in AUC between the AI algorithm and the human readers.

[‡] In these subanalyses, all control patients (n = 242) were included in addition to the nodule cases in each class.

[§] CI could not be computed using U-statistics because of a negative estimate of variance. The nonparametric maximum likelihood estimate, also included in iMRMC, was used instead.

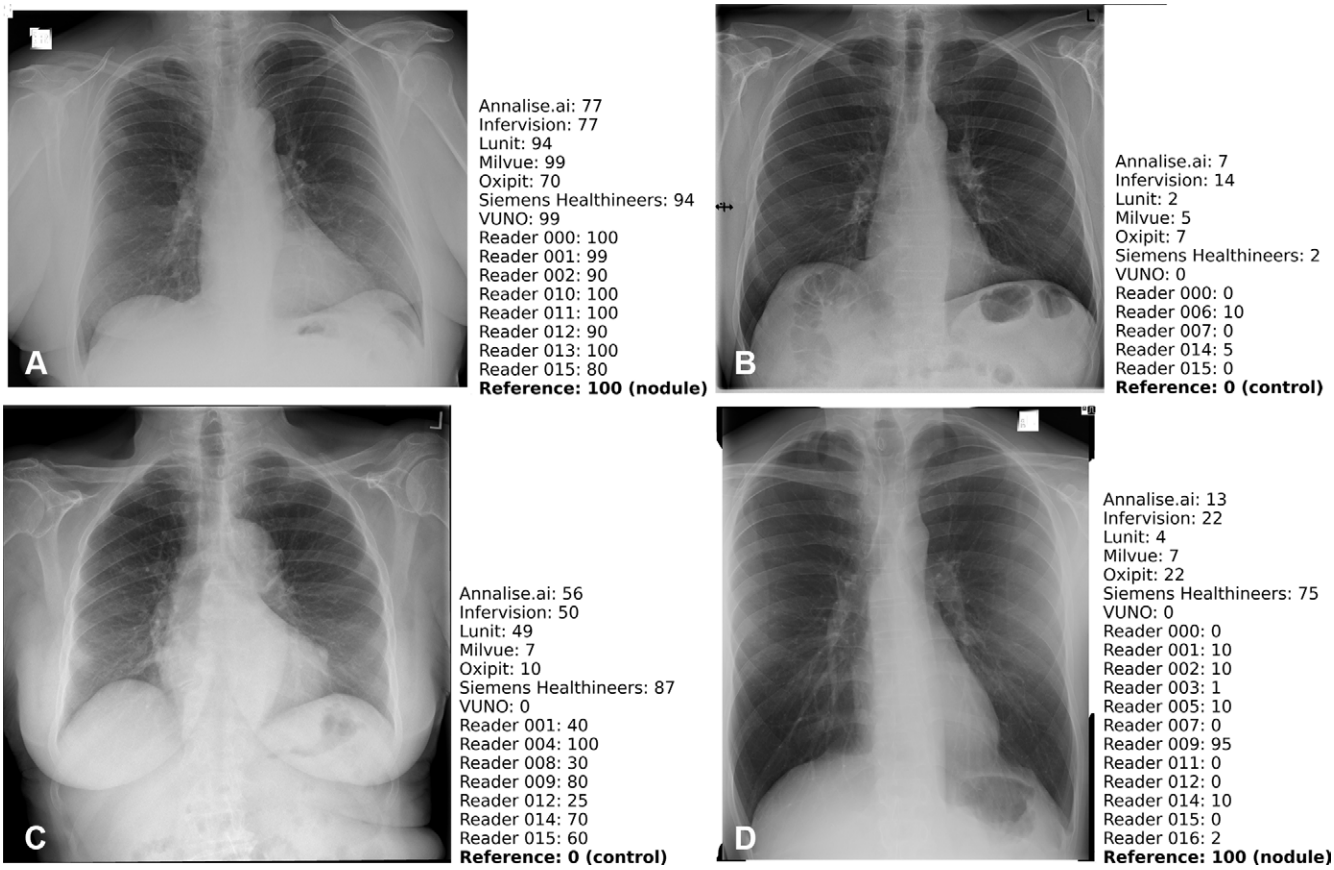

Figure 5: Example chest radiographs (posteroanterior projection) from the public test set illustrate algorithm and reader similarities to and discrepancies from the reference standard. A specialized radiologist determined the reference standard score (0, no nodule present; 100, one or more nodules present), and algorithms and human readers provided a probability score between 0 and 100 for each patient of the likelihood that the patient was a nodule case. **(A)** Radiograph in a man (age, 72 years) with a nodule present (reference standard score, 100) shows a true-positive result based on the average algorithm scores. **(B)** Radiograph in a man (age, 68 years) without a nodule present (reference standard score, 0) shows a true-negative result based on the average algorithm scores. **(C)** Radiograph in a woman (age, 64 years) without a nodule present (reference standard score, 0) shows a false-positive result based on the average algorithm scores. **(D)** Radiograph in a man (age, 37 years) with a nodule present (reference standard score, 100) shows a false-negative result based on the average algorithm scores. Corresponding lateral images and CT scans (when available) for these patients are presented in Figure S2. The images shown in this figure were part of a public subset and not part of the set on which metrics are reported, which remains confidential for reevaluation in the future. Algorithm scores provided for the images are raw, uncalibrated scores and cannot be directly compared to each other; they are provided for indicative purposes only.

and the test set came from the same distribution as the training set, which may have favored the performance.

The literature analysis demonstrates the difficulty of comparing product performance through individual scientific studies. The use of unrevealed benchmark data sets, as in Project AIR, enables more uniform, objective, and repeatable evaluation of commercial products. Nevertheless, our study has several limitations. First, it should be noted that several vendors of eligible products chose not to participate (yet) in our study. The practice of directly comparing AI products and publicly disclosing the results, including the product names, is still relatively uncommon in the field of AI in radiology. The potential impact on a vendor's market traction may have contributed to the reluctance to participate.

Second, the Project AIR methodology assesses the stand-alone performance of AI products, while most commercial products today are intended to be used as decision support tools in clinical practice. We also acknowledge that there are important factors other than stand-alone performance that should be considered when making purchasing decisions, such as additional functionalities, the user interface, integration into the workflow, and service and support. Conducting comparable prospective validation in a decision support setting on a repeated basis would be technically challenging, as well as resource intensive, as it would require integration of all the products in a diverse set of working environments and much more dedicated time from radiologists, which is currently scarce. Therefore, the authors feel this study presents the best available evidence.

Third, the data and reading circumstances differed from those in clinical practice. Images were read on the Grand Challenge platform instead of a picture archiving and communication system. We asked the readers to optimize their reading environment by using a suitable monitor and minimizing ambient light; however, there was no control over this. No clinical information was provided, the task was clearly defined, and the data distribution differed from that in clinical practice to different extents. Reader performance should therefore be regarded

only in the context of the study and cannot be generalized to performance in clinical practice.

In conclusion, we have shown the feasibility of the Project AIR methodology for external validation of commercial artificial intelligence (AI) products in medical imaging. For lung nodule detection on chest radiographs, four AI algorithms showed better performance than human readers, and three AI algorithms showed no evidence of a difference in performance compared with human readers. For bone age prediction, no difference was observed in the performance of the two algorithms tested and human readers. The Project AIR protocol, by allowing repetition and extension to new use cases in the future, can address the dynamic nature of AI products and increase the transparency of the AI market. It is conceivable that in the future, radiology departments will require vendors to participate in transparent and comparative evaluations as a prerequisite for purchasing AI products. Similarly, health care insurers might base reimbursement decisions on performance comparisons. For the vendors of AI products, the data may be useful as postmarket clinical follow-up data to demonstrate regulatory compliance. These factors may ultimately encourage more vendors to participate in such evaluations.

Acknowledgments: The authors thank all participating vendors for their trust and efforts to make Project AIR possible. Visiana, VUNO, Annalise.ai, Infervision, Lunit, Milvue, Oxipit, and Siemens Healthineers provided algorithms free of charge for the purpose of this study only. The vendors had no control of the data and information submitted for publication. We thank Pim de Jong, MD, PhD, for his contribution as data provider.

Author contributions: Guarantors of integrity of entire study, **B.v.G., C.F.v.D., E.L.K., F.M.t.B., J.M.L., M.M.**; 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; agree to ensure any questions related to the work are appropriately resolved, all authors; literature research, **K.G.v.L., S.S., M.J.C.M.R., C.M.S.P., B.v.G., B.M., C.F.v.D., I.A.H.v.d.B., J.M.L., M.M., O.V., P.A.**; clinical studies, **S.S., B.v.G., B.M., B.H.J.G., C.F.v.D., E.L.K., E.V.H., E.L.V., E.M.J.M., F.M.t.B., F.v.H., F.v.d.W., J.M.L., J.H., J.I.M.L.V., L.C.M.L., M.N., M.M., M.V., M.M.V., P.A., S.A., S.M.B., T.S.**; experimental studies, **K.G.v.L., S.S., B.v.G., C.F.v.D., F.A.A.M.H., F.v.d.W., J.M.L., J.M., M.M., M.F.B., P.A.**; statistical analysis, **K.G.v.L., M.J.C.M.R., C.F.v.D., J.M.L., J.H., M.M., P.A.**; and manuscript editing, **K.G.v.L., S.S., M.J.C.M.R., M.H., C.M.S.P., M.d.R., B.v.G., C.F.v.D., E.L.K., E.L.V., F.A.A.M.H., I.A.H.v.d.B., J.M.L., J.N., K.M.E.M., L.N.D., L.C.M.L., M.M., M.F.B., M.M.V., O.V., P.A., Y.H.G.v.B.F.**

Disclosures of conflicts of interest: **K.G.v.L.** Honoraria for educational presentations from Contextflow, and Siemens Healthineers; and support for speaking at the European Congress of Radiology 2023 from Bayer Pharmaceuticals. **S.S.** No relevant relationships. **M.J.C.M.R.** No relevant relationships. **M.H.** Speakers honoraria from Bayer, DeepC, and MedicalPHIT; and board or committee member for EuSoMII, *Radiology: Artificial Intelligence* Editorial Board, ESR eHealth and Informatics Subcommittee, and ECR 2025 Scientific Subcommittee Imaging Informatics/Artificial Intelligence and Machine Learning. **C.M.S.P.** Royalties from Elsevier and Thieme; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from Canon, Boehringer, and Philips; associate editor for *Radiology*. **M.d.R.** No relevant relationships. **B.v.G.** Royalties or licenses from Thirona, Delft Imaging, and MeVis Medical Solutions and stock or stock options from Thirona. **B.M.** No relevant relationships. **B.H.J.G.** Stock or stock options from AMD, NVIDIA, Broadcom, ORACLE, ASML, ASMI, Microsoft, and Super Micro Computer. **C.F.v.D.** No relevant relationships. **E.L.K.** No relevant relationships. **E.V.H.** No relevant relationships. **E.L.V.** No relevant relationships. **E.M.J.M.** No relevant relationships. **F.A.A.M.H.** No relevant relationships. **F.M.t.B.** No relevant relationships. **F.v.H.** No relevant relationships. **F.v.d.W.** No relevant relationships. **I.A.H.v.d.B.** No relevant relationships. **J.M.L.** No relevant relationships. **J.M.** Compute credits from Amazon Web Services. **J.H.** Board member for the Radiological Society of the Netherlands (NVvR). **J.I.M.L.V.** No relevant relationships. **J.N.** No relevant relationships. **K.M.E.M.** No relevant relationships. **L.N.D.** No relevant relationships. **L.C.M.L.** No relevant relationships. **M.N.** No relevant relationships. **M.M.** No relevant relationships. **M.F.B.** No relevant relationships. **M.V.** No relevant relationships. **M.M.V.** No relevant relationships. **O.V.** No relevant relationships. **P.A.** No relevant relationships. **S.A.** No relevant relationships. **S.M.B.** No relevant relationships. **T.S.** No relevant relationships. **Y.H.G.v.B.F.** No relevant relationships.

References

1. Diagnostic Imaging Analysis Group. AI for radiology. Radboud University Medical Center. https://grand-challenge.org/aiforradiology/. Updated 2023. Accessed January 15, 2023.
2. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol 2021;31(6):3786–3796.
3. Beheshtian E, Putman K, Santomartino SM, Parekh VS, Yi PH. Generalizability and bias in a deep learning pediatric bone age prediction model using hand radiographs. Radiology 2023;306(2):e220505.
4. Khunte M, Chae A, Wang R, et al. Trends in clinical validation and usage of US Food and Drug Administration-cleared artificial intelligence algorithms for medical imaging. Clin Radiol 2023;78(2):123–129.
5. Larson DB. Openness and transparency in the evaluation of bias in artificial intelligence. Radiology 2023;306(2):e222263.
6. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27(4):582–584.
7. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021;31(6):3797–3804.
8. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. J Am Coll Radiol 2021;18(3 Pt A):413–424.
9. Tariq A, Purkayastha S, Padmanaban GP, et al. Current clinical applications of artificial intelligence in radiology and their best supporting evidence. J Am Coll Radiol 2020;17(11):1371–1381.
10. Astley SM, Harkness EF, Sergeant JC, et al. A comparison of five methods of measuring mammographic density: a case-control study. Breast Cancer Res 2018;20(1):10.
11. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep 2019;9(1):15000.
12. Qin ZZ, Barrett R, Ahmed S, et al. Comparing different versions of computer-aided detection products when reading chest X-rays for tuberculosis. PLOS Digit Health 2022;1(6):e0000067.
13. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020;6(10):1581–1588.
14. Daye D, Wiggins WF, Lungren MP, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022;305(3):555–563.
15. Panch T, Pollard TJ, Mattie H, Lindemer E, Keane PA, Celi LA. "Yes, but will it work for *my* patients?" Driving clinically relevant research with benchmark datasets. NPJ Digit Med 2020;3(1):87.
16. Diagnostic Imaging Analysis Group. Grand Challenge: a platform for end-to-end development of machine learning solutions in biomedical imaging. Radboud University Medical Center. https://grand-challenge.org/. Updated 2023. Accessed January 15, 2023.
17. van Leeuwen KG, de Rooij M, Rutten MJCM, Schalekamp S, van Ginneken B. Project AIR - general study protocol. Zenodo. https://doi.org/10.5281/zenodo.7573175. Published January 26, 2023. Accessed January 26, 2023.
18. Schalekamp S, van Ginneken B, Meiss L, et al. Bone suppressed images improve radiologists' detection performance for pulmonary nodules in chest radiographs. Eur J Radiol 2013;82(12):2399–2405.
19. Schalekamp S, van Ginneken B, Koedam E, et al. Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. Radiology 2014;272(1):252–261.
20. Schalekamp S, van Ginneken B, Heggelman B, et al. New methods for using computer-aided detection information for the detection of lung nodules on chest radiographs. Br J Radiol 2014;87(1036):20140015.
21. van Leeuwen KG, de Rooij M, Rutten MJCM, van Ginneken B, Schalekamp S. Project AIR - lung nodule detection x-ray study protocol. Zenodo. https://doi.org/10.5281/zenodo.7573186. Published January 26, 2023. Accessed January 26, 2023.

22. van Leeuwen KG, Schalekamp S, de Rooij M, van Ginneken B, Rutten MJCM. Project AIR - bone age prediction hand x-ray study protocol. Zenodo. https://doi.org/10.5281/zenodo.7573224. Published January 26, 2023. Accessed January 26, 2023.
23. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd ed. Stanford, Calif: Stanford University Press, 1999.
24. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. Commun Stat Theory Methods 2009;38(15):2586–2603.
25. Gallas BD. DIDSR/iMRMC. https://github.com/DIDSR/iMRMC. Updated 2023. Accessed January 17, 2023.
26. Martin DD, Calder AD, Ranke MB, Binder G, Thodberg HH. Accuracy and self-validation of automated bone age determination. Sci Rep 2022;12(1):6388.
27. Kim JR, Shim WH, Yoon HM, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am J Roentgenol 2017;209(6):1374–1380.
28. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 2021;3(8):e496–e506.
29. Ahn JS, Ebrahimian S, McDermott S, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022;5(8):e2229289.
30. Plesner LL, Müller FC, Nybing JD, et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology 2023;307(3):e222268.
31. Homayounieh F, Digumarthy S, Ebrahimian S, et al. An artificial intelligence-based chest x-ray model on human nodule detection accuracy from a multicenter study. JAMA Netw Open 2021;4(12):e2141096.
32. Niehoff JH, Kalaitzidis J, Kroeger JR, Schoenbeck D, Borggrefe J, Michael AE. Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. Sci Rep 2023;13(1):3680.
33. Park S, Lee SM, Lee KH, et al. Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. Eur Radiol 2020;30(3):1359–1368.