

Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation


Anna Majkowska, PhD* • Sid Mittal, BS* • David F. Steiner, MD, PhD • Joshua J. Reicher, MD • Scott Mayer McKimney, MS • Gavin E. Duggan, PhD • Krish Eswaran, PhD • Po-Hsuan Cameron Chen, PhD • Yun Liu, PhD • Sreenivasa Raju Kalidindi, MBBS • Alexander Ding, MD, MS • Greg S. Corrado, PhD • Daniel Tse, MD • Shrawya Shetty, MS

From Google Health, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (A.M., S.M., D.F.S., S.M.M., G.E.D., K.E., P.H.C.C., Y.L., G.S.C., D.T., S.S.); Stanford Healthcare and Palo Alto Veterans Affairs, Palo Alto, Calif (J.J.R.); Apollo Radiology International, Hyderabad, India (S.R.K.); and California Advanced Imaging, Novato, Calif (A.D.). Received June 10, 2019; revision requested July 29; revision received October 8; accepted October 14. Address correspondence to S.S. (e-mail: sshetty@google.com).

*A.M. and S.M. contributed equally to this work.

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

See also the editorial by Chang in this issue.

Radiology 2019; 00:1–11 • <https://doi.org/10.1148/radiol.2019191293> • Content code: 

Background: Deep learning has the potential to augment the use of chest radiography in clinical radiology, but challenges include poor generalizability, spectrum bias, and difficulty comparing across studies.

Purpose: To develop and evaluate deep learning models for chest radiograph interpretation by using radiologist-adjudicated reference standards.

Materials and Methods: Deep learning models were developed to detect four findings (pneumothorax, opacity, nodule or mass, and fracture) on frontal chest radiographs. This retrospective study used two data sets. Data set 1 (DS1) consisted of 759 611 images from a multicity hospital network and ChestX-ray14 is a publicly available data set with 112 120 images. Natural language processing and expert review of a subset of images provided labels for 657 954 training images. Test sets consisted of 1818 and 1962 images from DS1 and ChestX-ray14, respectively. Reference standards were defined by radiologist-adjudicated image review. Performance was evaluated by area under the receiver operating characteristic curve analysis, sensitivity, specificity, and positive predictive value. Four radiologists reviewed test set images for performance comparison. Inverse probability weighting was applied to DS1 to account for positive radiograph enrichment and estimate population-level performance.

Results: In DS1, population-adjusted areas under the receiver operating characteristic curve for pneumothorax, nodule or mass, airspace opacity, and fracture were, respectively, 0.95 (95% confidence interval [CI]: 0.91, 0.99), 0.72 (95% CI: 0.66, 0.77), 0.91 (95% CI: 0.88, 0.93), and 0.86 (95% CI: 0.79, 0.92). With ChestX-ray14, areas under the receiver operating characteristic curve were 0.94 (95% CI: 0.93, 0.96), 0.91 (95% CI: 0.89, 0.93), 0.94 (95% CI: 0.93, 0.95), and 0.81 (95% CI: 0.75, 0.86), respectively.

Conclusion: Expert-level models for detecting clinically relevant chest radiograph findings were developed for this study by using adjudicated reference standards and with population-level performance estimation. Radiologist-adjudicated labels for 2412 ChestX-ray14 validation set images and 1962 test set images are provided.

© RSNA, 2019

Online supplemental material is available for this article.

Despite being one of the most common and well-established imaging modalities, chest radiography is subject to significant interreader variability and suboptimal sensitivity for important clinical findings. Recent advances in deep learning offer promise for improving chest radiograph interpretation (1–4) and there are several recent reports of machine learning models achieving radiologist-level performance for different chest radiograph findings (5–7).

A critical aspect of developing clinically relevant diagnostic models involves evaluation in representative test sets with carefully defined ground truth labels. Interreader variability in establishing reference standard image labels can significantly impact performance evaluation (8–14). Previous work in deep learning for radiologic

image analysis has generally used a single-reader or a majority-vote approach across multiple independent readers to provide reference-standard labels (5,6,15). However, because of errors or inconsistencies in the resulting labels, such approaches may lead to overestimation of model performance. For example, challenging but critical findings may be under recognized and thus mislabeled by a majority-vote approach if they are only identified by a minority of the independent readers. This can result in the inability for a model to detect these findings (because of incorrect training labels), and also the inability to measure these errors (because of incorrect reference standard labels), resulting in a false sense of model accuracy. Therefore, the use of more rigorous approaches to generating

Abbreviations

CI = confidence interval, DS1 = data set 1

Summary

Four deep learning models identified pneumothorax, fractures, opacity, and nodule or mass on frontal chest radiographs with similar performance to radiologists.

Key Result

- Deep learning models achieved parity to chest radiography interpretations from board-certified radiologists for the detection of pneumothorax, nodule or mass, airspace opacity, and fracture on a diverse multicenter chest radiography data set (areas under the receiver operative characteristic curve, 0.95, 0.72, 0.91, and 0.86 respectively).

reference standard labels, such as multiphase review (16), expert adjudication, or confirmatory imaging, is critical for high-quality algorithm development and evaluation.

The purpose of this work was to develop deep learning models and evaluate their potential to accurately detect clinically meaningful findings at chest radiography. Additional goals of this work were to underscore the importance of developing and validating diagnostic models by using thoughtfully assembled data sets with reliable reference standards to help standardize comparisons across studies in this growing field. Our image labeling data for ChestX-ray14 (17,18) are also made available for use by other researchers.

Materials and Methods

Data Sets

Institutional ethics committee approvals were obtained from all participating institutions in this retrospective study and all data were deidentified. Two independent data sets were used for model development and evaluation. Data set 1 (DS1) consisted of 759 611 deidentified frontal chest radiographs (digital and scanned) with reports from 538 390 patients (Table 1). This data set consists of all consecutive inpatient and outpatient images in DICOM format obtained from five regional centers across a large hospital group in India (Bangalore, Bhubaneswar, Chennai, Hyderabad, and New Delhi) between November 2010 and January 2018. The second data set was the publicly available data set from the National Institutes of Health (ChestX-ray14) (17,18) and consisted of 112 120 frontal chest radiograph images in 30 805 patients (Table 1). Because DS1 includes all chest radiographs from multiple different hospitals, the abnormalities in this data set reflect the natural population prevalence of different abnormalities in these populations. However, ChestX-ray14 is enriched for various thoracic abnormalities relative to the general population (17,18).

For DS1, patients were randomly assigned to training, validation, or testing sets (Fig 1). In ChestX-ray14, we preserved the original test set of 25 596 images from 2797 patients. The remaining 86 524 images from 28 008 patients were randomly split into training (68 801 images) and validation sets (17 723) (Fig 1). For both data sets, we ensured that images from the same

Table 1: Data and Patient Characteristics

Characteristic	DS1*	ChestX-ray14†
No. of patients	538 390	30 805
Median age (y)‡	50 (1 to >90)	49 (1 to >90)
Sex		
Women	205 762 (38.2)	14 175 (46.0)
Men	332 184 (62)	16 630 (54.0)
Unknown	444 (>0.1)	NA
No. of images	759 611	112 120
AP images	133 876 (17.6)	44 810 (39.9)
PA images	625 735 (82.4)	67 310 (60.0)
Final test set	1818	1962
Images with findings positive for pneumothorax, opacity, nodule or mass, or fracture in the final test set (%)		
Pneumothorax	88 (4.8)	195 (9.9)
Nodule or mass	322 (17.7)	295 (15.0)
Opacity	444 (24.4)	1135 (57.8)
Fracture	257 (14.1)	72 (3.7)
Image resolution range		
Width (pixels)	512–4400	1143–3827
Height (pixels)	512–4784	966–4715

Note.—Unless otherwise indicated, data in parentheses are percentages. AP = anteroposterior, DS1 = data set 1, NIH = National Institutes of Health, NA = not applicable, PA = postero-anterior.

* Data are from five clusters of hospitals from five cities in India

† Data are from the National Institutes of Health Clinical Center (17,18).

‡ Data in parentheses are range.

patient remained in the same split to avoid training and testing on the same patient.

Validation and Test Set Image Selection

To provide a sufficient number of diverse and high-quality labeled images with findings positive for pneumothorax, opacity, nodule or mass, or fracture, we selected approximately 2000 images from both DS1 and ChestX-ray14. Because ChestX-ray14 is already enriched for radiographs positive for pneumothorax, opacity, nodule or mass, or fracture, images were selected at random from the available images. For DS1, images were selected on the basis of radiology reports to enrich with radiographs positive for pneumothorax, opacity, nodule or mass, or fracture while maintaining radiograph diversity and also allowing for population adjustment at analysis by inverse probability weighting (19). Enrichment details are in Appendix E1 (online). Although the radiology reports were used to facilitate enrichment, the reference standard labels for each image were provided by an adjudicated radiologist image review.

Reference Standard Image Annotation

We sought to identify four chest radiographic findings: pneumothorax, opacity, nodule or mass (as a specific sub-

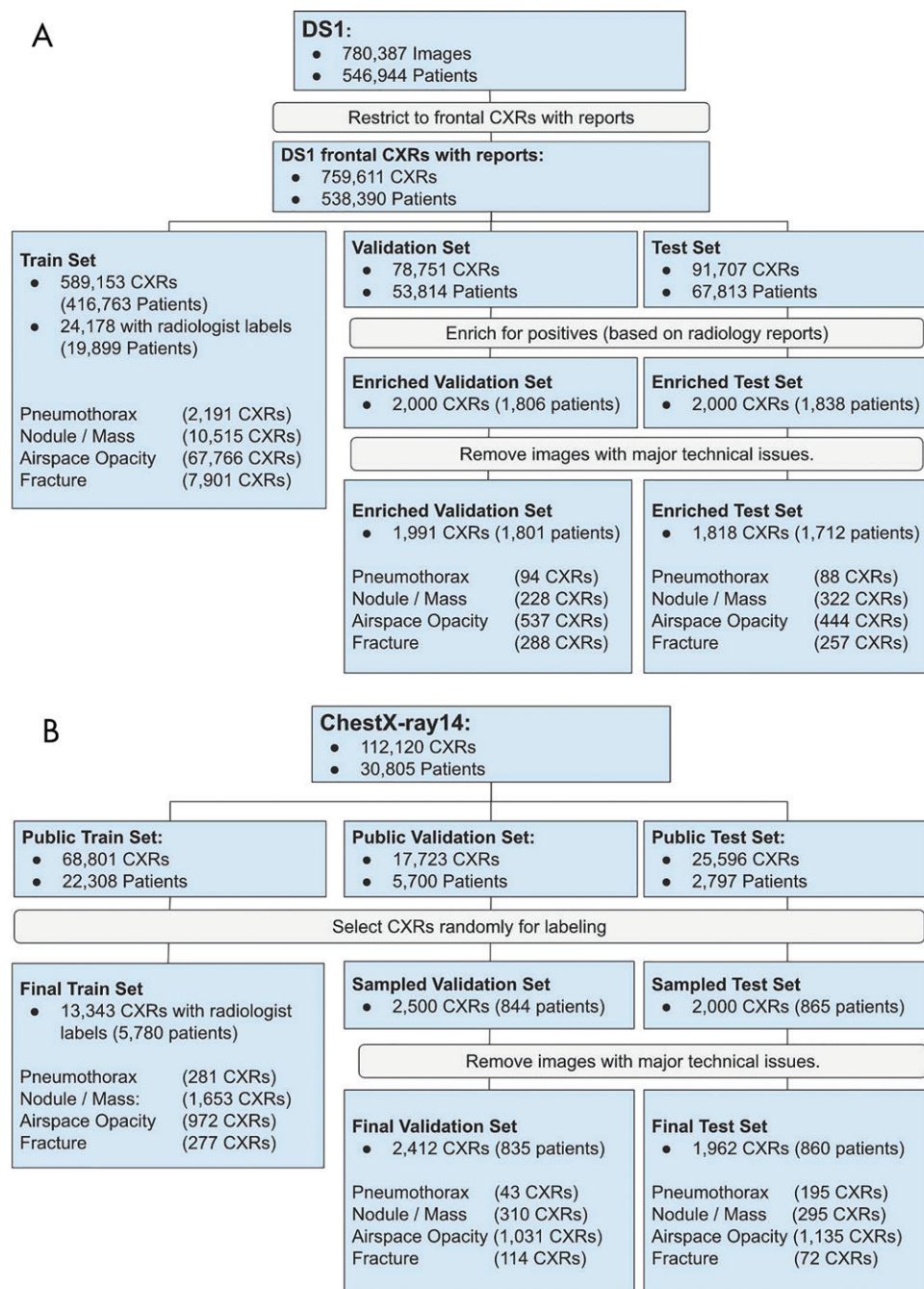


Figure 1: Flowchart of images used from, A, data set 1 (DS1) and, B, ChestX-ray14 data set. The final validation and test sets do not include images found to have technical issues (such as markings or poor image quality) or incomplete baseline reviews. CXR = chest radiograph.

type of opacity), and fracture. Clinical definitions for these categories were on the basis of the Fleischner Society Glossary of Terms for Thoracic Imaging, except for osseous fracture, which was defined as visible rib, clavicle, humeral, or vertebral body fractures (20). For example, a nodule was defined as smaller than 3 cm and mass as 3 cm or larger. The presence or absence of each of these findings were labeled at the image level. Chest tube and fracture acuity labels were also collected to facilitate planned subanalyses by using these labels.

Reference-standard labels for the final validation and test set images were assigned by an adjudicated review by three radiologists (Fig 2, Table E1 [online]). For each image in the test set, three readers were assigned from a cohort of 11 board-certified radiologists (range of experience, 3–21 years in general radiology with no thoracic experts; A.D.). The three readers for each image of the validation set were selected from a cohort of 13 individuals, consisting of both board-certified radiologists (no thoracic experts) and residents (Table E1 [online]). Briefly, images were independently evaluated by three readers and allowing disagreements to be resolved by up to five rounds of asynchronous anonymous discussion by the same readers, but not by enforcing consensus (Appendix E1 [online]). In cases where consensus was not reached, the majority vote was used. All readers had access to the patient age and image view (posteroanterior vs anteroposterior), but not to additional clinical or patient data. Nodule or mass and pneumothorax were adjudicated as present, absent, or hedge (ie, uncertain if present or absent), and opacity and fracture as present or absent. For evaluation, hedge was considered to be a positive result with the rationale that a clinical hedge would prompt additional read, action, and/or clinical follow up.

These expertly adjudicated labels for the ChestX-ray14 data set are provided and include 2412 development set images and 1962 test set images (https://cloud.google.com/healthcare/docs/resources/public-datasets/nih-chest#additional_labels).

Training Set Annotation

To optimize use of the entire DS1 test set for training, images were labeled by using two approaches: expert image annotation and natural language processing (Fig 2; Appendix E1 [online]).

Natural Language Processing Model for Training Set Labels

To label DS1 training images, we developed a natural language processing model to predict image labels from original radiology reports by using approximately 35 000 reports (Appendix E1, Table E2 [online]). Briefly, a one-dimensional deep convolutional neural network (21) was trained and performance was evaluated against human-labeled reports. The train, validation, and test sets for natural language processing model development were subsets of the corresponding data splits used for image modeling.

Model Development

Four separate deep learning models were trained and optimized to distinguish the presence or absence of fracture, nodule or mass, pneumothorax, or opacity, respectively. All models were convolutional neural networks trained with the combined set of training images from both DS1 and ChestX-ray14 training sets. We used Xception (22) as the convolutional neural network architecture that was pretrained on 300 000 000 natural images (23). The models for creating an ensemble were selected on the basis of the area under the precision-recall curve in the validation set. The final models were an ensemble of multiple models trained on the same data set and the final model predictions were calculated as an average of the predictions of the ensemble (Appendix E1 [online]).

Statistical Analysis

Model performance was evaluated by calculating the area under the receiver operating curve by using the per-image model prediction as the decision variable. Planned sub-analyses included radiographs with pneumothorax without chest tubes and radiographs with fracture by acuity. Noninferiority comparison of models with radiologists was an exploratory analysis. Model performance was compared with radiologist performance on the test sets at two operating points: the average radiologist sensitivity and the average radiologist specificity.

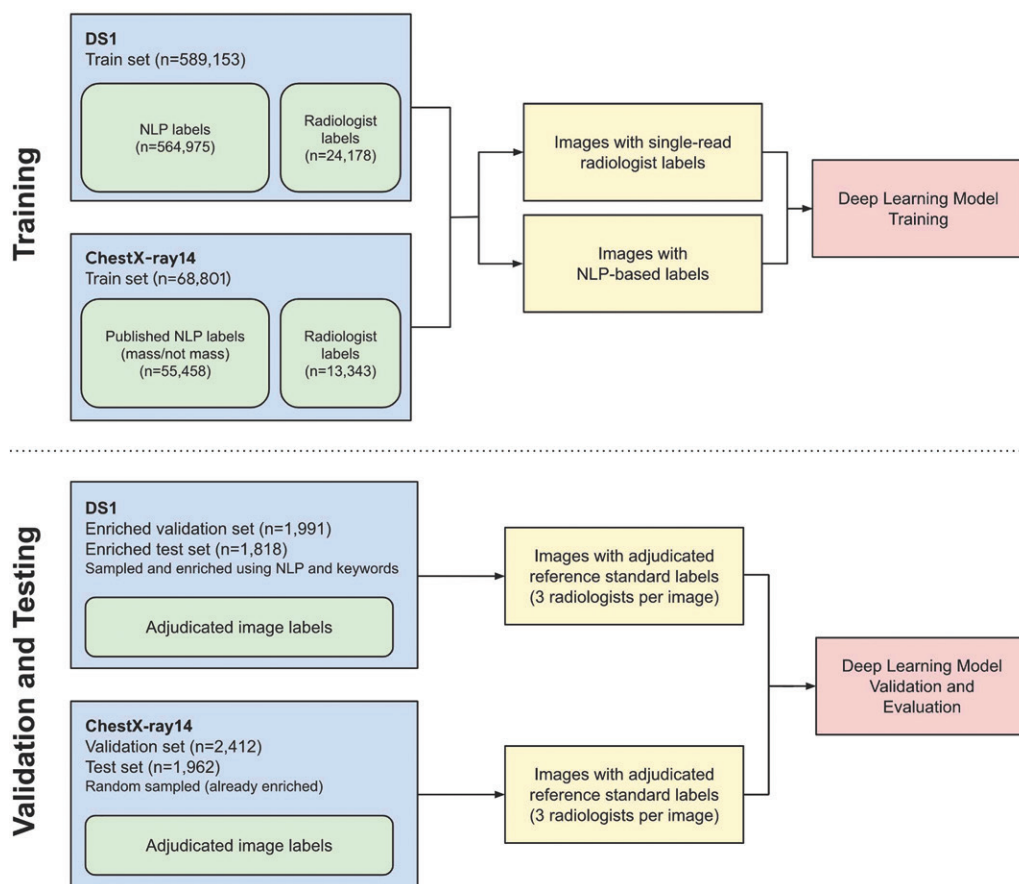


Figure 2: Schematic of labeling strategy used for training and validation and testing in the deep learning models. Training labels were provided by a mix of radiologist image interpretations and natural language processing (NLP) to maximize both high-quality and high-quantity training data. Test set labels were on the basis of adjudicated labels from a panel of three board-certified radiologists per image. Validation set labels were also adjudicated by a panel of three readers, including residents and radiologists. DS1 = data set 1.

To compare our model with radiologists, we collected additional radiologist interpretations for all test set images. All test set images were read independently by four nonthoracic specialty trained American Board of Radiology–certified radiologists (four radiologists for DS1 and four different radiologists for ChestX-ray14). These radiologists did not overlap with the radiologists who contributed to the reference standard labels. No clinical or patient data were available other than age and image view (anteroposterior vs posteroanterior).

To estimate both model and radiologist performance on the original population-level distribution for this particular data set, performance analysis was performed by using inverse probability weighting on the basis of degree of enrichment (known in DS1 but not ChestX-ray14) for each image (Appendix E1, Table E3 [online]).

Model and radiologist performance 95% confidence intervals (CIs) were calculated by using the nonparametric bootstrap method with 1000-fold resampling at the image level. Model performance was compared against radiologists by using the Obuchowski-Rockette-Hillis procedure (24,25). Originally for comparing imaging modalities, this analysis has been adapted for comparison of radiologist performance to that of a stand-alone algorithm (26). For this analysis, the model threshold was

established by using the operating point corresponding to the average radiologist sensitivity (when comparing specificity) and average radiologist specificity (when comparing sensitivity), and binarized agreement (ie, correct vs incorrect) was used for both model and radiologist (27). Noninferiority was assessed by incorporating the margin parameter (5%) into the numerator of the test statistic (28). A small *P* value indicated that the null hypothesis (ie, radiologists perform better than the model by 5% or more) was rejected. The jackknife method was used to estimate the covariance terms for the test.

Results

Test Set Reference Standards

After the adjudication process, the final DS1 test set consisted of 1818 images (88 images were positive for pneumothorax, 322 images were positive for nodule or mass, 444 images were positive for opacity, and 257 images were positive for fracture). The ChestX-ray14 test set consisted of 1962 images (195 images were positive for pneumothorax, 295 images were positive for nodule or mass, 1135 images were positive for opacity, and 72 images were positive for fracture) (Table 1, Table E4 [online]). We also compared the adjudicated labels to the corresponding natural language processing labels for ChestX-ray14 and to the so-called first-round majority vote labels for DS1. Notable differences were observed across conditions, and the adjudicated labels were consistently more sensitive, identifying findings positive for pneumothorax, opacity, nodule or mass, or fracture more than the other methods (Table 2). Example ChestX-ray14 images that were negative for pneumothorax, opacity, nodule or mass, or fracture by natural language processing labels and majority vote labels, but positive by adjudication are in Figure E1 (online).

Model and Radiologist Performance

Results for model and radiologist performance are summarized in Figure 3 and Table 3. For the ChestX-ray14 test set, the models demonstrated area under the receiver operating characteristic curve of 0.94 for pneumothorax (95% CI: 0.93, 0.96), 0.91 for nodule or mass (95% CI: 0.89, 0.93), 0.94 for airspace opacity (95% CI: 0.93, 0.95), and 0.81 for fracture (95% CI: 0.75, 0.86). For DS1, to estimate performance on the population-level distribution, each image was weighted by using inverse probability weighting to account for enrichment. For the population-adjusted analysis of the DS1 test set, the models demonstrated areas under the receiver operating characteristic curve of 0.95 for pneumothorax (95% CI: 0.91, 0.99), 0.72 for nodule or mass (95% CI: 0.66, 0.77), 0.91 for airspace opacity (95% CI: 0.88, 0.93), and 0.86 for fracture (95% CI: 0.79, 0.92).

Sensitivity, specificity, and positive predictive value comparisons were performed between the models and the radiologists (Table 3, Fig E4 [online]). Inverse probability weighting enabled more meaningful estimation of positive predictive values in the unenriched data set for DS1. The average radiologist sensitivity or specificity was used to fix the operating point of the model for these comparisons (ie, model sensitivity was compared with

radiologists at the operating point where the specificity for the model was equal to that of the radiologists). The unweighted areas under the receiver operating characteristic curve for DS1 and areas under the precision-recall curve are provided in Figures E2 and E3 (online).

For all conditions, the model demonstrated performance on par with radiologists. The models trended toward higher-than-average radiologist sensitivity for some findings including fracture for DS1 (absolute increase of 18.6%), pneumothorax for DS1 (absolute increase of 7.3%), nodule or mass for both data sets (absolute increases for DS1 and ChestX-ray14, 5.3% and 2.7%, respectively), and airspace opacity for both data sets (absolute increases for DS1 and ChestX-ray14, 2.9% and 4.4%, respectively). Figure 1 and Table 1 show the total number of radiographs evaluated to calculate these percentages. CIs are shown in Table 3. We also performed noninferiority testing by using a 5% noninferiority margin for each of these metrics (sensitivity, positive predictive value, and specificity). The *P* values for noninferiority comparisons are provided in Table 3. *P* values are provided but significance claims are not made because of the exploratory nature of our analysis involving multiple comparisons.

To provide additional insights into the accuracy of the model compared with the radiologists, we also evaluated the overlap of true-positive findings. This comparison was performed at the model operating point corresponding to the average radiologist specificity, thus keeping the number of false-positive findings fixed. Notably, a substantial number of nonoverlapping true-positive findings were identified. The overall percentage of true-positive findings unique to only the models or the radiologists was 25.1% for ChestX-ray14 and 43.7% for DS1 (Fig 4).

Several radiographs were identified in which the model correctly identified a finding of interest that was missed by the four independent radiologist reviews (*n* = 42). Figure 5a shows examples of these radiographs, with regions that were important for model predictions highlighted. Images with incorrect model predictions were also reviewed by our radiologists at study completion in an effort to identify any clear trends in model errors. Two common error types were noted, including the model highlighting granulomas as nodules (considered false-positive findings in this study design; Fig 5b), and borderline findings. The latter consisted of radiographs adjudicated as hedge or as findings positive for pneumothorax, opacity, nodule or mass, or fracture but missed by the model, and those scored as findings negative for pneumothorax, opacity, nodule or mass, or fracture by the adjudication panel but with possible pneumothorax, opacity, nodule or mass, or fracture manifest at retrospective review.

Evaluation of Condition Subsets

For many radiographic findings, certain subsets of images may be significantly easier or harder to read correctly because of variation in manifestation and/or co-occurring findings. To better understand the model performance on specific subsets of images for different conditions, we performed planned subgroup analyses for the images positive for pneumothorax and fracture.

First, the presence of a chest tube is certainly highly correlated with the manifestation of pneumothorax and may influence

Table 2: Comparison of Test Set Adjudicated Labels versus Natural Language Processing Labels or Majority Vote Labels for ChestX-ray14 and Data Set 1

Parameter	Pneumothorax		Nodule/Mass		Fracture		Airspace Opacity	
	Adjudication Positive	Adjudication Negative	Adjudication Positive	Adjudication Negative	Adjudication Positive	Adjudication Negative	Adjudication Positive	Adjudication Negative
ChestX-ray14: Adjudicated versus NLP								
NLP positive	97	103	134	103
NLP negative	98	1664	161	1564
NLP sens/spec (%)	49.7	94.2	45.4	93.8
DS1: Adjudicated versus majority vote								
Majority positive	77	12	253	38	161	9	392	34
Majority negative	11	1718	69	1458	96	1552	52	1340
Majority vote sens/spec (%)	87.5	99.3	78.6	97.5	62.6	99.4	92.0	96.3

Note.—For the ChestX-ray14 images, data represent the publicly available NLP labels. (Only pneumothorax and nodule or mass were compared because fracture and airspace opacity do not have immediately comparable NLP labels available.) For DS1, the majority vote labels from the initial round of image review are compared with the final adjudicated labels for all conditions. “Adjudication positive” means that adjudication showed findings positive for pneumothorax, opacity, nodule or mass, or fracture. “Adjudication negative” means that adjudication showed findings negative for pneumothorax, opacity, nodule or mass, or fracture. DS1 = data set 1, NLP = natural language processing, sens = sensitivity, spec = specificity.

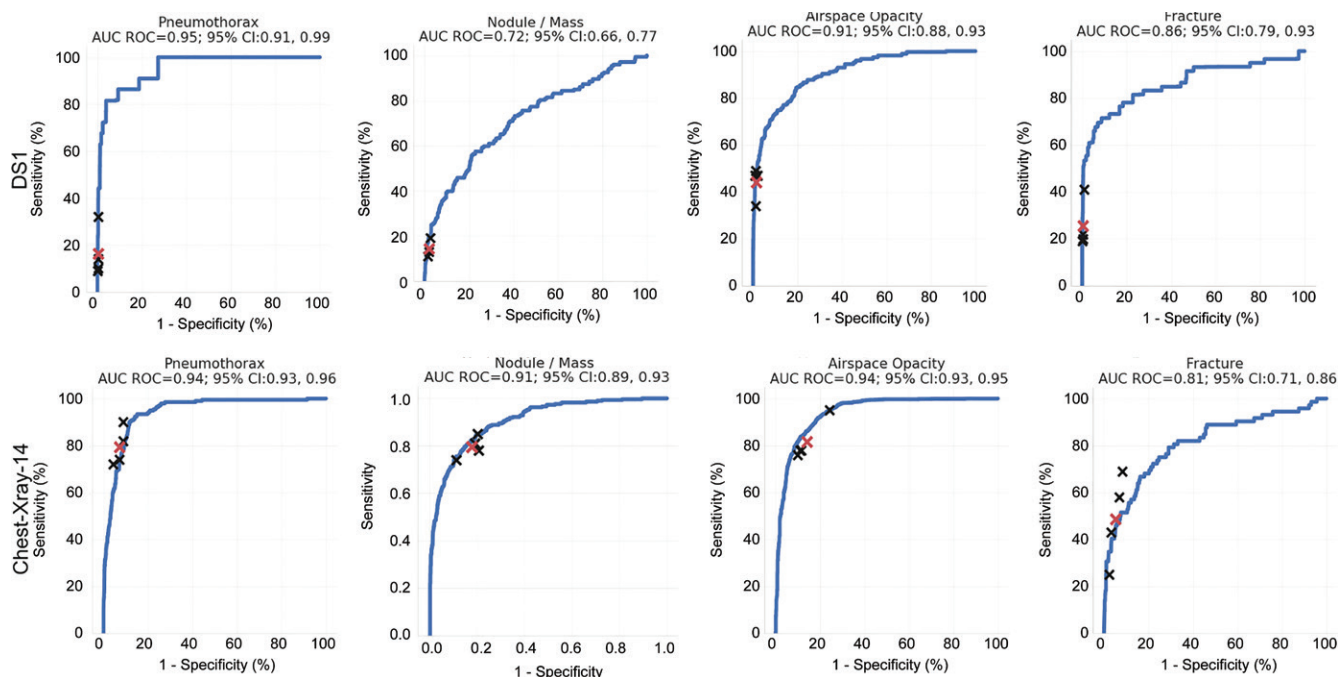


Figure 3: Receiver operating characteristic (ROC) curves of performance evaluation. Performance of the model (blue curves) and individual radiologists (black crosses) across the four findings on the test sets of data set 1 (DS1) and ChestX-ray14. The average reader performance is indicated by red crosses. For DS1, the ROC curves and individual reader operating points represent population-adjusted analysis. Hedge response counts and analysis on a 3-point scale by using these responses are provided in Figure E5 [online]. AUC = area under the curve, CI = confidence interval.

identification of this finding. Thus, we analyzed the subset of images corresponding to untreated pneumothorax (ie, pneumothorax present, but no ipsilateral chest tube). Although only a limited number of radiographs with findings positive for pneumothorax, opacity, nodule or mass, or fracture met this criteria (41 for DS1 and 53 for ChestX-ray14), the model demonstrated

a trend toward lower performance for these untreated pneumothorax radiographs relative to the radiologist average, with absolute decreases in sensitivity of 2.5% and 13.2% for DS1 and ChestX-ray14, respectively (Table E5 [online]).

Images positive for fracture included clavicle, shoulder, rib, and spine fractures and spanned acute, subacute, and chronic

Table 3: Model Noninferiority Performance in Data Set 1 and ChestX-ray14

Performance Metric	Pneumothorax			Nodule or Mass			Airspace Opacity			Fracture		
	Pneumothorax Model	Radiologist	P Value	Difference (Radiologist vs Model)	Nodule or Mass Model	Radiologist	P Value	Difference (Radiologist vs Model)	Opacity Model	Radiologist	P Value	Difference (Radiologist vs Model)
Sensitivity												
CXR14	72.8 (64.1, 81.0)	79.2 (75.5, 83.0)	.61	-6.4 (-18.3, 5.5)	82.4 (77.6, 87.1)	79.7 (76.4, 82.8)	.01	2.7 (-3.9, 9.3)	86.1 (83.3, 88.9)	81.6 (80.1, 83.1)	.06	4.4 (-9.7, 18.6)
DS1	64.8 (47.7, 78.4)	51.7 (43.2, 60.2)	<.001	13.1 (3.4, 22.7)	44.1 (38.2, 50.9)	40.1 (35.9, 44.4)	.02	4.0 (-4.7, 12.6)	53.4 (46.8, 60.8)	55.1 (51.3, 58.9)	.16	-1.7 (-9.2, 5.9)
DS1*	23.4 (6.7, 51.6)	16.1 (8.3, 27.6)	.08	7.3 (-10.9, 25.5)	19.4 (12.3, 26.0)	14.0 (10.3, 18.6)	.001	5.3 (-0.6, 11.3)	47.3 (37.9, 54.5)	44.4 (39.5, 50.0)	.05	2.9 (-6.9, 12.7)
Specificity												
CXR14	90.8 (88.9, 93.1)	92.8 (92.0, 93.7)	.03	-2.0 (-5.2, 1.2)	84.9 (80.3, 89.4)	82.3 (81.1, 83.4)	.02	2.6 (-4.1, 9.3)	89.7 (87.3, 92.4)	85.8 (84.3, 87.4)	.04	3.9 (-6.4, 14.2)
DS1	99.7 (99.3, 100)	99.5 (99.3, 99.7)	<.001	0.2 (-0.1, 0.6)	97.5 (96.1, 98.7)	96.7 (96.2, 97.2)	<.001	0.8 (-0.7, 2.3)	97.6 (96.5, 98.6)	97.9 (97.4, 98.3)	<.001	-0.3 (-1.1, 0.5)
DS1*	99.9 (99.5, 100)	99.8 (99.7, 99.9)	<.001	0.1 (0, 0.3)	98.8 (97.3, 99.5)	98.0 (97.4, 98.5)	<.001	0.8 (0, 1.6)	98.8 (97.8, 99.6)	98.7 (98.2, 99.1)	<.001	0.2 (-0.6, 0.9)
PPV												
CXR14	48.7 (43.8, 55.8)	54.8 (51.8, 58.2)	.70	-6.1 (-11.1, 1.0)	49.2 (42.1, 57.6)	44.3 (42.4, 46.3)	.03	4.8 (-2.2, 13.3)	91.6 (89.8, 93.6)	88.8 (87.7, 89.9)	.005	2.8 (1.0, 4.9)
DS1	90.0 (78.9, 100)	83.5 (77.5, 89.4)	.004	6.5 (-4.5, 16.5)	77.7 (68.4, 88.6)	72.4 (68.4, 76.1)	.001	5.3 (-4.0, 14.2)	88.1 (83.6, 92.8)	89.3 (87.0, 91.5)	.002	-1.2 (-5.8, 3.5)
DS1*	73.8 (21.1, 100)	50.7 (30.8, 71.3)	.16	23.0 (-31.1, 77.0)	54.7 (33.9, 72.5)	42.0 (32.4, 51.9)	.06	12.6 (-1.1, 26.3)	86.8 (78.8, 94.7)	85.4 (81.0, 89.5)	.04	1.3 (-5.8, 8.4)
												24.4)

Note.—Unless otherwise indicated, data are percentages; data in parentheses are 95% confidence intervals. Sensitivity value is computed at average radiologist specificity, and specificity and positive predictive value are computed at average radiologist sensitivity. P values are for model noninferiority test with 5% margin by using Obuchowski-Rockette-Hillis method as described in statistical analysis methods section. CXR14 = ChestX-ray14, DS1 = data set 1, PPV = positive predictive value.

* Population adjusted; DS1 includes weighted analysis to better estimate performance on the natural radiograph mix for this data set.

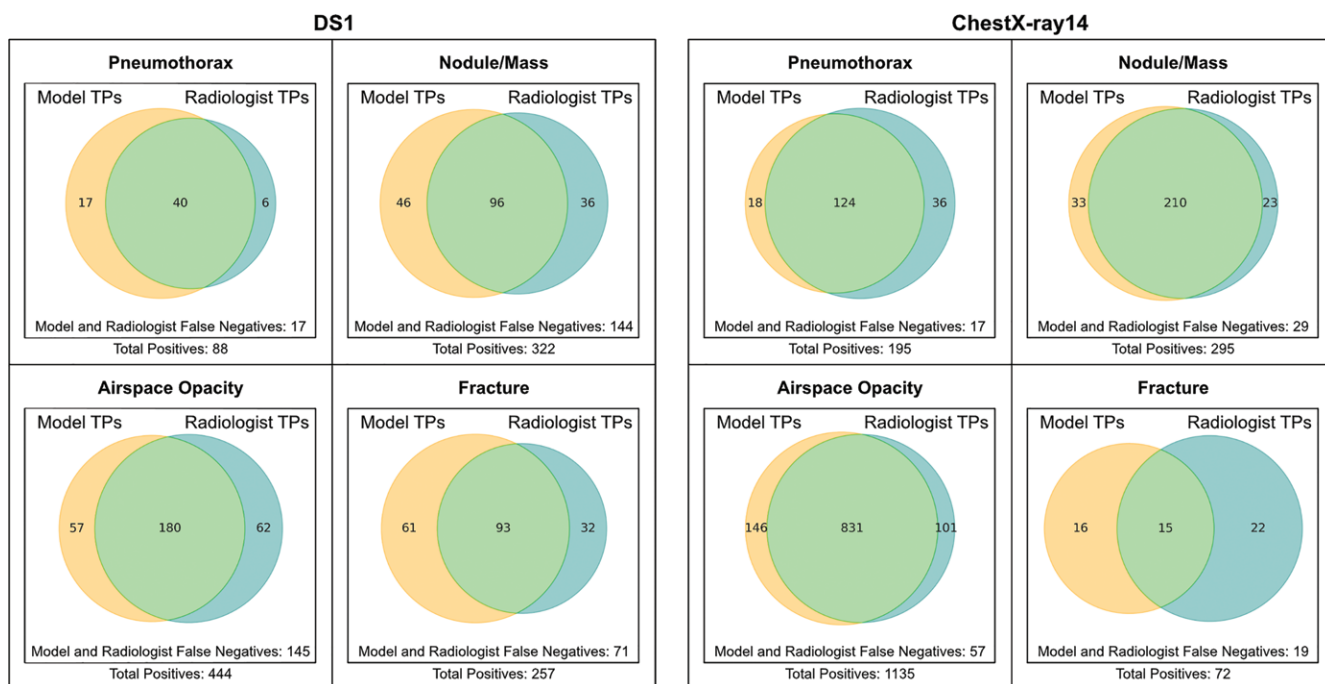


Figure 4: Venn diagrams show comparison of true positives identified by deep learning models and radiologists. For each image, one of the four radiologists' responses was selected randomly to approximate an average response across readers. Data represent comparison of all images with true-positive classifications by the model (yellow) or the radiologist (blue). Nonoverlapping regions thus represent true-positive findings identified by only the radiologists or only the models. DS1 = data set 1, TP = true positive.

radiographs. Given the clinical implications of fracture acuity, we evaluated performance on these subsets. When evaluating just the acute fractures, the model demonstrated higher sensitivity relative to the average radiologist's sensitivity, with a difference of 9.0% (population-adjusted DS1; $n = 65$; Table E6 [online]). For ChestX-ray14, the image set was not curated to include fractures. Therefore, the subset of acute fractures was too small for meaningful analysis ($n = 15$).

Discussion

We developed and validated deep learning models for chest radiograph interpretation by using adjudicated labels as a rigorous reference standard and by using a clinically representative data set to produce more generalizable and comparable results. The models performed on par with board-certified radiologists. In data set 1 (DS1) from five hospitals in India, the model demonstrated population-adjusted areas under the receiver operating characteristic curve of 0.95 (pneumothorax), 0.72 (nodule or mass), 0.91 (opacity), and 0.86 (fracture). This performance was on par with the performance of radiologists. With ChestX-ray14, the models demonstrated areas under the receiver operating characteristic curve of 0.94 (pneumothorax), 0.91 (nodule or mass), 0.94 (opacity), and 0.81 (fracture).

Ground truth in establishing the accuracy of the training sets for artificial intelligence in chest radiography is critical. In our study there was no contemporaneous CT chest examination data to independently confirm the presence or absence of the four abnormalities. Regarding the value of adjudicated labels, extensive work has been performed on the value of multiple interpretations in mammography (30) and other clinical settings (31,32).

Such work supports the value of agreement-based approaches in obtaining accurate diagnostic interpretations. Here, adjudication led to increased expert consensus of the labels used for model tuning and performance evaluation. Adjudication increased the overall consensus from 41.8% (1580 of 3780 images) after the initial read to 96.8% (3660 of 3780 images) (Table E7 [online]). This may in part explain the relatively low sensitivity observed for both radiologists and models across the diverse set of images represented by DS1. There are indeed notable differences in both the labels (Table 2) and the model performance depending on the reference standard labeling strategy, with absolute area under the receiver operating characteristic curve differences of up to 0.05 (ChestX-ray14) and 0.04 (DS1) (Table E8 [online]). By providing the adjudicated labels for publicly available ChestX-ray14 validation and test set images, we hope to facilitate further development, comparison, and evaluation of algorithms for the detection of these key findings.

Whereas area under the receiver operating characteristic curve and other common performance metrics can be useful, reporting these metrics on enriched data sets can fail to reflect expected real-world performance because of issues of prevalence and/or underrepresentation of infrequent yet critical findings. Some metrics reported in the literature might appear high, but additional analysis such as population-adjusted positive predictive value gives a more complete picture in assessing model performance. In this study, although radiologist specificity was greater than 98% across all four findings, population-adjusted positive predictive values ranged from 42% to 85% (DS1; Table E4 [online]), underscoring the importance of considering and reporting prevalence-dependent metrics that adjust for enrichment. This

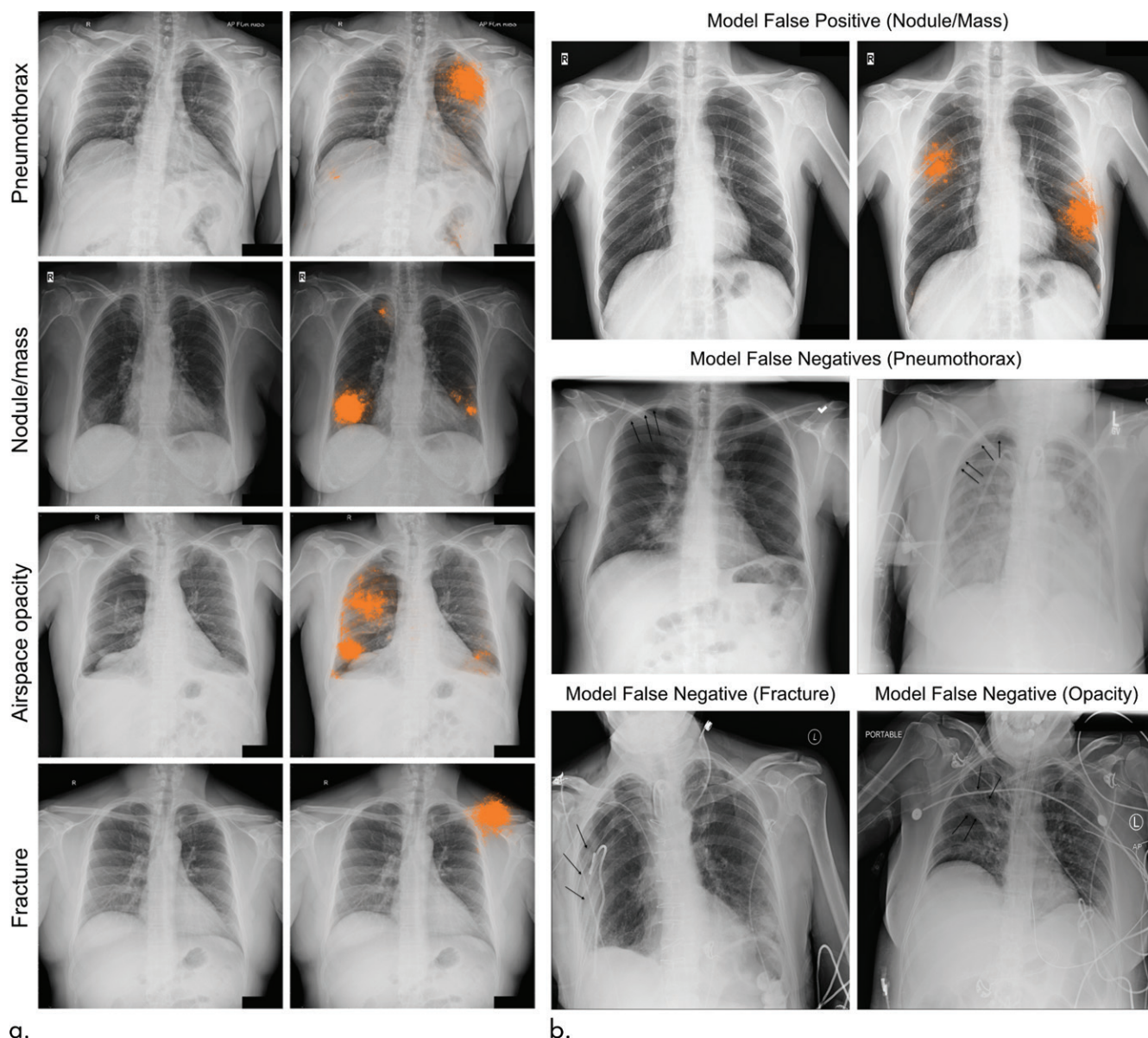


Figure 5: Radiographic images with discordant interpretations between deep learning models and radiologists. **(a)** Examples for the four classes of abnormality (pneumothorax, nodule or mass, airspace opacity, and fracture) classified correctly by the deep learning model (ie, concordant with the adjudicated reference standard), but not identified by any of the four radiologists in the performance comparison cohort. The highlighted areas (orange) indicate the regions with the greatest influence on image-level model predictions, as identified by using SmoothGrad (29). **(b)** Examples for the four classes of abnormality incorrectly classified by the deep learning model (ie, model discordant with the adjudicated reference standard), and correctly classified by at least 2 radiologists in the performance comparison cohort. Top, an example of a granuloma that was classified incorrectly by our model as a nodule. Middle and bottom, additional findings not identified by the model at the threshold corresponding to average radiologist specificity. Black arrows indicate the finding of interest for each radiograph.

work combined enrichment methods that provide representative data sets and evaluation methods with population-based adjustment to improve the thoroughness of the reported performance results.

Diversity among both positive and negative radiographs and in both training and evaluation is a key component of diagnostic model development. By beginning with a broad, hospital-based clinical image set, and then sampling a diverse set of radiographs for expert labeling, we believe the training and evaluation data in this work more accurately represent the spectrum for these conditions than many prior efforts. This was done largely to mitigate the risk of selecting only the radiographs with findings that were

the most obviously positive and negative for pneumothorax, opacity, nodule or mass, or fracture for training, which can fail to address the important challenge of learning to interpret more difficult images.

Data set selection is an important element of machine learning approaches in radiology. Enrichment for pneumothorax, opacity, nodule or mass, or fracture is a common strategy in creating data sets because it can provide requisite examples for training and evaluation with efficient use of labeling resources. However, because such data do not necessarily reflect real-world prevalence or diversity (33,34), such enrichment can also prevent meaningful clinical interpretation of diagnostic performance. Taken

together, issues of enrichment and poor diversity can degrade the meaningfulness of commonly reported performance metrics. In this work, we selected a diverse set of images from a large set of consecutive clinical images across multiple hospitals for evaluation, enabling population-adjusted analysis as a strategy to mitigate spectrum bias and to estimate prevalence-dependent metrics.

We showed that our deep learning models performed on par with radiologists across two independent test sets for four radiographic findings. Whereas the overall performance was comparable, the model identified radiographs that were consistently missed by radiologists and radiologists also identified findings that were missed by the model, suggesting that machine learning–based assistive tools could potentially improve performance over either models or radiologists alone. In addition, we conducted subgroup analysis to evaluate the model performance specifically on the most critically relevant subset of radiographs for pneumothorax (ie, radiographs without chest tubes present) and fracture (ie, acute fracture). Whereas the number of radiographs in these subgroups are small, the lower overall performance by both radiologists and models on these subsets highlights the importance of evaluating and reporting on clinically relevant radiographs. For the untreated pneumothoraces, the lower model performance relative to radiologists suggests the possibility that computer models may be optimized for the detection of chest tubes rather than pneumothoraces themselves, and that reported results for such findings should be considered carefully. For the fracture subanalysis, trends toward performance benefits for acute fracture detection relative to radiologists suggests the potential utility of computer assistance for this particular category of findings, though more work remains to further improve sensitivity.

We acknowledge the limitations of this work. First, the main limitation was lack of ground truth labels that would otherwise be provided at CT. Establishing reference standard annotations with additional clinical information such as discharge diagnosis, confirmation at CT, or clinical outcome would provide more accurate labels. Second, the data assembled for this work represented a broad spectrum of clinical diversity but were not all encompassing. In this regard, we noted differences in performance for the same findings across our two different data sets, likely stemming from case-mix differences. For example, the ChestX-ray14 images were not labeled for fracture or fractures and thus were likely to include a different spectrum of findings for this condition. Additionally, this study did not evaluate the models by using external data sets that were completely independent from those used for model training and did not establish the model prediction thresholds that would be optimal for specific clinical settings. Testing of these models on additional external data sets, prospectively, with predetermined operating points, and in specific clinical settings, is needed. Third, some pneumothorax, opacity, nodule or mass, or fractures were identified only by the model and others only by the radiologists, and therefore we acknowledge that future work is warranted to explore an assisted read application because this hybrid approach may be the best way to implement artificial intelligence in chest radiography.

In conclusion, we developed and evaluated clinically relevant artificial intelligence models for chest radiograph interpretation that performed similar to radiologists by using a diverse set of images. The population-adjusted performance analyses reported here along with the release of adjudicated labels for the publicly available ChestX-ray14 images can provide a useful resource to facilitate the continued development of clinically useful artificial intelligence models for chest radiography.

Acknowledgments: The authors thank the members of the Google Health Radiology team for software infrastructure support and logistical support. Sincere appreciation also goes to the radiologists who enabled this work with their image interpretation and annotation efforts throughout the study. Jonny Wong, BA, coordinated the imaging annotation work. Diego Ardila, PhD, and Zvika Ben-Haim, PhD, helped with modeling work. Rory Sayres, PhD, and Shahar Janshy, MS, contributed to data processing and analysis. Shabir Adeel, BS, and Mikhail Fomitchev, MS, assisted with releasing the adjudicated labels. Quang Duong, PhD, William Chen, BA, and Sahar Kazemzadeh, BS, contributed to labeling infrastructure. Akinori Mitani, MD, PhD, reviewed the paper.

Author contributions: Guarantors of integrity of entire study, A.M., S.M., K.E., G.S.C., S.S.; 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, A.M., D.F.S., J.J.R., G.E.D., P.H.C.C., Y.L., S.R.K., A.D., S.S.; clinical studies, J.J.R., S.R.K., A.D.; experimental studies, A.M., S.M., S.M.M., G.E.D., K.E., P.H.C.C., A.D., G.S.C., D.T., S.S.; statistical analysis, A.M., S.M., D.F.S., S.M.M., G.E.D., K.E., P.H.C.C., Y.L., S.S.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: A.M. disclosed no relevant relationships. S.M. disclosed no relevant relationships. D.F.S. Activities related to the present article: disclosed that Google funded this work. Activities not related to the present article: disclosed money paid to author for employment by Google. Other relationships: disclosed no relevant relationships. J.J.R. Activities related to the present article: disclosed money paid to author for consultant fees; writing or reviewing the manuscript; and provision of writing assistance, medicines, equipment, or administrative support from Google. Activities not related to the present article: disclosed money paid to author for manuscript preparation and travel/accommodations/meeting expenses from Google. Other relationships: disclosed no relevant relationships. S.M.M. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed patent pending for Google related to portions of this work. G.E.D. disclosed no relevant relationships. K.E. disclosed no relevant relationships. P.H.C.C. disclosed no relevant relationships. Y.L. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed payment to author for employment from Google. Other relationships: disclosed no relevant relationships. S.R.K. disclosed no relevant relationships. A.D. disclosed no relevant relationships. G.S.C. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for employment and stock/stock options from Google. Other relationships: disclosed patent pending related to this work. D.T. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed salary and stock from Google. Other relationships: disclosed patent pending related to this work. S.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for salary from Google. Other relationships: disclosed no relevant relationships.

References

1. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019;49(4):939–954.
2. McBee MP, Awan OA, Colucci AT, et al. Deep Learning in Radiology. *Acad Radiol* 2018;25(11):1472–1480.
3. Auffermann WF, Gozansky EK, Tridandapani S. Artificial Intelligence in Cardiothoracic Radiology. *AJR Am J Roentgenol* 2019 Feb 19:1–5 [Epub ahead of print] <https://doi.org/10.2214/AJR.18.20771>.
4. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triage of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* 2019;291(1):196–202.

5. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15(11):e1002686.
6. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. *PLoS Med* 2018;15(11):e1002697.
7. Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology* 2019;290(1):218–228.
8. Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psychophys* 2010;72(5):1205–1217.
9. Raykar VC, Yu S, Zhao LH, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Canada, June 14–18, 2009. New York, NY: ACM, 2009; 889–896 <https://doi.org/10.1145/1553374.1553488>.
10. Potchen EJ, Cooper TG, Sierra AE, et al. Measuring performance in chest radiography. *Radiology* 2000;217(2):456–459 <https://doi.org/10.1148/radiology.217.2.r00nv14456>.
11. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *J Natl Cancer Inst* 1998;90(23):1801–1809.
12. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. *J Med Imaging Radiat Oncol* 2012;56(2):173–178.
13. Rosenkrantz AB, Duszak R Jr, Babb JS, Glover M, Kang SK. Discrepancy Rates and Clinical Impact of Imaging Secondary Interpretations: A Systematic Review and Meta-Analysis. *J Am Coll Radiol* 2018;15(9):1222–1231.
14. Pinto A, Caranci F, Romano L, Carrafiello G, Fonio P, Brunese L. Learning from errors in radiology: a comprehensive review. *Semin Ultrasound CT MR* 2012;33(4):379–382.
15. Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* 2019;2(3):e191095.
16. Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 2011;38(2):915–931.
17. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 3462–3471.
18. Summers RM. NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories. <https://nihcc.app.box.com/v/ChestXray-NIHCC/file/220660789610>. Accessed May 2019.
19. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;352:i189.
20. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2008;246(3):697–722.
21. Conneau A, Schwenk H, Barrault L, Lecun Y. Very Deep Convolutional Networks for Text Classification. *arXiv [cs.CL]*. <http://arxiv.org/abs/1606.01781>. Published 2016. Accessed October 23, 2019.
22. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 1800–1807.
23. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, October 22–29, 2017. Piscataway, NJ: IEEE, 2017; 843–852 <https://doi.org/10.1109/ICCV.2017.97>.
24. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Commun Stat Simul Comput* 1995;24(2):285–308.
25. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat Med* 2007;26(3):596–619.
26. Chakraborty DP. Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples. Boca Raton, Fla: CRC, 2017.
27. Chen W, Wunderlich A, Petrick N, Gallas BD. Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing. *J Med Imaging (Bellingham)* 2014;1(3):031011.
28. Chen W, Petrick NA, Sahiner B. Hypothesis testing in noninferiority and equivalence MRM ROC studies. *Acad Radiol* 2012;19(9):1158–1165.
29. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. *arXiv [cs.LG]*. <http://arxiv.org/abs/1706.03825>. Published 2017. Accessed October 23, 2019.
30. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. *Radiology* 2018;287(3):749–757.
31. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. *JAMA Netw Open* 2019;2(3):e190096.
32. Krause J, Gulshan V, Rahimy E, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* 2018;125(8):1264–1272.
33. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286(3):800–809.
34. Park SH. Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. *Radiology* 2019;290(1):272–273.