UJEAS

UNIZIK JOURNAL OF ENGINEERING AND APPLIED SCIENCES

UNIZIK Journal of Engineering and Applied Sciences 6(5), December (2026), 4254-4265
Journal homepage: https://journals.unizik.edu.ng/index.php/ujeas
PRINT ISSN: 2992-4383 || ONLINE ISSN: 2992-4391

Local Data Adaptation of a Deep Learning Model for Pneumonia Detection: A Controlled Study Using Chest Radiographs from a Nigerian Hospital

Nkiru Celine Mokwe*
Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra State
Prof. K. A. Akpado
Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra State
Prof. A. N. Isizoh
Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Anambra State

Corresponding Author's E-mail: nknokwe@gmail.com

Abstract
Deep learning models for pneumonia detection that are trained on public chest radiograph collections lose accuracy on radiographs from African hospitals, and the usual recommendation is to add local data. How little local data is enough to change a model's behaviour, and what exactly it changes, has rarely been measured. This paper reports a controlled, paired test at very small scale. A DenseNet121 classifier was trained on 13,891 public radiographs (RSNA and Kermany collections) and evaluated on radiographs photographed from printed films at a hospital in Awka, Nigeria. It was then retrained with 74 of those local radiographs added, each presented ten times per epoch so that local images formed about 5% of every epoch, and with greyscale input to remove a colour cast that tracked the class label. Both versions were scored on the same 33 held-out local radiographs. For the deployed models, pneumonia sensitivity rose from 6 of 19 (31.6%) to 12 of 19 (63.2%) (exact McNemar test, $p = 0.031$), and the same gain was found when the output head and decision rule were held identical across the two versions (5 of 19 to 11 of 19, $p = 0.031$). Neither result survives a Bonferroni correction for the five comparisons made. Specificity fell from 13 of 14 to 9 of 14 ($p = 0.125$) and the area under the ROC curve rose from 0.688 to 0.741 (DeLong test, $p = 0.49$). Balanced accuracy changed by 0.015 (95% confidence interval −0.145 to 0.175). Accuracy on two public test sets of 2,688 and 2,862 radiographs was unchanged. Part of the gain came from the greyscale step alone. A small number of local radiographs therefore moved the model towards detecting pneumonia on local images at no cost elsewhere, but did not demonstrably improve its overall discrimination, and sensitivity remained far below the 95% for which the decision threshold had been calibrated. Local radiographs should be used to calibrate the threshold as well as to train the model.
Keywords: Pneumonia detection, chest radiograph, local data, domain shift, external validation, low-resource setting, DenseNet121, McNemar test.

1 Introduction

Chest radiography is the principal imaging test for pneumonia, but its interpretation depends on radiologists, who are scarce in much of sub-Saharan Africa. Idowu and Okedere (2020) reported 250 to 300 practising radiologists in Nigeria, about 1.5 per million population. Deep learning classifiers trained on large public collections of chest radiographs have been proposed as decision support for settings of this kind (Rajpurkar et al., 2017; Kermany et al., 2018).

The difficulty is that such classifiers do not travel well. Zech et al. (2018) showed that a pneumonia classifier performed worse at hospitals other than those that supplied its training data, and Togunwa et al. (2025) found that a model reaching 86% accuracy on open-source paediatric radiographs reached 58% on radiographs from a Nigerian hospital. The remedy usually proposed is local data. Togunwa et al. (2025) called for locally relevant datasets, and studies elsewhere have shown that training or fine-tuning on local radiographs improves performance at the local site (Chamveha et al., 2020; Mohn et al., 2023).

Those studies had local data in abundance: the Thai study of Chamveha et al. (2020) drew on 421,859 local radiographs. A district or private hospital in Nigeria is in a different position. Its radiographs may exist only as printed films, its records may yield about a hundred labelled examinations, and every image used for training is

one fewer available for testing. For such a hospital the practical question is not whether local data helps in principle, but whether a few dozen radiographs, added to a large public training set, change anything that can be measured.

This paper answers that question for one hospital. A pneumonia classifier was trained on public data alone and then retrained with 74 local radiographs added, and the two versions were compared on the same 33 held-out local radiographs. The contributions are as follows.

(1) A paired comparison of two versions of one classifier, trained without and with a small set of local radiographs, analysed with exact paired tests on radiographs that neither version was trained on.
(2) A separation of the effect of the local training data from that of a greyscale conversion step, introduced at the same time to remove a colour cast in the photographed films that was correlated with the class label.
(3) Evidence that the gain on local radiographs carried no cost on two public test sets, and that it did not extend to the decision threshold, which remained calibrated for the public data.

The novelty of the study lies in the combination of three features that have not been brought together in earlier work. The first is the quantity of local data, which is deliberately small: 74 radiographs, about 0.5% of the public training split, against the 4,510 local radiograph studies used for fine-tuning by Mohn et al. (2023) and the 421,859 local radiographs of Chamveha et al. (2020). The second is the paired design, in which the same held-out radiographs are scored by the model trained without and with the local images, so that every change can be traced to individual radiographs and tested exactly, rather than inferred from two accuracy figures obtained on different test sets. The third is the treatment of photographed films. The study shows that a colour cast in the photographs is correlated with the diagnostic label, a potential shortcut of the kind described by DeGrave et al. (2021), and removes it before training, whereas earlier work on photographed radiographs treated photography only as a source of image degradation (Phillips et al., 2020; Kuo et al., 2021).

The gap in the literature that this study covers is therefore specific. It is known that models trained on public data lose accuracy on Nigerian radiographs (Togunwa et al., 2025), that large local datasets improve performance at the local site (Chamveha et al., 2020; Mohn et al., 2023), and that a fixed model can be recalibrated with a modest number of labelled radiographs from a new site (Nazim-E-Alam, 2026). It has not been reported whether a few dozen local radiographs, added to training at a Nigerian hospital and tested with paired statistics on held-out radiographs from the same hospital, change the behaviour of the model, or which part of that behaviour they change. The innovative aspect of the study is that it answers this question at the scale at which a typical Nigerian hospital can supply labelled data, that it separates the effect of the local images from that of a label-correlated artefact of photographing films, and that it shows that adapting the weights of a model does not adapt its decision threshold.

2.0 Related Work

2.1 The Generalisation Gap

Zech et al. (2018) trained pneumonia classifiers on radiographs from three hospital systems and found that performance at an external hospital was often lower than at the training hospitals. They also found that the networks could identify which hospital an image came from, which allowed site-specific cues to stand in for disease. Togunwa et al. (2025) provide the closest precedent for the present setting. Their model reached an accuracy of 86% and an area under the ROC curve (AUC) of 0.93 on its internal test data, and 58% and 0.65 on 190 radiographs from a Nigerian hospital.

2.2 Adaptation with Local Data

Chamveha et al. (2020) developed a thoracic disease classifier with 421,859 chest radiographs from Thailand and reported that incorporating local data improved its accuracy. Mohn et al. (2023) fine-tuned a model that screens chest radiographs for 14 labels on data from their own institution and reported that local data enhanced its performance. Nazim-E-Alam (2026) approached the problem from the other side. A DenseNet121 ensemble trained on paediatric radiographs from one country kept an AUC of 0.74 to 0.80 on radiographs from two others, but at its frozen decision threshold sensitivity fell to 6.2% and 0%. Recalibrating the output with 163 labelled radiographs from the target site restored sensitivity to 88.3%, at a specificity of 47.9%. That study recalibrated a fixed model and did not add target-site images to training.

2.3 Photographed Radiographs

Where radiographs exist only as printed films, they reach a model as photographs. Phillips et al. (2020) released photographs of chest radiographs to measure the effect of the artefacts this introduces, and Kuo et al. (2021) proposed a recalibration method so that models trained on digital radiographs could be applied to radiographs captured with a smartphone. These studies treat photography as a source of degradation. They do not address the possibility, examined here, that the photographs of one class differ systematically from those of the other.

In summary, local adaptation has been demonstrated with large local datasets and with post-hoc recalibration. The effect of adding a very small number of local radiographs to training, tested on held-out radiographs from the same hospital with paired statistics, has not been reported for a Nigerian setting.

3.0 Materials and Methods

3.1 Study Design

The study compares two versions of one system (Figure 1). Version A was trained on public radiographs alone, with colour input. Version B was trained on the same public radiographs with 74 local radiographs added, and with every image converted to greyscale. The backbone, classification head, loss function, training schedule, model-selection procedure, threshold-calibration procedure and random seed were the same in both. Both versions were scored on the same 33 local radiographs, which neither was trained on, so that every comparison is paired. Because Version A was trained with colour input, it was scored on those radiographs twice, with colour input as originally deployed and with greyscale input, to separate the effect of the local training data from that of the greyscale step. Every public image is already grey, so the greyscale conversion leaves the public training and validation images unchanged. Retraining Version A with greyscale input would therefore present it with exactly the same images, and Version A scored with greyscale input serves as the greyscale-matched baseline against which the effect of the local training data is measured.

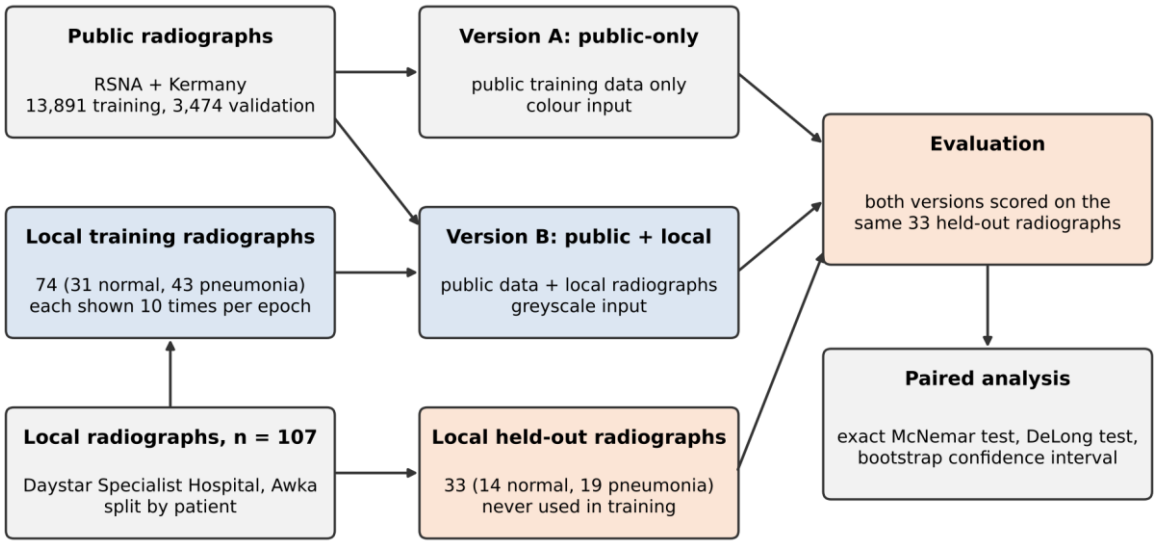


Figure 1: Study design. Both versions are evaluated on the same 33 held-out local radiographs.

3.2 Public Data

Table 1 lists the data. The primary public source is the RSNA Pneumonia Detection Challenge dataset (Shih et al., 2019), from which radiographs labelled Normal and Lung Opacity were taken as the NORMAL and PNEUMONIA classes. Its training portion was combined with the paediatric collection of Kermany et al. (2018) and divided 80:20, grouped by patient, into training and validation splits. Two public test sets were used: a primary test partition of 2,688 RSNA radiographs, and 2,862 radiographs from the NIH ChestX-ray14 database (Wang et al., 2017), whose labels are derived from radiology reports.

3.3 Local Radiographs

The local collection comprises 107 chest radiographs from Daystar Specialist Hospital, Awka, Anambra State, Nigeria: 45 NORMAL and 62 PNEUMONIA, one radiograph per patient. Each label was taken from the hospital's clinical record for the examination. The radiographs were not re-read by a radiologist for this study, so an unknown share of the labels may be in error. The images were captured by photographing the printed films and were de-identified before use.

The collection was divided by patient, stratified by class, into 74 radiographs for training (31 NORMAL and 43 PNEUMONIA) and 33 held out for testing (14 NORMAL and 19 PNEUMONIA), 30% of each class. The assignment of every image was written to a file when the split was first made and was never redrawn, so that no radiograph used in training could later enter the held-out set. The 33 held-out radiographs were not used for training, validation, model selection or threshold calibration.

Table 1: Data used by the two versions

Partition	Source	NORMAL	PNEUMONIA	Total	Role	Used by
Training split	RSNA + Kermany	7,094	6,797	13,891	Model fitting	A and B
Local training images	Daystar Hospital, Awka	31	43	74	Model fitting (×10 per epoch)	B only
Validation split	RSNA + Kermany	1,769	1,705	3,474	Early stopping; threshold calibration	A and B
Local held-out test	Daystar Hospital, Awka	14	19	33	Paired comparison	A and B
Primary test	RSNA	1,328	1,360	2,688	Check for loss on public data	A and B
Online held-out test	NIH ChestX-ray14	1,431	1,431	2,862	Check for loss on public data	A and B

3.4 Adding Local Radiographs to Training

The 74 local radiographs are few beside the 13,891 public images in the training split, about 0.5% of it, and would carry almost no weight if each were seen once per epoch. Each local training image was therefore repeated ten times in every epoch, so that local images made up 740 of the 14,631 images presented per epoch, about 5%. Training-time augmentation was applied independently to every presentation, so the ten copies of an image were not identical when the model saw them. The local images were added after the public data had been divided into training and validation splits. The validation split therefore contains public images only, and is the same in both versions.

3.5 Greyscale Conversion

The photographed films carry a faint colour cast, and its direction differs between the classes, the NORMAL images leaning towards magenta. A simple index of the cast, the mean of the red and blue channels minus the green channel, separated the two classes across the 107 local images with an AUC of 0.67, without any information about the lungs. A model trained on these images in colour could learn the cast in place of the radiographic findings, and held-out images photographed in the same way would not reveal that it had done so. In Version B every image was therefore converted to greyscale by averaging its three channels, in training, in evaluation and in deployment. The public images are already grey and are unchanged by this step.

3.6 Model and Training

Both versions use DenseNet121 (Huang et al., 2017) pretrained on ImageNet, followed by global average pooling, dropout at a rate of 0.3 and a two-unit output layer. Images were resized to 224 × 224 pixels. Training followed a two-phase schedule: five epochs with the backbone frozen at a learning rate of 0.001, then up to 30 epochs with the top 25% of the backbone unfrozen at a learning rate of 0.00001, with early stopping on validation loss. The loss was a class-balanced focal loss (Lin et al., 2017) with a focusing parameter of 2.0, the batch size was 32, and the random seed was 42.

Each version was trained with two alternative output activations, sigmoid and softmax, and one of the two was selected for deployment by a composite score that combines balanced accuracy on the two public test sets with accuracy under simulated image degradation. The score uses no local data. It selected the softmax head in Version A and the sigmoid head in Version B, by margins of 0.0008 and 0.0026. Results are reported for both heads and, separately, for the head deployed in each version. Because the deployed heads differ, comparisons in which the head is the same in both versions are reported alongside the deployed comparison, so that the effect of the local data can also be read with the architecture held identical. Table 2 summarises the differences between the versions.

Table 2: Differences between the two versions

Setting	Version A	Version B
Public training images	13,891	13,891
Local training images	None	74, each ×10 per epoch
Images presented per epoch	13,891	14,631
Input	Colour	Greyscale
Deployed head	Softmax	Sigmoid
Calibrated threshold, sigmoid	0.4606	0.4307
Calibrated threshold, softmax	0.4934	0.4280

3.7 Decision Rules

Two decision rules were applied. The first is the conventional highest-probability (argmax) rule. The second is a threshold calibrated for screening: among all thresholds at which sensitivity on the validation split was at least 0.95, the one giving the highest specificity was selected. Because the validation split contains public images only, the thresholds of both versions were calibrated without any local radiograph. Results at the calibrated threshold of the deployed head describe what the system does in use.

3.8 Statistical Analysis

Proportions are reported with 95% Wilson score confidence intervals. Because both versions were scored on the same 33 radiographs, comparisons between them are paired. For sensitivity, specificity and accuracy, each radiograph was classed as correctly or incorrectly reported by each version, and the exact McNemar test was applied to the discordant radiographs, those reported correctly by one version and incorrectly by the other. AUC values were compared with DeLong's test for correlated ROC curves. The change in balanced accuracy was given a 95% confidence interval by bootstrap resampling of radiographs within each class, with 20,000 resamples. A difference is described as statistically significant when $p < 0.05$. Five comparisons are reported, and the effect of a Bonferroni correction (a threshold of $p < 0.01$) is stated. The tests used per-radiograph predictions recomputed from the saved models of both versions, which reproduce the recorded confusion matrices exactly.

3.9 Ethical Considerations

The public collections are de-identified datasets released for research. The local radiographs were acquired in the course of routine care before the study began, none was requested for research, and they were supplied de-identified with a class label only. No patient was contacted.

4.0 Results

4.1 The Public-Only Model on Local Radiographs

Version A was first evaluated on all 107 local radiographs, none of which it had been trained on. Under the argmax rule the sigmoid head reached an accuracy of 62.62% and the softmax head 60.75%, with AUC values of 0.758 and 0.754. These summary figures conceal a one-sided failure. The sigmoid head detected 25 of the 62 pneumonia cases (40.3%, 95% confidence interval 29.0 to 52.7) and the softmax head 27 (43.5%, 31.9 to 55.9), while clearing 42 and 38 of the 45 normal radiographs. At its calibrated threshold of 0.4934, chosen to give at least 95% sensitivity on public validation data, the deployed softmax head still detected only 27 of 62. The model was not guessing. It was systematically reluctant to report pneumonia on local radiographs.

4.2 Paired Comparison on the Held-Out Radiographs

Table 3 reports every condition on the 33 held-out radiographs, and Figure 2 plots sensitivity and specificity at the calibrated thresholds. For the models actually deployed, sensitivity rose from 6 of 19 (31.6%, 15.4 to 54.0) in Version A to 12 of 19 (63.2%, 41.0 to 80.9) in Version B. Specificity fell from 13 of 14 (92.9%, 68.5 to 98.7) to 9 of 14 (64.3%, 38.8 to 83.7), and accuracy rose from 19 to 21 of 33. Figure 3 shows the corresponding confusion matrices. Under the argmax rule, which removes the influence of the threshold, both heads of Version B detected 11 of 19 pneumonia cases, against 5 and 6 for the two heads of Version A with colour input.

Table 3: Performance on the 33 held-out local radiographs (14 NORMAL, 19 PNEUMONIA)

Training data	Input	Head	Decision rule	Accuracy	Balanced accuracy	Specificity	Sensitivity	AUC
Public-only (A)	Colour	Sigmoid	Argmax	57.58%	0.6316	14/14 (100.0%)	5/19 (26.3%)	0.7068
Public-only (A)	Colour	Sigmoid	Calibrated (0.4606)	54.55%	0.5959	13/14 (92.9%)	5/19 (26.3%)	0.7068
Public-only (A)	Colour	Softmax	Argmax	57.58%	0.6222	13/14 (92.9%)	6/19 (31.6%)	0.6880
Public-only (A)	Colour	Softmax*	Calibrated (0.4934)	57.58%	0.6222	13/14 (92.9%)	6/19 (31.6%)	0.6880
Public-only (A)	Greyscale	Sigmoid	Argmax	57.58%	0.6128	12/14 (85.7%)	7/19 (36.8%)	0.7143
Public-only (A)	Greyscale	Sigmoid	Calibrated (0.4606)	54.55%	0.5677	10/14 (71.4%)	8/19 (42.1%)	0.7143
Public-only (A)	Greyscale	Softmax	Argmax	57.58%	0.5940	10/14 (71.4%)	9/19 (47.4%)	0.7105
Public-only (A)	Greyscale	Softmax	Calibrated (0.4934)	57.58%	0.5940	10/14 (71.4%)	9/19 (47.4%)	0.7105
Public + local (B)	Greyscale	Sigmoid	Argmax	63.64%	0.6466	10/14 (71.4%)	11/19 (57.9%)	0.7406
Public + local (B)	Greyscale	Sigmoid*	Calibrated (0.4307)	63.64%	0.6372	9/14 (64.3%)	12/19 (63.2%)	0.7406
Public + local (B)	Greyscale	Softmax	Argmax	63.64%	0.6466	10/14 (71.4%)	11/19 (57.9%)	0.7744
Public + local (B)	Greyscale	Softmax	Calibrated (0.4280)	69.70%	0.6898	9/14 (64.3%)	14/19 (73.7%)	0.7744

*Note: * marks the model deployed in each version.*

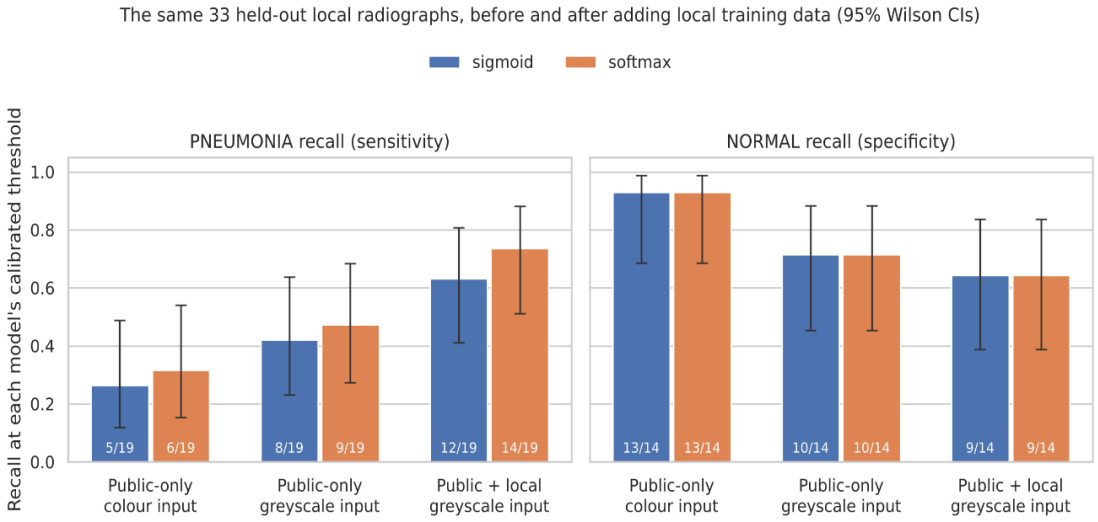


Figure 2: Sensitivity and specificity on the 33 held-out local radiographs at each model's calibrated threshold, with 95% Wilson confidence intervals.

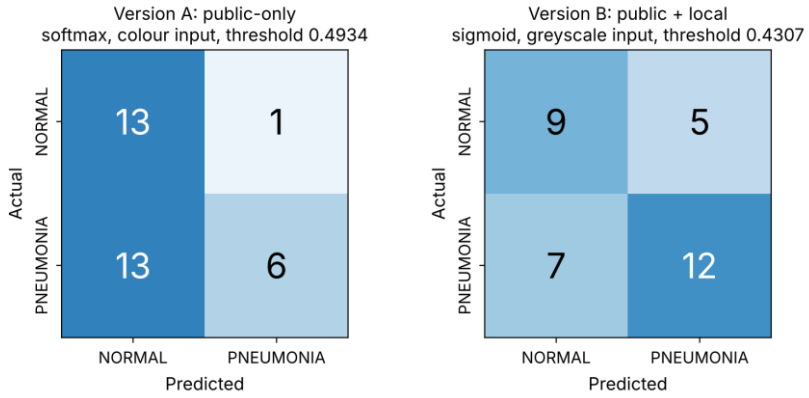


Figure 3: Confusion matrices of the deployed models on the 33 held-out radiographs: Version A (left) and Version B (right).

4.3 Paired Tests

Table 4 sets out the paired tests. The six pneumonia cases newly detected by the deployed model of Version B had all been missed by Version A, and no case detected by Version A was missed by Version B. With six discordant radiographs all favouring one version, the exact McNemar test gives $p = 0.031$. The same one-directional pattern holds in all five comparisons in Table 4: in none of them was a previously detected pneumonia case lost.

The other changes are not statistically significant. The fall in specificity for the deployed models, from 13 to 9 of 14, gives $p = 0.125$. The rise in accuracy, from 19 to 21 of 33, gives $p = 0.754$. The AUC rose from 0.688 to 0.741 ($p = 0.490$). Balanced accuracy, which weights the two classes equally, rose by only 0.015, with a 95% bootstrap interval from −0.145 to 0.175.

Of the five comparisons, the gain in sensitivity reaches significance in two, the deployed models and the sigmoid head under the argmax rule ($p = 0.031$ in both). Neither would survive a Bonferroni correction for five comparisons ($0.031 \times 5 = 0.16$). The gain in sensitivity is therefore the only result that reaches nominal significance, and it should be read as provisional until it is tested on a larger sample.

The deployed comparison changes the output head as well as the training data, because the model-selection score chose softmax in Version A and sigmoid in Version B. The other four rows of Table 4 hold the head fixed, so that the two versions being compared have identical architectures. With the sigmoid head under the argmax rule, sensitivity rose from 5 to 11 of 19, with six cases gained and none lost ($p = 0.031$), the same result as for the deployed models. With the softmax head under the argmax rule it rose from 6 to 11 of 19 ($p = 0.063$), and at the calibrated thresholds, against the greyscale baseline, it rose from 8 to 12 of 19 for the sigmoid head ($p = 0.125$) and from 9 to 14 of 19 for the softmax head ($p = 0.063$). The change of head is therefore not the source of the gain: with the architecture held identical, every comparison moves in the same direction and none loses a previously detected case.

Table 4: Paired comparisons on the 33 held-out local radiographs, Version A to Version B

Comparison	Sensitivity	Specificity	Accuracy	AUC	Change in balanced accuracy [95% CI]
Deployed models: softmax, colour, 0.4934 → sigmoid, greyscale, 0.4307	6/19 → 12/19 (6 gained, 0 lost); **p = 0.031**	13/14 → 9/14; p = 0.125	19/33 → 21/33; p = 0.754	0.688 → 0.741; p = 0.490	0.015 [−0.145, 0.175]
Sigmoid, greyscale input, calibrated (0.4606 → 0.4307)	8/19 → 12/19 (4 gained, 0 lost); p = 0.125	10/14 → 9/14; p = 1.000	18/33 → 21/33; p = 0.508	0.714 → 0.741; p = 0.725	0.070 [−0.109, 0.256]
Softmax, greyscale input, calibrated (0.4934 → 0.4280)	9/19 → 14/19 (5 gained, 0 lost); p = 0.063	10/14 → 9/14; p = 1.000	19/33 → 23/33; p = 0.344	0.711 → 0.774; p = 0.360	0.096 [−0.090, 0.282]
Sigmoid, argmax, colour → greyscale + local	5/19 → 11/19 (6 gained, 0 lost); **p = 0.031**	14/14 → 10/14; p = 0.125	19/33 → 21/33; p = 0.754	0.707 → 0.741; p = 0.655	0.015 [−0.145, 0.175]
Softmax, argmax, colour → greyscale + local	6/19 → 11/19 (5 gained, 0 lost); p = 0.063	13/14 → 10/14; p = 0.250	19/33 → 21/33; p = 0.727	0.688 → 0.774; p = 0.225	0.024 [−0.126, 0.165]

Note: sensitivity, specificity and accuracy were compared with the two-sided exact McNemar test, and AUC with the DeLong test. The balanced-accuracy interval is a 95% bootstrap interval. Gained and lost refer to pneumonia cases detected by one version and not the other. Bold marks $p < 0.05$.

4.4 Contribution of the Greyscale Step

Version B differs from Version A in two respects, and the greyscale rows of Table 3 show that the second matters. Given greyscale input, the two heads of Version A detected 7 and 9 of the 19 pneumonia cases under the argmax rule, against 5 and 6 with colour input, and cleared 12 and 10 of the 14 normal radiographs, against 14 and 13. Removing the colour cast at inference alone therefore moved the public-only model in the same direction as retraining did.

Measured against this greyscale baseline, the addition of local training data still raised sensitivity at the calibrated thresholds, from 8 to 12 of 19 for the sigmoid head and from 9 to 14 of 19 for the softmax head, and raised the AUC from 0.71 to between 0.74 and 0.77, while specificity fell by one radiograph. None of these greyscale-matched changes is statistically significant on its own ($p = 0.125$ and 0.063 for sensitivity). The share of the gain that belongs to the local training data, as distinct from the greyscale step, cannot be fixed at this sample size. Because the public training and validation images are unchanged by the greyscale conversion, retraining Version A with greyscale input would reproduce its training data exactly, apart from the run-to-run variation that any retraining introduces. This comparison therefore holds the input format fixed and isolates the effect of the local training data; what limits it is the sample size rather than the design.

4.5 Discrimination

Figure 4 shows the ROC curves of both heads on the held-out radiographs under the three conditions. The AUC of the sigmoid head rose from 0.707 to 0.741 and that of the softmax head from 0.688 to 0.774, but the curves lie close together and neither change is significant ($p = 0.655$ and 0.225). With 33 radiographs, a handful of re-ordered cases is enough to move the AUC by this amount.

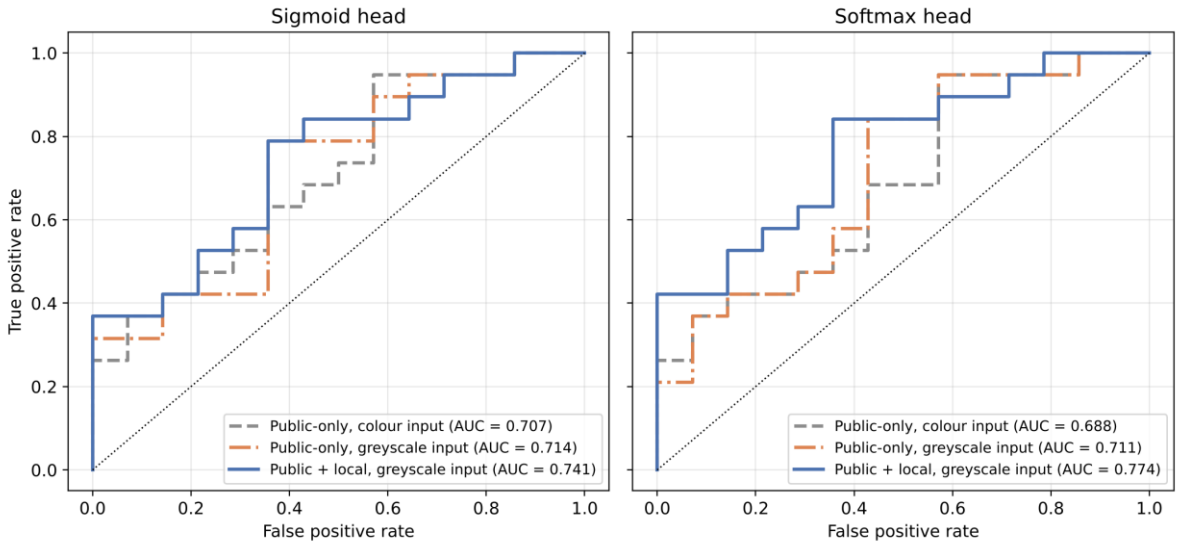


Figure 4: ROC curves on the 33 held-out local radiographs for the sigmoid head (left) and the softmax head (right): Version A with colour input, Version A with greyscale input, and Version B.

4.6 Effect on the Public Test Sets

The gain on local radiographs cost nothing on the public test sets (Table 5). Accuracy on the RSNA primary test partition and on the NIH set is within half a percentage point of its earlier value for both heads, and the AUC within 0.004. The 74 local radiographs, about 5% of each training epoch, changed the model's behaviour on local radiographs without measurably changing it elsewhere.

Table 5: Public test sets before and after adding local training data (argmax rule)

Partition	Head	A: accuracy	A: sensitivity	A: AUC	B: accuracy	B: sensitivity	B: AUC
RSNA primary, n = 2,688	Sigmoid	92.97%	89.78%	0.9814	93.08%	90.81%	0.9814
RSNA primary, n = 2,688	Softmax	93.38%	91.47%	0.9806	93.38%	91.76%	0.9806
NIH online, n = 2,862	Sigmoid	71.45%	63.94%	0.7788	71.87%	65.41%	0.7809
NIH online, n = 2,862	Softmax	72.15%	65.76%	0.7806	71.80%	66.11%	0.7772

4.7 The Decision Threshold

Adding local radiographs to training did not calibrate the system for local use. The deployed threshold of Version B, 0.4307, was selected on the public validation split, where it gave a sensitivity of 95.0%. At that threshold sensitivity was 93.8% on the RSNA primary test partition, 70.8% on the NIH set and 63.2% on the local held-out radiographs. The local figure is double that of Version A, but it remains far below the level the threshold was chosen to deliver, because the split on which the threshold is calibrated contains no local radiograph in either version.

4.8 Comparison with a Previous Nigerian Evaluation

Table 6 places these results beside those of Togunwa et al. (2025). The like-for-like comparison is with Version A, since neither that model nor theirs had seen Nigerian radiographs in training. Version A reached 60.7% to 62.6% accuracy on 107 radiographs, against 58% on 190. With 107 radiographs the 95% confidence interval for an accuracy of 60.7% runs from 51.3% to 69.5%, so the difference is well within the uncertainty of the figures. The two studies agree: a model trained on public data alone performs poorly on Nigerian radiographs.

Table 6: Performance on Nigerian radiographs in two studies

Model	Local test set	Accuracy	AUC
Togunwa et al. (2025)	190	58%	0.65
Version A, sigmoid	107	62.6%	0.758
Version A, softmax	107	60.7%	0.754
Version B, sigmoid	33 held out	63.6%	0.741
Version B, softmax	33 held out	63.6%	0.774

Note: figures for this study are under the argmax rule. Version B was trained with 74 radiographs from the same hospital as its test set; the other models were trained without Nigerian radiographs.

5.0 Discussion

5.1 The Generalisation Gap in Context

The public-only model behaved on local radiographs as models in earlier studies have behaved on data from new sites. Its AUC was 0.98 on the RSNA primary test partition, 0.78 on the NIH set and 0.69 to 0.71 on the 33 held-out local radiographs (0.75 to 0.76 on all 107). Zech et al. (2018) reported a fall from 0.931 on internal test data to 0.815 at an external hospital, Sathitratanacheewin et al. (2020) a fall from 0.8502 to 0.7054 when a tuberculosis classifier developed on the Shenzhen collection was tested on the ChestX-ray8 collection, and Togunwa et al. (2025) a fall from 0.93 to 0.65 when a paediatric model was applied to radiographs from a Nigerian hospital. The gap measured here is of the same order. It took the form of missed pneumonia cases more than of false alarms: Version A detected 25 to 27 of the 62 pneumonia cases among all local radiographs while clearing 38 to 42 of the 45 normal ones.

5.2 What the Local Radiographs Changed

The clearest effect of the 74 local radiographs was on how readily the model reported pneumonia. Under the argmax rule the two heads of Version A, given colour input, called only 5 and 7 of the 33 held-out radiographs pneumonia, in a set where 19 were pneumonia. After retraining, both heads called 15. Every pneumonia case that had been detected before was still detected. The sigmoid head found six further cases and the softmax head five, and four and three normal radiographs respectively were newly reported as abnormal.

This is the pattern of a model whose outputs have moved along the trade-off between sensitivity and specificity, more than of a model that separates the classes better. Balanced accuracy for the deployed models was almost unchanged, at 0.62 and 0.64, and the rise in AUC was within what chance could produce. The honest summary is that the model became better at finding pneumonia on local radiographs, and that whether it became better overall cannot be established from 33 radiographs. For a screening aid, in which a missed case is the more costly error, the movement is in the right direction, but it has a price: 5 of 14 normal radiographs were flagged by the deployed model.

The size of the effect can be set against earlier adaptation studies, which used far more local data. Mohn et al. (2023) fine-tuned an open-source model on 4,510 local radiograph studies, tested it on 802 local radiographs and found a statistically significant overall improvement, and Chamveha et al. (2020), with 421,859 local radiographs, concluded that incorporating local images into training was key to the performance of their model. The present study used 74 radiographs, nearly two orders of magnitude fewer than Mohn et al. (2023), and obtained a gain in sensitivity but no demonstrable gain in overall discrimination. The pattern resembles that reported by Nazim-E-Alam (2026), where recalibration with 163 labelled radiographs from a new site raised sensitivity from 6.2% to 88.3% while specificity fell to 47.9%. In both studies a small amount of target-site data moved the operating point towards detecting disease, and the price was paid in false alarms. Here specificity fell to 64.3% (9 of 14), but with 14 normal radiographs the size of that fall cannot be estimated with any precision.

5.3 Why So Few Images Had an Effect

Seventy-four images are 0.5% of the training split. Two features of the design gave them weight. Repetition raised their share of each epoch to about 5%, and independent augmentation of each presentation kept the repeated images from being identical. The result suggests that a hospital need not assemble thousands of radiographs before local adaptation becomes worthwhile. It does not show how the benefit grows with more images, since only one quantity of local data was tested.

Other studies have also found small amounts of target data to be useful. Kuo et al. (2021) recalibrated models for photographed chest radiographs with fewer than 200 photographs and obtained a significant improvement in performance, and Sellergren et al. (2022) reached an AUC of 0.95 for tuberculosis in external validation with small models trained on only 45 radiographs, using embeddings from a network pretrained on chest radiographs. The present result is weaker than either. Two differences in design may account for this: here the local images were mixed into the training of a network pretrained on natural images rather than used to adapt a network already specialised for chest radiographs, and the evaluation rested on 33 radiographs, on which only a large change can reach significance.

5.4 The Colour Cast

The colour cast is a caution for any study that builds a local dataset by photographing films. A crude index of the cast separated the classes with an AUC of 0.67, which is close to the AUC of 0.69 to 0.71 that Version A achieved on the held-out radiographs from the images themselves. Had the local radiographs been added to training in colour, an apparent gain could have come from the cast, and a test set photographed in the same way would have confirmed it. Converting to greyscale removed the cue. It also changed the behaviour of a model that had never been trained on local images, which shows that the cast influenced predictions even without local training. Zech et al. (2018) described the same hazard at the level of hospitals: their networks identified the hospital system of

99.95% of NIH radiographs and 99.98% of Mount Sinai radiographs, so that the site could stand in for the disease. DeGrave et al. (2021) showed that deep learning models for detecting COVID-19 on chest radiographs relied on confounding factors of this kind rather than on pathology. Here the hazard arose within a single hospital's films and was weaker, an AUC of 0.67 rather than near-perfect identification, but it was strong enough to rival the discrimination that the model achieved from the lungs.

5.5 The Decision Threshold

The failure of the threshold to transfer is consistent with findings in other settings. Nazim-E-Alam (2026) found that a model which kept an AUC of 0.74 to 0.80 on radiographs from two other countries detected only 6.2% and 0% of positive cases at its original threshold, a more extreme form of the shortfall seen here, where a threshold chosen for 95% sensitivity on public validation data gave 70.8% on the NIH set and 63.2% on the local radiographs. In tuberculosis screening, where computer-aided detection of chest radiographs is in routine use, Qin et al. (2021) evaluated five commercial systems in Bangladesh and stressed that the threshold score must be selected for the setting in which a system is used. The present study adds that training on local radiographs does not remove this need. The threshold of Version B was calibrated after the local images had been added to training, but on a validation split that contained none of them, and it inherited the operating point of the public data.

5.6 Practical Implications

- A hospital with about a hundred labelled radiographs can use most of them to adapt a public model, provided some are held out by patient before any training and the split is fixed.
- Local images should be inspected for acquisition differences between the classes before they are used, and removed or neutralised where found.
- Local data should also be used to set the decision threshold. Adding local images to training left the threshold calibrated for public data, and sensitivity at 63% against a target of 95%.
- Specificity should be monitored when a model is adapted, since a gain in sensitivity may be partly a shift in operating point.

5.7 Limitations

The comparison rests on 33 radiographs, so every estimate is imprecise, and only the largest change, the gain in sensitivity, reached significance, without surviving correction for multiple comparisons. The local collection held 107 radiographs, and every radiograph added to the held-out set would have been one fewer for training, so a larger test set requires more local data rather than a different split. The local labels come from hospital records and were not re-read by a radiologist for this study. Errors in the held-out labels could either inflate or conceal a difference between the versions, so an independent reading of these radiographs is the first priority for further work. The training and held-out radiographs come from one hospital and were photographed in the same way, so the result shows adaptation to that hospital and not generalisation to Nigerian facilities at large. Two things changed between the versions, the training data and the greyscale input, and their effects were separated by scoring Version A with greyscale input, a step that relies on the public images being unchanged by the conversion. The heads deployed in the two versions differ, although the same-head comparisons in Table 4 show the same pattern. Each version was trained once, and one quantity of local data was tested. The decision threshold was not recalibrated on local radiographs, because none could be spared from training or testing.

6.0 Conclusion

This study set out to measure what a small number of local chest radiographs, added to a large public training set, changes in a deep learning model for pneumonia detection when the model is used on radiographs from a Nigerian hospital. It compared two versions of one DenseNet121 classifier, trained without and with 74 radiographs from Daystar Specialist Hospital, Awka, on the same 33 radiographs from that hospital that neither version had seen in training.

The argument that guided the study was that local data, the remedy usually proposed for the poor performance of public-data models in African hospitals, should be tested at the scale at which such a hospital can actually supply it, with paired statistics on held-out radiographs, and with attention to which part of the model's behaviour it changes rather than to a single accuracy figure.

The main findings are as follows. A model trained on public data alone detected only 25 to 27 of 62 pneumonia cases among all 107 local radiographs, confirming the generalisation gap reported for Nigerian radiographs by Togunwa et al. (2025). Adding the 74 local radiographs, about 5% of each training epoch, doubled the detection of pneumonia on the held-out radiographs, from 6 to 12 of 19 for the deployed models, a gain that was significant at the 5% level ($p = 0.031$) but not after correction for multiple comparisons, and that was reproduced when the output head was held identical. Specificity fell from 13 to 9 of 14, the AUC rose from 0.688 to 0.741 and balanced accuracy changed by only 0.015, none of these changes being significant. Performance on two public test sets was unchanged. Part of the gain came from converting the photographed films to greyscale, which removed a colour

cast that by itself separated the classes with an AUC of 0.67. The decision threshold, calibrated on public data, gave a local sensitivity of 63.2% against a target of 95%.

These results matter for hospitals that hold only a few dozen labelled radiographs, as many Nigerian facilities do. They show that such a collection is worth assembling, since it can move a public model towards detecting pneumonia on local images without cost on public data, but that it should not be expected to improve overall discrimination on its own. The study contributes a paired, small-scale measurement of local adaptation in a Nigerian setting; evidence that photographing printed films can introduce an artefact correlated with the diagnosis, which must be removed before local images are used for training; and a demonstration that adapting the weights of a model leaves its decision threshold calibrated for the data it came from. In practice, local radiographs should be divided three ways, for training, for calibrating the threshold and for testing.

Further work should, first, have certified radiologists independently re-read the local radiographs, beginning with the held-out set; second, collect a larger set of radiographs from several Nigerian hospitals, so that adaptation can be distinguished from site-specific fitting and the changes in specificity and AUC can be resolved; third, reserve part of the local data for calibrating the decision threshold; fourth, repeat the comparison with several quantities of local data, using repeated cross-validation over all local radiographs and several random seeds, to establish how the benefit grows with the amount of local data; and fifth, compare the simple addition of oversampled local images used here with fine-tuning and domain-adaptation methods. No clinical use should precede these steps.

References

Chamveha, I., Tongdee, T., Saiviroonporn, P., Chaisangmongkon, W., 2020. Local adaptation improves accuracy of deep learning model for automated X-ray thoracic disease detection: A Thai study. arXiv preprint arXiv:2004.10975.

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3, 7, 610 – 619.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21 – 26 July, 2261 – 2269.

Idowu, B.M., Okedere, T.A., 2020. Diagnostic radiology in Nigeria: A country report. Journal of Global Radiology, 6, 1, Article 1072.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M.Y.L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A.N., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172, 5, 1122 – 1131.e9.

Kuo, P.-C., Tsai, C.C., López, D.M., Karargyris, A., Pollard, T.J., Johnson, A.E.W., Celi, L.A., 2021. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. npj Digital Medicine, 4, 1, Article 25.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, Venice, 22 – 29 October, 2999 – 3007.

Mohn, S.F., Law, M., Koleva, M., Lee, B., Berg, A., Murray, N., Nicolaou, S., Parker, W.A., 2023. Machine learning model for chest radiographs: Using local data to enhance performance. Canadian Association of Radiologists Journal, 74, 3, 548 – 556.

Nazim-E-Alam, 2026. Cross-dataset transportability of pediatric chest X-ray deep learning across three countries: Discrimination, calibration, operating-point failure, and limited-label recovery. arXiv preprint arXiv:2609.05140.

Phillips, N.A., Rajpurkar, P., Sabini, M., Krishnan, R., Zhou, S., Pareek, A., Phu, N.M., Wang, C., Jain, M., Du, N.D., Truong, S.Q.H., Ng, A.Y., Lungren, M.P., 2020. CheXphoto: 10,000+ photos and transformations of chest X-rays for benchmarking deep learning robustness. Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR 136, 318 – 327.

Qin, Z.Z., Ahmed, S., Sarker, M.S., Paul, K., Adel, A.S.S., Naheyan, T., Barrett, R., Banu, S., Creswell, J., 2021. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: An evaluation of five artificial intelligence algorithms. The Lancet Digital Health, 3, 9, e543 – e554.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2017. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

Sathitratanacheewin, S., Sunanta, P., Pongpirul, K., 2020. Deep learning for automated classification of tuberculosis-related chest X-ray: Dataset distribution shift limits diagnostic performance generalizability. Heliyon, 6, 8, e04614.

Sellergren, A.B., Chen, C., Nabulsi, Z., Li, Y., Maschinot, A., Sarna, A., Huang, J., Lau, C., Kalidindi, S.R., Etemadi, M., Garcia-Vicente, F., Melnick, D., Liu, Y., Eswaran, K., Tse, D., Beladia, N., Krishnan, D., Shetty, S., 2022. Simplified transfer learning for chest radiography models using less data. Radiology, 305, 2, 454 – 465.

Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., Gill, R.R., Godoy, M.C.B., Hobbs, S., Jeudy, J., Laroia, A., Shah, P.N., Vummidi, D., Yaddanapudi, K., Stein, A., 2019. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence, 1, 1, e180041.

Togunwa, T.O., Babatunde, A.O., Fatade, O.E., Olatunji, R., Ogbole, G., Falade, A., 2025. Detection of pneumonia in children through chest radiographs using artificial intelligence in a low-resource setting: A pilot study. PLOS Digital Health, 4, 9, e0000713.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21 – 26 July, 3462 – 3471.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15, 11, e1002683.