Research article Medical Imaging Open access

Demo article. This is a design sample, not real research. The authors, institutions, data and results are fictional, and the dates only illustrate the layout.

External validation of a deep learning pneumothorax detector on chest radiographs: accuracy with and without chest drains at four hospitals

Sofia Petrakis-Lund[1*], Kwame Asante-Boateng[1], Claire Ducharme[2], Noor Al-Haddad[3], Viktor Halvorsen[4] and Yuki Tanabe-Ross[1]

[1] Department of Radiology, Demo University Medical School, Minneapolis, MN, USA

[2] Division of Biostatistics, Demo University School of Public Health, Minneapolis, MN, USA

[3] Department of Radiology, Demo Northern Health, St Paul, MN, USA

[4] Department of Emergency Medicine, Demo Northern Health, St Paul, MN, USA

*** Corresponding author:** Sofia Petrakis-Lund, email withheld for this demo.

Received	Revised	Accepted	Published
June 8, 2026	July 20, 2026	August 3, 2026	August 18, 2026

Volume	DOI	License	Peer review
1 (2026), article 1	Not yet assigned	CC BY 4.0, copyright the authors	Not peer reviewed (demo)

Journal	Publisher
Augmented Medicine	Eldenhall Sciences Inc., Harrisonville, MO, United States

Demo dates: the journals opened on November 2, 2026. These dates only show the layout of an article's history.

Cite this article: Petrakis-Lund S, Asante-Boateng K, Ducharme C, Al-Haddad N, Halvorsen V, Tanabe-Ross Y. External validation of a deep learning pneumothorax detector on chest radiographs: accuracy with and without chest drains at four hospitals. Augment. Med. 2026;1:1. Available from: https://eldenhall.org/articles/pneumothorax-detection-chest-drain-external-validation/ [Demo article]

DEMO

Abstract

Background. Deep learning models for pneumothorax on chest radiographs report high accuracy, but some may detect the chest drains used to treat pneumothorax rather than the pneumothorax itself. We externally validated a pneumothorax detector and examined its accuracy in radiographs with and without a chest drain.

Methods. In this retrospective diagnostic accuracy study, we assembled 5,412 adult frontal chest radiographs from 4,630 patients at four hospitals with four scanner vendors (January to June 2025), enriched for pneumothorax. The reference standard was the consensus of two thoracic radiologists, with computed tomography where available. We applied PTX-Net (version 2.0), a convolutional neural network trained on public datasets, at its prespecified threshold, and analyzed prespecified subgroups by chest drain, pneumothorax size, projection, sex, age and hospital.

Results. Pneumothorax was present on 438 radiographs (8.1%). The area under the receiver operating characteristic curve was 0.927 (95% CI 0.913 to 0.940), sensitivity 84.2% (80.5 to 87.4) and specificity 93.1% (92.4 to 93.8). Sensitivity was 94.7% (90.3 to 97.2) with a chest drain and 77.5% (72.2 to 82.1) without one, and 63.6% (54.1 to 72.1) for small pneumothoraces without a drain. Specificity was 81.4% in radiographs with a drain and 94.8% without. Accuracy was similar across sex, age groups and hospitals.

Conclusions. Overall accuracy overstated the model's ability to find untreated pneumothorax, the cases in which detection matters most. Pneumothorax detectors should be evaluated and reported separately for radiographs without chest drains.

Keywords: pneumothorax; radiography; thoracic; deep learning; diagnostic accuracy; hidden stratification; chest tubes

Evidence in context

Before this study

Retrospective work showed that pneumothorax models can rely on chest drains, but few external validations report accuracy by drain status.

This study

At four hospitals, sensitivity fell from 94.7% with a drain to 77.5% without one, and to 63.6% for small untreated pneumothoraces.

Implications

Buyers, regulators and journals should ask for accuracy in radiographs without drains before relying on pneumothorax detection.

1. Introduction

Pneumothorax is a common finding on chest radiographs in emergency departments and after procedures, and a missed pneumothorax can enlarge quickly in patients who are ventilated or who have underlying lung disease. Its management depends on its size and on the patient's symptoms [1]. Deep learning models can detect pneumothorax on frontal radiographs [2], and a model covering many findings, including pneumothorax, improved the accuracy of radiologists who used it [3]. Such models are now proposed for prioritizing reading worklists and for reporting normal radiographs without a radiologist, which would leave a missed pneumothorax unreviewed [4].

High accuracy in development data does not guarantee accuracy elsewhere. Reviews have found that few diagnostic imaging studies of artificial intelligence test models on external data [5,6] and that performance usually falls when they do [7]. Models can also learn features that are associated with the

label but are not the disease itself, such as the hospital in which an image was taken [8]. For pneumothorax, the most important of these features is the chest drain. In public datasets, many radiographs labeled as showing pneumothorax were taken after a drain had been inserted, and a model can score these images highly by detecting the drain. Oakden-Rayner and colleagues showed that this hidden stratification led to much lower performance in images without drains [9], which are the images in which an undetected pneumothorax could harm a patient. Performance can also differ between patient groups [10].

We externally validated a pneumothorax detector at four hospitals that had not contributed to its training, and examined its accuracy in radiographs with and without a chest drain, by pneumothorax size and in other prespecified subgroups.

2. Methods

2.1 Design and reporting

This was a retrospective diagnostic accuracy study, reported according to STARD 2015 [11], its extension for artificial intelligence, STARD-AI [12], and the 2024 update of CLAIM [13]. The protocol, including the subgroups and the threshold, was written before the model was run on the study data.

2.2 Setting and radiographs

The four hospitals of Demo Northern Health, in Minnesota, each use a different radiography vendor (labeled A to D). We included frontal chest radiographs of adults aged 18 years or older taken between January 1 and June 30, 2025, in emergency, inpatient and outpatient settings. Because pneumothorax is uncommon on unselected radiographs, we enriched the sample: we included all 1,206 radiographs whose original report mentioned pneumothorax in any way (present, absent, resolving or suspected) and a simple random sample of 4,206 of the remaining radiographs. Radiographs were excluded if they were lateral views, if the lung apices or costophrenic angles were not included, or if they were of pediatric patients. Images were exported in DICOM format, and identifying text burned into the images was removed.

2.3 Reference standard

Two fellowship-trained thoracic radiologists, with 9 and 16 years of experience, independently reviewed each radiograph with the clinical request and with any chest computed tomography (CT) performed within 24 hours, which was available for 612 radiographs. They did not see the original report or the model's output. They recorded the presence of pneumothorax, its size, measured as the interpleural distance at the level of the hilum and classified as large at 2 cm or more [1], and the presence of a chest drain or pleural catheter. Agreement on the presence of pneumothorax was high (kappa 0.88), and a third thoracic radiologist resolved the 97 disagreements.

2.4 The model

PTX-Net (version 2.0) is a DenseNet-121 convolutional neural network developed by the Demo Institute for Imaging AI. It was trained on two public datasets, CheXpert [14] and ChestX-ray14 [15], and on radiographs from a hospital outside Demo Northern Health. It takes a frontal radiograph resized to 512 by 512 pixels and returns a probability of pneumothorax. The developer set the operating threshold

DEMO

at 0.50 before this study. The model was run once on every study radiograph by an analyst who did not take part in the reference standard, and the developer had no role in the analysis.

2.5 Statistical analysis

We calculated the area under the receiver operating characteristic curve (AUROC), and sensitivity, specificity and predictive values at the 0.50 threshold. Confidence intervals for proportions used the Wilson method [16]; confidence intervals for the AUROC used 2,000 bootstrap samples drawn by patient, because some patients contributed several radiographs. Prespecified subgroups were chest drain status, pneumothorax size, projection (posteroanterior or anteroposterior), sex, age group (under 65 years, or 65 years or older) and hospital. We tested whether sensitivity differed by chest drain status with a logistic model fitted by generalized estimating equations clustered by patient. Because the sample was enriched, we also report predictive values at a prevalence of 2%, calculated from sensitivity and specificity. As a sensitivity analysis, we repeated the main analysis using one randomly selected radiograph per patient. Analyses used R 4.3.2.

3. Results

3.1 Radiographs

The 5,412 radiographs came from 4,630 patients with a median age of 66 years (interquartile range 52 to 77); 2,528 radiographs (46.7%) were of women. Of the radiographs, 2,459 (45.4%) were anteroposterior, mostly portable images of inpatients, and 783 (14.5%) showed a chest drain. Pneumothorax was present on 438 radiographs (8.1%): 251 were large and 187 small, and 171 showed a chest drain.

3.2 Overall accuracy

The AUROC was 0.927 (95% CI 0.913 to 0.940). At the prespecified threshold, PTX-Net detected 369 of 438 pneumothoraces, a sensitivity of 84.2% (80.5 to 87.4), with a specificity of 93.1% (92.4 to 93.8; Table 1). In this enriched sample, the positive predictive value was 51.8% (369 of 712) and the negative predictive value 98.5% (4,631 of 4,700). At a prevalence of 2%, the positive predictive value would be 19.9% and the negative predictive value 99.7%. Using one radiograph per patient gave a sensitivity of 83.9% and a specificity of 93.3%.

3.3 Chest drains and pneumothorax size

Sensitivity was 94.7% (90.3 to 97.2) in radiographs with a chest drain and 77.5% (72.2 to 82.1) in radiographs without one ($P < 0.001$ for the difference; Figure 1). Without a drain, sensitivity was 86.9% (80.8 to 91.3) for large and 63.6% (54.1 to 72.1) for small pneumothoraces. The AUROC was 0.952 with a drain and 0.894 without. Drains also caused false-positive results: specificity was 81.4% (78.1 to 84.3) in radiographs with a drain and no pneumothorax, against 94.8% (94.0 to 95.4) in radiographs with neither.

DEMO

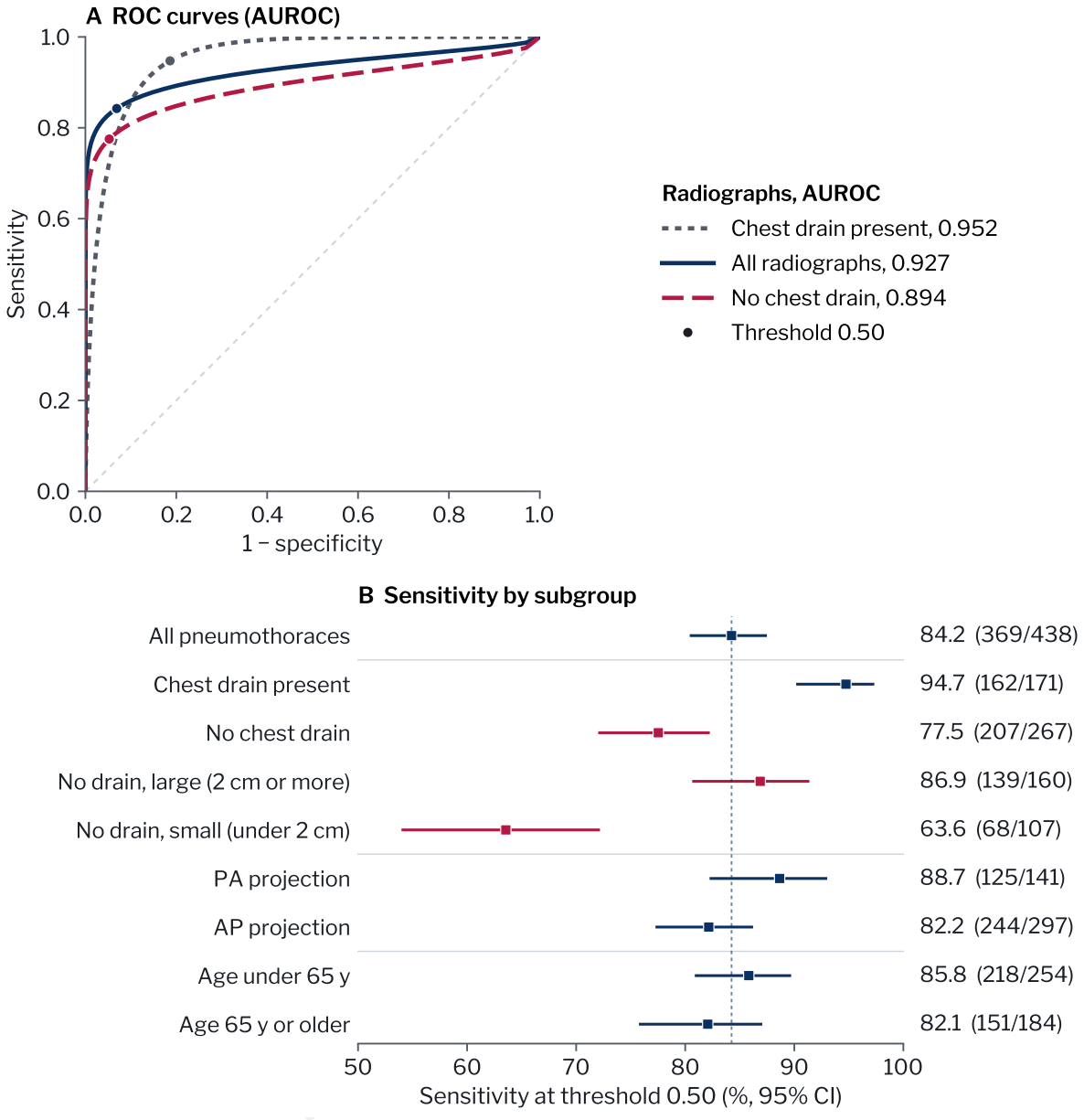


Figure 1. Accuracy of PTX-Net by chest drain status. (A) Receiver operating characteristic curves for all 5,412 radiographs and for radiographs with (783) and without (4,629) a chest drain; dots mark the prespecified threshold of 0.50. (B) Sensitivity at the threshold with Wilson 95% confidence intervals; the dotted line shows overall sensitivity, and subgroups without a chest drain are shown in red. AUROC, area under the receiver operating characteristic curve.

3.4 Other subgroups

Sensitivity was higher on posteroanterior than anteroposterior radiographs (88.7% against 82.2%), and specificity was lower on anteroposterior radiographs (89.6% against 95.8%), which more often showed lines, tubes and skin folds. Accuracy was similar in men and women and in the two age groups. Across the four hospitals, sensitivity ranged from 81.7% to 85.6%, specificity from 92.5% to 93.5% and the AUROC from 0.905 to 0.941, with overlapping confidence intervals (Table 1).

Table 1. Diagnostic accuracy of PTX-Net at the prespecified threshold of 0.50, overall and by subgroup. Sensitivity and specificity are percentages with Wilson 95% confidence intervals; AUROC intervals are from 2,000 bootstrap samples drawn by patient. Size subgroups apply only to radiographs with pneumothorax. AUROC, area under the receiver operating characteristic curve.

Subgroup	With pneumothorax, n	Without pneumothorax, n	Sensitivity, % (95% CI)	Specificity, % (95% CI)	AUROC (95% CI)
All radiographs	438	4,974	84.2 (80.5 to 87.4), 369 of 438	93.1 (92.4 to 93.8), 4,631 of 4,974	0.927 (0.913 to 0.940)
Chest drain present	171	612	94.7 (90.3 to 97.2), 162 of 171	81.4 (78.1 to 84.3), 498 of 612	0.952 (0.934 to 0.970)
No chest drain	267	4,362	77.5 (72.2 to 82.1), 207 of 267	94.8 (94.0 to 95.4), 4,133 of 4,362	0.894 (0.873 to 0.915)
Large pneumothorax (2 cm or more)	251	Not applicable	90.8 (86.6 to 93.8), 228 of 251	Not applicable	Not applicable
Small pneumothorax (under 2 cm)	187	Not applicable	75.4 (68.8 to 81.0), 141 of 187	Not applicable	Not applicable
Small pneumothorax, no chest drain	107	Not applicable	63.6 (54.1 to 72.1), 68 of 107	Not applicable	Not applicable
Posteroanterior projection	141	2,812	88.7 (82.4 to 92.9), 125 of 141	95.8 (95.0 to 96.5), 2,694 of 2,812	0.945 (0.924 to 0.966)
Anteroposterior projection	297	2,162	82.2 (77.4 to 86.1), 244 of 297	89.6 (88.2 to 90.8), 1,937 of 2,162	0.908 (0.889 to 0.927)
Men	286	2,598	85.0 (80.4 to 88.6), 243 of 286	92.8 (91.7 to 93.7), 2,410 of 2,598	0.928 (0.911 to 0.945)
Women	152	2,376	82.9 (76.1 to 88.1), 126 of 152	93.5 (92.4 to 94.4), 2,221 of 2,376	0.925 (0.900 to 0.950)
Age under 65 years	254	2,301	85.8 (81.0 to 89.6), 218 of 254	94.1 (93.1 to 95.0), 2,166 of 2,301	0.934 (0.915 to 0.953)
Age 65 years or older	184	2,673	82.1 (75.9 to 86.9), 151 of 184	92.2 (91.1 to 93.2), 2,465 of 2,673	0.919 (0.897 to 0.941)
Hospital 1 (vendor A)	152	1,642	85.5 (79.1 to 90.2), 130 of 152	93.2 (91.9 to 94.4), 1,531 of 1,642	0.941 (0.918 to 0.964)
Hospital 2 (vendor B)	118	1,370	85.6 (78.1 to 90.8), 101 of 118	93.5 (92.1 to 94.7), 1,281 of 1,370	0.932 (0.905 to 0.959)
Hospital 3 (vendor C)	97	1,104	82.5 (73.7 to 88.8), 80 of 97	92.5 (90.8 to 93.9), 1,021 of 1,104	0.918 (0.886 to 0.950)
Hospital 4 (vendor D)	71	858	81.7 (71.2 to 89.0), 58 of 71	93.0 (91.1 to 94.5), 798 of 858	0.905 (0.866 to 0.944)

3.5 Errors

Of the 69 false-negative results, 39 were small pneumothoraces without a drain, 21 were large pneumothoraces without a drain (most of them anterior or basal on supine anteroposterior radiographs) and 9 were in radiographs with a drain. Of the 343 false-positive results, 114 were in radiographs with a drain. Among the 229 without a drain, the radiologists attributed 97 to skin folds, 42 to the medial border of the scapula, 31 to bullous lung disease and 59 to other causes or no identifiable cause.

4. Discussion

DEMO

At four hospitals that had not contributed training data, a pneumothorax detector had a high overall AUROC, but its sensitivity depended strongly on whether a chest drain was present. It missed nearly one in four pneumothoraces without a drain and more than one in three small pneumothoraces without a drain, and it flagged nearly one in five radiographs that showed a drain but no pneumothorax. These findings confirm in external data the hidden stratification described in retrospective work [9], and they show that overall accuracy is a poor guide to the model's value.

The subgroup in which the model performed worst is the one that matters most. A radiograph that shows a chest drain belongs to a patient whose pneumothorax is already known and treated. The clinical case for a detector rests on finding pneumothoraces that nobody has yet recognized, and in that group its sensitivity was 77.5%. A model used to prioritize worklists or to report radiographs as normal without review [4] would therefore miss more untreated pneumothoraces than its headline figures suggest. The lower specificity in radiographs with drains matters less clinically but adds to alert burden on wards where drains are common.

The likely explanation is the composition of the training data. Large public chest radiograph datasets were labeled automatically from radiology reports [14,15], and in at least one of them a large share of the images labeled as pneumothorax show a chest drain [9]. A model can reach high accuracy on such data by learning the drain. Our results agree with evidence that models generalize poorly when they learn features of the data source rather than of the disease [8]. The difference between posteroanterior and anteroposterior radiographs probably reflects the same problem, since supine anteroposterior images of inpatients show pneumothorax less clearly and contain more devices.

The study has limitations. The sample was enriched for pneumothorax, so predictive values in the sample do not apply to unselected radiographs; we therefore also report them at a prevalence of 2%. CT was available for only a minority of radiographs, and small pneumothoraces visible only on CT were not counted. We evaluated one model at one threshold and did not study how radiologists would use it. The hospitals belong to one health system in one region.

For practice and research, the implication is direct. Studies of pneumothorax detection should report accuracy separately for radiographs with and without chest drains, as recommended for hidden stratification [9] and in line with the subgroup reporting asked for by STARD-AI and CLAIM [12,13]. Health systems evaluating such tools should ask vendors for these figures, and developers should train and test on radiographs without drains or remove the drain shortcut from their data.

5. Conclusions

A pneumothorax detector with an overall AUROC of 0.927 missed 22.5% of pneumothoraces in radiographs without a chest drain and 36.4% of small ones. Accuracy in radiographs without drains should be the primary measure of performance for these tools.

Declarations

DEMO

Ethics approval and consent to participate	The study was approved by the Demo University Institutional Review Board (approval DEMO-IRB-2025-0388, August 14, 2025), which waived the requirement for informed consent because only de-identified retrospective images and records were used. Images were de-identified under the HIPAA Safe Harbor method, including removal of DICOM identifiers and burned-in text. The study followed the Declaration of Helsinki. Registration: not applicable. (Demo statement: no such approval exists.)
Data availability	The radiographs cannot be shared because the institutional review board approval does not permit release of images outside Demo Northern Health. The reference standard labels, model outputs and subgroup variables for each radiograph, without images, are available from the corresponding author under a data use agreement. Code and model availability: the analysis code is deposited in the Demo University research repository under the MIT License. PTX-Net is not publicly available; the developer provided it for this evaluation under a research agreement that gave the authors full control of the analysis and publication.
Use of generative AI	Not applicable to the fictional study. This demonstration article was drafted with the help of a generative AI tool so that the journal's article design could be previewed; it reports no real research.
Funding	This research received no specific funding.
Competing interests	The authors declare no competing interests.
Author contributions (CRediT)	Conceptualization: SPL, YTR; Investigation: KAB, NAH, YTR; Formal analysis: CD; Methodology: SPL, CD; Data curation: VH; Writing, original draft: SPL; Writing, review and editing: all authors. CD and VH accessed and verified the underlying data; neither has any relationship with the model's developer. All authors approved the final version and agree to be accountable for the work.
Acknowledgments	This is a demonstration article created to preview the journal's design. The study, its data and its authors are fictional.
Peer review	Not peer reviewed: this is a demo article. A published article names its review model here (double-blind, external reviewers) and its handling editor.

DEMO

References

1. MacDuff A, Arnold A, Harvey J. Management of spontaneous pneumothorax: British Thoracic Society pleural disease guideline 2010. Thorax. 2010;65(Suppl 2):ii18-ii31. doi:10.1136/thx.2010.136986

2. 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. doi:10.1371/journal.pmed.1002697

3. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health. 2021;3(8):e496-e506. doi:10.1016/S2589-7500(21)00106-0

4. Plesner LL, Müller FC, Nybing JD, et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology. 2023;307(3):e222268. doi:10.1148/radiol.222268

5. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

6. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20(3):405-410. doi:10.3348/kjr.2019.0025

7. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. 2022;4(3):e210064. doi:10.1148/ryai.210064

8. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683

9. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. Association for Computing Machinery; 2020:151-159. doi:10.1145/3368555.3384468

10. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0

11. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. doi:10.1136/bmj.h5527

12. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-3289. doi:10.1038/s41591-025-03953-8

13. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. doi:10.1148/ryai.240300

14. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol 33. AAAI Press; 2019:590-597. doi:10.1609/aaai.v33i01.3301590

15. 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). IEEE; 2017:3462-3471. doi:10.1109/CVPR.2017.369

16. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209-212. doi:10.1080/01621459.1927.10502953

Open access. Copyright 2026 the authors. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.

Published by Eldenhall Sciences Inc., 117 South Lexington Street, Suite 100, Harrisonville, MO 64701, United States. editorial@eldenhall.org.
Online version: https://eldenhall.org/articles/pneumothorax-detection-chest-drain-external-validation/

DEMO