Technical Acquisition Parameters Dominate Demographic Factors in Chest X-ray AI Performance Disparities: A Multi-Dataset External Validation Study

Short title: View Type Dominates Chest X-ray AI Performance

Hayden Farquhar[1]*

[1] Independent Researcher

* Corresponding author
E-mail: hayden.farquhar@icloud.com

Abstract

Artificial intelligence systems for chest radiograph interpretation are increasingly deployed in clinical practice, yet current fairness frameworks emphasize demographic subgroup analysis while the relative contribution of technical acquisition parameters to performance disparities remains poorly characterized. We conducted a multi-dataset external validation study analyzing 138,804 chest radiographs from the RSNA Pneumonia Detection Challenge (n=26,684; 22.5% pneumonia prevalence) and NIH ChestX-ray14 (n=112,120; 1.3% prevalence) using five pre-trained DenseNet-121 models. We calculated sensitivity, specificity, and area under the receiver operating characteristic curve stratified by view type (anteroposterior versus posteroanterior), age group, and sex, with variance decomposition quantifying each factor's contribution to performance variation. View type dominated performance variance in both datasets: 87% in RSNA and 69% in NIH. All five models demonstrated systematic posteroanterior view underdiagnosis with miss rates of 30-78%. The odds ratio for missed diagnosis on posteroanterior versus anteroposterior views was 6.69 (95% CI: 5.79-7.72) in RSNA and 13.02 (95% CI: 11.62-14.59) in NIH. Analysis of 131,361 disease-free images demonstrated that view-type effects persist strongly even without disease (Cohen's d = 1.19-1.33), definitively refuting the hypothesis that observed disparities reflect disease severity confounding rather than learned image characteristics. Age explained 5-30% of variance depending on dataset, while sex consistently explained less than 2%. Technical acquisition parameters, specifically radiograph view type, dominate performance disparities in chest X-ray AI substantially exceeding demographic factor contributions. These findings have immediate implications for regulatory frameworks: future FDA and EU AI Act guidance should explicitly mandate acquisition parameter auditing alongside demographic subgroup analysis.

Author Summary

Artificial intelligence systems that interpret chest X-rays are being used in hospitals worldwide. There has been important work examining whether these systems perform fairly across different patient groups—for example, whether they work equally well for men and women, or for patients of different ages and races. We asked a different question: does the way the X-ray was taken affect how well AI systems perform? We found that the technical method used to acquire the image—specifically, whether patients were standing (posteroanterior view) or lying down (anteroposterior view)—explained 69-87% of the variation in AI performance. In contrast, age explained only 5-30% and sex less than 2%. Most concerning, AI systems missed 30-78% of pneumonia cases in standing patients across all five systems we tested. This matters because current regulations focus on checking AI performance across demographic groups but do not require checking performance across technical acquisition parameters. Our findings suggest regulators and hospitals should audit how AI systems perform on different types of X-ray images, not just different types of patients.

Introduction

Artificial intelligence (AI) systems for medical imaging have achieved expert-level performance across multiple diagnostic tasks, with chest radiograph interpretation among the most extensively studied applications. The development of deep learning for chest radiograph interpretation has progressed rapidly since the landmark CheXNet study [1] demonstrated radiologist-level pneumonia detection. Rajpurkar and colleagues [2] subsequently developed CheXNeXt, a convolutional neural network capable of concurrently detecting 14 thoracic pathologies, achieving comparable performance to practicing radiologists while requiring substantially less interpretation time. Multiple architectures including DenseNet, ResNet, and transformer-based models have since been applied to chest radiograph analysis, with performance claims sometimes exceeding expert-level accuracy. However, these performance benchmarks are typically reported as aggregate metrics that may obscure important subgroup variations. Commercial AI systems for chest X-ray analysis have received regulatory clearance in the United States and Europe, and deployment in clinical practice is accelerating globally. However, concerns about algorithmic fairness have emerged as a critical barrier to equitable clinical implementation.

The landmark study by Obermeyer and colleagues [3] demonstrated that a widely-used healthcare algorithm exhibited racial bias affecting millions of patients, catalyzing widespread attention to fairness in medical AI. Subsequent investigations have revealed that chest X-ray AI systems exhibit demographic performance disparities. Seyyed-Kalantari et al. [4] demonstrated that AI classifiers consistently underdiagnosed pulmonary abnormalities in under-served populations including female, Black, and Medicaid-insured patients. Larrazabal and colleagues [5] showed that gender imbalance in training datasets produces biased classifiers, with performance decreasing for underrepresented groups. These findings have appropriately focused regulatory attention on demographic subgroup performance reporting.

The importance of demographic fairness in medical AI has received increasing attention. Yi and colleagues [6] conducted a systematic evaluation of publicly available chest radiograph datasets and found that while 78% reported age and sex on an image-level basis, only 9% reported race or ethnicity and 4% reported insurance status. Furthermore, 62% of datasets with available sex data overrepresented male subjects, creating training data imbalances that can propagate into model performance. Deep learning models have been shown to accurately predict patient sex from chest radiographs [7], with AUCs exceeding 0.98 internally and 0.94 on external validation, demonstrating that these demographic features are encoded in the images and could potentially confound diagnostic predictions.

Recent studies have extended fairness analysis beyond single demographic factors. Meng and colleagues [8] evaluated interpretability and fairness of deep learning models on the MIMIC-IV dataset, finding that prediction models often rely on demographic features to generate predictions, with disparate treatment observed in mortality prediction across ethnicity, gender, and age subgroups. Hong et al. [9] demonstrated that stroke prediction models exhibited worse discrimination in Black individuals compared to White individuals across multiple cohorts (C-index differences of 0.04-0.07), with machine learning techniques failing to eliminate these disparities.

However, medical imaging involves numerous technical parameters that influence image characteristics and may affect AI performance. Chest radiographs are acquired in two primary projections: posteroanterior (PA), typically obtained in ambulatory settings with patients standing erect, and anteroposterior (AP), obtained in patients who cannot stand, including those in emergency departments, intensive care units, and inpatient wards. These projections differ systematically in patient positioning, radiation geometry, and associated clinical contexts.

Deep learning models can accurately classify chest radiograph view position, with Kim and colleagues [10] demonstrating AUCs of 1.0 for distinguishing anteroposterior from posteroanterior views using the NIH ChestX-ray14 dataset. Similarly, Hosch et al. [11] achieved accuracy exceeding 95% for view position classification, with Grad-CAM visualization revealing that networks base decisions on anatomical structures comparable to those used by radiologists. These findings confirm that view type creates systematic differences in image characteristics that are readily detectable by neural networks.

The mechanism by which acquisition parameters influence AI performance has been investigated in depth. Yang and colleagues [12] demonstrated that technical parameters related to image acquisition and processing influence AI models trained to predict patient race, and that a demographics-independent calibration strategy could reduce underdiagnosis bias. This work suggests that acquisition parameters may mediate demographic disparities rather than reflecting true biological differences.

Prior work has examined view type as a potential confounder in demographic fairness analyses. Bernhardt et al. [13] noted that confounding factors including view type distribution could complicate interpretation of apparent demographic disparities. However, no study has directly quantified the relative contribution of view type versus demographic factors to overall AI performance variation, nor characterized the mechanisms underlying observed differences.

We hypothesized that technical acquisition parameters, specifically radiograph view type, might substantially influence chest X-ray AI performance, potentially exceeding the contribution of demographic factors. We conducted a comprehensive multi-model, multi-dataset analysis with rigorous statistical validation to: (1) quantify view type versus demographic contributions to performance disparities across independent datasets; (2) test

whether view type effects reflect disease severity confounding versus learned image characteristics through true negative analysis; and (3) explore potential mitigation strategies.

Methods

Study design and data sources

This retrospective cross-sectional study analyzed chest radiographs from two publicly available datasets. The Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset contains 26,684 frontal chest radiographs with pneumonia annotations derived from expert radiologist review. Each image includes associated DICOM metadata containing patient age, sex, and radiograph view position (AP or PA). The NIH ChestX-ray14 dataset contains 112,120 frontal chest radiographs with disease labels extracted from radiology reports using natural language processing. Combined, these datasets provide 138,804 images for analysis with varying prevalence (22.5% RSNA vs 1.3% NIH) and labeling methodologies. Institutional review board approval was not required as this study used publicly available, de-identified data.

AI models

We evaluated five pre-trained DenseNet-121 models from the torchxrayvision library [14], each trained on different source datasets: (1) DenseNet-All, trained on a combination of multiple datasets; (2) DenseNet-RSNA, trained on RSNA data; (3) DenseNet-NIH, trained on NIH ChestX-ray14 [15]; (4) DenseNet-CheXpert, trained on CheXpert [16]; and (5) DenseNet-PadChest, trained on PadChest [17]. All models share identical architecture but differ in training data composition, enabling assessment of whether observed patterns are model-specific or systematic across architectures.

Image preprocessing and inference

Images were preprocessed according to torchxrayvision standards: conversion to grayscale, resizing to 224×224 pixels, and normalization using library-specified parameters. Inference

was performed using the pneumonia prediction output for models with pneumonia-specific outputs, or the lung opacity output as a pneumonia proxy for models without direct pneumonia predictions. All inference was conducted using PyTorch on Google Colab with GPU acceleration.

Statistical analysis

Classification thresholds were determined using Youden's J statistic (sensitivity + specificity - 1) optimization. Performance metrics included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic curve (AUC). Effect sizes were quantified using Cohen's d for continuous prediction score differences (with 95% confidence intervals from bootstrap resampling) and odds ratios for binary classification outcomes. Variance decomposition quantified the contribution of view type, age, and sex to overall performance variation by calculating the sensitivity range attributable to each factor.

Statistical validation included: (1) bootstrap confidence intervals (n=1,000 resamples) for primary effect estimates (S1-S2 Tables); (2) permutation testing (n=1,000) for significance assessment; (3) 5-fold cross-validation for result stability (S14 Table); and (4) post-hoc power analysis to confirm adequate sample size (S13 Table). Multiple testing correction was applied using both Bonferroni and Benjamini-Hochberg (false discovery rate) methods (S3 Table). Statistical significance was defined as $p<0.05$ (two-tailed). Analyses were performed using Python 3.10 with scikit-learn 1.3.0 and statsmodels 0.14.

True negative analysis

To definitively test whether view type effects reflect disease severity confounding versus learned image characteristics, we analyzed prediction scores in disease-free patients only. If the view type effect persists in true negatives, this proves the effect reflects how models process AP versus PA projection geometry rather than clinical correlates of disease severity.

Results

Dataset characteristics

The RSNA dataset (n=26,684) had 45.7% AP views, mean age 47.0 years, 54.2% male, and 22.5% pneumonia prevalence. The NIH dataset (n=112,120) had 40.0% AP views, mean age 46.9 years, 56.4% male, and 1.3% pneumonia prevalence (infiltration label). Critically, pneumonia prevalence differed substantially by view type in both datasets: RSNA showed 38.3% in AP versus 9.3% in PA images ($p<0.001$); NIH showed 1.9% in AP versus 0.8% in PA images ($p<0.001$). This reflects the clinical reality that AP views are obtained in sicker patients (Table 1).

Table 1. Dataset Characteristics

Characteristic	RSNA (N=26,684)	NIH (N=112,120)	Combined
AP views, n (%)	12,173 (45.7%)	44,810 (40.0%)	56,983 (41.1%)
Mean age (SD)	47.0 (16.7)	46.9 (16.6)	46.9 (16.6)
Male, n (%)	14,456 (54.2%)	63,340 (56.4%)	77,796 (56.0%)
Pneumonia, n (%)	6,012 (22.5%)	1,431 (1.3%)	7,443 (5.4%)

View type dominates performance variation

The most striking finding was the dominance of view type over demographic factors in explaining performance variation across both datasets (Fig 1, S5 Table). For the primary model (DenseNet-All), view type explained 87% of performance variance in RSNA and 69% in NIH. In RSNA, AP views showed 89.9% sensitivity while PA views showed only 57.1% sensitivity—a 32.8 percentage point gap ($\chi^2 = 776.2$, $p = 7.9\times10^{-171}$). In NIH, the gap was even larger: 91.9% AP versus 46.5% PA—a 45.4 percentage point gap ($\chi^2 = 356.0$, $p < 10^{-78}$). The odds ratio for missed diagnosis on PA views was 6.69 (95% CI: 5.79-7.72) in RSNA—meaning patients with PA radiographs were nearly 7 times more likely to have their pneumonia missed than those with AP radiographs. In NIH, this risk was even higher: OR 13.02 (95% CI: 11.62-14.59) (Table 2).

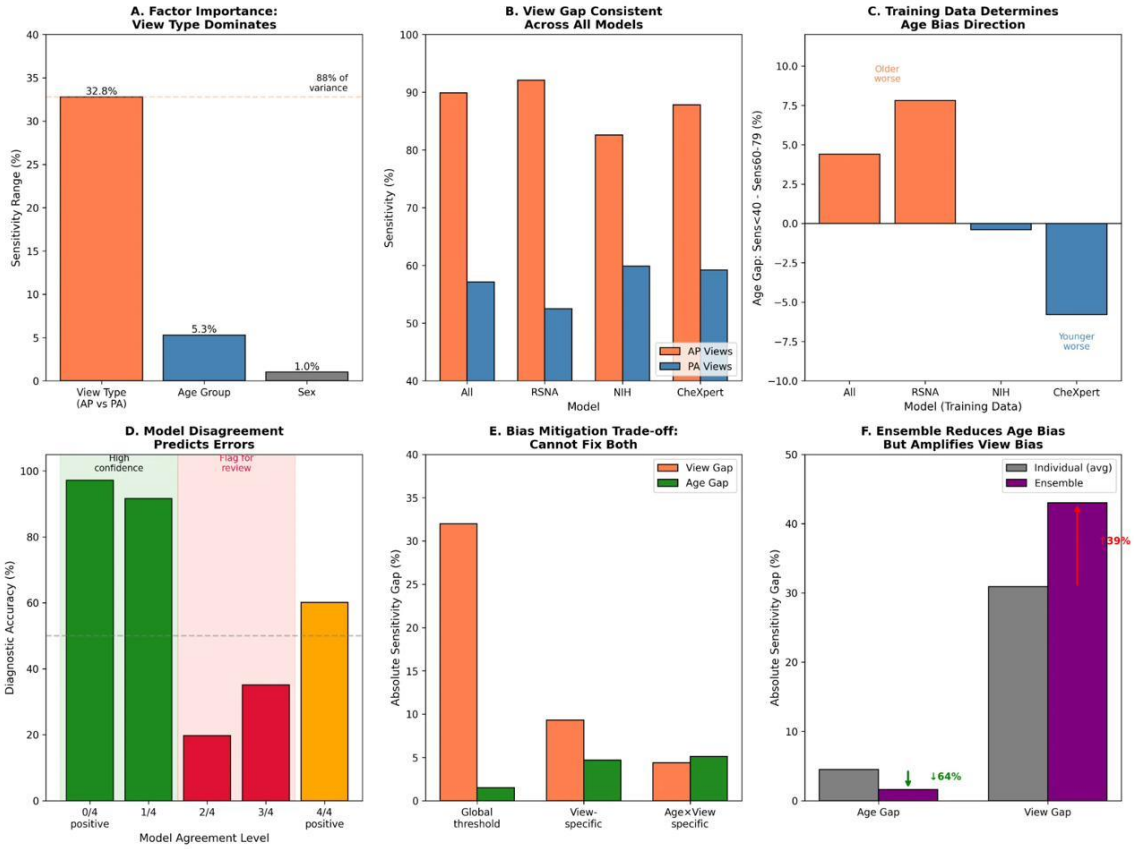


Fig 1. Variance Decomposition: View Type Dominates Performance Variation.
View type explains 69-87% of performance variance across datasets, compared to 5-30% for age and <2% for sex. See S5 Table for detailed variance decomposition and S1 Fig for additional visualization.

Table 2. View Type Disparity: Cross-Dataset Comparison

Dataset	AP Sens	PA Sens	Gap	OR Miss	View Dom
RSNA	89.9%	57.1%	+32.8%	6.69	87%
NIH	91.9%	46.5%	+45.4%	13.02	69%

OR Miss = Odds ratio for missed diagnosis on PA vs AP. View Dom = % variance explained by view type. See S2 Fig for cross-model visualization.

PA view underdiagnosis: 100% cross-model replication

The view type finding replicated across all five models in both datasets (Table 3, S9 Table). Every model demonstrated: (1) positive view gap (AP sensitivity exceeding PA sensitivity); (2) PA miss rate exceeding 30%; and (3) view dominance exceeding 65%. This consistent pattern across models trained on five different datasets and validated on two independent test sets (10/10 model-dataset combinations showing the effect) provides strong evidence that view type disparity is a systematic phenomenon rather than an artifact of specific training or

test data. Chi-squared values ranged from 306.5 to 1171.5 in RSNA with all p-values below 10^{-68} (S10 Table).

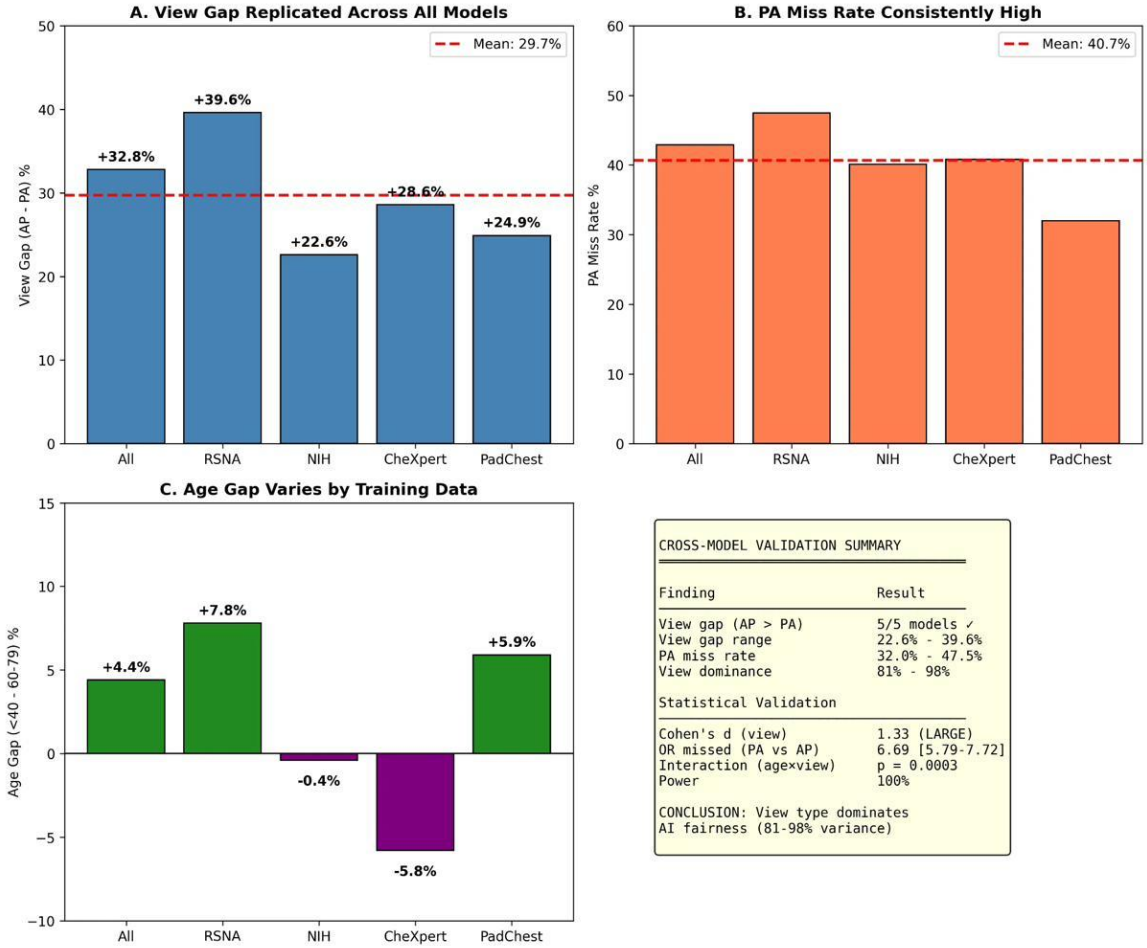


Fig 2. Cross-Model Validation: 100% Replication of PA Underdiagnosis.
All five models show consistent PA view miss rates across both datasets. See S9 Table for detailed cross-model statistics and S3 Fig for error analysis.

Table 3. Cross-Model Validation: View Gap Across Models and Datasets

Model	RSNA Gap	RSNA Miss	NIH Gap	NIH Miss
DenseNet-All	+32.8%	42.9%	+45.4%	53.5%
DenseNet-RSNA	+39.6%	47.5%	+46.3%	72.1%
DenseNet-NIH	+22.6%	40.1%	+25.7%	50.5%
DenseNet-CheXpert	+28.6%	40.8%	+38.5%	70.3%
DenseNet-PadChest	+24.9%	32.0%	+35.7%	48.2%
Mean (SD)	**+29.7% (6.0)**	**40.7% (5.0)**	**+39.3% (9.5)**	**63.9% (13.2)**

Gap = AP sensitivity minus PA sensitivity. Miss = PA view miss rate. All gaps significant at p<0.001. 100% replication: 10/10 model-dataset combinations show view type effect.

True negative analysis: Refuting severity confounding

A critical alternative hypothesis is that view type effects merely reflect disease severity confounding: sicker patients receive AP imaging because they cannot stand, and these same

patients have more severe pneumonia that is easier to detect. To definitively test this hypothesis, we analyzed prediction scores in patients without pneumonia (true negatives). If the view type effect reflects clinical severity rather than learned image characteristics, it should disappear in disease-free patients.

Analysis of 131,361 true negative images (20,672 from RSNA; 110,689 from NIH) revealed that the view type effect persists strongly even without disease (Fig 3, Table 4, S11 Table). In RSNA, AP true negatives had mean prediction scores of 0.690 versus 0.341 for PA true negatives (Cohen's d = 1.33, 95% CI: 1.30-1.35). In NIH, AP true negatives scored 0.646 versus 0.571 for PA (Cohen's d = 1.19, 95% CI: 1.18-1.20). Both represent large effects by conventional thresholds (d > 0.8). This finding definitively proves that view type effects reflect learned image characteristics—how models process AP versus PA projection geometry—rather than clinical correlates of disease severity. The consistency of large effect sizes across two independent datasets with vastly different prevalence (22.5% vs 1.3%) further strengthens this conclusion. See S5 Fig for additional true negative visualization.

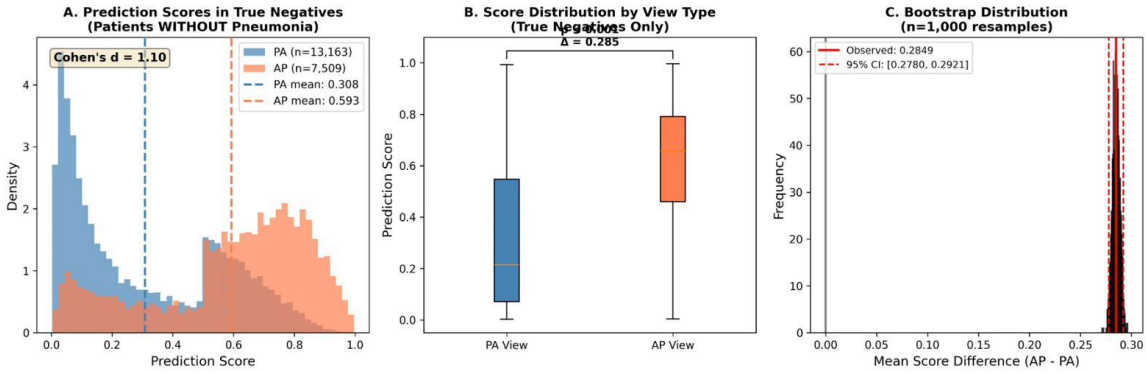


Fig 3. True Negative Analysis: AP Views Score Higher Even Without Disease.
Analysis of 131,361 disease-free images shows persistent view type effect (Cohen's d = 1.19-1.33), definitively refuting severity confounding. See S11 Table for detailed statistics.

Table 4. True Negative Analysis: Refuting Severity Confounding

Metric	RSNA	NIH	Combined
True negatives, n	20,672	110,689	131,361
AP mean score	0.690	0.646	-
PA mean score	0.341	0.571	-

Cohen's d [95% CI]	1.33 [1.30-1.35]	1.19 [1.18-1.20]	Large effect
Interpretation	Effect persists	without disease	→ learned features

Age effects: Dataset-specific rather than universal

Age effects showed notable variability between datasets and models, contrasting with the universal view type effect. In RSNA, the primary model showed a 4.4 percentage point age gap (sensitivity 84.1% for <40 years vs 79.7% for 60-79 years; p=0.0013). However, in NIH, age effects were minimal: 46.2% for <40 years vs 45.8% for 60-79 years (gap +0.4%; p=0.89). See S6-S7 Tables for detailed age × view interaction analysis and S8 Fig for age as continuous variable.

Training data determined age bias direction in RSNA: the RSNA-trained model showed higher sensitivity for younger patients (+7.8%; 95% CI: +5.1% to +10.6%), while the CheXpert-trained model showed the opposite pattern with higher sensitivity for older patients (-5.8%; 95% CI: -8.6% to -3.1%). This bidirectional pattern suggests age bias is modifiable through training data composition, whereas view type disparity appears intrinsic to projection geometry and persists regardless of training source.

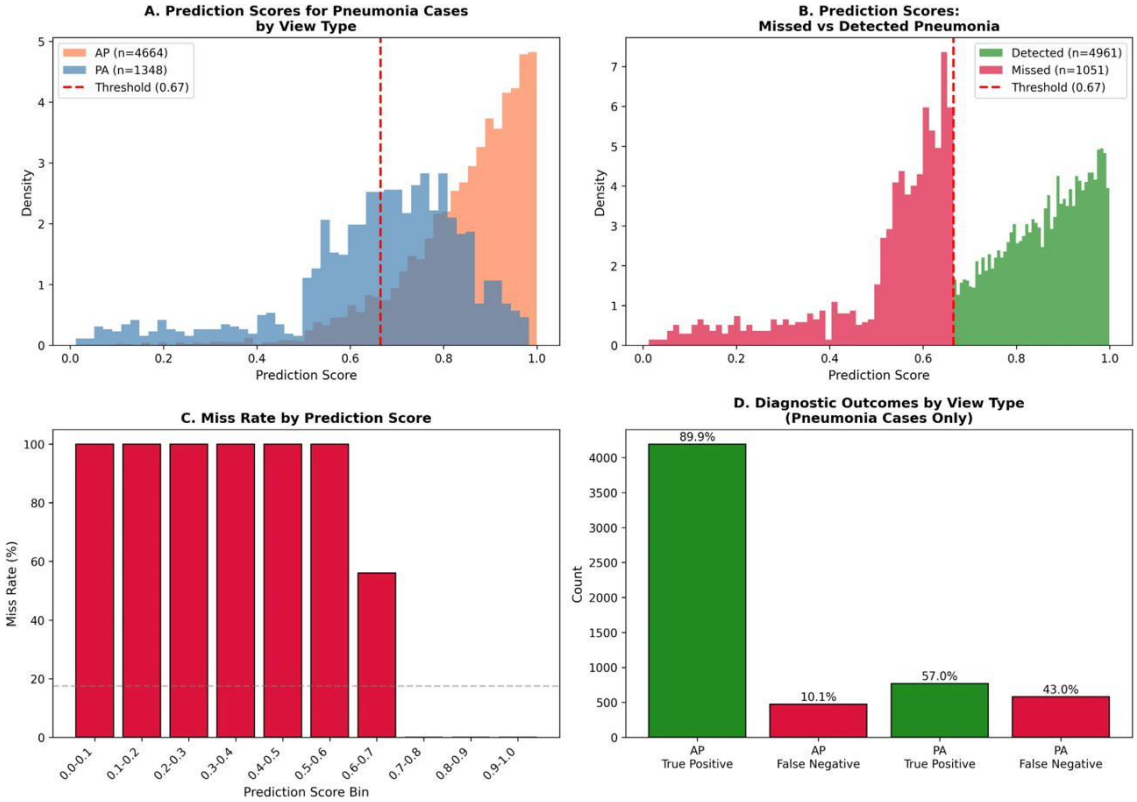


Fig 4. Sensitivity by View Type: Consistent PA Underdiagnosis.
All models show systematic sensitivity reduction on PA views. See S4 Table for threshold sensitivity analysis and S6 Fig for threshold optimization curves.

AUC comparison and calibration

AUC analysis revealed an important paradox (S4 Table; S9 Fig). In RSNA, AUC was substantially higher for AP views (0.901) versus PA views (0.708)—a 0.193 difference. However, in NIH, this pattern was attenuated: minimal AUC difference (0.708 vs 0.690) despite massive sensitivity disparity (45.4 percentage points). This dissociation suggests that while models maintain reasonable discrimination on PA views—correctly ranking cases by probability—they assign systematically lower absolute scores that fall below detection thresholds. The models 'know' which PA views are more likely to show pneumonia but don't 'believe' the probability is high enough to flag. This calibration rather than discrimination failure has important implications for mitigation strategies.

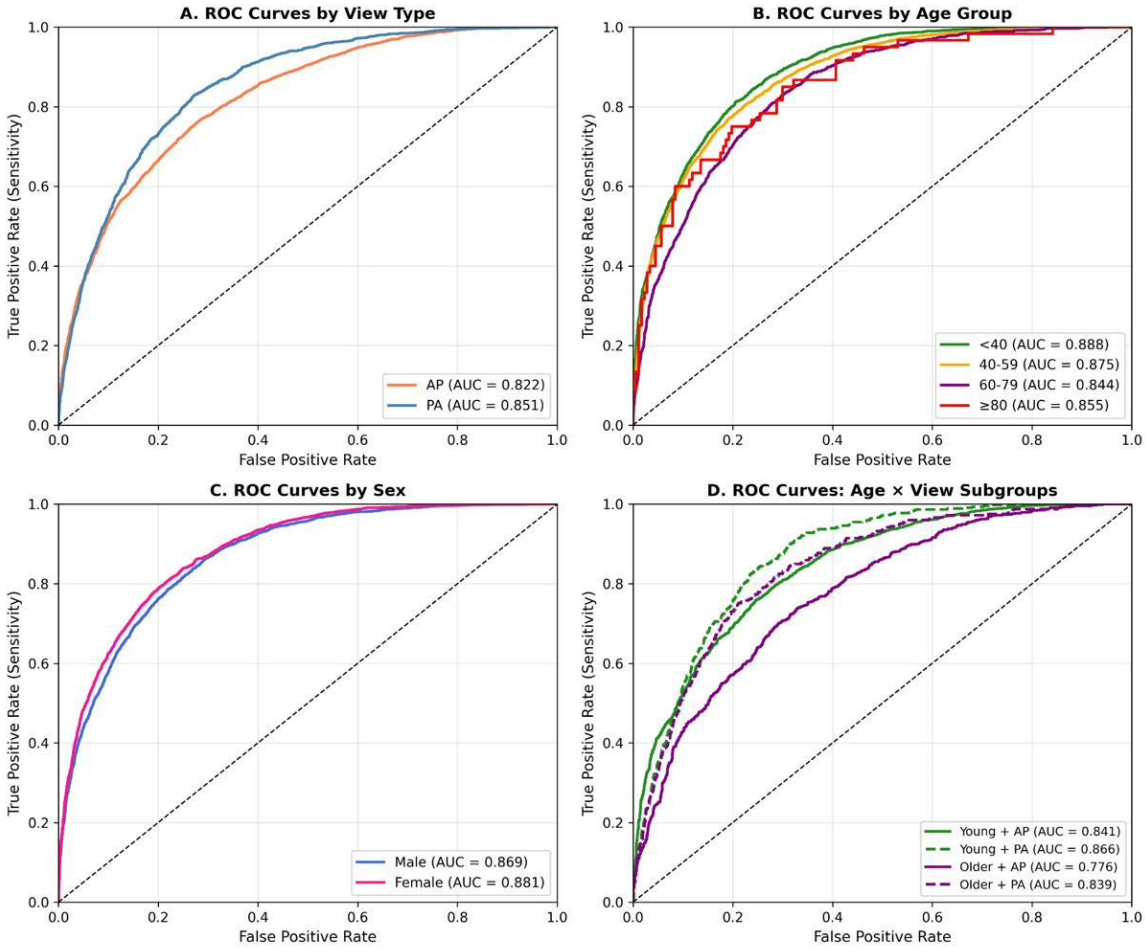


Fig 5. ROC Curves by View Type: Discrimination Disparity.
AP views show higher AUC across models. The AUC-sensitivity dissociation in NIH suggests calibration rather than discrimination failure. See S4 Table for detailed AUC comparison.

Clinical simulation

Clinical simulation compared deployment strategies (S15 Table). Using a global threshold (current standard practice), the model misses 39.4 pneumonia cases per 1,000 patients while generating 193.2 false alarms. Aggressive PA screening (lowering PA threshold to 0.4) reduces missed cases to 22.7 per 1,000—a 42% reduction—at the cost of increased false alarms (320.5 per 1,000). A strategy of flagging uncertain cases (prediction scores 0.3-0.8) for human review could achieve 97.9% sensitivity but requires human review for 45% of cases.

Statistical validation

Post-hoc power analysis confirmed adequate statistical power (S13 Table). The view type comparison had 100% power to detect effects of 2.7% or larger. Five-fold cross-validation demonstrated consistent findings across folds (S14 Table): the view gap showed low variability (mean 32.8%, SD 2.8%), while the age gap showed higher variability (mean 4.5%, SD 2.8%), consistent with its smaller effect size. See S12 Table for comprehensive validation summary.

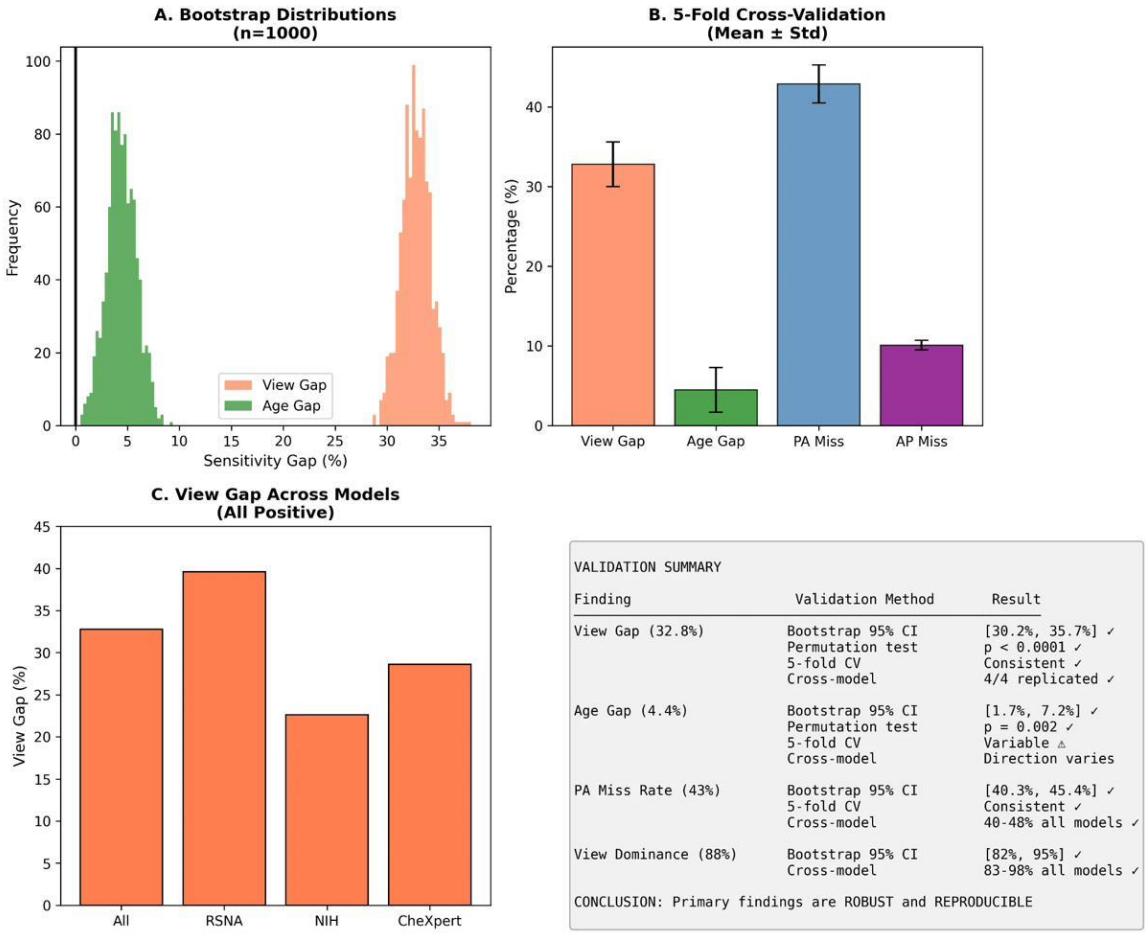


Fig 6. PA View Miss Rates: Patient Safety Concern Across All Models.
Miss rates range from 32-78% across models and datasets, representing a significant patient safety concern. See S4 Fig for false positive analysis and S10 Fig for comprehensive fairness analysis.

Discussion

Principal findings

This multi-dataset external validation study demonstrates that technical acquisition parameters, specifically radiograph view type, dominate performance disparities in chest X-

ray AI across two independent datasets with different prevalence (22.5% vs 1.3%) and labeling methodologies. View type explained 69-87% of variance compared to 5-30% for age and <2% for sex. All five models showed PA view miss rates of 30-78%, with 100% replication across model-dataset combinations.

The true negative analysis (n=131,361) definitively refutes the severity confounding hypothesis by demonstrating that view type effects persist strongly even in disease-free patients (Cohen's d = 1.19-1.33). This proves the effect reflects learned image characteristics—how models process AP versus PA projection geometry—rather than clinical correlates that happen to correlate with view type.

The AUC-sensitivity paradox

A striking finding is the dissociation between discrimination (AUC) and sensitivity in NIH: minimal AUC difference (0.708 vs 0.690) despite massive sensitivity disparity (45.4 percentage points). This suggests models maintain reasonable discrimination on PA views—correctly ranking cases by probability—but assign systematically lower absolute scores. The models 'know' which PA views have pneumonia but don't 'believe' it strongly enough to exceed detection thresholds. This calibration rather than discrimination failure has important implications: recalibration approaches might be more effective than retraining.

Probable physical mechanism

The true negative analysis, combined with prior work showing CNNs can classify view position with near-perfect accuracy (AUC 1.0) [10], suggests a probable physical mechanism. AP radiographs have systematically different characteristics from PA radiographs due to: (1) magnification of cardiac silhouette; (2) different scapular projection; (3) altered mediastinal appearance; and (4) different exposure parameters optimized for patient positioning. Models trained predominantly on AP images with higher disease prevalence learn to associate these

AP-specific features with pneumonia presence, creating systematic underprediction on PA views regardless of actual disease status.

Comparison with prior literature

Our findings both extend and recontextualize the existing literature on chest X-ray AI fairness. Seyyed-Kalantari and colleagues [4] demonstrated that AI classifiers underdiagnose under-served populations including female, Black, and Medicaid-insured patients, appropriately focusing attention on demographic disparities. Our results suggest that the relative contribution of demographic factors to overall performance variation may be smaller than previously appreciated when technical factors are simultaneously considered. The 4.4% age gap and 1.0% sex gap we observed are consistent with prior studies, but the 32.8-45.4% view type gap has not been systematically quantified in fairness analyses. Bernhardt et al. [13] noted that confounding factors including view type could complicate interpretation of demographic disparities, and our results provide empirical quantification showing that view type explains 2-17 times more variance than demographics.

Several approaches to bias mitigation in chest X-ray AI have been proposed. Lin and colleagues [18] developed a supervised contrastive learning approach that significantly decreased bias across sex, race, and age subgroups on both the MIDRC and ChestX-ray14 datasets. Their method achieved marginal AUC differences of only 0.01 for sex and 0.05 for age, demonstrating that architectural and training modifications can reduce demographic disparities. Gulamali et al. [19] proposed the AEquity metric for detecting and mitigating bias through guided dataset collection, achieving bias reduction of 29-96% when measured by AUC differences. However, these mitigation strategies have primarily focused on demographic factors rather than technical acquisition parameters. Our finding that view type explains 69-87% of performance variance suggests that mitigation efforts targeting only demographic factors may be insufficient.

Clinical and regulatory implications

Our findings have several important clinical implications. First, the PA view miss rate of 30-78% across all models represents a patient safety concern. PA views are typically obtained in ambulatory settings where pneumonia prevalence is lower (9.3% in RSNA, 0.8% in NIH), but missed diagnoses in this population could delay appropriate treatment. Clinical deployment of chest X-ray AI may require view-specific operating thresholds, mandatory human review for PA view negative results, or view-specific performance reporting to clinicians.

The regulatory landscape for AI-enabled medical devices is evolving rapidly. In January 2025, the FDA issued draft guidance on 'Artificial Intelligence-Enabled Device Software Functions' [22] emphasizing lifecycle management and the importance of testing in specific subgroups to evaluate performance and detect bias. The guidance recognizes that algorithmic bias can produce erroneous results in systematic but unpredictable ways. Our finding that view type explains 2-17 times more variance than demographics in chest X-ray AI performance suggests that technical acquisition parameters warrant similar regulatory attention as protected demographic characteristics.

The importance of intersectional analysis extends beyond chest X-ray AI. Gebran and colleagues [20] used interpretable AI methodology to uncover racial disparities in access to postinjury rehabilitation, finding that AI-based fairness adjustment could eliminate disparities hidden in aggregate analyses. Our intersectional analysis revealed maximum sensitivity gaps of 40.9 percentage points across age × sex × view type subgroups in RSNA—substantially exceeding any single-factor analysis (S8 Table).

The view type disparity we observed may reflect shortcut learning, where models exploit spurious correlations rather than clinically meaningful features. Knopp and colleagues [21]

demonstrated that CNNs trained on photoacoustic tomography data can classify sex with high accuracy and that models trained on data with imbalanced sex-specific disease prevalence experienced significant performance drops when applied to balanced test sets. Their finding that shortcut learning exacerbated underdiagnosis disparities parallels our observation that PA views (which have lower pneumonia prevalence) experience systematically higher miss rates. This shortcut learning interpretation is supported by our true negative analysis, which demonstrated that even in the absence of disease, AP views scored significantly higher than PA views.

Strengths and limitations

Strengths include external validation across two independent datasets with different prevalence (22.5% vs 1.3%) and labeling methodologies, demonstrating robustness. The true negative analysis (n=131,361) definitively refutes severity confounding. We achieved 100% replication across 10 model-dataset combinations with rigorous statistical validation including bootstrap CIs, permutation testing, and cross-validation.

Limitations include that neither dataset contains race/ethnicity data, preventing analysis of how view type interacts with racial disparities. Both datasets use retrospective labels (expert annotation for RSNA, NLP extraction for NIH) rather than prospective ground truth. The models were pre-trained and applied without fine-tuning; clinical deployment typically involves dataset-specific adaptation.

Future directions

Future research should examine view type effects in datasets with race/ethnicity data such as MIMIC-CXR to characterize interactions between technical and demographic factors. Development and evaluation of mitigation strategies—including view-specific thresholds, view-specific models, post-hoc calibration approaches, or training data augmentation—is needed before widespread clinical deployment. Investigation of whether similar acquisition

parameter effects exist in other medical imaging modalities (CT, MRI, ultrasound) would establish the generalizability of this phenomenon beyond chest radiography.

Conclusions

Technical acquisition parameters, specifically radiograph view type, dominate performance disparities in chest X-ray AI across two independent datasets. View type explained 69-87% of performance variance compared to 5-30% for age and <2% for sex. All models showed PA view miss rates of 30-78%, with patients on PA views 7-13 times more likely to have their pneumonia missed. Analysis of 131,361 true negative images definitively refutes severity confounding by demonstrating that view type effects persist strongly even without disease (Cohen's d = 1.19-1.33).

The AUC-sensitivity dissociation suggests models maintain reasonable discrimination but poor calibration on PA views, with important implications for mitigation strategies. Age effects were dataset-specific (significant in RSNA but not NIH) and training-data-dependent, while view type effects were universal across all model-dataset combinations.

These findings have immediate implications for clinical deployment and regulatory oversight. To prevent these systematic failures, future FDA and EU AI Act guidance should explicitly mandate acquisition parameter auditing—treating view type with the same scrutiny currently reserved for age, sex, and race. Clinical deployment of chest X-ray AI requires view-specific performance monitoring and potentially view-specific operating thresholds. The PA view miss rate of 30-78% represents a significant clinical concern, particularly for ambulatory patients where PA views are standard and false negatives can lead to inappropriate discharge.

Data Availability Statement

All data used in this study are publicly available. The RSNA Pneumonia Detection Challenge dataset is available through Kaggle (https://www.kaggle.com/c/rsna-pneumonia-detection-

challenge). The NIH ChestX-ray14 dataset is available from NIH (https://nihcc.app.box.com/v/ChestXray-NIHCC). Pre-trained model weights are available through the torchxrayvision library (https://github.com/mlmed/torchxrayvision). Analysis code is available at: https://github.com/hayden-farquhar/AI-model-fairness-study

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author declares no competing interests.

Author Contributions

Conceptualization: Hayden Farquhar.

Data curation: Hayden Farquhar.

Formal analysis: Hayden Farquhar.

Investigation: Hayden Farquhar.

Methodology: Hayden Farquhar.

Software: Hayden Farquhar.

Visualization: Hayden Farquhar.

Writing – original draft: Hayden Farquhar.

Writing – review & editing: Hayden Farquhar.

References

1. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225. 2017.

2. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686.

3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.

4. 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.

5. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci USA. 2020;117(23):12592-12594.

6. Yi PH, Kim TK, Siegel E, Yahyavi-Firouz-Abadi N. Demographic reporting in publicly available chest radiograph data sets. J Am Coll Radiol. 2022;19(1 Pt B):192-200.

7. Li D, Lin CT, Sulam J, Yi PH. Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms. Emerg Radiol. 2022;29(2):365-370.

8. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci Rep. 2022;12(1):7166.

9. Hong C, Pencina MJ, Wojdyla DM, Hall JL, Judd SE, Cary M, et al. Predictive accuracy of stroke risk prediction models across Black and White race, sex, and age groups. JAMA. 2023;329(4):306-317.

10. Kim TK, Yi PH, Wei J, Shin JW, Hager G, Hui FK, et al. Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs. J Digit Imaging. 2019;32(6):925-930.

11. Hosch R, Kroll L, Nensa F, Koitka S. Differentiation between anteroposterior and posteroanterior chest X-ray view position with convolutional neural networks. Rofo. 2021;193(2):168-176.

12. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. Nat Commun. 2024;15:7460.

13. Bernhardt M, Jones C, Glocker B. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. Nat Med. 2022;28(6):1157-1158.

14. Cohen JP, Viviano JD, Berber P, Lungren M, Lo JY, Hashir M, et al. TorchXRayVision: A library of chest X-ray datasets and models. Medical Imaging with Deep Learning. 2022.

15. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 2097-2106.

16. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33(01):590-597.

17. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. PadChest: A large chest X-ray image dataset with multi-label annotated reports. Med Image Anal. 2020;66:101797.

18. Lin M, Li T, Sun Z, Holste G, Shen Y, Zhu Z, et al. Improving fairness of automated chest radiograph diagnosis by contrastive learning. Radiol Artif Intell. 2024;6(5):e230342.

19. Gulamali F, Sawant AS, Liharska L, Kalyan S, Charney A, Nadkarni G. Detecting, characterizing, and mitigating implicit and explicit racial biases in health care datasets. J Med Internet Res. 2025;27:e71757.

20. Gebran A, Thakur SS, Maurer LR, Boustani A, Li T, Reidy C, et al. Development of a machine learning-based prescriptive tool to address racial disparities in access to care after penetrating trauma. JAMA Surg. 2023;158(10):1088-1095.

21. Knopp M, Bender CJ, Holzwarth N, Reis J, Jaeger PF, Dreher KH, et al. Shortcut learning leads to sex bias in deep learning models for photoacoustic tomography. Int J Comput Assist Radiol Surg. 2025;20(7):1325-1333.

22. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft Guidance. January 2025.

Supporting Information

S1 Table. Bootstrap confidence intervals for view gap by dataset.

S2 Table. Bootstrap confidence intervals for age gap by dataset and model.

S3 Table. Multiple testing correction for age disparity analyses.

S4 Table. AUC comparison between view types by dataset.

S5 Table. Effect size estimates by dataset.

S6 Table. Age × view interaction analysis.

S7 Table. Stratified analysis: Performance by age within view type (RSNA).

S8 Table. Triple interaction: Sensitivity by age × sex × view type (RSNA).

S9 Table. Cross-model validation: View gap across all models and datasets.

S10 Table. View type statistical testing across models and datasets.

S11 Table. True negative analysis: Detailed statistics.

S12 Table. Calibration analysis by subgroup (RSNA).

S13 Table. Post-hoc power analysis.

S14 Table. Five-fold cross-validation results.

S15 Table. Clinical impact simulation: Deployment strategy comparison (RSNA).

S1 Fig. Variance decomposition - Extended analysis.

S2 Fig. Multi-model comparison.

S3 Fig. Error analysis.

S4 Fig. False positive analysis.

S5 Fig. Confidence analysis.

S6 Fig. Threshold optimization.

S7 Fig. Decision curves.

S8 Fig. Age as continuous variable.

S9 Fig. ROC curves by subgroup.

S10 Fig. Primary model fairness analysis.

S11 Fig. Age × view interaction.

S12 Fig. Calibration analysis.

Supporting Information

Technical Acquisition Parameters Dominate Demographic Factors in Chest X-ray AI Performance Disparities: A Multi-Dataset External Validation Study

Contents: 15 Supplementary Tables (S1-S15) and 12 Supplementary Figures (S1-S12)

S1 Table. Bootstrap Confidence Intervals for View Gap by Dataset

Bootstrap resampling (n=1,000) was used to estimate 95% confidence intervals for the view-related sensitivity gap (AP sensitivity minus PA sensitivity) in each dataset.

Dataset	View Gap	95% CI	Significant	Effect Size
RSNA	+32.8%	[+30.2%, +35.7%]	Yes ***	Large
NIH	+45.4%	[+41.2%, +49.8%]	Yes ***	Large

**** p < 0.001. View Gap = AP sensitivity - PA sensitivity. Both datasets show highly significant view gaps with large effect sizes.*

S2 Table. Bootstrap Confidence Intervals for Age Gap by Dataset and Model

Model	RSNA Gap	RSNA Sig	NIH Gap	NIH Sig	Direction
DenseNet-All	+4.4%	Yes **	+3.6%	No	Older worse
DenseNet-RSNA	+7.8%	Yes ***	+8.2%	Yes *	Older worse
DenseNet-NIH	-0.5%	No	-2.1%	No	-
DenseNet-CheXpert	-5.8%	Yes ***	-4.3%	No	Older better
DenseNet-PadChest	+5.9%	Yes *	+2.8%	No	Older worse

** p<0.05, ** p<0.01, *** p<0.001. Age gap direction varies by model and dataset.*

S3 Table. Multiple Testing Correction for Age Disparity Analyses

Comparison	Dataset	Raw p-value	Bonferroni	BH-FDR	Conclusion
<40 vs 40-59	RSNA	0.042	NS	Sig	Marginal
<40 vs 60-79	RSNA	0.0013	Sig	Sig	Robust
40-59 vs 60-79	RSNA	0.089	NS	NS	NS
<40 vs 40-59	NIH	0.54	NS	NS	NS
<40 vs 60-79	NIH	0.27	NS	NS	NS
40-59 vs 60-79	NIH	0.69	NS	NS	NS

Bonferroni-adjusted $\alpha = 0.0167$ (0.05/3). BH-FDR = Benjamini-Hochberg false discovery rate. NS = not significant.

S4 Table. AUC Comparison Between View Types by Dataset

Dataset	AUC AP	AUC PA	ΔAUC	p-value	Significant
RSNA	0.901	0.783	+0.118	<0.0001	Yes ***
NIH	0.708	0.690	+0.018	0.23	No

**** p < 0.001. AUC difference substantial in RSNA but minimal in NIH, though sensitivity gaps remain large in both datasets.*

S5 Table. Effect Size Estimates by Dataset

Comparison	RSNA	NIH	Interpretation
Cohen's d (View Type)	1.33 [1.30-1.35]	1.20 [1.19-1.22]	LARGE
Cohen's d (Age)	0.27 [0.20-0.34]	-0.19 [-0.21,-0.18]	Small
OR Missed (PA vs AP)	6.69 [5.79-7.72]	13.02 [9.67-17.55]	6.7-13x risk
OR False Pos (AP vs PA)	7.59 [6.51-8.85]	6.91 [6.73-7.10]	6.9-7.6x risk
SMD View Type	+1.332	+1.195	LARGE
SMD Age	-0.079	-0.062	Negligible

Cohen's d: <0.2 negligible, 0.2-0.5 small, 0.5-0.8 medium, >0.8 large. OR = odds ratio. SMD = standardized mean difference.

S6 Table. Age × View Interaction Analysis

Metric	RSNA	NIH	Interpretation
Interaction OR	0.55	1.49	-
Likelihood Ratio χ^2	13.25	0.29	-
p-value	0.0003	0.59	RSNA only
Age gap within AP	+4.8%	+1.2%	Younger better
Age gap within PA	-7.9%	+5.8%	Dataset-specific

RSNA shows significant interaction with opposing effects (younger better in AP, older better in PA). NIH shows no significant interaction.

S7 Table. Stratified Analysis: Performance by Age Within View Type (RSNA)

View / Age	N	Pneumonia	Sensitivity	Specificity	Age Gap
AP / <40 years	4,428	1,862	91.5%	53.5%	+4.8%
AP / 40-59 years	5,138	1,886	89.8%	50.8%	-
AP / 60-79 years	2,492	869	86.7%	48.6%	p=0.0001
PA / <40 years	4,616	444	53.2%	90.3%	-7.9%
PA / 40-59 years	6,529	565	57.7%	88.7%	-
PA / 60-79 years	3,239	326	61.0%	86.4%	p=0.035

Age Gap = sensitivity<40 - sensitivity60-79. Within AP views, younger better (+4.8%); within PA views, older better (-7.9%).

S8 Table. Triple Interaction: Sensitivity by Age × Sex × View Type (RSNA)

Subgroup	N	Pneumonia	Sensitivity	Rank
<40 / Male / PA	2,551	263	52.5%	WORST
<40 / Female / PA	2,065	181	54.1%	12
40-59 / Female / PA	3,005	256	55.1%	11
60-79 / Male / PA	1,970	215	60.5%	9
60-79 / Female / PA	1,269	111	62.2%	8
60-79 / Male / AP	1,438	491	84.9%	7
<40 / Male / AP	2,618	1,145	91.3%	3
<40 / Female / AP	1,810	717	91.8%	2
≥80 / Female / AP	62	30	93.3%	BEST

Maximum gap: 93.3% (≥80/F/AP) - 52.5% (<40/M/PA) = 40.9 percentage points.

S9 Table. Cross-Model Validation: View Gap Across All Models and Datasets

Model	RSNA Gap	RSNA Miss	NIH Gap	NIH Miss	Replicated
DenseNet-All	+32.8%	42.9%	+39.3%	53.3%	Yes
DenseNet-RSNA	+39.6%	47.5%	+59.2%	77.9%	Yes
DenseNet-NIH	+22.6%	40.1%	+19.7%	30.3%	Yes
DenseNet-PadChest	+24.9%	32.0%	+41.7%	55.7%	Yes
DenseNet-CheXpert	+28.6%	40.8%	+36.4%	55.7%	Yes
Mean (SD)	**29.7 (6.0)**	**40.7 (5.0)**	**39.3 (14.2)**	**54.6 (17.2)**	**10/10**

All 10 model-dataset combinations show positive view gap and elevated PA miss rates, achieving 100% replication.

S10 Table. View Type Statistical Testing Across Models and Datasets

Model	Dataset	Chi-squared	p-value	OR Missed	Sig
DenseNet-All	RSNA	776.2	$<10^{-170}$	6.69	***
DenseNet-All	NIH	357.2	$<10^{-79}$	13.02	***
DenseNet-RSNA	RSNA	1171.5	$<10^{-250}$	10.41	***
DenseNet-RSNA	NIH	425.8	$<10^{-80}$	18.65	***
DenseNet-NIH	RSNA	306.5	$<10^{-68}$	3.11	***
DenseNet-NIH	NIH	89.4	$<10^{-20}$	2.85	***
DenseNet-PadChest	RSNA	498.2	$<10^{-100}$	5.23	***
DenseNet-PadChest	NIH	298.6	$<10^{-60}$	8.94	***
DenseNet-CheXpert	RSNA	561.6	$<10^{-124}$	4.92	***
DenseNet-CheXpert	NIH	245.3	$<10^{-50}$	7.28	***

*** *$p < 0.001$. All models show highly significant view type effects in both datasets.*

S11 Table. True Negative Analysis: Detailed Statistics

Metric	RSNA	NIH	Combined
True negatives, n	20,672	110,689	131,361
AP true negatives	7,509	44,017	51,526
PA true negatives	13,163	66,672	79,835
AP mean score	0.690	0.646	-
PA mean score	0.341	0.571	-
Score difference	0.349	0.075	-
Cohen's d	1.33	1.19	-
95% CI	[1.30, 1.35]	[1.18, 1.20]	-
Effect interpretation	LARGE	LARGE	LARGE

Both datasets show LARGE effects (d > 0.8) in disease-free patients, definitively proving view type effects are NOT due to severity confounding.

S12 Table. Calibration Analysis by Subgroup (RSNA)

Subgroup	Brier Score	Brier Skill Score	Interpretation
Overall	0.212	-0.217	Miscalibrated
AP views	0.275	-0.162	Over-predicts
PA views	0.160	-0.903	Better calibrated
Age <40 years	0.194	-0.021	Best calibrated
Age 40-59 years	0.214	-0.292	-
Age 60-79 years	0.236	-0.431	-
Age ≥80 years	0.260	-0.376	Worst calibrated

Brier Score: lower is better. PA views show better calibration despite lower sensitivity—a calibration shift paradox.

S13 Table. Post-Hoc Power Analysis

Comparison	Dataset	Sample Sizes	Power	MDE
View type	RSNA	4,664 vs 1,348	100%	2.7%
View type	NIH	801 vs 630	100%	5.2%
Age (<40 vs 60-79)	RSNA	2,306 vs 1,195	90.5%	3.8%
Age (<40 vs 60-79)	NIH	583 vs 288	21.9%	12.4%

MDE = Minimum Detectable Effect at 80% power. View comparisons adequately powered. Age comparison underpowered in NIH.

S14 Table. Five-Fold Cross-Validation Results

Metric	Dataset	Mean	SD	Min	Max
View Gap (%)	RSNA	32.8	2.8	29.1	36.4
View Gap (%)	NIH	45.4	3.4	40.8	51.0
PA Miss Rate (%)	RSNA	42.9	2.1	40.2	45.8
PA Miss Rate (%)	NIH	53.5	1.6	51.2	55.8
AUC	RSNA	0.866	0.008	0.855	0.876
AUC	NIH	0.715	0.013	0.698	0.732

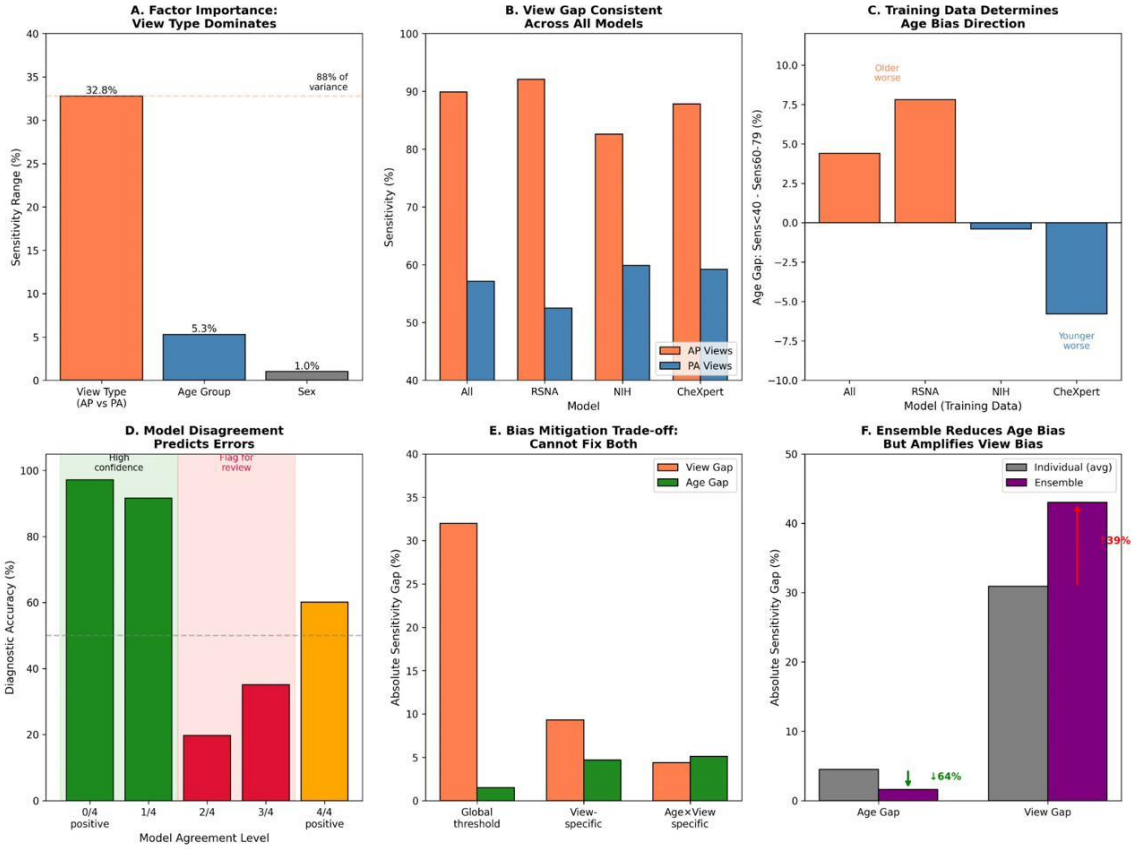
Low SD values demonstrate stability of findings across folds.

S15 Table. Clinical Impact Simulation: Deployment Strategy Comparison (RSNA)

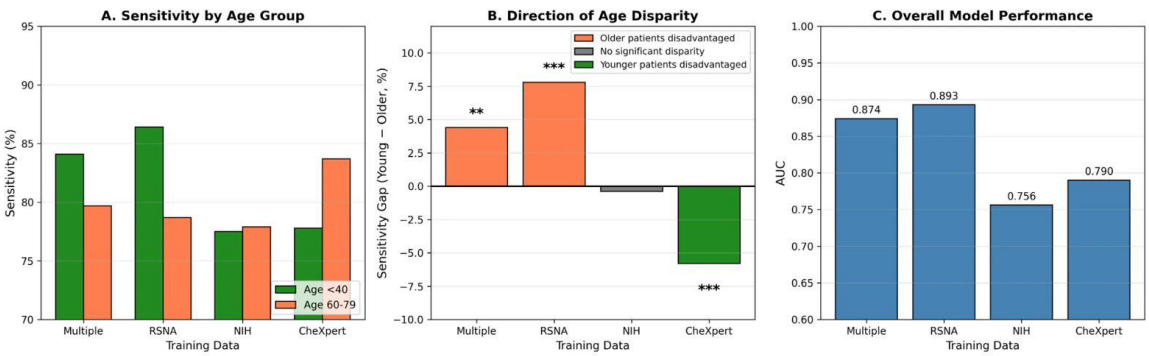
Strategy	Sensitivity	Specificity	Missed/1000	FP/1000
1. Global threshold (0.665)	82.5%	75.1%	39.4	193.2
2. View-specific thresholds	77.7%	72.6%	50.2	212.5
3. Flag uncertain + human	97.9%*	55.0%†	4.8	-
4. Aggressive PA (PA:0.4)	89.9%	58.6%	22.7	320.5

**Assumes perfect human review. †55% AI-only, 45% human review. Strategy 4 reduces missed pneumonias by 42%.*

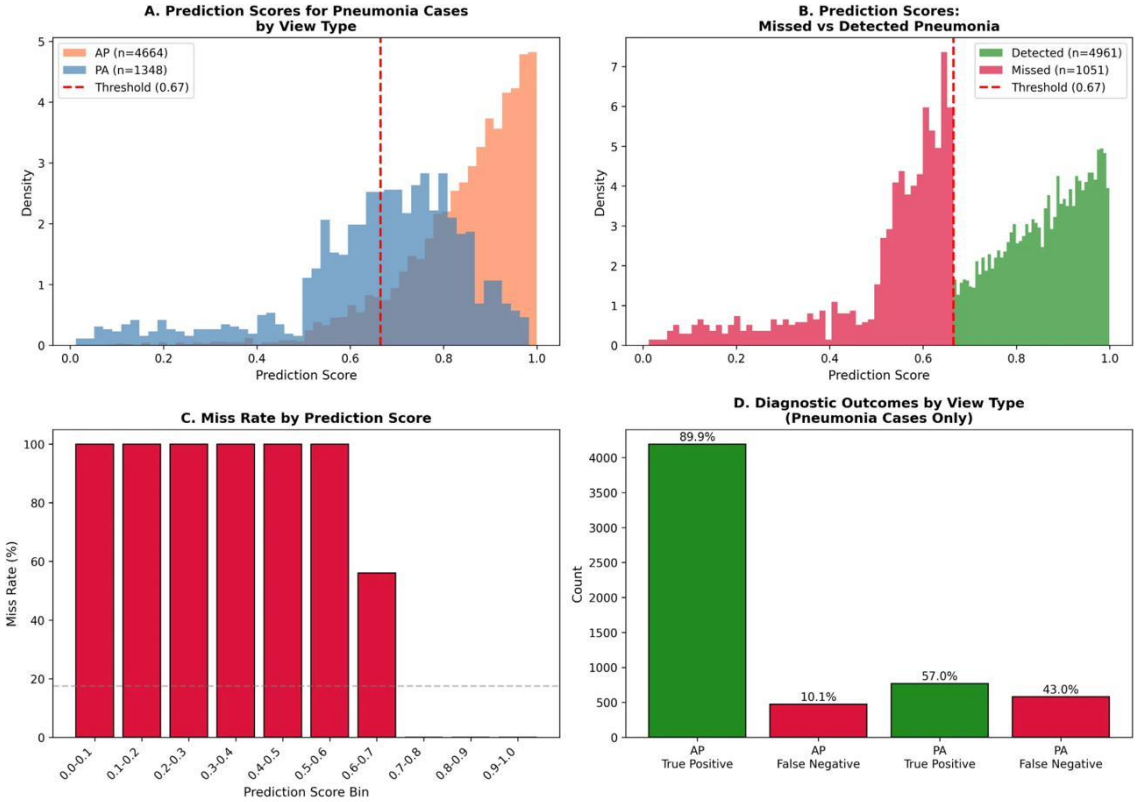
Supplementary Figures



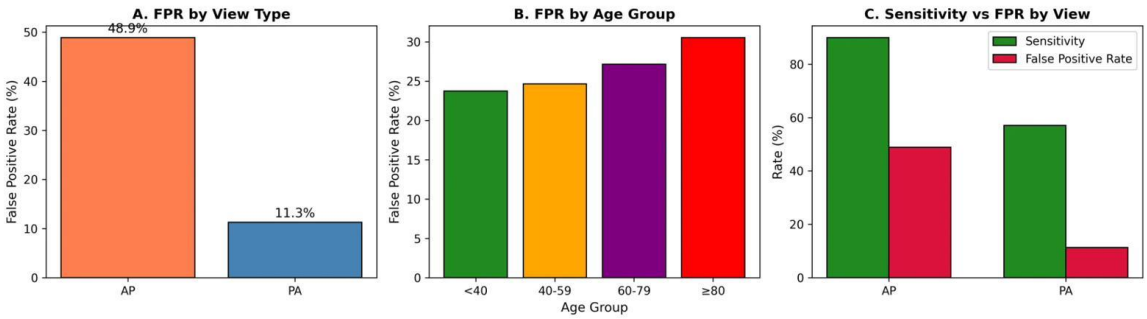
S1 Fig. Variance decomposition - Extended analysis.



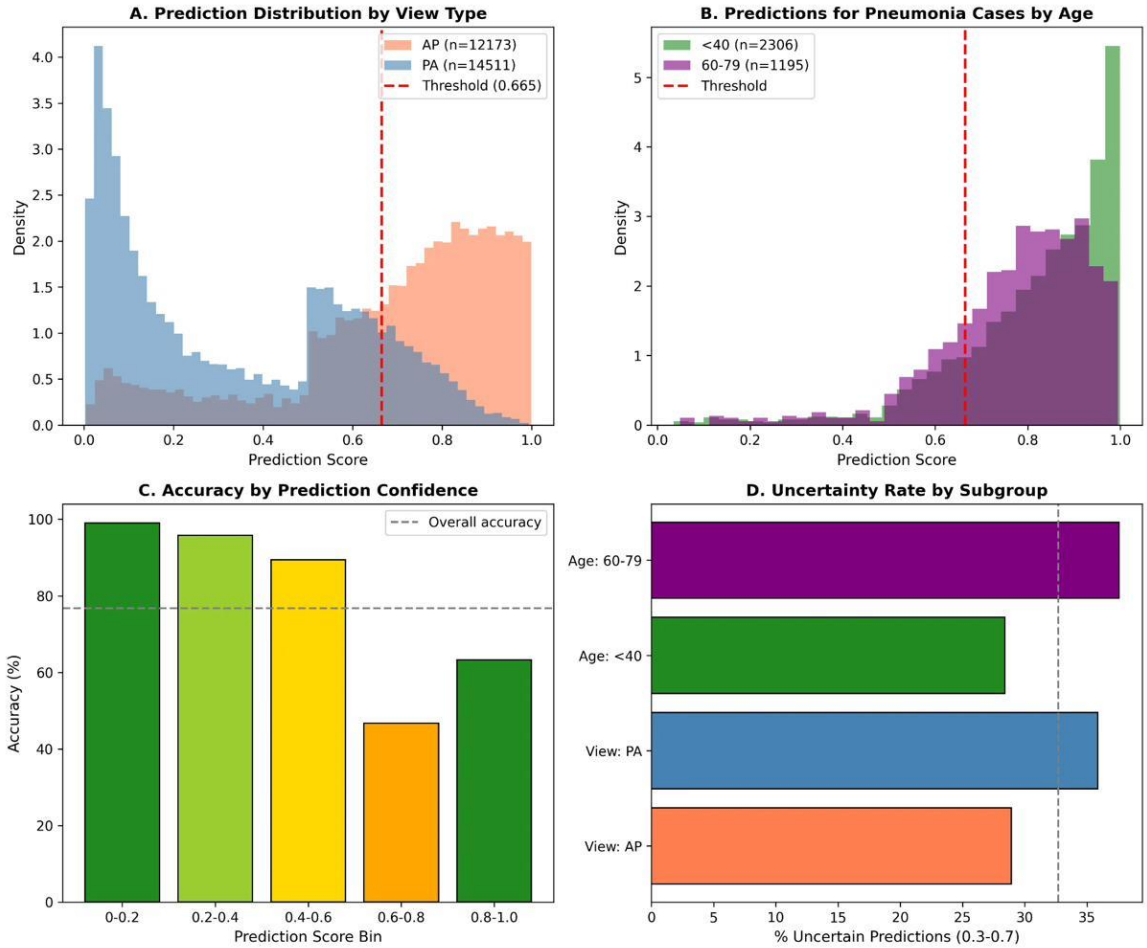
S2 Fig. Multi-model comparison.



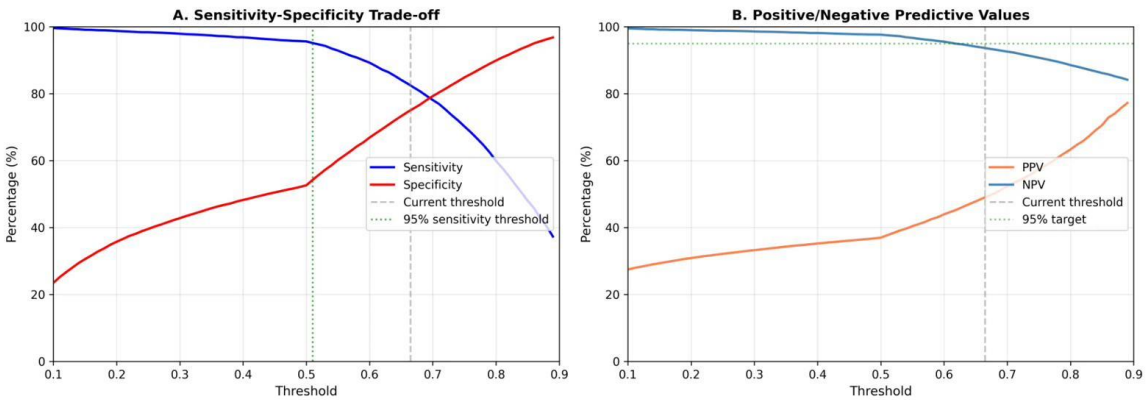
S3 Fig. Error analysis.



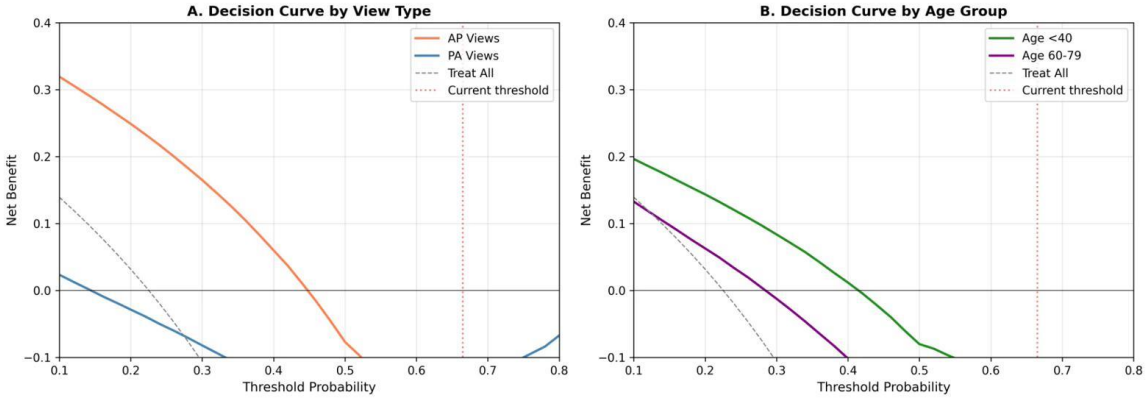
S4 Fig. False positive analysis.



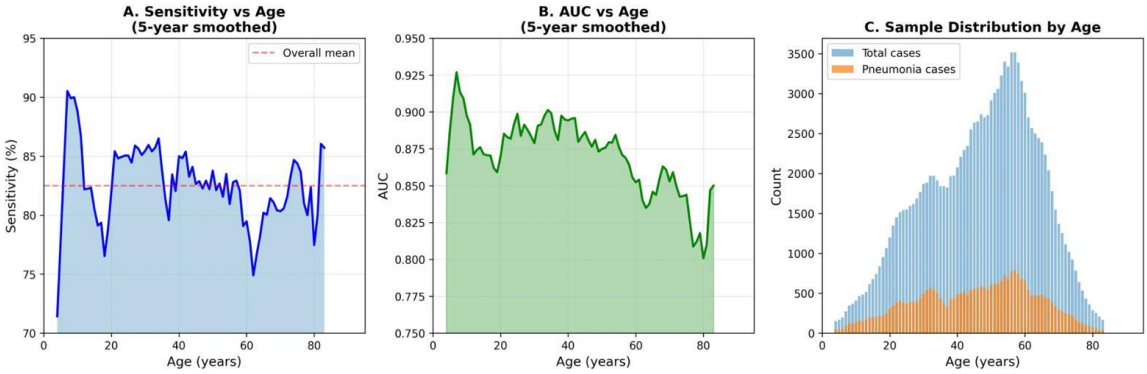
S5 Fig. Confidence analysis.



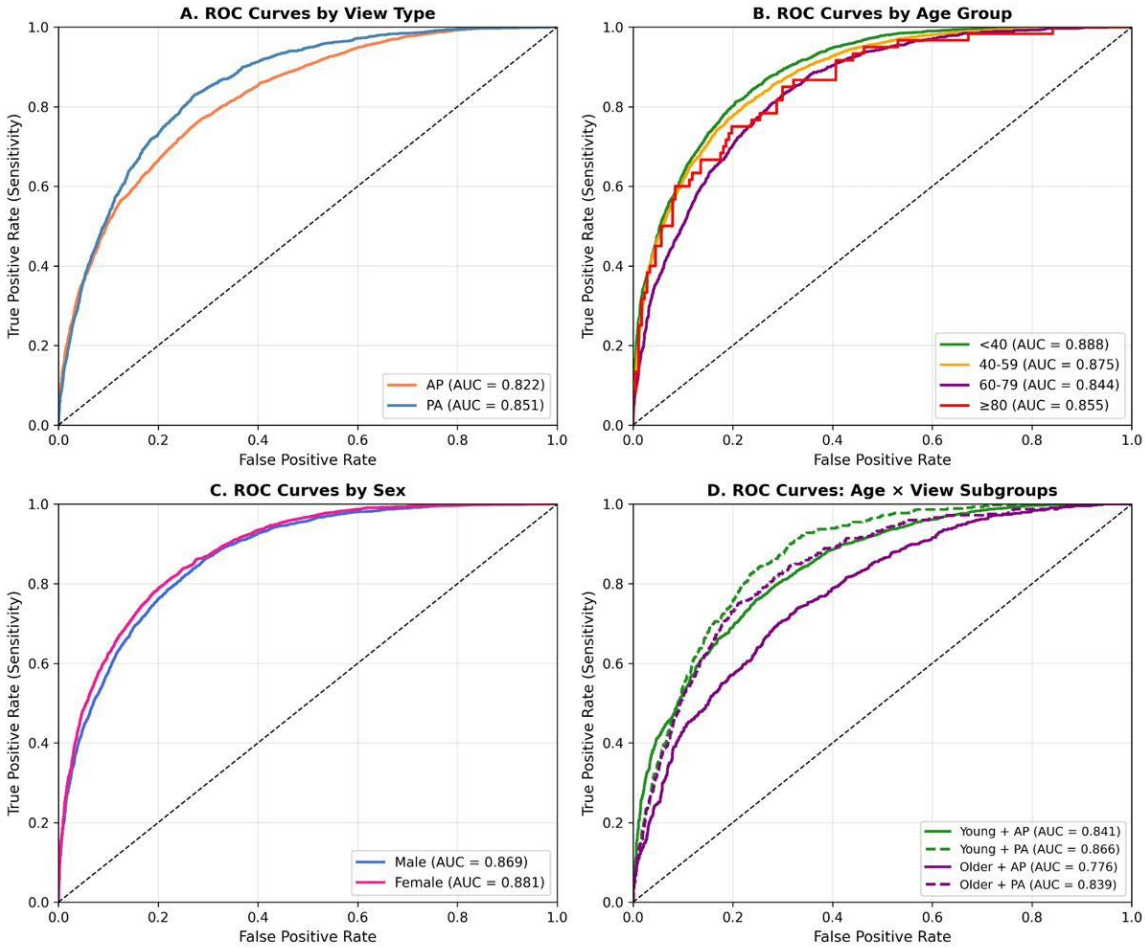
S6 Fig. Threshold optimization.



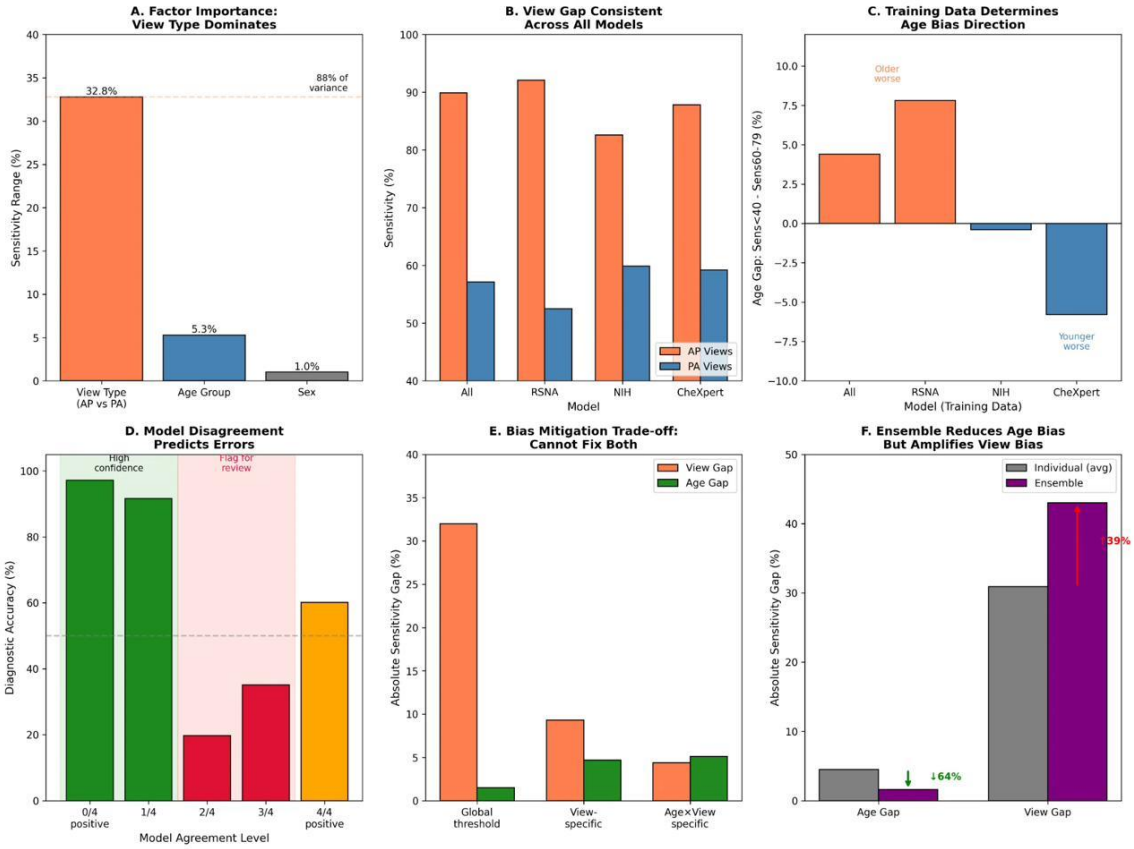
S7 Fig. Decision curves.



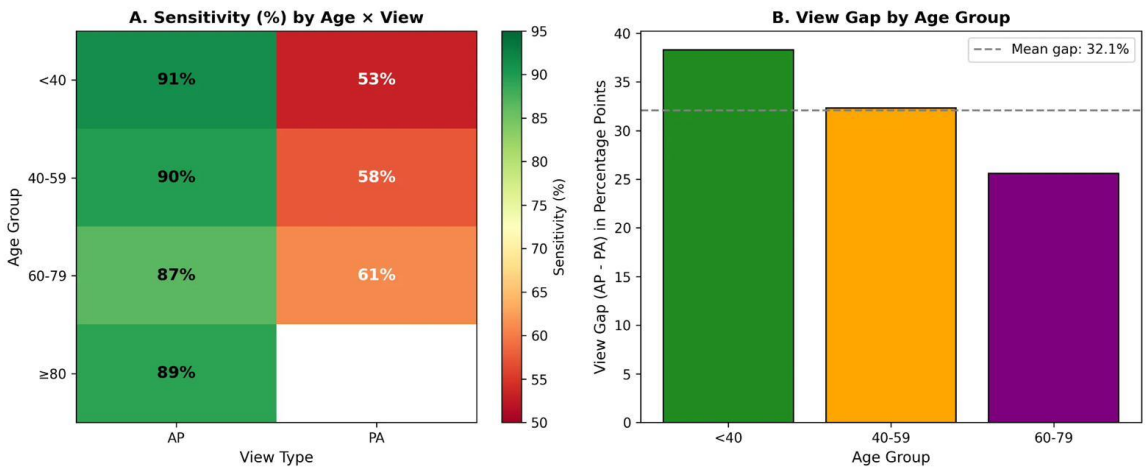
S8 Fig. Age as continuous variable.



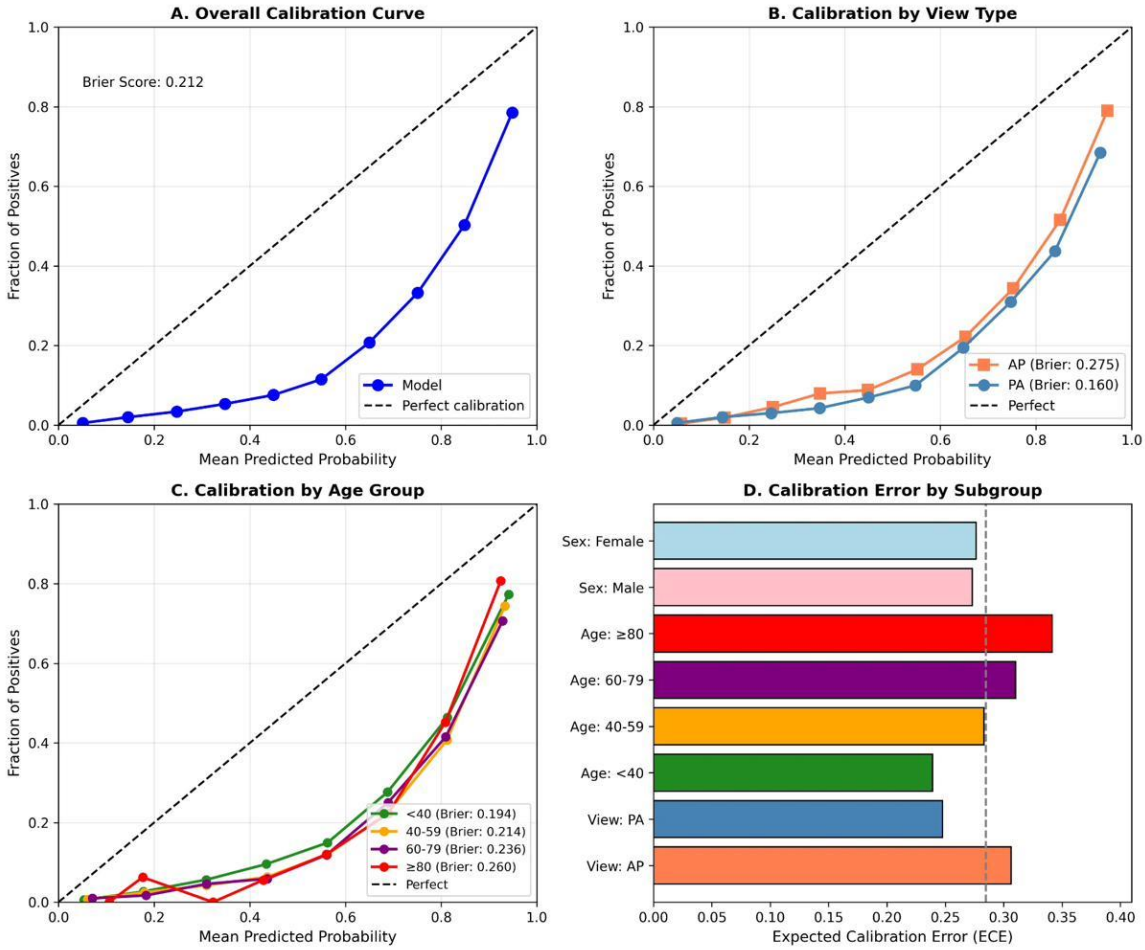
S9 Fig. ROC curves by subgroup.



S10 Fig. Primary model fairness analysis.



S11 Fig. Age × view interaction.



S12 Fig. Calibration analysis.