Cureus
Part of SPRINGER NATURE

Open Access Original Article

Cross-Dataset Reliability of Chest Radiograph Models: A Provenance-Aware Evaluation of Performance and Thresholds

Review began 10/20/2026
Review ended 11/10/2026
Published 11/17/2026

© **Copyright** 2026
Sakib et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119832

Sadman Haque Sakib [1], Sharmin Akter Poppy [2], Murtaza Hussain Shaikh [3], Najmuddin . [4], Md Anik Hosan [5], Irin Jahan Sorna [6], Sk Rakib Ul Islam Rahat [7]

1. Department of Global Business, Kyungsung University, Busan, KOR 2. Department of Business Analytics, Trine University, Detroit, USA 3. Department of Global Business, Dong-Eui University, Busan, KOR 4. Department of Mechatronics, Kyungsung University, Busan, KOR 5. Department of Global IT Engineering, Kyungsung University, Busan, KOR 6. Department of Microbiology, University of Rajshahi, Rajshahi, BGD 7. Department of Computer Science and Engineering, Wright State University, Dayton, USA

Corresponding author: Sadman Haque Sakib, sadmanhaque.official@gmail.com

Abstract

Background

Public chest radiograph datasets attach different label meanings and sources to findings such as pneumonia and lung opacity. A shared label name does not ensure that models are being evaluated against the same target.

Objective

The objective of this study was to compare discrimination, calibration, high-confidence errors, and source-threshold flag volume across four public chest radiograph tasks.

Methods

Four dataset-defined tasks were evaluated: Kaggle/Kermany pneumonia, Radiological Society of North America (RSNA) lung opacity, CheXpert pneumonia, and CheXpert lung opacity. For each image, probabilities from five saved DenseNet-121 seeds per source task were averaged. Each resulting ensemble was evaluated on all four held-out target splits, producing 16 comparisons. Performance measures included area under the receiver operating characteristic curve (AUROC), area under precision-recall curve (AUPRC), balanced accuracy, Brier score, expected calibration error with 15 equal-width bins (ECE15), negative log-likelihood, high-confidence errors, and flags per 1000 images at source-validation Youden thresholds. Percentile intervals were estimated from 500 target-set bootstrap replicates, resampling patients where identifiers were available and images otherwise.

Results

Same-label pneumonia transfer produced ensemble AUROCs of 0.699 (95% CI, 0.664-0.732) from Kaggle/Kermany to CheXpert and 0.758 (0.718-0.797) in the reverse direction. Opacity transfer produced AUROCs of 0.824 (0.808-0.841) from RSNA to CheXpert and 0.835 (0.820-0.847) in the reverse direction. ECE15 ranged from 0.053 to 0.533. For Kaggle pneumonia to RSNA opacity, 28.4% (1096/3853) of all images were both incorrect and at least 0.90 confident. On CheXpert opacity, Kaggle- and RSNA-derived thresholds flagged 959.5 and 348.0 images per 1000, respectively.

Conclusion

Model ranking, probability error, confident mistakes, and threshold-based flag volume varied across dataset-defined tasks. These results support reporting complementary measures and checking label provenance when evaluating image models across datasets. They do not establish clinical effectiveness or validate a clinical workflow.

Categories: Radiology, Healthcare Technology, Infectious Disease
Keywords: artificial intelligence, business analytics, calibration, chest radiography, dataset shift, knowledge engineering, machine learning, medical imaging, model validation

Introduction

Public chest radiograph datasets vary in their label definitions and in how well a model generalizes beyond its development data [1-4]. The pediatric Kaggle/Kermany collection uses folder labels for normal and pneumonia [1]; the Radiological Society of North America (RSNA) pneumonia challenge uses expert annotations of possible opacity [2]; and CheXpert derives findings from radiology reports while retaining

uncertainty in its labels [3]. These targets are related, but they are not the same reference standard. Prior work has also shown that apparent radiographic performance can be affected by site-specific shortcuts, subgroup differences, and avoidable study-design errors [5-7].

This is both a business-analytics and knowledge-engineering problem. The label's meaning, source, and limits need to remain attached to model outputs, while measures such as calibration and flags per 1000 images add information that a ranking score alone cannot provide. A service cannot infer workload from area under the receiver operating characteristic curve (AUROC) alone.

Recent work has examined the transport of pediatric chest radiograph models across datasets and countries [8]. The present analysis compares four dataset-defined pneumonia and opacity tasks, including folder-derived, expert-annotated, and report-derived labels. Its contribution is a provenance-aware evaluation that reports discrimination, probability error, confident mistakes, and the flag volume produced when thresholds selected on source-validation data are carried to target datasets. It does not propose a new transport algorithm or treat the tasks as clinically interchangeable. Reporting follows the principle of making task definitions, data provenance, and evaluation limits explicit [9].

The primary objective was to assess how chest radiograph models trained on one dataset-defined task performed when evaluated on another. We compared discrimination across tasks and also examined calibration, high-confidence errors, and the number of flags produced when source-validation thresholds were applied to target data. Pneumonia and lung opacity were evaluated as separate targets, with each dataset's label definition retained.

Materials And Methods

Study design and datasets

This retrospective benchmark analysis compared four binary tasks assembled from three public chest radiograph resources: the pediatric Kaggle/Kermany pneumonia dataset [1], the RSNA pneumonia challenge dataset [2], and the CheXpert dataset [3]. Each task retained its own label definition and provenance; pneumonia and opacity were not merged into a single clinical endpoint. Table *1* summarizes the held-out target test sets. The four task splits contain 20,095 image-label rows in total, but they do not represent 20,095 independent patients. The two CheXpert tasks draw from the same source cohort and may include the same studies. Counts are therefore reported by task and are not pooled across datasets.

Task	Positive label and provenance	Test n	Positive n (%)
Kaggle/Kermany pneumonia	Folder-derived pneumonia versus normal; pediatric set; no patient identifier in the supplied split	640	398 (62.2%)
RSNA lung opacity	Expert-marked opacity under the project's binary mapping	3853	849 (22.0%)
CheXpert pneumonia	Radiology-report-derived pneumonia finding; uncertain labels excluded	991	688 (69.4%)
CheXpert lung opacity	Radiology-report-derived lung-opacity finding; uncertain labels excluded	14611	13,866 (94.9%)

TABLE 1: Task definitions and held-out test composition

The two CheXpert tasks use the same source cohort and are not independent samples. Counts should not be added as though they represented distinct people. Labels are dataset-defined and are not microbiology results.

RSNA - Radiological Society of North America

Task definitions and data handling

The two CheXpert tasks were split separately from the same source cohort. For each cross-task direction, we checked whether target test patient IDs or image IDs also appeared in source training or validation. We repeated both evaluations on target images from patients absent from those source partitions, using the same saved ensemble and source threshold. The 16 primary comparisons use the full predefined target splits. We report the restricted comparisons separately as point estimates.

An analysis map recorded each task's dataset, target concept, label source, uncertainty handling, and split unit. The Kaggle/Kermany positive class came from the supplied pneumonia folder category [1]. RSNA positives were examinations labeled for lung opacity under the project's binary mapping [2]. CheXpert tasks

used frontal-view, report-derived labels, with uncertain labels excluded from the primary manifests [3]. These similarly named labels come from different annotation processes.

For Kaggle/Kermany [1], the original training folder was stratified by label into 4433 training and 783 source-validation images. The original test and validation folders were combined to form the held-out test split of 640 images. No usable patient identifiers were available. RSNA and the two CheXpert tasks used stratified 70/15/15 splits grouped by the available patient or study identifier. The split audit found no overlap of available group identifiers across partitions. The predefined held-out target split for each task was used in all reported comparisons.

Model development and saved predictions

A separate DenseNet-121 model with ImageNet-pretrained weights [10] was trained for each source task using five random seeds (42, 1337, 2025, 7, and 99), for 20 source-model training runs. Images were converted to 224 × 224 RGB and normalized with ImageNet statistics. The documented training implementation uses batches of 32, random horizontal flips (probability 0.5), and random rotations of up to 7° for training images. Validation and test images use resizing and normalization without random augmentation. The model was optimized with AdamW (learning rate, 1×10^{-4}; weight decay, 1×10^{-4}) and class-weighted binary cross-entropy, with the positive class weighted by the ratio of negative to positive training images. Source-validation AUROC was used to select the best checkpoint and monitor early stopping. Training ran for up to five epochs, with early stopping after three epochs without improvement. We used saved predictions from the source models to evaluate target images without retraining or tuning thresholds on target labels. The 20 source-model runs were executed in Google Colab with GPU acceleration. The checkpoints selected by source-validation AUROC came from epochs 1 to 5, with five of the 20 runs selecting a checkpoint at epoch 5.

Evaluation and analytics

For each of the four source models, each of the four held-out target test splits was evaluated, producing 16 source-to-target comparisons. For each image, predicted probabilities from the five saved seeds were averaged to form a five-seed ensemble estimate. Area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) were used as ranking measures; AUPRC was interpreted alongside target prevalence. Balanced accuracy was calculated at a probability cutoff of 0.50. Probability quality was summarized by the Brier score, expected calibration error with 15 equal-width bins (ECE15), and negative log-likelihood (NLL). ECE15 is the weighted absolute difference between predicted probability and observed frequency within bins [11].

Target-set uncertainty was quantified with 500 percentile-bootstrap replicates for every comparison. Where patient identifiers were available, patients were sampled with replacement, including all target images belonging to each sampled patient. When a target split had no usable patient identifier, images were sampled with replacement; this image-level fallback applied to comparisons involving Kaggle/Kermany as the target. The 2.5th and 97.5th percentiles are reported. These intervals describe uncertainty from the observed target sample conditional on the saved five-seed ensemble. Companion per-seed summaries provide a separate view of training-seed variation. Neither interval captures hospital-to-hospital variation or prospective clinical performance.

For each seed, the Youden threshold, which maximizes sensitivity plus specificity minus one, was selected using source-validation predictions only [12]. The five seed-specific thresholds were then averaged for each source task. This source threshold was applied to the five-seed mean probabilities on every target test set. Positive flags per 1000 images are reported with target-bootstrap intervals. A flag is a hypothetical model output, not an observed service action, staffing estimate, cost, or patient benefit. Thresholds were not tuned on target labels, and results were not pooled across non-equivalent tasks.

High-confidence predictions were defined as predictions for which $\max(p, 1 - p) \geq 0.90$. The fraction of images covered, the fraction of all target images that were both incorrect and covered, and the error fraction among covered images are reported. These measures distinguish how often the models made high-confidence predictions from the error rate within that subset.

The split audit found no duplicate image identifiers in the Kaggle/Kermany manifests. The released prediction audit also checks agreement of target image identifiers and labels across seeds. These identifier checks cannot detect repeated patients or confirm whether identical or near-identical radiographs occur under different identifiers.

We averaged the seed-specific Youden thresholds to use one source-validation operating point per source task with the five-seed mean probabilities. For ECE15, empty bins contributed zero, and each nonempty bin was weighted by its share of test images, including bins with few observations. The bootstrap intervals quantify uncertainty in the held-out target sample for the saved ensemble. We summarized variation across the five training seeds separately.

The public repository listed in the Acknowledgments includes the post-training analysis notebook, the split audit summary, and the metric tables for individual seeds and the five-seed ensembles. The notebook requires saved per-image predictions and source-validation thresholds as inputs. The release provides aggregate results and the expected input format. Raw images, per-image predictions, and model checkpoints are outside the public release.

Results

Shared patients across CheXpert tasks

The two CheXpert cross-task comparisons use the full target splits, which include patients from source training or validation. For pneumonia to opacity, 2443 target images from 594 such patients were identified, including 536 repeated image IDs. For opacity to pneumonia, the corresponding counts were 685 images from 573 patients, including 505 repeated image IDs.

After excluding these patients, the pneumonia-to-opacity evaluation included 12,168 images (11,509 positive). AUROC changed from 0.827 to 0.825, AUPRC from 0.986 to 0.985, and ECE15 from 0.331 to 0.333. At the 0.50 cutoff, balanced accuracy changed from 0.756 to 0.754. High-confidence errors accounted for 0.4% of all target images in both evaluations. The unchanged source threshold flagged 677.4 per 1000 images, compared with 686.9 on the full split.

For opacity to pneumonia, 306 images (153 positive) remained. AUROC changed from 0.765 to 0.746, AUPRC from 0.859 to 0.724, and ECE15 from 0.172 to 0.160. Balanced accuracy changed from 0.701 to 0.663, high-confidence errors from 4.1% to 5.9% of all images, and flags from 563.1 to 343.1 per 1000. Positive prevalence fell from 69.4% to 50.0%, which affects the interpretation of AUPRC and flag volume. These restricted results are point estimates.

Transfer comparisons and variation across seeds

The pneumonia-to-pneumonia and opacity-to-opacity comparisons evaluate transfer between similarly named findings with different reference labels. Comparisons between pneumonia and opacity are cross-concept evaluations. Table *2* includes both types of comparison, along with each model's performance on its own held-out task.

Source → target	n (positive)	AUROC (95% bootstrap CI)	AUPRC (95% bootstrap CI)	Balanced accuracy at 0.50 (95% bootstrap CI)
Kaggle pneumonia → Kaggle pneumonia	640 (398)	0.980 (0.970–0.988)	0.987 (0.980–0.993)	0.819 (0.786–0.851)
Kaggle pneumonia → RSNA opacity	3853 (849)	0.813 (0.798–0.828)	0.505 (0.467–0.538)	0.672 (0.662–0.683)
Kaggle pneumonia → CheXpert pneumonia	991 (688)	0.699 (0.664–0.732)	0.829 (0.796–0.863)	0.573 (0.547–0.599)
Kaggle pneumonia → CheXpert opacity	14,611 (13,866)	0.775 (0.759–0.795)	0.983 (0.980–0.985)	0.576 (0.563–0.592)
RSNA opacity → Kaggle pneumonia	640 (398)	0.906 (0.879–0.931)	0.937 (0.912–0.958)	0.655 (0.628–0.680)
RSNA opacity → RSNA opacity	3853 (849)	0.896 (0.885–0.906)	0.720 (0.687–0.748)	0.810 (0.794–0.824)
RSNA opacity → CheXpert pneumonia	991 (688)	0.768 (0.735–0.801)	0.869 (0.842–0.895)	0.652 (0.625–0.680)
RSNA opacity → CheXpert opacity	14,611 (13,866)	0.824 (0.808–0.841)	0.986 (0.984–0.988)	0.654 (0.642–0.665)
CheXpert pneumonia → Kaggle pneumonia	640 (398)	0.758 (0.718–0.797)	0.802 (0.758–0.851)	0.540 (0.523–0.557)
CheXpert pneumonia → RSNA opacity	3853 (849)	0.818 (0.805–0.832)	0.507 (0.474–0.544)	0.746 (0.731–0.762)
CheXpert pneumonia → CheXpert pneumonia	991 (688)	0.780 (0.746–0.811)	0.863 (0.835–0.892)	0.734 (0.704–0.763)
CheXpert pneumonia → CheXpert opacity	14,611 (13,866)	0.827 (0.810–0.843)	0.986 (0.984–0.988)	0.756 (0.742–0.771)
CheXpert opacity → Kaggle pneumonia	640 (398)	0.945 (0.928–0.960)	0.968 (0.957–0.977)	0.792 (0.769–0.817)
CheXpert opacity → RSNA opacity	3853 (849)	0.835 (0.820–0.847)	0.547 (0.512–0.583)	0.758 (0.742–0.775)
CheXpert opacity → CheXpert pneumonia	991 (688)	0.765 (0.732–0.797)	0.859 (0.826–0.891)	0.701 (0.673–0.729)
CheXpert opacity → CheXpert opacity	14,611 (13,866)	0.870 (0.853–0.884)	0.990 (0.988–0.991)	0.810 (0.796–0.824)

TABLE 2: Five-seed ensemble discrimination and fixed-cutoff classification for all 16 source-to-target comparisons

Each point estimate uses the image-wise mean probability from five saved random-seed models. Parenthetical intervals are 95% percentile intervals from 500 target-set bootstrap replicates, resampling patients where identifiers were available and images otherwise. The Kaggle/Kermany task has no usable patient identifiers. AUPRC should be read alongside the positive prevalence in Table 1.

AUPRC - area under the precision-recall curve; RSNA - Radiological Society of North America

Across the five individual seeds, AUROC ranged from 0.682 to 0.698 for Kaggle/Kermany to CheXpert pneumonia and from 0.656 to 0.822 in the reverse direction. For opacity, the ranges were 0.799 to 0.819 from RSNA to CheXpert and 0.814 to 0.840 in the reverse direction. For RSNA to CheXpert opacity, ECE15 ranged from 0.417 to 0.668 across seeds. The public per-seed table gives the individual values for all 16 comparisons.

For pneumonia transfer, mean individual-seed AUROC was 0.690 (SD 0.007) from Kaggle/Kermany to CheXpert and 0.707 (SD 0.067) in the reverse direction. These standard deviations describe variation across the five runs.

We compared the 95% percentile intervals from the first 500 resamples with those from 2000 resamples for four full-target comparisons: both CheXpert cross-task directions, CheXpert to Kaggle/Kermany pneumonia, and CheXpert to RSNA opacity. The resampling unit was the recorded patient or study identifier when available and the image for Kaggle/Kermany. Across these four comparisons, the largest change in either interval endpoint was 0.003 for AUROC, 0.004 for AUPRC, 0.002 for ECE15, and 1.6 flags per 1000 images.

Target-set composition

The held-out test sets ranged from 640 images in Kaggle/Kermany to 14,611 frontal radiographs in CheXpert opacity (Table *1*). Positive prevalence ranged from 22.0% (849/3853) for RSNA opacity to 94.9% (13,866/14,611) for CheXpert opacity. This spread matters when interpreting AUPRC, predictive values, and the volume of positive model flags. CheXpert pneumonia and opacity come from the same broad cohort, so their task counts are not independent patient totals.

Discrimination and fixed-cutoff classification

The five-seed ensemble AUROC on each model's own held-out task was 0.980 (95% CI, 0.970-0.988) for Kaggle pneumonia, 0.896 (0.885-0.906) for RSNA opacity, 0.780 (0.746-0.811) for CheXpert pneumonia, and 0.870 (0.853-0.884) for CheXpert opacity. Same-label pneumonia transfer yielded AUROCs of 0.699 (0.664-0.732) from Kaggle/Kermany to CheXpert and 0.758 (0.718-0.797) in the reverse direction. Opacity-to-opacity transfer yielded 0.824 (0.808-0.841) from RSNA to CheXpert and 0.835 (0.820-0.847) in the reverse direction. Across all 16 comparisons, ensemble AUROC ranged from 0.699 to 0.980 (Table *2*; Figure *1*). AUPRC results are shown in Figure *2*. A high AUROC across different labels does not show that the two tasks are clinically interchangeable.

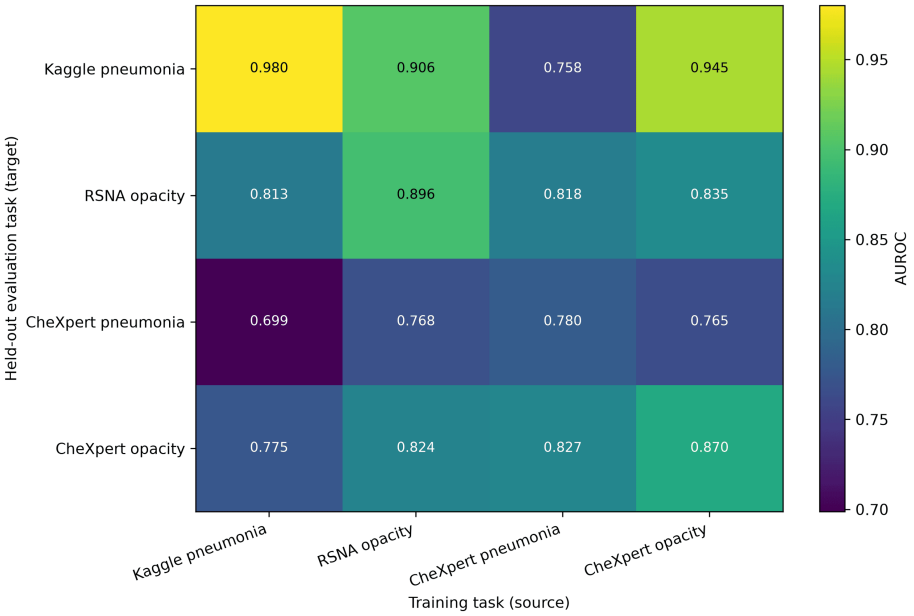


FIGURE 1: Five-seed ensemble AUROC across source-to-target comparisons

AUROC for the five-seed ensemble across all 16 held-out source-to-target comparisons. Higher values indicate better discrimination against each target task's dataset-defined labels.

AUROC - area under the receiver operating characteristic curve; RSNA - Radiological Society of North America

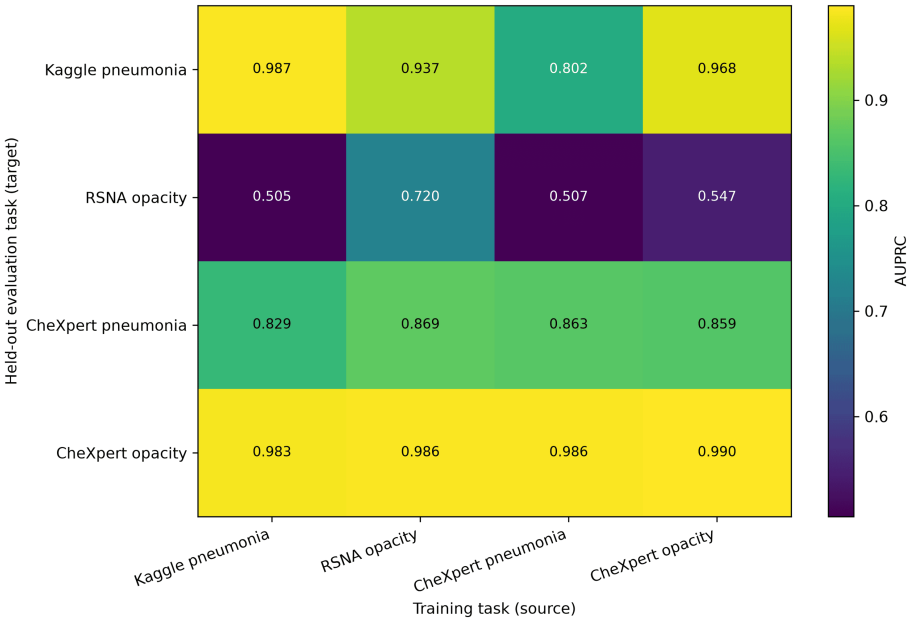


FIGURE 2: Five-seed ensemble AUPRC across source-to-target comparisons

AUPRC for all 16 comparisons. Because AUPRC depends on target prevalence, interpret it alongside the positive prevalence reported in Table 1.

AUPRC - area under the precision-recall curve; RSNA - Radiological Society of North America

Calibration and probability error

Across the 16 comparisons, ECE15 ranged from 0.053 to 0.533 (Table *3*; Figure *3*). RSNA opacity to CheXpert opacity had ECE15 0.533 (0.527-0.540) despite AUROC 0.824 (0.808-0.841). Kaggle pneumonia to CheXpert opacity had ECE15 0.053 (0.048-0.057); its target split was 94.9% (13,866/14,611) positive, and its source threshold flagged nearly all images. ECE therefore needs to be read with prevalence, Brier score, NLL, and threshold behavior rather than as a standalone sign of clinical reliability.

Source → target	n	Brier score	ECE15 (95% bootstrap CI)	NLL
Kaggle pneumonia → Kaggle pneumonia	640	0.100	0.134 (0.112–0.156)	0.349
Kaggle pneumonia → RSNA opacity	3853	0.400	0.475 (0.462–0.487)	1.546
Kaggle pneumonia → CheXpert pneumonia	991	0.242	0.211 (0.190–0.239)	0.970
Kaggle pneumonia → CheXpert opacity	14,611	0.063	0.053 (0.048–0.057)	0.236
RSNA opacity → Kaggle pneumonia	640	0.233	0.288 (0.263–0.321)	0.650
RSNA opacity → RSNA opacity	3853	0.130	0.138 (0.127–0.148)	0.400
RSNA opacity → CheXpert pneumonia	991	0.287	0.321 (0.293–0.347)	0.830
RSNA opacity → CheXpert opacity	14,611	0.360	0.533 (0.527–0.540)	0.994
CheXpert pneumonia → Kaggle pneumonia	640	0.315	0.340 (0.310–0.379)	0.840
CheXpert pneumonia → RSNA opacity	3853	0.187	0.215 (0.201–0.225)	0.561
CheXpert pneumonia → CheXpert pneumonia	991	0.185	0.141 (0.120–0.167)	0.559
CheXpert pneumonia → CheXpert opacity	14,611	0.175	0.331 (0.325–0.337)	0.526
CheXpert opacity → Kaggle pneumonia	640	0.180	0.249 (0.225–0.277)	0.534
CheXpert opacity → RSNA opacity	3853	0.155	0.126 (0.115–0.137)	0.460
CheXpert opacity → CheXpert pneumonia	991	0.215	0.172 (0.151–0.202)	0.639
CheXpert opacity → CheXpert opacity	14,611	0.155	0.289 (0.283–0.295)	0.491

TABLE 3: Probability error and calibration for all 16 comparisons

All values are computed from the five-seed mean probability. Lower values indicate smaller probability errors for the Brier score, ECE15, and negative log-likelihood (NLL). ECE15 uses 15 equal-width probability bins. Bootstrap intervals are conditional on the saved five-seed ensemble.

ECE15 - expected calibration error with 15 equal-width bins; NLL - negative log-likelihood; RSNA - Radiological Society of North America

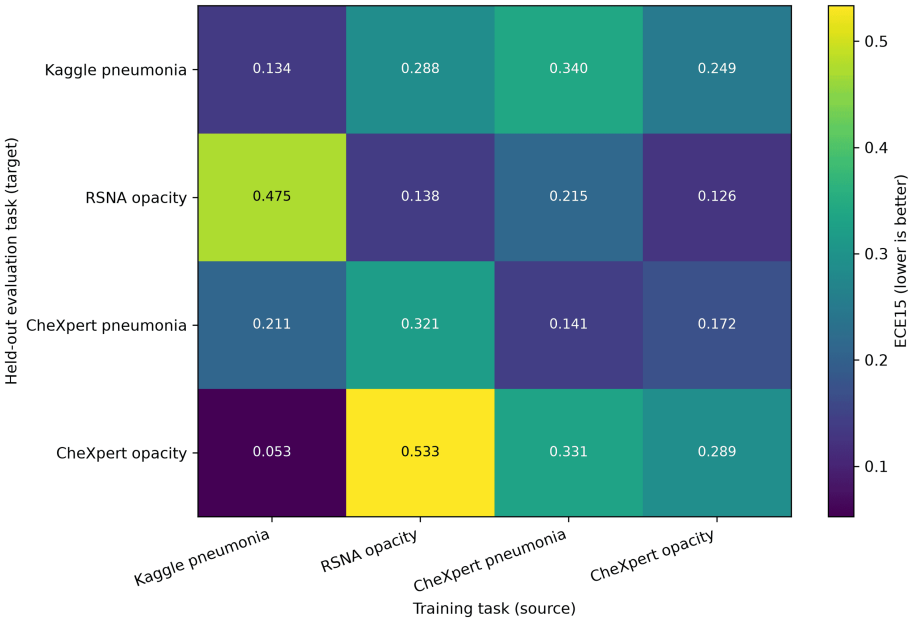


FIGURE 3: Calibration error across source-to-target comparisons

ECE15 for the five-seed ensembles. Lower values indicate smaller average calibration error.

ECE15 - expected calibration error using 15 equal-width probability bins; RSNA - Radiological Society of North America

Confident errors

For Kaggle pneumonia to RSNA opacity, 52.9% (2039/3853) of images met the 0.90 confidence cutoff. Of all target images, 28.4% (1096/3853) were both incorrect and covered; among covered predictions, 53.8% (1096/2039) were wrong (Table *4*). Across all comparisons, coverage and error among covered cases varied independently.

Source → target	n	Covered (confidence ≥0.90), n/N (%)	Wrong and covered, all images, n/N (%)	Wrong among covered, n/N (%)
Kaggle pneumonia → Kaggle pneumonia	640	493/640 (77.0%)	34/640 (5.3%)	34/493 (6.9%)
Kaggle pneumonia → RSNA opacity	3853	2039/3853 (52.9%)	1096/3853 (28.4%)	1096/2039 (53.8%)
Kaggle pneumonia → CheXpert pneumonia	991	702/991 (70.8%)	162/991 (16.3%)	162/702 (23.1%)
Kaggle pneumonia → CheXpert opacity	14,611	12,801/14,611 (87.6%)	529/14,611 (3.6%)	529/12,801 (4.1%)
RSNA opacity → Kaggle pneumonia	640	37/640 (5.8%)	2/640 (0.3%)	2/37 (5.4%)
RSNA opacity → RSNA opacity	3853	1500/3853 (38.9%)	48/3853 (1.2%)	48/1500 (3.2%)
RSNA opacity → CheXpert pneumonia	991	201/991 (20.3%)	62/991 (6.3%)	62/201 (30.8%)
RSNA opacity → CheXpert opacity	14,611	1235/14,611 (8.5%)	834/14,611 (5.7%)	834/1235 (67.5%)
CheXpert pneumonia → Kaggle pneumonia	640	35/640 (5.5%)	1/640 (0.2%)	1/35 (2.9%)
CheXpert pneumonia → RSNA opacity	3853	571/3853 (14.8%)	79/3,853 (2.1%)	79/571 (13.8%)
CheXpert pneumonia → CheXpert pneumonia	991	138/991 (13.9%)	23/991 (2.3%)	23/138 (16.7%)
CheXpert pneumonia → CheXpert opacity	14,611	1041/14,611 (7.1%)	54/14,611 (0.4%)	54/1041 (5.2%)
CheXpert opacity → Kaggle pneumonia	640	117/640 (18.3%)	3/640 (0.5%)	3/117 (2.6%)
CheXpert opacity → RSNA opacity	3853	1410/3853 (36.6%)	32/3853 (0.8%)	32/1410 (2.3%)
CheXpert opacity → CheXpert pneumonia	991	173/991 (17.5%)	41/991 (4.1%)	41/173 (23.7%)
CheXpert opacity → CheXpert opacity	14,611	1244/14,611 (8.5%)	321/14,611 (2.2%)	321/1244 (25.8%)

TABLE 4: High-confidence coverage and errors for the five-seed ensembles

A prediction is covered when max(p, 1−p) ≥ 0.90. Cells report n/N (%). "Wrong and covered, all images" divides incorrect covered predictions by the full target set; "wrong among covered" divides by covered predictions only. Values are point estimates for the five-seed ensemble.

RSNA - Radiological Society of North America

Source-selected thresholds and target flag volume

Source-validation thresholds produced substantially different flag volumes on the same target split (Table 5; Figure 4). On CheXpert opacity, the Kaggle pneumonia threshold produced 959.5 (95% CI, 955.6-963.0) positive flags per 1000 images; the RSNA opacity threshold produced 348.0 (334.4-361.2). The first count is about 2.8 times the second. These counts are a benchmark workload proxy, not observed test orders or evidence that either operating point is appropriate for a clinical workflow.

Source → target	Mean source threshold	Flags per 1000 (95% bootstrap CI)
Kaggle pneumonia → Kaggle pneumonia	0.415	768.8 (735.9–802.4)
Kaggle pneumonia → RSNA opacity	0.415	740.5 (727.0–752.0)
Kaggle pneumonia → CheXpert pneumonia	0.415	877.9 (855.3–898.2)
Kaggle pneumonia → CheXpert opacity	0.415	959.5 (955.6–963.0)
RSNA opacity → Kaggle pneumonia	0.513	192.2 (159.4–221.1)
RSNA opacity → RSNA opacity	0.513	322.9 (308.6–337.8)
RSNA opacity → CheXpert pneumonia	0.513	336.0 (300.2–367.0)
RSNA opacity → CheXpert opacity	0.513	348.0 (334.4–361.2)
CheXpert pneumonia → Kaggle pneumonia	0.525	56.2 (42.2–73.4)
CheXpert pneumonia → RSNA opacity	0.525	387.7 (371.3–403.2)
CheXpert pneumonia → CheXpert pneumonia	0.525	570.1 (534.0–600.3)
CheXpert pneumonia → CheXpert opacity	0.525	686.9 (675.1–697.0)
CheXpert opacity → Kaggle pneumonia	0.475	376.6 (339.1–415.6)
CheXpert opacity → RSNA opacity	0.475	357.4 (342.2–373.6)
CheXpert opacity → CheXpert pneumonia	0.475	563.1 (526.0–595.6)
CheXpert opacity → CheXpert opacity	0.475	773.4 (763.1–783.2)

TABLE 5: Positive flags per 1000 target images at source-validation thresholds

Each threshold is the mean of five Youden thresholds selected on source-validation predictions and is applied unchanged to the five-seed mean probabilities on the target set. A flag is a model output, not an observed laboratory order, staffing requirement, cost, or patient benefit.

RSNA - Radiological Society of North America

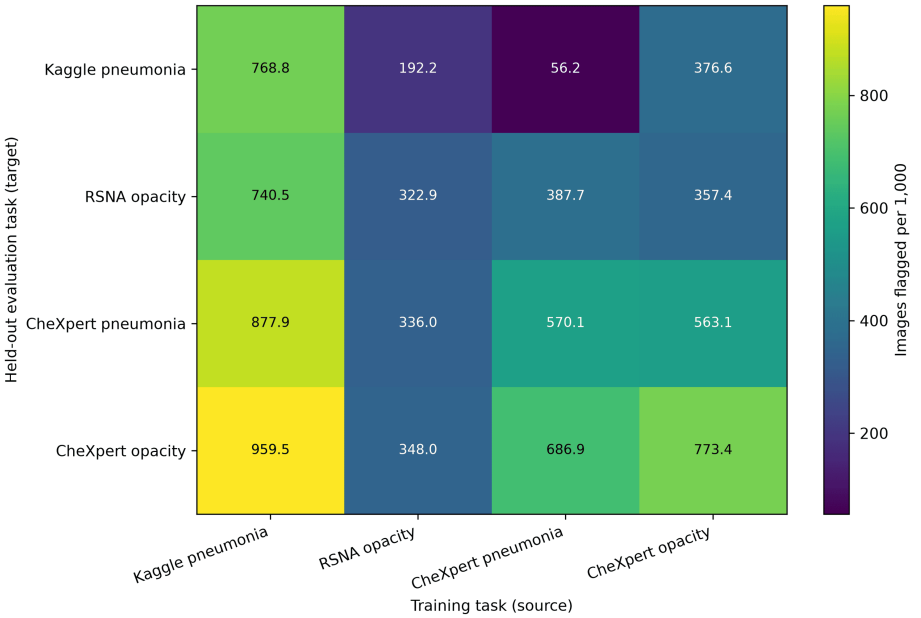


FIGURE 4: Positive flags at source-selected thresholds

Positive flags per 1000 target images using thresholds selected on source-validation predictions and applied unchanged to target predictions. A flag is a model output, not an observed laboratory order or evidence of clinical benefit.

RSNA - Radiological Society of North America

Figure *5* shows the reliability diagrams and image counts by probability bin for pneumonia transfer from CheXpert to Kaggle/Kermany and lung opacity transfer from CheXpert to RSNA.

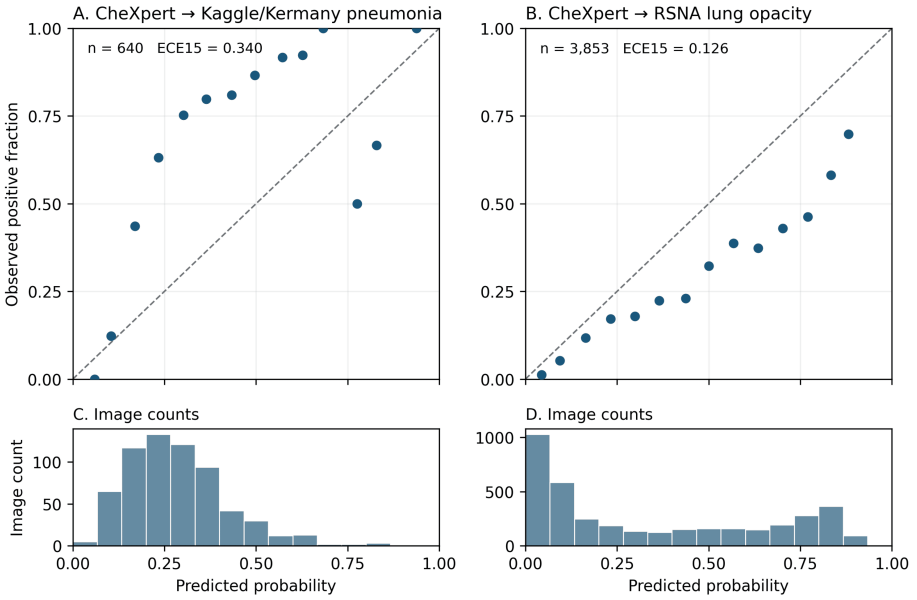


FIGURE 5: Reliability diagrams for pneumonia and lung opacity transfer

The left panel shows CheXpert pneumonia → Kaggle/Kermany pneumonia transfer (640 images, ECE15 = 0.340). The right panel shows CheXpert → RSNA lung opacity transfer (3853 images, ECE15 = 0.126). Each point compares the mean predicted probability from the five-model ensemble with the observed positive fraction in one of 15 equal-width probability bins. The dashed line indicates perfect calibration. The lower panels show the number of images in each bin.

ECE15 - expected calibration error with 15 equal-width bins; RSNA - Radiological Society of North America

Discussion

This benchmark shows that ranking, probability quality, confident errors, and threshold flag volume describe different parts of cross-dataset behavior. Same-label pneumonia transfer was not symmetric: the five-seed AUROC was 0.699 (95% CI, 0.664-0.732) from Kaggle/Kermany to CheXpert and 0.758 (0.718-0.797) in the reverse direction. Opacity-to-opacity transfer was higher in both directions, at 0.824 (0.808-0.841) and 0.835 (0.820-0.847). These differences cannot be attributed to label provenance alone, consistent with prior evidence that dataset composition and confounding can affect external performance [4,5]. The datasets also differ in patient mix, age, prevalence, image acquisition, processing, and annotation, and the two CheXpert tasks draw on the same underlying cohort [1-3].

Several cross-task AUROCs were higher than some same-label transfers. That does not establish that pneumonia and opacity are equivalent, or that a score trained on one task is suitable for another [1-3]. AUROC measures ranking against the target labels used here; it does not establish a shared clinical meaning, a valid target probability, or improved care [7,11]. AUPRC also depends on prevalence: the CheXpert opacity set with 94.9% positive images (13,866/14,611) has a high baseline, so a high AUPRC must be read against that prevalence [13].

Calibration further separates ranking from probability reliability [11]. RSNA opacity to CheXpert opacity had AUROC 0.824 but ECE15 0.533. In the reverse opacity direction, AUROC was 0.835, and ECE15 was 0.126. The minimum ECE15 in the matrix occurred for Kaggle pneumonia to CheXpert opacity, a highly imbalanced target on which the source threshold flagged most images. ECE15 should therefore be interpreted with prevalence, Brier score, negative log-likelihood, and threshold behavior rather than as a standalone measure of model usefulness [9,11].

The confident-error result for Kaggle pneumonia to RSNA opacity is also relevant: 53.8% (1096/2039) of predictions meeting the high-confidence cutoff were wrong. At a source-derived threshold, the Kaggle model generated about 959 positive flags per 1000 images on CheXpert opacity, compared with 348 from the RSNA model on the same target. These counts make threshold differences visible in a common unit, but the study did not observe a service queue, downstream action, staffing decision, or cost. Youden's index weights sensitivity and specificity equally [12]; it does not represent the consequences of errors in a particular service. Prior transport work likewise shows that ranking and source-defined operating-point behavior need separate evaluation [8].

The study overlaps with the 2026 pediatric transport preprint cited above [8], which also examines

discrimination, calibration, and operating-point behavior across datasets. This analysis keeps four dataset-defined tasks separate, includes folder-derived pneumonia, expert-marked opacity, and report-derived findings, and presents all 16 source-target comparisons with target-sample uncertainty. Its contribution is a focused comparison of label provenance and multiple evaluation measures, not a new transport method, a first demonstration of dataset shift, or evidence of clinical superiority [4,5,8].

For business analytics, positive flags per 1000 images provide a more operationally readable comparison than AUROC alone, but remain hypothetical until connected to a defined workflow and measured downstream actions and costs [7,8]. For knowledge engineering, the task map records each label concept, source, uncertainty handling, and split unit alongside model outputs. This makes the assumptions behind each comparison easier to inspect and reuse, consistent with transparent reporting of data provenance and task definitions [9]. The present data do not establish how these measures would affect clinical decisions or patient outcomes.

Strengths

The analysis has several strengths beyond the number of source-target comparisons. It preserves four distinct task definitions rather than merging pneumonia and opacity into one endpoint, and it makes the label source and uncertainty handling explicit. Five saved random-seed models were evaluated for each task, with the ensemble examined using ranking, fixed-cutoff classification, calibration, confidence-error, and threshold-volume measures. Thresholds were selected on source-validation data and held unchanged during target testing, reducing the risk of choosing operating points from target labels [7,9]. Patient-cluster bootstrap intervals were used when identifiers were available, and the image-level fallback for Kaggle/Kermany is stated explicitly. Finally, the manuscript separates benchmark predictions from clinical actions and outcomes, while the public code, aggregate tables, and figures support reproducibility, in line with reporting recommendations [9].

Limitations

Some target patients in both CheXpert cross-task comparisons also appeared in source training or validation. The patient-disjoint sensitivity analysis excluded all target images from those patients. This exclusion also changed the case mix, particularly when pneumonia was the target, so the changes in performance reflect both the exclusion and the resulting change in the target set.

This retrospective analysis used public benchmark datasets rather than external hospital cohorts. Label definition is entangled with population, prevalence, age, equipment, acquisition, preprocessing, and annotation, so the independent contribution of each factor cannot be isolated; earlier work has demonstrated related confounding and shortcut concerns [4,5]. Only one architecture family and five saved seeds per source task were studied. The bootstrap intervals describe target-sample variation conditional on the saved ensemble; they do not measure variation across hospitals or establish prospective performance. The two CheXpert tasks also come from the same source cohort.

Patient-cluster resampling was possible when identifiers were available. The Kaggle/Kermany split lacks a usable patient identifier, so comparisons using it as a target relied on image-level resampling and may not account for repeated-patient overlap. CheXpert labels are report-derived, and uncertain labels were excluded; none of the dataset labels is an independent clinical outcome [1-3]. The source threshold used Youden's index, which weights sensitivity and specificity equally and does not encode the consequences of errors in a specific workflow [12]. Flag counts therefore cannot be treated as observed orders, workload, or cost.

No linked laboratory testing, treatment, or patient-outcome data were available. The study does not establish clinical benefit, diagnostic impact, safe changes to care, or performance in a prospective setting. No subgroup or clinician-interaction evaluation was performed. Prior work on methodological limitations and subgroup underdiagnosis supports evaluating these dimensions before drawing clinical-use conclusions [6,7,9]. These boundaries should be considered when interpreting the benchmark results.

The Kaggle/Kermany folder split leaves two distinct uncertainties. A patient could appear in more than one folder, which could affect the within-dataset estimate. Repeated examinations in the test folder could also make an image-level interval narrower than an interval that accounts for clustering. We could not quantify either effect without patient identifiers. The study used one architecture and public benchmark cohorts, so the transfer results may differ for other model families or hospital populations.

The 94.9% positive prevalence in the CheXpert opacity test set raises the baseline for AUPRC and changes the mix of cases that contributes to the calibration metrics. Positive flag volume also depends on that mix and on the threshold. We therefore interpret these measures alongside the prevalence of each target split. The comparison with prior pediatric transport work [8] concerns the use of multiple label sources and distinct target definitions across four tasks. Because label provenance varies with the dataset, the study cannot estimate its independent effect.

Conclusions

Across four public chest radiograph tasks, model behavior varied by source, target, and label definition. The direction-dependent results show why discrimination, calibration, confident errors, and threshold-based flag volume should be reported together. A high ranking score on one task did not guarantee reliable probabilities or a similar operating point on another.

The task map and public analysis materials make the label assumptions behind these comparisons visible. The reported flag counts are hypothetical outputs, not measured clinical workload or benefit. These results support careful benchmark interpretation and local external validation before operational use; they do not establish clinical effectiveness.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sk Rakib Ul Islam Rahat, Sadman Haque Sakib

Acquisition, analysis, or interpretation of data: Sk Rakib Ul Islam Rahat, Najmuddin ., Md Anik Hosan, Irin Jahan Sorna, Sharmin Akter Poppy, Murtaza Hussain Shaikh, Sadman Haque Sakib

Critical review of the manuscript for important intellectual content: Sk Rakib Ul Islam Rahat, Najmuddin ., Md Anik Hosan, Irin Jahan Sorna, Sharmin Akter Poppy, Murtaza Hussain Shaikh, Sadman Haque Sakib

Supervision: Sk Rakib Ul Islam Rahat, Murtaza Hussain Shaikh

Drafting of the manuscript: Sharmin Akter Poppy

Disclosures

Human subjects: Informed consent for treatment and open access publication was obtained or waived by all participants in this study. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The source datasets are available from their respective providers under the providers' terms and are not redistributed here. The analysis notebook, code, aggregate result tables, and figures are publicly available at https://github.com/skrakibulislamrahat/chest-xray-microbiology-analytics. Per-image predictions and trained model checkpoints are not publicly shared. The authors used ChatGPT (OpenAI, San Francisco, California) to assist with manuscript preparation and language editing. All final content was reviewed and approved by the authors, who take responsibility for the article

References

1. Kermany DS, Goldbaum M, Cai W, et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018, 172:1122-31. 10.1016/j.cell.2018.02.010
2. Shih G, Wu CC, Halabi SS, et al.: Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell. 2019, 1:e180041. 10.1148/ryai.2019180041
3. Irvin J, Rajpurkar P, Ko M, et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell. 2019, 33:590-7. 10.1609/aaai.v33i01.3301590
4. 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:e1002683. 10.1371/journal.pmed.1002683
5. DeGrave AJ, Janizek JD, Lee SI: AI for radiographic COVID-19 detection selects shortcuts over signal . Nat Mach Intell. 2021, 3:610-9. 10.1038/s42256-021-00338-7
6. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021, 27:2176-82. 10.1038/s41591-021-01595-0
7. Roberts M, Driggs D, Thorpe M, et al.: Common pitfalls and recommendations for using machine learning to

detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021, 3:199-217. 10.1038/s42256-021-00307-0

8. Nazim-E-Alam: Cross-dataset transportability of pediatric chest X-ray deep learning across three countries: discrimination, calibration, operating-point failure, and limited-label recovery [PREPRINT]. Electr Eng Syst Sci. 2026, 10.48550/arXiv.2609.05140
9. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr: Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024, 6:e240300. 10.1148/ryai.240300
10. Huang G, Liu Z, van der Maaten L, Weinberger KQ: Densely connected convolutional networks . Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 4700-8. 10.1109/CVPR.2017.243
11. Guo C, Pleiss G, Sun Y, Weinberger KQ: On calibration of modern neural networks . Proceedings of the 34th International Conference on Machine Learning. 2017, 70:1321-30.
12. Youden WJ: Index for rating diagnostic tests . Cancer. 1950, 3:32-5. 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
13. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015, 10:e0118432. 10.1371/journal.pone.0118432