17th Conference on Data Analysis Methods for Software Systems
DAMSS 2026 · November 26–28, 2026 · Druskininkai, Lithuania
Poster I-02 · Session I: AI in Health and Life Sciences

# Cross-Institution Reliability of Multi-Label Chest X-Ray Classification Software:

## Discrimination, Calibration and Subgroup Fairness under Domain Shift

**Evaluation outputs and analysis code**
github.com/PauleNajus/mcads-model-evaluations

**Paulius BUNDZA, Justas TRINKŪNAS**
Department of Information Systems, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania
paulius.bundza@vilniustech.lt, justas.trinkunas@vilniustech.lt

## Introduction

Chest X-ray classifiers are benchmarked mostly on data from their training institutions (in-distribution, ID). Elsewhere they meet new populations, equipment and labelling conventions (out-of-distribution, OOD): in a systematic review of 86 algorithms, 81% performed worse on external data [1]. Yet public model weights [2] come with ID benchmarks only.

## Aim

To quantify how reliably seven public, pre-trained multi-label chest radiograph classifiers perform, without fine-tuning, at an institution they were not trained on: discrimination, calibration, default operating points and subgroup performance.

## Methods

**Software under test**
Seven pre-trained classifiers from TorchXRayVision 1.4.0 [2]:
- six DenseNet-121 (224 × 224 px)
- one ResNet-50 (512 × 512 px)

Fixed weights, no fine-tuning.

**ID reference: NIH ChestX-ray14, USA [3]**
Three models trained on NIH data, tested on patients held out from training: 59,303 images (DenseNet-121), 22,256 (ResNet-50).

**OOD test: BRAX v1.1.0, São Paulo, Brazil [4]**
All seven models: 18,424 studies of 15,112 patients, one frontal image per study; labels from Portuguese reports.

**Evaluation on 7 common labels:** atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumonia, pneumothorax
**Discrimination:** AUROC; primary endpoint: macro AUROC
**Calibration:** Brier score, ECE, slope; Platt / isotonic recalibration
**Operating points:** sensitivity, specificity at default thresholds
**Subgroups:** sex, age, view position (AP/PA), manufacturer
**Statistics:** patient-level bootstrap (2000 replicates); Holm adjustment

## Results · Discrimination under domain shift

**In-distribution benchmarks overestimated external performance: macro AUROC fell by 6.0–11.4 percentage points.**

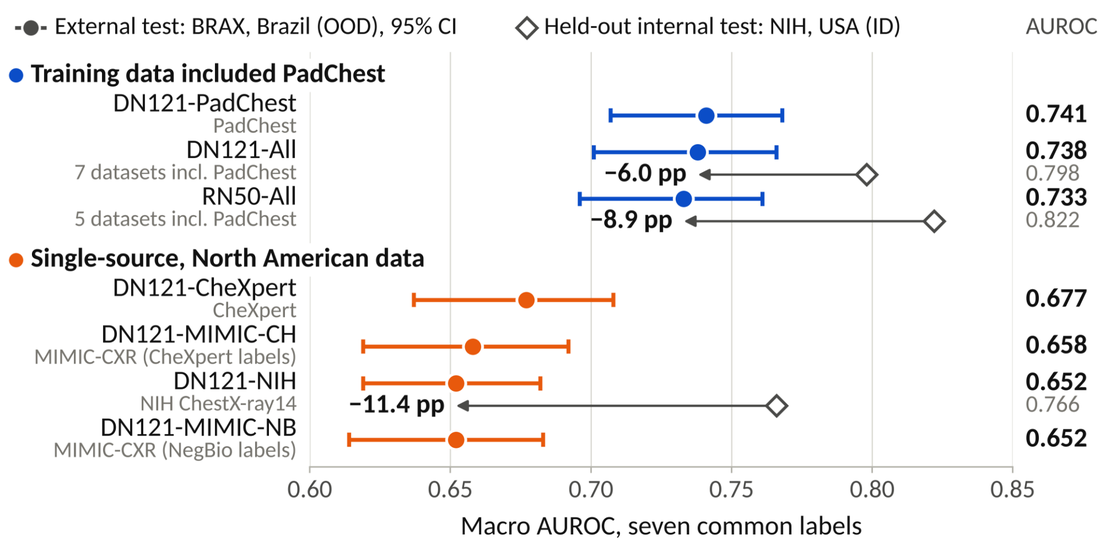


**Figure 1.** Macro AUROC on the seven common labels in the external BRAX cohort (OOD) and in held-out NIH patients (ID); arrows show the change in percentage points (pp).

- External macro AUROC: 0.733–0.741 for models trained with PadChest [5] vs 0.652–0.677 for single-source models (North American data).
- All 12 between-group differences were significant (Holm-adjusted $p \le 0.002$); none of the nine within-group differences was.
- Not attributable to PadChest itself: multi-source training, architecture and label definitions are confounded.

## Results · Calibration and operating points

**Default scores and thresholds did not transfer.**

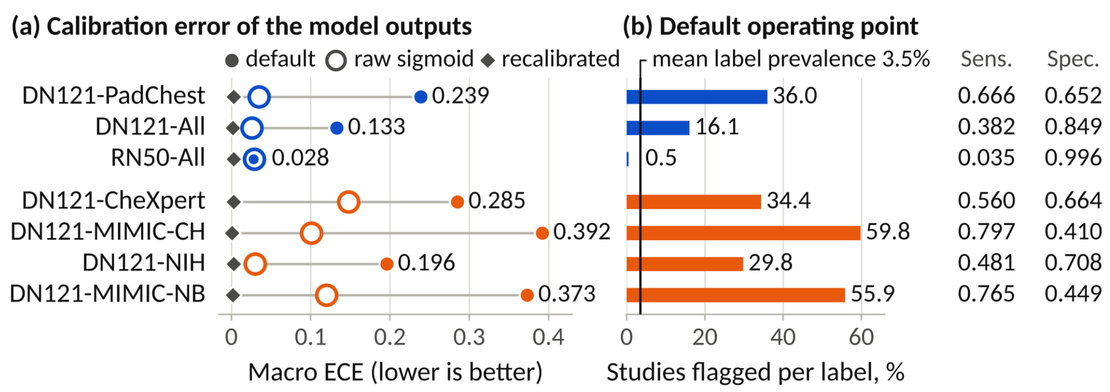


**Figure 2.** BRAX cohort, macro averages over the seven labels; colours as in Figure 1. ECE, expected calibration error; recalibration by Platt scaling.

- Default ECE (0.028–0.392) largely reflected the library's score normalization; local recalibration reduced it to 0.001–0.003.
- Default thresholds flagged 0.5–59.8% of studies per label at a mean prevalence of 3.5%; sensitivity ranged from 0.035 to 0.797.

## Results · Subgroup fairness

**The largest gaps followed projection and age, not sex.**

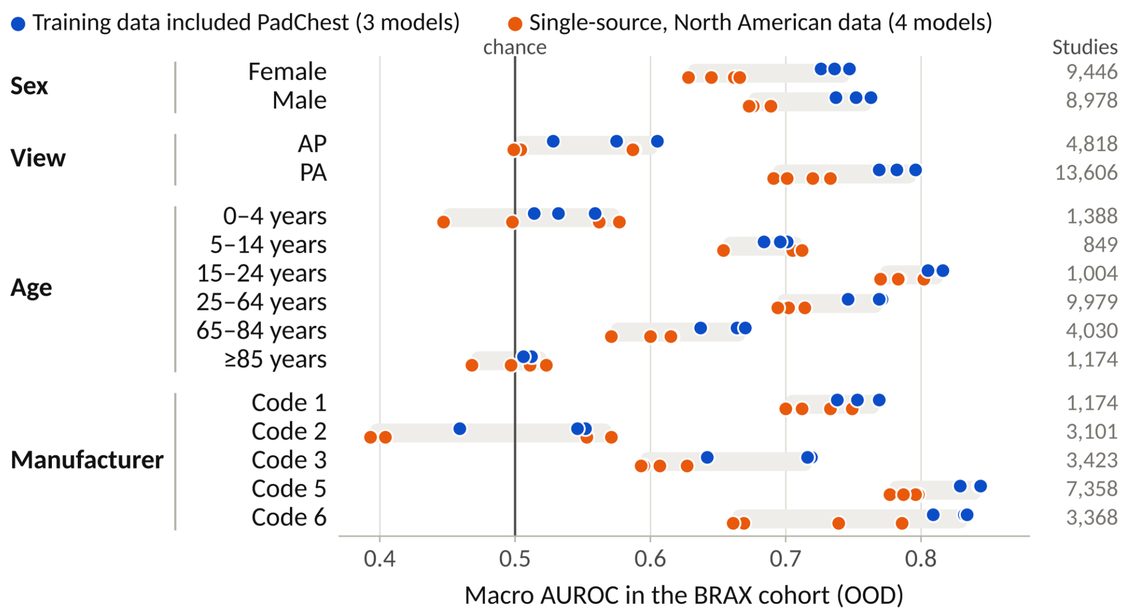


**Figure 3.** Macro AUROC in BRAX subgroups; each dot is one model. Labels averaged: six (sex, view), four (age), five (manufacturer; anonymized codes).

- AP vs PA radiographs: 0.499–0.605 vs 0.691–0.796, a gap of 10.3–26.9 pp that persisted within age groups.
- Age: best at 15–24 years (0.770–0.816), worst at 0–4 years (0.447–0.577) and ≥85 years (0.468–0.523).
- Manufacturer code 2: lowest for every model (0.393–0.571), partly explained by its case mix (96.2% AP radiographs).
- Sex: differences of at most 0.045; all 95% CIs include zero.

## Conclusions

1. ID benchmarks overestimated external performance: macro AUROC fell by 6.0–11.4 pp in the three models with an ID reference.
2. Models trained with PadChest generalized significantly better, but this cannot be attributed to PadChest itself.
3. Default outputs and operating points did not transfer; local recalibration was effective (ECE ≤ 0.003).
4. Discrimination was markedly lower for AP radiographs, the youngest and oldest patients and one manufacturer.

**Limitations:** one external hospital; report-derived labels; descriptive subgroup analyses (associations, not causal effects).

**Next steps:** further external cohorts, including prospective data from Lithuanian hospitals; site-specific recalibration; domain adaptation and foundation models.

**References**
[1] Yu A.C., Mohajer B., Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol. Artif. Intell. 2022, 4, e210064.
[2] Cohen J.P., et al. TorchXRayVision: a library of chest X-ray datasets and models. Proc. MIDL 2022, 231–249.
[3] Wang X., et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks. Proc. CVPR 2017, 3462–3471.
[4] Reis E.P., de Paiva J.P.Q., da Silva M.C.B., et al. BRAX, Brazilian labeled chest x-ray dataset. Sci. Data 2022, 9, 487.
[5] Bustos A., Pertusa A., Salinas J.-M., de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. 2020, 66, 101797.

**Abbreviations:** AP/PA, anteroposterior/posteroanterior; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; ECE, expected calibration error; DN121, DenseNet-121; RN50, ResNet-50.