

When Repository Labels Are Not Image-Level Truth: A Supervision Auditing Framework for Chest Radiograph AI

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Large public chest radiograph repositories have become the primary source of supervision for developing medical imaging AI. However, repository labels are typically derived automatically from radiology reports rather than direct image review and are widely used as image-level ground truth despite limited validation. We introduce *Repository Supervision Auditing (RSA)*, a framework for validating repository-derived labels against expert image-level annotations before model development. Using cardiomegaly as a representative case study, we audit the report-derived labels distributed with MIMIC-CXR, characterize the sources of disagreement, and construct an expert-curated cohort for deployment-oriented evaluation. Repository-derived cardiomegaly labels exhibited near-zero agreement with expert image-level assessment ($\kappa = 0.011$), despite a high positive predictive value (0.875). Among 1,080 expert-confirmed positive studies, only 14 received a positive repository label. Discordance arose predominantly from routine non-mention rather than explicit report negation, with expert-confirmed cardiomegaly identified in nearly half of studies assigned a repository-derived *No Finding* label. Using the curated cohort, DenseNet121 achieved an ROC-AUC of 0.853 (95% CI 0.803–0.893). These findings establish Repository Supervision Auditing as a practical framework for validating repository-derived learning targets before developing trustworthy medical imaging AI.

Keywords: CXR; Repository Supervision Auditing; Trustworthy AI.

1 Introduction

Large-scale public chest radiograph repositories have become the foundation of modern AI for medical imaging. Datasets such as MIMIC-CXR, CheXpert, ChestX-ray14, and PadChest provide millions of radiographs paired with reports or automatically derived labels and now underpin the development, pretraining, and evaluation of deep learning and vision-language models for chest radiograph interpretation [8, 6, 16, 3, 2, 15]. As these repositories increasingly support foundation models and deployment-oriented AI systems, the validity of the supervision used during model development has become increasingly important.

Most repository labels are not obtained through direct image review but are automatically extracted from radiology reports using natural language processing systems such as CheXpert and NegBio [6, 12]. Although this approach enables scalable dataset construction, report-derived labels reflect clinical documentation rather than a comprehensive image-level assessment. Findings that are stable, morphometric, or clinically secondary may remain undocumented despite being visually apparent, causing repository-derived labels to encode reporting practice as much as radiographic evidence. Previous work has improved report label extraction [12, 6], characterized hidden stratification in medical imaging [11], and reported disagreement between report-derived labels and expert image annotations [7]. However, repository-derived labels are still routinely treated as image-level ground truth during model development. As foundation models trained on public repositories move toward clinical deployment, this assumption becomes increasingly consequential because benchmark performance depends on the validity of the supervisory signal itself. Auditing repository-derived supervision should therefore precede model development rather than serve as a downstream data quality exercise.

In this work, we introduce a *repository supervision auditing framework* (RSA) that treats the supervisory signal itself as the object of evaluation. Rather than proposing a new model architecture, we compare the repository-derived labels distributed with MIMIC-CXR against expert-curated image-level reference annotations obtained through structured radiologist review. Using cardiomegaly as a representative case study, we demonstrate near-zero agreement between repository-derived and expert image-level labels, show that the observed discordance arises predominantly from routine non-mention rather than explicit report negation, and construct a balanced, quality-controlled cohort for deployment-oriented model development and evaluation. Together, these results motivate supervision auditing as a prerequisite for developing clinically trustworthy AI from public imaging repositories.

2 Repository Supervision Auditing

The RSA methodology first conducts a retrospective study to evaluate whether report-derived labels distributed with a public chest radiograph repository constitute appropriate image-level supervision for machine learning. RSA comprised four stages: (1) construction of an expert-curated image-level reference, (2) auditing repository derived labels against the expert reference, (3) construction of an expert-curated cohort for downstream model development, and (4) deployment-oriented model evaluation (see Figure. 1).

Expert-curated image-level reference. We audited 2,160 studies sampled from MIMIC-CXR [8]. Board-certified radiologists independently reviewed each study using both the original DICOM image and its associated radiology report. They annotated findings through a custom web-based interface using terminology standardized to RadLex, SNOMED CT, and Fleischner Society recommendations [13, 9, 1], producing the image-level reference standard. We com-

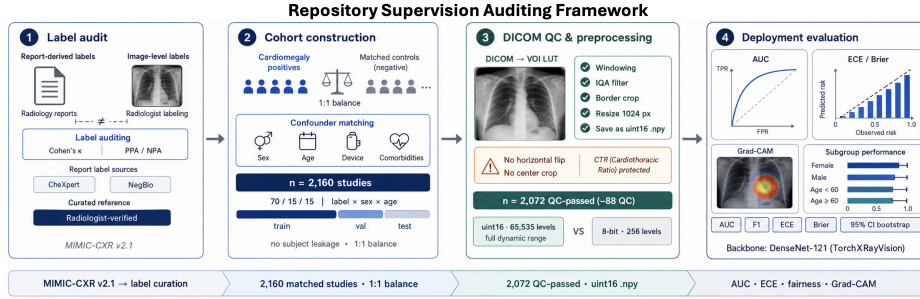


Fig. 1. Overview of the proposed repository supervision auditing framework. Repository-derived labels are validated against expert image-level annotations to establish a curated reference standard, which is used to construct a balanced, quality-controlled cohort for deployment-oriented model development and evaluation.

pared these annotations against the CheXpert and NegBio labels distributed with MIMIC-CXR [6, 12]. Following the original CheXpert protocol, we mapped *positive* and *uncertain* labels to the positive class and treated *negative* and *not mentioned* labels as negative. We quantified agreement using Cohen’s κ , positive percent agreement (PPA), negative percent agreement (NPA), positive predictive value (PPV), and negative predictive value (NPV). We selected cardiomegaly as the index finding because its diagnosis depends primarily on image-level morphology of the cardiac silhouette. We additionally evaluated pleural effusion and atelectasis to compare agreement across radiographic findings.

Characterization of label discordance. For each discordant cardiomegaly case, we reviewed the original radiology report and categorized the report-derived negative label as arising from (1) explicit negation, (2) assignment to the repository *No Finding*, or (3) non-mention despite documentation of other abnormalities. This analysis distinguished disagreement arising from report interpretation from disagreement arising from routine reporting practice.

Expert-curated cohort construction. We constructed the downstream cohort using expert image-level annotations rather than repository-derived labels. We selected all expert-confirmed cardiomegaly studies from a pool of 14,200 expert-reabeled examinations and matched negative studies on ten demographic and imaging covariates, including age, sex, imaging-device category, and age-related thoracic findings. Matching reduced the largest standardized mean difference from 0.87 to 0.005 and achieved balance across all matching covariates ($|\text{SMD}| < 0.10$). We performed quality control directly on DICOM images. The preprocessing pipeline applied modality LUT transformation, VOI windowing, photometric normalization, polarity correction, border cropping, and resizing of the longest image dimension to 1,024 pixels while preserving 16-bit pixel precision [5]. Quality control excluded non-frontal examinations, annotated images, and studies with inadequate exposure, yielding a final cohort of 2,072 studies (1,020 positive and 1,052 negative). We partitioned the curated cohort into

subject-independent training ($n = 1,453$), validation ($n = 311$), and held-out test ($n = 308$) sets using stratified sampling on the joint distribution of label, sex, and age group to preserve class balance while preventing patient-level information leakage.

Deployment-oriented evaluation. We fine-tuned a DenseNet121 backbone implemented in TorchXRayVision [4] using multi-dataset pretrained weights to avoid initialization from cardiomegaly models trained on repository-derived MIMIC labels. Images were resized to the network’s native input resolution while preserving the cardiothoracic ratio, and horizontal flipping was disabled because it reverses cardiac laterality.

The expert-curated cohort was partitioned into subject-independent training ($n = 1,453$), validation ($n = 311$), and held-out test ($n = 308$) sets using stratified sampling by label, sex, and age group, ensuring no patient-level leakage. Operating thresholds were selected exclusively on the validation cohort and fixed for all subsequent analyses. To validate the expert-audited cohort under deployment-oriented conditions [10], we evaluated discrimination (ROC-AUC and PR-AUC), threshold-dependent performance (sensitivity, specificity, precision, and F1-score), calibration (Brier score and expected calibration error), subgroup performance across age and sex, robustness across clinically relevant evaluation cohorts, and qualitative Grad-CAM visualizations [14]. Unlike conventional benchmark evaluation, this protocol combines predictive performance, calibration, robustness, fairness, and explainability to provide a comprehensive assessment of clinical readiness. Confidence intervals were estimated using a class-stratified bootstrap with 1,000 resamples.

3 Repository Supervision Audit Findings

Repository-derived labels do not reliably recover image-level findings.

We first evaluated whether repository-derived labels agreed with expert image-level annotations. Agreement for the index finding, cardiomegaly, was minimal (Table 1). Cohen’s κ was 0.011 (95% CI 0.004–0.019), with a positive percent agreement (PPA) of 1.3% and a negative percent agreement (NPA) of 99.8%. Among 1,080 expert-confirmed cardiomegaly studies, the repository-derived labels identified only 14 as positive. In contrast, repository-derived positive labels

Table 1. Agreement between repository-derived labels and expert image-level annotations on the matched MIMIC-CXR audit cohort ($N = 2,160$). Results for cardiomegaly were identical across CheXpert, NegBio, and their union. PPA: positive percent agreement; NPA: negative percent agreement; PPV: positive predictive value.

Finding	Cohen’s κ [95% CI]	PPA	NPA	PPV
Cardiomegaly (index)	0.011 [0.004, 0.019]	1.3%	99.8%	0.875
Pleural effusion	0.319 [0.190, 0.447]	21.7%	99.7%	0.682
Atelectasis	0.120 [0.067, 0.171]	7.9%	99.5%	0.655

Table 2. Characteristics of the expert-curated cohort after matching and DICOM quality control ($N = 2,072$). Matching achieved covariate balance across all ten variables ($|\text{SMD}| < 0.10$), reducing the largest imbalance (age) from 0.87 to 0.005.

Characteristic	Cardiomegaly	No Cardiomegaly	Overall
Studies, n	1,020	1,052	2,072
Age (years)	64.8 ± 15.2	65.1 ± 15.3	65.0 ± 15.2
Female, n (%)	629 (61.7)	655 (62.3)	1,284 (62.0)
Cardiac device, n (%)	103 (10.1)	101 (9.6)	204 (9.8)

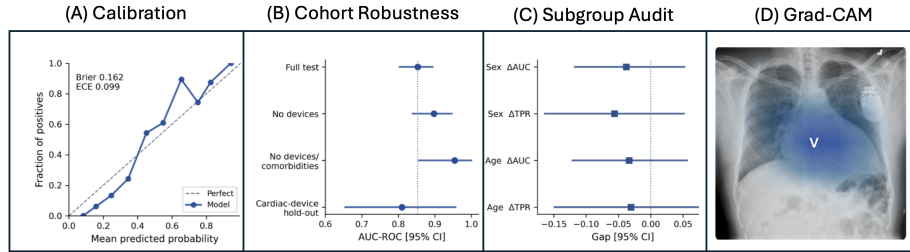


Fig. 2. Deployment-oriented evaluation using the expert-curated cohort. (A) Reliability diagram showing predicted versus observed probabilities (Brier score = 0.162, ECE = 0.099). (B) Robustness analysis across clinically relevant evaluation cohorts. (C) Fairness analysis reporting differences in discrimination (ΔAUC) and sensitivity (ΔTPR) across age and sex; error bars denote 95% confidence intervals. (D) Representative Grad-CAM visualization demonstrating that model attention localizes to the cardiac silhouette rather than implanted cardiac devices.

remained highly concordant with the expert reference ($\text{PPV} = 0.875$). CheXpert, NegBio, and their union produced identical results. Agreement differed across radiographic findings. Pleural effusion showed the highest agreement ($\kappa = 0.319$, 95% CI 0.190–0.447), followed by atelectasis ($\kappa = 0.120$, 95% CI 0.067–0.171), both substantially exceeding cardiomegaly (Table 1).

Repository-derived discordance predominantly reflects non-mention.

We next examined the origin of discordant cardiomegaly labels. Among the 1,080 expert-confirmed cardiomegaly studies, 1,066 (98.7%) received a negative repository-derived label, whereas only two repository-derived positive labels occurred in expert-negative studies. Only 55 discordant studies (5.2%) explicitly negated cardiomegaly in the radiology report. The remaining 1,011 studies (94.8%) omitted the finding entirely. Most omitted cases (979/1,011; 96.8%) were assigned the repository *No Finding* label, while only 32 studies (3.2%) contained other reported abnormalities without mention of cardiomegaly. Overall, expert review identified cardiomegaly in 979 of the 1,998 studies assigned the repository *No Finding* label.

Expert curation produces a balanced deployment-ready cohort. Following the audit, we constructed a matched expert-curated cohort for downstream

Table 3. Downstream performance on the held-out test cohort ($N = 308$). The operating threshold (0.436) was selected on the validation set using Youden’s index. Confidence intervals were estimated using 1,000 class-stratified bootstrap resamples.

Metric	Value [95% CI]	Operating point	Value
ROC-AUC	0.853 [0.803, 0.893]	Sensitivity	0.862
PR-AUC	0.824 [0.769, 0.878]	Specificity	0.705
Brier score	0.162	Precision	0.740
ECE (5 bins)	0.099 [0.062, 0.138]	F1-score	0.796

evaluation. Matching reduced the largest standardized mean difference from 0.87 to 0.005 and achieved balance across all matching covariates ($|\text{SMD}| < 0.10$). After DICOM quality control, the final cohort comprised 2,072 studies (1,020 cardiomegaly-positive and 1,052 cardiomegaly-negative examinations). Subject-independent train, validation, and test partitions preserved cohort characteristics while preventing patient-level leakage (Table 2).

Expert image-level supervision supports trustworthy downstream learning. Figure 2 illustrates the downstream evaluation (Rightmost block). Table 3 summarizes the downstream performance. Using the curated cohort, DenseNet121 achieved an ROC-AUC of 0.853 (95% CI 0.803–0.893) and a PR-AUC of 0.824 (95% CI 0.769–0.878) on the held-out test cohort. The model demonstrated moderate calibration (Brier score = 0.162, ECE = 0.099), maintained consistent discrimination across progressively cleaner evaluation cohorts, and showed no statistically significant differences in discrimination (ΔAUC) or sensitivity (ΔTPR) across age and sex. Grad-CAM visualizations localized model attention primarily to the cardiac silhouette rather than implanted cardiac devices, suggesting that model predictions were driven primarily by cardiac-silhouette morphology rather than device-related shortcuts (See example in Figure. 2).

4 From Repository Labels to Trustworthy AI

This work introduces *Repository Supervision Auditing (RSA)*, a framework for validating repository-derived learning targets before model development. Applied to MIMIC-CXR, RSA revealed that report-derived labels for cardiomegaly exhibit near-zero agreement with expert image-level assessment despite their widespread use for supervised learning. In contrast, pleural effusion and atelectasis showed substantially higher agreement, indicating that the reliability of repository-derived learning targets is finding-dependent rather than uniform. This dependence is interpretable: report-derived supervision degrades most for findings that are visual, morphometric, or chronic and thus under-documented in routine reporting (e.g., cardiomegaly), while remaining usable for acute findings that are consistently dictated (e.g., pleural effusion). Report-derived schemas are moreover confined to a small fixed label set (14 categories in CheXpert/NegBio),

whereas our expert reference spans 52 finding categories; findings outside the report schema cannot be supervised from reports at all. Repository Supervision Auditing can therefore be prioritized both for findings where report supervision is unreliable and for those it cannot represent.

RSA further showed that this disagreement stems primarily from routine reporting practice rather than report extraction errors. Most expert-confirmed cardiomegaly studies assigned a negative repository label were not explicitly negated in the radiology report but simply omitted. Consequently, repository-derived labels reflect clinical documentation rather than a complete inventory of radiographic findings. For visually defined abnormalities such as cardiomegaly, treating these labels as image-level ground truth can therefore introduce systematic supervision error and bias models toward reporting conventions instead of the underlying visual phenotype.

Beyond identifying label discordance, RSA provides a practical workflow for developing deployment-oriented datasets. Expert image-level auditing, matched cohort construction, DICOM-level quality control, and reproducible evaluation produced a balanced reference cohort that supported reliable downstream learning using a standard DenseNet121 backbone. More broadly, our results suggest that RSA fills a missing stage in the medical AI development pipeline by validating repository-derived learning targets before model training. Rather than replacing advances in model architecture, optimization, or foundation model pre-training, RSA complements them by ensuring that the supervisory signal itself is appropriate for the intended learning task.

Limitations and Future Work. This study has several limitations. First, the analysis was retrospective and limited to a single public repository (MIMIC-CXR). Second, cardiomegaly served as the primary case study; although our expert reference spans 52 finding categories and our cross-finding comparison shows supervision reliability to be finding-dependent—near-absent for cardiomegaly yet substantially higher for pleural effusion and atelectasis—we report three representative findings, and a comprehensive per-finding map across the label space shared by the expert and report schemas remains future work. Third, downstream evaluation used a single subject-independent train/validation/test split without external or prospective validation. Finally, the held-out test cohort was modest in size, limiting statistical power for subgroup and calibration analyses. Future work should extend RSA across additional repositories, imaging modalities, and disease phenotypes to determine where repository-derived learning targets are sufficient and where expert auditing provides the greatest benefit. External and prospective validation will further establish the generalizability of RSA and quantify the independent contributions of expert supervision, cohort curation, and image quality control to downstream model performance.

5 Conclusion

Repository-derived labels should not be assumed to represent image-level learning targets without verification. We introduced *Repository Supervision Audit-*

ing (RSA), a practical framework for auditing repository-derived supervision against expert-curated image-level reference annotations before model development. Applied to chest radiography, RSA revealed substantial disagreement between repository-derived and expert image-level labels for cardiomegaly, demonstrated that this discordance primarily reflects routine reporting practices rather than report extraction errors, and enabled the construction of a balanced, quality-controlled cohort for deployment-oriented model development and evaluation.

As public imaging repositories increasingly underpin foundation models and clinically deployed AI systems, Repository Supervision Auditing establishes supervision validation as a distinct stage of the medical AI development lifecycle. By validating repository-derived learning targets before model training, benchmarking, and deployment, RSA complements advances in model architecture, optimization, and foundation model pretraining, providing a practical pathway toward more trustworthy medical imaging AI. We envision Repository Supervision Auditing as a bedrock to trustworthy AI where it serves as a pre-requisite supervision validation stage in future medical AI development pipelines.

References

1. Bankier, A.A., MacMahon, H., Colby, T., Gevenois, P.A., Goo, J.M., Leung, A.N., Lynch, D.A., Schaefer-Prokop, C.M., Tomiyama, N., Travis, W.D., et al.: Fleischner society: glossary of terms for thoracic imaging. *Radiology* **310**(2), e232558 (2024)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: *European conference on computer vision*. pp. 1–21. Springer (2022)
3. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2020)
4. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: *International Conference on Medical Imaging with Deep Learning*. pp. 231–249. PMLR (2022)
5. Dapamede, T., Li, F., Khosravi, B., Purkayastha, S., Trivedi, H., Gichoya, J.: Dicom lut is a key step in medical image preprocessing towards ai generalizability. *Journal of Imaging Informatics in Medicine* pp. 1–9 (2025)
6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 590–597 (2019)
7. Jain, S., Smit, A., Truong, S.Q.H., Nguyen, C.D.T., Huynh, M.T., Jain, M., Young, V.A., Ng, A.Y., Lungren, M.P., Rajpurkar, P.: VisualCheXbert: Addressing the discrepancy between radiology report labels and image labels. In: *Proceedings of the Conference on Health, Inference, and Learning*. pp. 105–115 (2021)
8. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019)

9. Lee, D., Cornet, R., Lau, F., De Keizer, N.: A survey of snomed ct implementations. *Journal of biomedical informatics* **46**(1), 87–96 (2013)
10. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: Recommendations for image analysis validation. *Nature Methods* **21**, 195–212 (2024). <https://doi.org/10.1038/s41592-023-02151-z>
11. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proceedings of the ACM conference on health, inference, and learning*. pp. 151–159 (2020)
12. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings* **2018**, 188 (2018)
13. Rubin, D.L.: Creating and curating a terminology for radiology: ontology modeling and analysis. *Journal of digital imaging* **21**(4), 355–362 (2008)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision* **128**(2), 336–359 (2020)
15. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering* **6**(12), 1399–1406 (2022)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2097–2106 (2017)