Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does the Model See the Disease or the Shortcut? Causal Lesion Grounding for Chest Radiograph Diagnosis

Ameer Tamoor Khan

atk@plen.ku.dk

University of Copenhagen https://orcid.org/0000-0001-6838-992X

Saman Kainat Akhter

International University of Kyrgyzstan

Abdul Rehman Khan

Chinese University of Hong Kong

Shuai Li

University of Oulu

Research Article

Keywords: Chest radiography, counterfactual intervention, explainable artificial intelligence, lesion localization, shortcut learning, weakly supervised segmentation.

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11676802/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Does the Model See the Disease or the Shortcut? Causal Lesion Grounding for Chest Radiograph Diagnosis

Ameer Tamoor Khan, Saman Kainat Akhter, Abdul Rehman Khan, and Shuai Li

***Abstract*—Chest radiograph classifiers can reach high discrimination while relying on cues unrelated to the pathology, such as burned-in markers, support devices, or site-specific image appearance; a correct label does not show that the lesion was used. We present RADAR, a lesion-grounded framework in which the disease probability is computed directly from class-specific lesion maps, so that diagnosis and localization share a spatial computation. Lesion extent is learned from multi-reader radiologist boxes, anatomical priors, and confidence-filtered teacher pseudo-labels, and controlled image interventions encourage lesion-specific prediction dependence during training: replacing the lesion with healthy tissue should remove the diagnosis, whereas equal-area off-lesion, outside-thorax, text-marker, and lesion-free edits should not. On held-out images from three public collections with radiologist boxes and pixel masks, and on the external pixel-mask benchmark in particular, RADAR reached a macro area under the receiver operating characteristic curve of 0.878, and at equal explanation area its lesion maps overlapped the radiologist masks more than gradient-based activation maps of a conventional classifier with the same encoder and training data (intersection-over-union 0.37 versus 0.32). After the annotated lesion was replaced with healthy tissue, 62% of the conventional classifier's confident positive predictions persisted, compared with 16% for RADAR. Replacing the lesion that RADAR marked lowered its disease probability by 0.78, against 0.29 for a matched off-lesion edit, and the gap persisted under an image operator never used in training. RADAR thus evaluates both localization and the sensitivity of diagnostic predictions to edits of the marked region, while acknowledging the limitations of synthetic image interventions.**

***Index Terms*—Chest radiography, counterfactual intervention, explainable artificial intelligence, lesion localization, shortcut learning, weakly supervised segmentation.**

I. Introduction

CHEST radiography is the most frequently performed imaging examination, and deep learning has brought automated interpretation close to radiologist-level discrimination for several lesion categories [1], [2], largely on the basis of public collections labelled at the image level [3]–[5]. Image-level labels tell a model *what* to predict but not *why*, and a network optimized for discrimination will use any feature that correlates with the label in its training distribution [6]. In chest radiography such shortcuts are well documented: pneumonia classifiers recognized the hospital and acquisition protocol [7], coronavirus disease detectors relied on markers and image borders rather than the lungs [8], and pneumothorax detectors were biased by chest tubes, detecting treated pneumothoraces more readily than untreated ones [9]. Aggregate metrics can also hide clinically important failures in subgroups [10], [11]. A high area under the receiver operating characteristic curve (AUROC) shows that a model predicts a label, not that it predicts the label from the pathology.

Saliency maps are the usual answer to this question, but a weak one. Class activation methods [12], [13] explain a classifier after training, can be insensitive to its parameters [14], and localize chest radiograph pathology poorly [15], [16]. More fundamentally, a map that overlaps the lesion shows where evidence is *displayed*, not that the prediction *depends* on it [17], [18]. Removing image regions tests dependence more directly [19], [20], but the edit itself shifts the image distribution [21], [22], so a drop in probability can reflect reliance on the removed pathology or sensitivity to the appearance of the replacement.

We argue that a trustworthy diagnostic model should predict the disease (Q1), mark the region a radiologist would mark (Q2), lose the diagnosis when that region is replaced with healthy tissue (Q3), and keep it when non-pathological content is changed (Q4), and that the four properties should be tested separately. We present RADAR, a lesion-grounded framework designed for all four. Its central element is an evidence bottleneck: a ConvNeXt encoder [23] and feature-pyramid decoder [24] predict one lesion map per label, and the disease probability is pooled from the strongest responses in that map, with no separate classifier head. Lesion extent is learned from multi-reader radiologist boxes [25], anatomical priors, and teacher pseudo-labels [26], and controlled counterfactual edits turn Q3 and Q4 into training objectives. The contributions are as follows.

- A diagnostic architecture in which the image-level probability is pooled from class-specific lesion maps, so that the localized region is the computational source of the diagnosis rather than a post-hoc explanation.
- A training scheme that learns lesion extent from boxes, anatomy, and pseudo-labels and uses lesion-removal, lesion-preservation, matched-control, and invariance interventions with a shared randomized edit appearance as explicit objectives.

Manuscript received XX XXXX 2026.

A. T. Khan is with the Department of Plant and Environmental Sciences, University of Copenhagen, 2630 Taastrup, Denmark (e-mail: atk@plen.ku.dk).

S. K. Akhter is with the International School of Medicine, International University of Kyrgyzstan, Bishkek, Kyrgyzstan.

A. R. Khan is with The Chinese University of Hong Kong, Hong Kong SAR, China.

S. Li is with the Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland.

- An evaluation protocol that combines calibrated and area-matched localization on held-out boxes and masks, interventions with matched controls and held-out donor tissue, an independent classifier, and an edit-operator audit separating dependence on removed content from sensitivity to edit appearance.
- Experiments on three public collections showing that, compared with a conventional classifier with the same encoder and data, RADAR improves equal-area localization and markedly reduces the fraction of diagnoses that survive removal of the annotated lesion.

II. Related Work

A. Weakly and Box-Supervised Localization

Class activation mapping [12] and pooling variants such as WILDCAT [27] derive localization from image-level labels by aggregating local class evidence, and ChestX-ray8 introduced this setting to chest radiography [3]. Li *et al.* [28] combined image labels with a few boxes in a multiple-instance formulation, and Ouyang *et al.* [29] constrained activation- and gradient-based attention hierarchically; fully supervised detection and segmentation require instance masks [30]. Box-supervised segmentation turns boxes into tightness and emptiness constraints [31] or projection and affinity losses [25], and mean-teacher training [26] propagates supervision to unlabelled images. RADAR uses these components for lesion extent but evaluates localization together with whether the localized region determines the diagnosis.

B. Explanation and Its Evaluation

Post-hoc attribution dominates explainability in medical image analysis [32], yet sanity checks [14] and benchmarks with radiologist ground truth [15], [16] showed that such maps can ignore model parameters and localize pathology poorly. Deletion tests [19], [20] measure the effect of removing regions, and remove-and-retrain [21] and generative infilling [22] address the distribution shift that removal introduces. For high-stakes use, models whose reasoning is inspectable by construction have been advocated [17], [18]; RADAR follows this direction by making the lesion map part of the decision, and adds matched controls for the edit itself.

C. Shortcut Learning and Its Mitigation

Shortcut learning [6] has been shown repeatedly in chest radiography [7]–[9], [33] and contributes to poor cross-site generalization [34]; shortcut testing quantifies the use of nuisance attributes [35]. Penalizing attribution on irrelevant regions [36] constrains the explanation, but correcting saliency toward expert masks did not reliably improve generalization [37], and studies of confounding thoracic tubes highlight the relevance of lesion extent and acquisition context [9]. RADAR constrains the decision path instead: the classification score is pooled from spatial evidence logits rather than a separate global head; because the encoder has nonlocal receptive fields, interventions are needed to test whether input-region dependence follows the displayed evidence.

D. Counterfactual and Causal Reasoning

Causal reasoning about data generation has been argued to be essential for robust medical imaging models [38], and deep structural causal models generate counterfactual images under explicit assumptions [39]. Counterfactual explainers remove or exaggerate a lesion category with adversarial networks [40], [41] or autoencoder latent shifts [42] to explain a fixed classifier. RADAR uses simpler region replacement with matched healthy tissue, but applies it during training as well as evaluation and pairs every lesion intervention with a matched control. *Causal* is used here in this interventional sense, without claiming a structural causal model of disease.

III. Materials and Methods

A. Datasets

Three public collections provided complementary supervision. VinDr-CXR [5], with image labels and multi-reader boxes, supplied the spatial training signal; its 15,000 official training/development images (not the separate official test set) were subdivided into 12,000 training, 1,500 validation, and 1,500 test images. NIH ChestX-ray14 [3] (112,120 images) provided image-labelled training data and a held-out box subset. Using the official patient-wise lists, patients in the box subset were excluded from training, a patient-wise validation set was drawn from the remaining training patients, and half of the training images labelled *No Lesion* were retained; with validation and non-box test sets capped at 5,000 and 10,000 images, this gave 55,149 training, 5,000 validation, and 10,880 test images, including all 880 box-annotated images. CheXlocalize [16] provides radiologist pixel masks on CheXpert [4] images; its 202 frontal validation and 518 test radiographs were used for validation-based selection and external testing, respectively, and its support device masks only for the shortcut audit.

The ontology contained seven labels: Cardiomegaly, Pleural effusion, Atelectasis, Pneumothorax, Consolidation, Opacity, and Nodule/mass (NIH *Infiltration* and CheXlocalize *Lung* or *Airspace Opacity* were mapped to Opacity; *Nodule*, *Mass*, and *Lung Lesion* to Nodule/mass). An image could contribute more than one annotated image–label pair, so the per-label box counts in Table I do not sum to the number of distinct box-annotated images. The 67,149-image training mixture was sampled with 40% of the probability mass on VinDr-CXR and 60% on NIH. Per-label counts (TABLE I) show that extent boxes were scarce for Atelectasis (51) and Pneumothorax (44) and that CheXlocalize contained only nine Pneumothorax and nine Nodule/mass localization pairs. VinDr-CXR reader boxes were rasterized after resizing, with a consensus region defined by at least two readers and the reader union as fallback (Fig. 1a). VinDr-CXR Cardiomegaly boxes mark a cardiothoracic-ratio strip rather than the heart and were excluded from extent supervision.

B. Preprocessing

Images were converted to grayscale, resized with preserved aspect ratio to a longest side of 384 pixels, histogram equalized, and zero padded to 384×384 (Fig. 1b); annotations

TABLE I
POSITIVE IMAGES AND LOCALIZATION PAIRS PER LABEL IN EACH SPLIT.

	Training		Test positives			Localization pairs		
Lesion Label	**Images**	**Boxes**	**VinDr-CXR**	**NIH**	**CheXlocalize**	**VinDr-CXR**	**NIH**	**CheXlocalize**
Cardiomegaly	3,005	–	177	571	154	177	146	154
Pleural effusion	8,357	505	62	1,983	105	62	153	105
Atelectasis	7,456	51	4	1,504	158	4	180	157
Pneumothorax	2,438	44	8	1,143	9	8	98	9
Consolidation	2,683	99	8	775	29	8	0	29
Opacity	13,102	622	76	2,632	272	76	123	271
Nodule/mass	7,688	326	37	1,362	9	37	164	9
Images	**67,149**	**1,202**	**1,500**	**10,880**	**518**	**1,500**	**880**	**518**

Images: positive training images. Boxes: VinDr-CXR training images with reader boxes (Cardiomegaly is supervised by the anatomical heart mask). Localization pairs: annotated image–label pairs used for evaluation.

followed the same geometry with nearest-neighbour interpolation. Augmentation comprised rotation within $\pm 7°$, scaling by about $\pm 10\%$, translation within $\pm 5\%$, and brightness, contrast, and gamma changes, and synthetic burned-in text (for example "AP", "PORTABLE", "L") was pasted near the border with probability 0.5, independently of the label. Anatomical zones at the 96×96 map resolution (a dilated thorax region, a dilated heart–mediastinum region, and an undilated heart mask) were obtained with a public organ-segmentation model from TorchXRayVision [43], used to derive anatomy regions rather than disease labels; when segmentation failed, the image content box replaced the thorax region and no heart target was created.

C. Lesion-Grounded Diagnostic Model

The overall pipeline is shown in Fig. 2. Let $x \in \mathbb{R}^{384\times384}$ be a preprocessed radiograph and $K = 7$ the number of lesion labels. A ConvNeXt-Tiny encoder [23] extracts multi-scale features, and a 128-channel feature-pyramid decoder [24] predicts class-specific lesion logits

$$S(x) = \{s_1, \ldots, s_K\}, \qquad s_k \in \mathbb{R}^{96\times96}, \tag{1}$$

with lesion probability maps $M_k(x) = \sigma(s_k)$, where σ is the logistic function. Let $\mathcal{C}_k$ contain the positions of the largest $q = 0.5\%$ of the logits of class k (46 positions). The image-level logit is the mean over this diagnostic core,

$$z_k(x) = \exp(a_k)\,\frac{1}{|\mathcal{C}_k|}\sum_{u\in\mathcal{C}_k} s_{k,u}(x) + b_k, \tag{2}$$

with learned scale a_k and bias b_k, and the disease probability is $p_k(x) = \sigma(z_k(x))$. The map M_k represents lesion extent and its strongest core determines the diagnosis, thus classification gradients act through spatial lesion logits. This pooling constraint does not mathematically ensure that a positive score corresponds to a thresholded lesion mask: the learned bias and nonlocal receptive fields can influence the output. Earlier training stages used a normalized log-sum-exp over the whole map,

$$z_k^{\mathrm{LSE}}(x) = \exp(a_k)\,\frac{1}{r}\log\Big(\frac{1}{|\Omega|}\sum_{u\in\Omega} e^{\,r\,s_{k,u}(x)}\Big) + b_k, \tag{3}$$

with $r = 5$ and Ω the set of map positions; the two rules are compared in Section VI. Classification used class-weighted binary cross-entropy,

$$\mathcal{L}_{img} = -\frac{1}{K}\sum_{k=1}^{K}\big[w_k y_k \log p_k + (1-y_k)\log(1-p_k)\big], \tag{4}$$

where y_k is the label and w_k the negative-to-positive ratio clipped at 10; because p_k is pooled from the map, classification gradients act directly on the spatial representation.

D. Spatial Supervision

Radiologist boxes. Boxes do not define the lesion contour, so the decoder is not trained to fill them. A projection loss inspired by BoxInst [25] matches the horizontal and vertical maximum projections P_x, P_y of the predicted map M to those of the consensus box B,

$$\mathcal{L}_{proj} = \mathcal{D}\big(P_x(M), P_x(B)\big) + \mathcal{D}\big(P_y(M), P_y(B)\big), \tag{5}$$

with a Dice-type discrepancy $\mathcal{D}$. Positions outside the union of all readers' boxes, Ω_{out}, are treated as negative, $\mathcal{L}_{out} = |\Omega_{out}|^{-1}\sum_{u\in\Omega_{out}}\mathrm{softplus}(s_u)$, and an intensity-affinity pairwise term $\mathcal{L}_{pair}$ encourages neighbouring pixels of similar intensity inside the box to share a label [25].

Anatomical priors. For positive Cardiomegaly images, the Cardiomegaly map is supervised by the heart mask H with $\mathcal{L}_{heart} = \mathcal{L}_{BCE}(M_{card}, H) + \mathcal{L}_{Dice}(M_{card}, H)$. For all labels, logits outside the class-appropriate zone Z_k (thorax for lung label, heart and mediastinum for Cardiomegaly) are penalized,

$$\mathcal{L}_{zone} = \frac{\sum_{k,u}\mathrm{softplus}(s_{k,u})(1-Z_{k,u})}{\sum_{k,u}(1-Z_{k,u}) + \epsilon}. \tag{6}$$

Teacher pseudo-labels. Most positive images have no boxes. An exponential-moving-average teacher [26] provides pseudo-labels on unboxed positives: teacher probabilities above 0.85 are positive, below 0.15 negative, and intermediate positions are ignored. Class-specific area caps prevent pseudo-masks from growing without bound. The resulting loss is $\mathcal{L}_{pl}$.

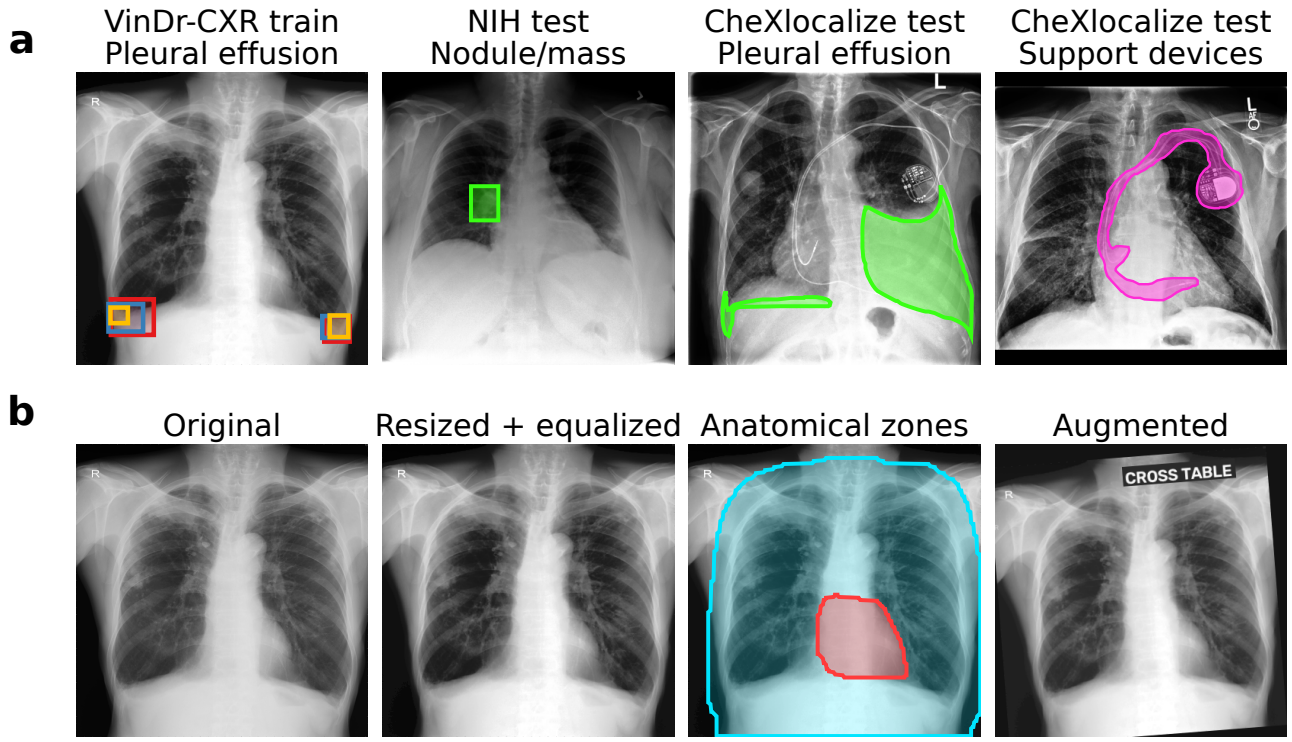


Fig. 1. Data and preprocessing. (a) Annotation regimes: VinDr-CXR training reader boxes (one colour per radiologist), an NIH held-out box, a CheXlocalize held-out pixel mask, and a CheXlocalize support device mask used only for the shortcut audit. (b) Preprocessing of one radiograph: original image, aspect-preserving resizing to 384×384 with histogram equalization and zero padding, anatomical zones (cyan: thorax; red: heart), and a training view with geometric and photometric augmentation and synthetic burned-in text.

E. Counterfactual Training

Spatial supervision does not ensure that the diagnosis depends on the localized region. For a selected positive class k, let E_k be the intervention mask, taken from the radiologist box when available and from the teacher lesion map otherwise. Healthy reference tissue r is drawn from a pool of negative radiographs (donor pool A) by nearest-neighbour matching of a low-resolution appearance descriptor and intensity matched to x. The lesion-removal and lesion-preserving images are

$$x_k^- = x \odot (1 - E_k) + r \odot E_k, \qquad x_k^+ = x \odot E_k + r \odot (1 - E_k). \tag{7}$$

The removal objective (where BCE on a logit denotes binary cross-entropy with logits) $\mathcal{L}_{nec} = \mathrm{BCE}(z_k(x_k^-), 0) + \lambda_{pix}\mathcal{L}_{removed}$ requires the diagnosis, and the lesion response inside E_k, to vanish, and the preservation objective $\mathcal{L}_{suff} = \mathrm{BCE}(z_k(x_k^+), 1)$ requires the lesion alone to support it. To separate lesion dependence from edit sensitivity, a matched-control objective $\mathcal{L}_{ctrl}$ replaces an equal-area region away from the lesion and requires the diagnosis to persist; lesion-shaped donor patches on negative images must remain negative.

The invariance objective $\mathcal{L}_{inv}$ replaces content outside the thorax or inserts lesion-free blobs and thin device-like strokes, and requires the probabilities and the maps in unedited regions to match the teacher's on the original image. In the final stage, donor contrast, gamma, noise, and blur were randomized and edit boundaries stochastically feathered, with the same appearance distribution for removal and invariance edits to reduce systematic edit-appearance cues. The training objective is the weighted sum $\mathcal{L} = \sum_j \lambda_j \mathcal{L}_j$ over all terms above.

F. Training and Inference

Training proceeded in three stages (TABLE II). The encoder was initialized from the conventional baseline classifier (Section IV-A), itself pretrained on ImageNet-22k. The lesion-grounding stage ran for 12 epochs of 32,000 sampled images (batch 16, encoder learning rate 10^{-4}, decoder rate five times higher, EMA decay 0.999) with $\lambda_{img} = 1$, $\lambda_{proj} = 2$, $\lambda_{out} = 1$, $\lambda_{pair} = 0.5$, $\lambda_{heart} = 1$, $\lambda_{zone} = 1$, and $\lambda_{pl} = 0.5$. A robustness refinement (5 epochs, learning rate 5×10^{-5}) introduced lesion-free invariance edits, stricter pseudo-label thresholds, and $\lambda_{nec} = 1$, $\lambda_{suff} = 0.1$, $\lambda_{ctrl} = 0.5$, $\lambda_{inv} = 1$. The final stage (4 epochs, 5×10^{-5}) started from the selected refinement checkpoint and added core pooling (2) and randomized edit appearance; the first two stages used (3). The core-pooled model was designated as final before the ablation in Section VI was run.

Checkpoints (EMA weights) were selected by validation macro AUROC on VinDr-CXR and NIH plus 0.5 times validation localization on VinDr-CXR boxes and CheXlocalize validation masks, and decision thresholds were F1-optimal on the same validation sets, clipped to $[0.30, 0.95]$. No test image was used for optimization or selection. At inference, a single forward pass returns p_k and M_k; interventions are used only for training and auditing.

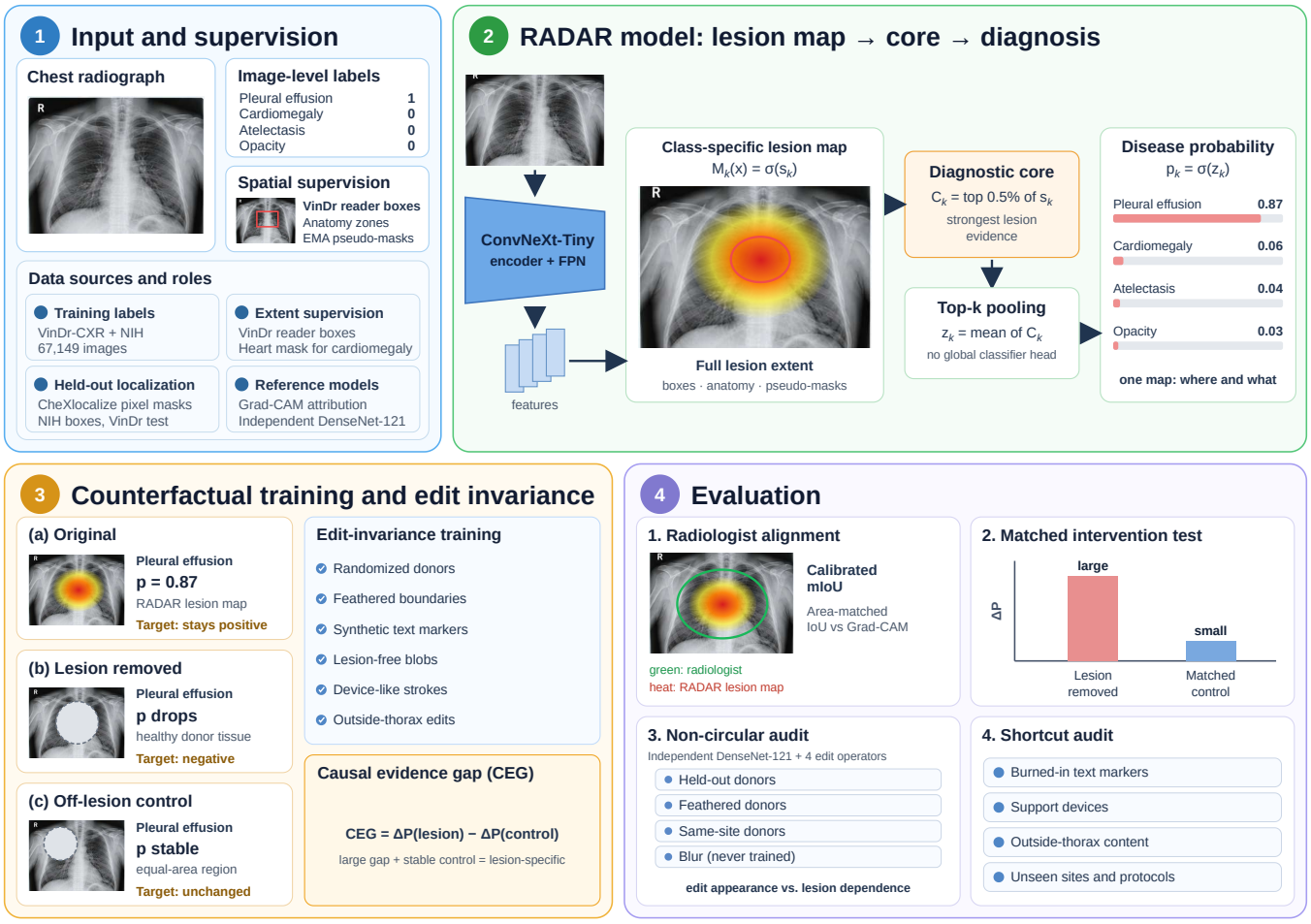


Fig. 2. Overview of RADAR. (1) Image labels, multi-reader boxes, anatomical zones, and teacher pseudo-masks supervise the model. (2) A ConvNeXt encoder and feature-pyramid decoder predict one lesion map per label, and the disease probability is pooled from the strongest lesion responses, with no separate classifier head. (3) Counterfactual training replaces the lesion, the lesion context, or matched non-lesion regions with healthy tissue, with randomized edit appearance. (4) Evaluation combines radiologist overlap, matched interventions, an independent classifier, and shortcut audits.

TABLE II
TRAINING AND EVALUATION CONFIGURATION OF RADAR.

Component	Configuration
Input / lesion-map resolution	384×384 / 96×96
Encoder / decoder	ConvNeXt-Tiny / feature pyramid, 128 channels
Training mixture	67,149 images; 40% VinDr-CXR, 60% NIH
Stages (epochs)	Grounding (12), robustness (5), final (4; epoch 2 selected)
Samples per epoch / batch	32,000 / 16
Learning rate	10^{-4}, then 5×10^{-5}; decoder $\times 5$
Diagnostic pooling	Top 0.5% (final); log-sum-exp, $r = 5$ (earlier)
Donor pools	A: 1,024 training negatives; B: 1,024 disjoint negatives

IV. EXPERIMENTAL SETUP

A. Baselines

Conventional classifier: ConvNeXt-Tiny with global average pooling and a linear head, trained for 8 epochs on the same data, preprocessing, and marker augmentation, with random healthy-tissue patches as augmentation, and explained with Grad-CAM [13] from its last encoder stage; this represents the standard practice of training a classifier and explaining it post hoc. *Grad-CAM on RADAR*: computed from RADAR's last encoder stage, separating intrinsic maps from post-hoc attribution of the same network. *Independent classifier*: DenseNet-121 [44], trained for 6 epochs on the same labels with a different architecture and seed; it never saw RADAR's maps and also tests whether the regions RADAR marks carry diagnostic information for another model.

B. Evaluation Metrics

Q1. Per-label and macro AUROC; for NIH, on the full 10,880-image test set unless stated. *Q2.* Threshold-selected mean intersection-over-union (mIoU; calibration here refers to segmentation thresholds, not calibration of disease probabilities), with one binarization threshold per label selected on CheXlocalize validation masks, separately for each map type, and fixed for all test sets. Area-matched IoU retains the same fraction of pixels (2.5%, 5%, 10%, 20%) from every map, removing any advantage of larger masks.

Q3. On image–label pairs predicted positive by the audited model, regions were replaced with tissue from donor pool B, disjoint in images and patients from the training pool.

With $x^-_{k,\mathrm{GT}}$, $x^-_{k,\mathrm{RADAR}}$, and x^{ctrl}_k denoting replacement of the radiologist lesion, the calibrated RADAR lesion, and an equal-area region away from both, ΔP_{GT}, $\Delta P_{\mathrm{RADAR}}$, and ΔP_{ctrl} denote the corresponding drops $p_k(x) - p_k(\cdot)$, and

$$\mathrm{CEG} = \Delta P_{\mathrm{RADAR}} - \Delta P_{\mathrm{ctrl}} \tag{8}$$

where the causal evidence gap (CEG) is positive when removing the localized evidence affects the diagnosis more than a matched non-lesion edit. The flip rate is the fraction of predicted positives falling below the decision threshold after $x^-_{k,\mathrm{GT}}$; the survival rate is the fraction of pairs with $p_k(x) \geq 0.5$ that remain at $p_k \geq 0.5$, which uses a common probability cutoff but does not ensure a paired comparison, because model-specific predicted-positive sets can differ.

Q4. Mean absolute probability change after inserting synthetic text, replacing content outside the thorax, and removing CheXlocalize support device pixels outside the annotated lesion. An edit-operator audit repeated the off-lesion control with four replacement operators: held-out donors with hard boundaries, the same donors with feathered boundaries, same-site donors (CheXlocalize validation images negative for all seven labels), and Gaussian blur, which was never used in training.

C. Statistical Analysis

Uncertainty was estimated with 1,000 bootstrap resamples clustered by image, so that several labels on one radiograph were resampled together, and summarized as 95% percentile intervals. Intervention metrics were computed on each model's own predicted positives; comparisons between models therefore refer to different, model-specific case sets and should not be interpreted as paired causal contrasts. All models were trained with a single seed, and differences smaller than the bootstrap interval widths (about ± 0.02 for mIoU) are not interpreted. No formal multiplicity-adjusted hypothesis tests were performed. These intervals capture test-image sampling variability, not training-seed variability.

V. Results

A. Comparison With Conventional Classifiers

TABLE III compares RADAR with the two conventional classifiers. Diagnostic discrimination was comparable: RADAR had the highest AUROC on CheXlocalize (0.878 versus 0.855 and 0.758) and was within 0.015 of the conventional ConvNeXt classifier on NIH and VinDr-CXR. At equal explanation area, RADAR lesion maps overlapped the annotations more than Grad-CAM of the conventional classifier on CheXlocalize (0.37 versus 0.32) and VinDr-CXR (0.40 versus 0.32), with a smaller difference on NIH (0.26 versus 0.25). The largest difference concerned Q3. When the radiologist lesion was replaced with healthy tissue, 62%, 80%, and 67% of the conventional classifier's confident positives persisted on CheXlocalize, NIH, and VinDr-CXR, and 70%, 75%, and 74% of the DenseNet-121 positives, compared with 16%, 25%, and 7% for RADAR.

B. Diagnostic Performance

On the full test sets (TABLE IV), RADAR reached a macro AUROC of 0.878 on CheXlocalize, 0.788 on the 10,880-image NIH test set, and 0.969 on VinDr-CXR. Per-label AUROC (TABLE V) was highest for Pleural effusion on CheXlocalize (0.930) and lowest for Opacity (0.771 on CheXlocalize and 0.669 on NIH), the most heterogeneous label in the ontology; the CheXlocalize Nodule/mass value (0.963) rests on nine positive images.

C. Radiologist-Aligned Localization

RADAR maps reached a calibrated mIoU of 0.38 on CheXlocalize, 0.26 on NIH boxes, and 0.51 on VinDr-CXR (TABLE IV). All annotated pairs, rather than predicted positives, gave 0.32 and 0.23 on CheXlocalize and NIH, so the result is not due to case selection. Grad-CAM of the same network reached 0.13–0.14 at every matched area (Fig. 3a); RADAR IoU peaked at 5–10% of the image, consistent with lesions occupying less than a fifth of the radiograph.

Localization depended strongly on the label (TABLE V). Cardiomegaly, supervised by the heart mask, reached 0.72 on CheXlocalize and 0.58 on NIH, and diffuse labels with several hundred training boxes were intermediate (Opacity 0.37 and 0.31; Pleural effusion 0.31 and 0.26). The weakest labels were those with the least extent supervision: Pneumothorax (0.12 on NIH, 44 training boxes) and Nodule/mass (0.17 on NIH).

D. Dependence on the Localized Lesion

Replacing the RADAR lesion lowered the disease probability by 0.78 on CheXlocalize, 0.84 on NIH, and 0.85 on VinDr-CXR, whereas the matched off-lesion edit lowered it by 0.29, 0.03, and 0.01 (Fig. 3b), giving CEGs of 0.49, 0.82, and 0.84 with intervals well above zero. Replacing the radiologist lesion, a region the model did not choose, lowered the probability by 0.69, 0.66, and 0.89 and moved 90%, 82%, and 95% of predicted positives below threshold.

Two analyses addressed whether these effects reflect sensitivity to the replacement procedure used in training. With Gaussian blur, never used in training, the CEG remained 0.69, 0.84, and 0.89. The independent DenseNet-121 also lost more confidence when the RADAR lesion was removed than when the matched control was removed (CEG 0.20, 0.34, and 0.31), which suggests that marked regions contain predictive information for a model that did not produce them; its smaller effect matches its high survival rate (TABLE III).

E. Robustness to Non-Pathological Edits

Synthetic text changed the probability by 0.012, 0.011, and 0.003, and replacing content outside the thorax by 0.042, 0.013, and 0.012 (Fig. 3c). Removing CheXlocalize support device pixels outside the lesion changed it by 0.118, compared with 0.025 for the DenseNet-121, and by 0.090 when only device pixels not touching the RADAR map were removed; this was the largest non-pathological sensitivity of the model.

The edit-operator audit (Fig. 3d) explains why the CheXlocalize control drop (0.29) exceeded that on NIH and VinDr-CXR. With donors from VinDr-CXR and NIH the drop was

TABLE III
COMPARISON OF RADAR WITH CONVENTIONAL CLASSIFIERS TRAINED ON THE SAME DATA.

	Q1: macro AUROC↑			Q2: IoU at 10% area↑			Q3: survival after lesion removal↓		
Model (localization map)	**CheXloc.**	**NIH**	**VinDr**	**CheXloc.**	**NIH**	**VinDr**	**CheXloc.**	**NIH**	**VinDr**
ConvNeXt-Tiny classifier (Grad-CAM)	0.855	**0.803**	**0.971**	0.32	0.25	0.32	0.62	0.80	0.67
DenseNet-121 classifier (–)	0.758	0.783	0.954	–	–	–	0.70	0.75	0.74
RADAR network (Grad-CAM)	–	–	–	0.14	0.13	0.16	–	–	–
RADAR (intrinsic lesion map)	**0.878**	0.789	0.969	**0.37**	**0.26**	**0.40**	**0.16**	**0.25**	**0.07**

NIH AUROC on the 880-image box subset for all models. IoU at 10%: top 10% of each map versus the annotation, predicted positives. Survival: fraction of each model's pairs with $p \geq 0.5$ that remain at $p \geq 0.5$ after the radiologist lesion is replaced with held-out healthy tissue (lower indicates stronger lesion dependence).

TABLE IV
DIAGNOSTIC, LOCALIZATION, AND LESION-DEPENDENCE RESULTS OF RADAR ON THE HELD-OUT TEST SETS.

	Q1	Q2: calibrated mIoU↑		Q3: lesion dependence				
Test set	**AUROC↑**	**RADAR**	**Grad-CAM**	$\Delta P_{\mathrm{GT}} \uparrow$	$\Delta P_{\mathrm{RADAR}} \uparrow$	$\Delta P_{\mathrm{ctrl}} \downarrow$	**CEG↑**	**Flip↑**
CheXlocalize	0.878	0.38 [0.36, 0.40]	0.13	0.69	0.78	0.29	0.49 [0.44, 0.54]	0.90
NIH boxes	0.788	0.26 [0.25, 0.28]	0.13	0.66	0.84	0.03	0.82 [0.79, 0.84]	0.82
VinDr-CXR	0.969	0.51 [0.49, 0.52]	0.14	0.89	0.85	0.01	0.84 [0.80, 0.87]	0.95

AUROC: macro over seven labels (NIH: full 10,880-image test set). mIoU: calibrated, predicted positives; Grad-CAM computed on the RADAR network. ΔP, CEG: see (8), held-out donor tissue. Flip: fraction of predicted positives below threshold after radiologist-lesion removal. Brackets: 95% bootstrap intervals. VinDr-CXR shares site and annotation protocol with the training boxes.

TABLE V
PER-LABEL AUROC AND LOCALIZATION OF RADAR AND GRAD-CAM.

	AUROC		CheXloc. mIoU			NIH mIoU		
Lesion Label	**CheX.**	**NIH**	**RADAR**	**GC**	n	**RADAR**	**GC**	n
Cardiomegaly	0.860	0.891	0.72	0.37	23	0.58	0.33	59
Pleural effusion	0.930	0.830	0.31	0.11	84	0.26	0.13	127
Atelectasis	0.849	0.766	0.39	0.08	75	0.21	0.05	101
Pneumothorax	0.893	0.859	0.14	0.14	2	0.12	0.11	60
Consolidation	0.878	0.723	0.41	0.17	16	–	–	0
Opacity	0.771	0.669	0.37	0.13	157	0.31	0.11	92
Nodule/mass	0.963	0.778	0.20	0.07	7	0.17	0.10	102

mIoU: calibrated, predicted positives. GC: Grad-CAM on the RADAR network. n: evaluated pairs.

0.29 with hard and 0.26 with feathered boundaries, but it fell to 0.07 with same-site donors and 0.12 with blur, while the DenseNet-121 changed by less than 0.01 under every operator. The decline in off-lesion sensitivity with same-site donors is consistent with a contribution from donor-domain appearance, although the experiment does not exclude changes in informative non-lesion content. Removing only the part of the RADAR map outside the annotation still lowered the probability (0.35 on CheXlocalize, 0.47 on NIH), a quantity that combines mislocalization with lesion label present but not annotated.

F. Qualitative Analysis

Cases were selected by the rule given in the caption of Fig. 4. The Cardiomegaly map traced the cardiac silhouette, and the bilateral Pleural effusion and basal Atelectasis maps covered both annotated regions; replacing them lowered the probability from 0.94–0.97 to 0.00, whereas Grad-CAM concentrated centrally, spread over the lower thorax, or highlighted the mediastinum. For the NIH Pneumothorax, Opacity, and Nodule/mass cases the segmentation matched the box with at most a small additional focus.

VI. ABLATION STUDIES

A 2×2 factorial refinement compared diagnostic pooling (normalized log-sum-exp versus top-k core pooling) and randomized versus fixed donor-edit appearance. All four branches used the same starting refinement checkpoint, seed, data, losses, and four-epoch schedule. The comparison in Table VI therefore tests whether the apparent robustness improvement is attributable to pooling, donor randomization, or their interaction.

Randomized edits markedly increased the causal evidence gap (CEG) under the Gaussian-blur audit, an operator excluded from training: from 0.143 to 0.731 with LSE pooling and from 0.132 to 0.688 with top-k pooling. Randomization also decreased the probability change after matched off-lesion donor edits (0.477 to 0.244 with LSE, and 0.454 to 0.293 with top-k). Top-k with randomization achieved the highest CheXlocalize macro AUROC (0.878), while LSE with randomization had a higher blur-based CEG (0.731 versus 0.688) and a smaller off-lesion control drop (0.244 versus 0.293). Thus, the principal edit-robustness effect was associated with randomized appearance, whereas core pooling entailed a trade-off between discrimination and intervention stability. These are single-seed comparisons; small differences should not be overinterpreted. Individual intervention-loss components were not removed separately.

VII. DISCUSSION

The baseline comparison supports treating discrimination and diagnostic reasoning as separate properties. Both conventional classifiers discriminated disease about as well as RADAR, and Grad-CAM of the ConvNeXt classifier overlapped the annotations reasonably well, yet both kept most diagnoses after the annotated lesion was removed: a plausible saliency map was not evidence of dependence [17], whereas

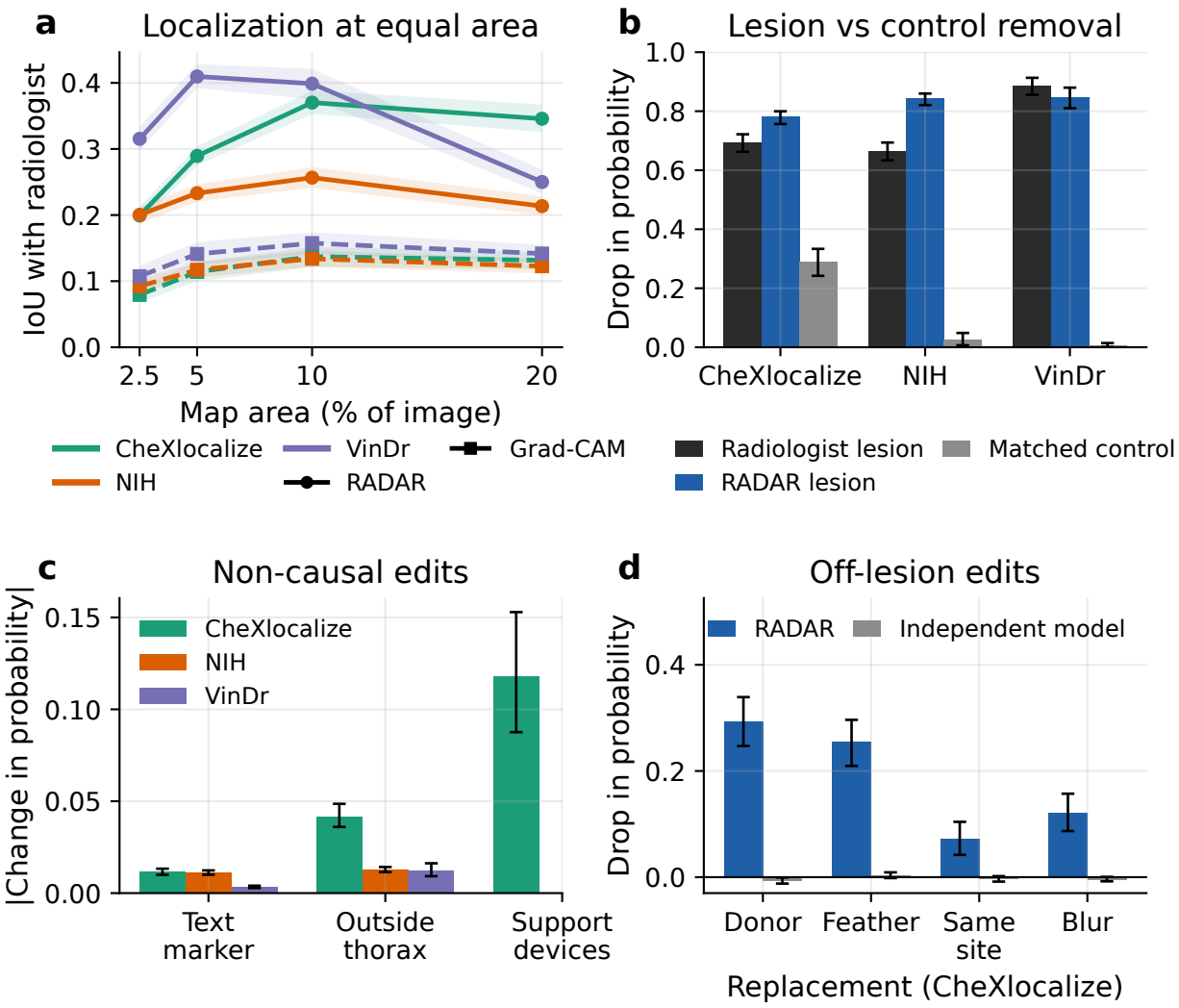


Fig. 3. Quantitative grounding and intervention analysis on the held-out sets (95% bootstrap intervals). (a) IoU with the radiologist annotation when RADAR lesion maps and Grad-CAM of the same network retain the same fraction of the image. (b) Drop in disease probability after replacing the radiologist lesion, the RADAR lesion, or an equal-area off-lesion region with held-out healthy tissue. (c) Absolute probability change after non-pathological edits. (d) Off-lesion control edits on CheXlocalize under four replacement operators, for RADAR and the independent DenseNet-121.

TABLE VI
FACTORIAL ABLATION OF DIAGNOSTIC POOLING AND DONOR-EDIT APPEARANCE ON CHEXLOCALIZE. ALL BRANCHES USE IDENTICAL FOUR-EPOCH REFINEMENT BUDGETS. LARGER AUROC AND BLUR CEG ARE BETTER; SMALLER OFF-LESION DONOR DROP IS BETTER.

Pooling / edit appearance	AUROC	Blur CEG	Control drop
LSE / fixed	0.853	0.143	0.477
LSE / randomized	0.863	**0.731**	**0.244**
Top-k / fixed	0.859	0.132	0.454
Top-k / randomized (RADAR)	**0.878**	0.688	0.293

pooling decisions from lesion maps, together with intervention training, strengthened empirical dependence on annotated regions. The factorial ablation indicates that randomized donor appearance contributes substantially to robustness under the held-out blur operator, while diagnostic pooling has a smaller and metric-dependent effect. The interventional analyses also show why perturbation tests need controls: without the matched control, the 0.78 drop on CheXlocalize would be indistinguishable from edit sensitivity, and the operator audit suggested that donor-domain appearance contributed to the remaining control drop. Matched controls and several replacement operators should accompany claims of reliance based on region removal [21], [22].

A. Clinical Implications

A diagnosis computed from a displayed lesion map gives the reader something to verify: if the marked region is not pathological, the diagnosis is not supported. This matters where devices and pathology co-occur, such as chest tubes and pneumothorax [9], and across sites with different acquisition cues [7], [34]. Survival after lesion removal and the CEG with matched controls can be computed on any locally annotated set and could complement AUROC and subgroup analyses in pre-deployment audits [10], [11], [35]. The maps are not diagnostic-grade segmentations, and their clinical value was not tested in a reader study. Cardiomegaly localization uses

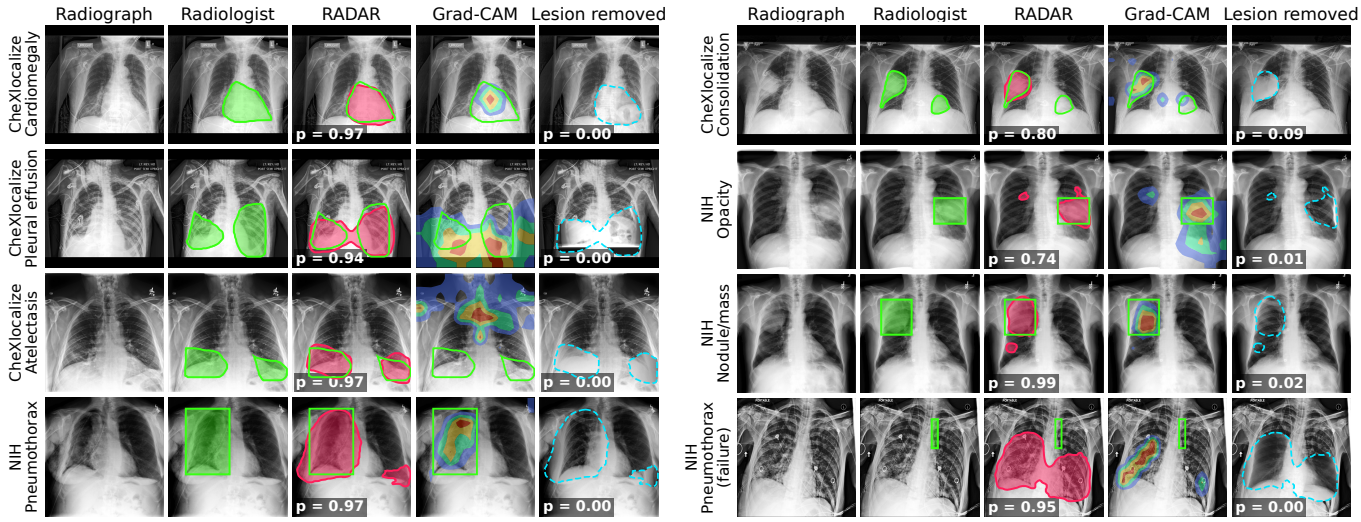


Fig. 4. Held-out cases, one per lesion label plus an additional failure case, selected by a fixed rule: the case with the highest IoU among predictions that flip after lesion removal, followed by the confident prediction with the least lesion-map energy inside the annotation (failure, last row). Green: radiologist annotation (filled, or outline for comparison); red: RADAR lesion segmentation; Grad-CAM (of the RADAR network) as colour bands; dashed cyan: region replaced with healthy tissue; p: disease probability before and after replacement.

a heart-anatomy surrogate rather than direct delineation of a pathological boundary.

B. Failure Cases

For an NIH Pneumothorax annotated by a narrow left upper-zone box, RADAR marked large parts of both lungs (Fig. 4, last row): the diagnosis depended on the marked region, but its extent was wrong, consistent with only 44 training boxes for this lesion label. In the Consolidation case, RADAR missed a smaller left-sided region, although removing the marked region lowered the probability from 0.80 to 0.09. When large regions were replaced, donor tissue could look anatomically implausible (Pleural effusion row), a limitation of patch-based counterfactuals relative to generative ones [40]–[42].

C. Limitations

Extent supervision was box-based and strongly class dependent; independent extent annotations such as MS-CXR [45] are the most direct route to better localization of small focal lesions. The Cardiomegaly target depends on an external anatomy model. Region replacement cannot reproduce the unobserved radiograph without the lesion, and the audit showed that the donor source matters. Device removal still changed the probability more than for the independent classifier; because devices can be associated with severity, this invariance should be read as robustness to direct dependence on device pixels. Removing map regions outside the annotation lowered the probability, and the annotations cannot separate unannotated labels from mislocalization without radiologist adjudication. Results come from single training runs, a factorial refinement compared pooling and donor randomization but did not isolate individual counterfactual losses, and the study is retrospective. Finally, *causal* denotes sensitivity under specified synthetic image interventions, not a structural causal model of disease, proof of exclusive pathology reliance, or a clinically valid no-disease counterfactual [38], [39].

VIII. Conclusion

Computing the diagnosis from class-specific lesion maps, supervising extent with boxes, anatomy, and pseudo-labels, and training with controlled lesion and non-lesion edits produced a model with discrimination comparable to conventional classifiers, better equal-area localization than Grad-CAM of a conventional classifier, and predictions that were more sensitive to annotated-lesion removal than those of the conventional baselines. An unseen edit operator and independent classifier provided additional, but not definitive, tests of regional diagnostic information. Extent annotations for focal lesion labels, same-site or generative counterfactuals, and radiologist adjudication are the next steps toward clinically meaningful lesion-grounded diagnosis.

Data and Code Availability

VinDr-CXR is available from PhysioNet under credentialed access, NIH ChestX-ray14 from the NIH Clinical Center, and CheXlocalize with CheXpert from the Stanford AIMI Center. Code release is planned upon publication; details of access and reproducibility will be provided with the final published version.

Acknowledgment

The authors thank the providers of the public datasets used in this study.

References

[1] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Med.*, vol. 15, no. 11, p. e1002686, 2018.

[2] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, p. 102125, 2021.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2097–2106.

[4] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[5] H. Q. Nguyen *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, no. 1, p. 429, 2022.

[6] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[7] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, p. e1002683, 2018.

[8] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 610–619, 2021.

[9] J. Rueckel *et al.*, "Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs," *Invest. Radiol.*, vol. 55, no. 12, pp. 792–798, 2020.

[10] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proc. ACM Conf. Health, Inference, Learn. (CHIL)*, 2020, pp. 151–159.

[11] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat. Med.*, vol. 27, no. 12, pp. 2176–2182, 2021.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2921–2929.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 9505–9515.

[15] N. Arun *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiol. Artif. Intell.*, vol. 3, no. 6, p. e200267, 2021.

[16] A. Saporta *et al.*, "Benchmarking saliency methods for chest X-ray interpretation," *Nat. Mach. Intell.*, vol. 4, no. 10, pp. 867–878, 2022.

[17] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[18] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[19] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, 2017.

[20] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 3429–3437.

[21] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 9737–9748.

[22] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 976–11 986.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[25] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5443–5452.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 1195–1204.

[27] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5957–5966.

[28] Z. Li *et al.*, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8290–8299.

[29] X. Ouyang *et al.*, "Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2698–2710, 2021.

[30] J. Lian *et al.*, "A structure-aware relation network for thoracic diseases detection and segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2042–2052, 2021.

[31] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. Ben Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," in *Proc. Int. Conf. Med. Imaging Deep Learn. (MIDL)*, ser. Proc. Mach. Learn. Res., vol. 121, 2020, pp. 365–381.

[32] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102470, 2022.

[33] S. Jabbour, D. Fouhey, E. Kazerooni, M. W. Sjoding, and J. Wiens, "Deep learning applied to chest X-rays: Exploiting and preventing shortcuts," in *Proc. Mach. Learn. Healthc. Conf. (MLHC)*, ser. Proc. Mach. Learn. Res., vol. 126, 2020, pp. 750–782.

[34] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction," in *Proc. Int. Conf. Med. Imaging Deep Learn. (MIDL)*, ser. Proc. Mach. Learn. Res., vol. 121, 2020, pp. 136–155.

[35] A. Brown, N. Tomasev, J. Freyberg, Y. Liu, A. Karthikesalingam, and J. Schrouff, "Detecting shortcut learning for fair medical AI using shortcut testing," *Nat. Commun.*, vol. 14, p. 4314, 2023.

[36] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2662–2670.

[37] J. D. Viviano, B. Simpson, F. Dutil, Y. Bengio, and J. P. Cohen, "Saliency is a possible red herring when diagnosing poor generalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[38] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nat. Commun.*, vol. 11, p. 3673, 2020.

[39] N. Pawlowski, D. C. Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 857–869.

[40] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using Wasserstein GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8309–8319.

[41] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, "Explaining the black-box smoothly—a counterfactual approach," *Med. Image Anal.*, vol. 84, p. 102721, 2023.

[42] J. P. Cohen *et al.*, "Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest X-rays," in *Proc. Int. Conf. Med. Imaging Deep Learn. (MIDL)*, ser. Proc. Mach. Learn. Res., vol. 143, 2021, pp. 74–104.

[43] ——, "TorchXRayVision: A library of chest X-ray datasets and models," in *Proc. Int. Conf. Med. Imaging Deep Learn. (MIDL)*, ser. Proc. Mach. Learn. Res., vol. 172, 2022, pp. 231–249.

[44] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.

[45] B. Boecking *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, ser. Lecture Notes in Computer Science, vol. 13696, 2022, pp. 1–21.