Supplementary Information for: Demographic shortcuts as marker-free backdoors: silent subgroup underdiagnosis that only operating-point audits detect

Saptarshi Purkayastha[1] Parvati Naliyatthaliyazchayil[1] Judy Wawira Gichoya[2]

[1]Indiana University Indianapolis, Indianapolis, IN, USA
[2]Emory University, Atlanta, GA, USA

This document contains Supplementary Notes S1–S4, Supplementary Tables S1–S22 and Supplementary Figures S1–S4. All Methods are reported in the main manuscript file.

S1 Supplementary Note: positioning relative to prior work

This study sits at the intersection of four literatures: semantic backdoors and subpopulation poisoning in machine-learning security, fairness-targeted poisoning, demographic shortcut learning in medical imaging, and medical foundation-model attacks. We do not claim to originate the attack class. What we characterize is a specific and clinically consequential instance of it: what it costs an adversary, how much of its measured behaviour is set by the operating point at which subgroup error is read, why removing the demographic–label correlation does not prevent it, and whether the defenses and audits available in deployment detect it.

Semantic backdoors: the class this attack belongs to. Bagdasaryan et al.[1] introduced semantic backdoors, in which a naturally occurring feature of a sample, such as an unusual car colour or a racing stripe, serves as the trigger, so the adversary need not modify the digital or physical input at inference time. Bagdasaryan and Shmatikov[2] extended this and noted that trigger-reversal scanners degrade when the trigger is a large natural feature rather than a small patch, which is the mechanism behind the Neural Cleanse and STRIP failures reported in the main text. Our attack is a semantic backdoor whose semantic feature is the patient's demographic identity. What differs is the instance. Where the triggers in that work are incidental object properties of everyday scenes, ours is a protected attribute of a patient. It is decodable from pixels at AUROC above 0.97 with no human-visible correlate[3], it is present in every image of the target group and not merely in a rare subset of it, and the failure it produces is masked by documented underdiagnosis bias[4].

Subpopulation poisoning: the same threat model, formalized. Jagielski et al.[5] define subpopulation attacks as compromising classifier performance on a naturally defined subpopulation while leaving the remainder of the data unaffected, explicitly without inference-time modification, and connect the setting to group fairness. They further show that under stated assumptions such attacks admit no general defense. Our threat model is an instance of theirs, restricted to label flips within a demographic-by-finding cell. Their impossibility result supplies the theoretical account for

our finding that five detectors fail, and our observation that representation-space detectors score at or below chance because the poisoned rows are the majority of their cell is the empirical mechanism behind it.

Fairness poisoning. Solans et al.[6] introduced gradient-based poisoning to widen classification disparities; Mehrabi et al.[7] added anchoring and influence attacks against group-fairness measures; Liu et al.[8] extended poisoning to fair representation learning; and Chan and Tong[9] give a theoretical treatment for naive Bayes classifiers. These operate by inserting or crafting data points, largely in tabular or representation-learning settings, and target a fairness *metric*. Our adversary cannot insert data, operates on images, and targets an operating-point error rate in a clinical subgroup.

Closest prior art in medical imaging. Kulkarni et al.[10] established the feasibility of the demographic-conditional label-flip primitive on chest radiographs. Table S1 gives a head-to-head comparison. Two points are easily misstated. First, their dose–response model is a *logistic* fit: their vulnerability index ν is the rate parameter β of a logistic regression of the group-versus-overall metric difference on the injected-bias rate, and a logistic is by construction sigmoidal. Their appendix further discusses crossover points between injection rates of 0.25 and 0.75 at which a group's performance passes the overall model. We therefore do not replace a linear description with a non-linear one. What we add is an *operational* installation point under criteria fixed before analysis, and the finding that this point is stable across axes once evaluation is anchored to a defensible operating point. Second, they anticipated the audit-metric asymmetry: they selected thresholded error rates over AUROC on the grounds that clinical decision-making is binary, and reported that AUROC has the potential to mask the induced bias. We quantify how much each audit moves, at matched false-positive rate, against a defined adversary and alongside five backdoor detectors; the observation itself is theirs. The originating group's subsequent work on demographic bias is defense-side, comprising generative counterfactual augmentation[11] and adversarial debiasing of three-dimensional CT foundation embeddings[12]; neither extends the attack, and our adversarial-debiasing diagnostic (Table S9) bears directly on the second.

Fairness-targeted backdoors with injected triggers. BadFair[13] in language, TrojFair[14] in images including dermoscopy, and Un-Fair Trojan[15] in faces and tabular federated settings install group-conditional failures through an injected synthetic trigger; the demographic attribute selects which samples receive the trigger, and an untriggered target-group image elicits no unfair behaviour. This is the inverse of our mechanism, in which the demographic feature is itself the trigger.

Clean-label and clean-image backdoors. Label-consistent[16] and hidden-trigger[17] attacks keep poisoned images plausibly labelled but still rely on a synthetic pixel pattern at inference. Clean-image backdoor attacks[18] select a naturally present feature via label falsification but operate on MNIST and CIFAR with no demographics or medicine. The GCB attack[19] needs only training-label manipulation but *discovers and optimizes* a maximally separable generic natural feature, installing at very low poison rates because the trigger is engineered for separability; FFCBA[20] is clean-label yet still injects an autoencoder-generated pixel trigger. The high within-cell poison fraction required in our setting follows from the trigger being given and entangled rather than optimized, and our PatchCamelyon and ISIC results indicate what happens at the separable end of that spectrum: the attack installs readily but destroys aggregate performance.

Medical foundation-model backdoors. BadMatch and BadDist on MedCLIP[21], BAPLe[22] and BadEncoder[23] backdoor medical and self-supervised encoders via injected triggers or target-image alignment, and are class-targeted with no demographic conditioning. Our foundation-model attack uses no trigger in either probe or fine-tune setting and shows that the installation point is not lowered by fine-tuning.

Healthcare-AI poisoning and shortcut-suppression defenses. A 2026 analytical review of data-poisoning vulnerabilities across healthcare AI[24] flags conceptually that triggers correlated with protected characteristics can present as bias rather than sabotage, and argues that attack success in this domain tracks absolute poisoned-sample count rather than rate. Our evidence is consistent with that argument without confirming it: the cell-size factorial of Table S13 could not separate the two quantities, while holding the within-cell rate fixed and varying the target cell across four demographic conditions (Table S21b) does exclude an account in terms of rate alone, though those conditions differ in the triggering attribute as well as in cell size. The one detection method built specifically for shortcut learning in clinical models is shortcut testing[25], which uses multitask training and access to the sensitive attribute to ask whether an observed subgroup disparity is driven by a shortcut. It is not directly applicable here, because our shortcut is known in advance and has been made label-predictive on purpose, so the question is not whether a shortcut exists but whether the resulting error is visible in deployment metrics. On the defense side, augmentation policies that reduce demographic decodability[26] and acquisition-parameter calibration[27] are the shortcut-suppressing interventions most directly implied by the threat model, and Supplementary Note S2 reports the outcome of our attempt to evaluate the former. We note that removing demographic information is not automatically the fair choice[28]; the argument here is a security one. Governance and transparency frameworks[29–31] motivate the operating-point subgroup-error reporting our results make testable, and equalized-odds auditing[32] provides its conceptual basis. Concurrent work makes the non-adversarial half of our operating-point argument directly: Pham et al.[33] report that rare-finding subgroup underdiagnosis in chest radiography depends jointly on the finding, the subgroup and the operating threshold rather than on ranking metrics, and Dominique et al.[34] show that fairness comparisons between skin-lesion classifiers can invert with the threshold and the calibration at which they are read. Shu et al.[35] reach the same conclusion from acquisition rather than from demographics, reporting exposure regimes under which overall AUROC is essentially unchanged while sensitivity falls by more than twenty points, and proposing a pre-training exposure-regimen audit on that basis. They also report that saliency maps for their biased models show no consistent anatomic focus, which matches our own attribution result (Fig. S4). None of this work studies an induced failure. The step taken here is to show that the same sensitivity governs the measured cost and the apparent axis-dependence of an installed attack. Reports that fairness interventions on chest-radiograph classifiers often fail to improve worst-group error[36] and that simple post-hoc demographic corrections are frequently ineffective[37] are consistent with our adversarial-debiasing result.

S2 Supplementary Note: what this study establishes, and two arms that did not inform it

Established. A demographic-conditional label flip installs a subgroup-targeted operating-point failure at a within-cell flip rate between 0.50 and 0.75. This holds in every one of the six chest-radiograph, histopathology and dermoscopy settings that installs at a clinically defensible operating point, but it is architecture-dependent: three of six backbones require the entire cell or never install.

The attack does not require the demographic–label correlation, and removing that correlation raises its visibility cost roughly fivefold without preventing installation. Five backdoor detectors fail, two of them scoring below chance, while a subgroup-FNR audit at the deployed threshold flags 95.8% of installed attacks against 73.2% for a rank audit at matched false-positive rate. The induced disparity is approximately 1.7 times the baseline underdiagnosis gap.

Established as a methodological caution. The measured behaviour of the attack depends on the operating point at which subgroup error is evaluated. Relative error-rate metrics are normalized by the clean model's sensitivity, so an insensitive operating point inflates them; on NIH sex the effect is large enough to move the apparent installation point from 0.75 to 0.10.

Two arms that did not inform the study, reported for completeness. We attempted a test-time counterfactual demographic audit, flipping the demographic of each input and flagging models whose target-label prediction moves more for the attacked subgroup. A CycleGAN race translator[38] proved too weak to support it, removing only about 1.6% of a held-out race decoder's between-group separability (race-decoder AUROC 0.987 on real versus 0.979 on counterfactual images), so the resulting inconsistency of roughly 0.004 is uninformative, and is not evidence that the attack evades the audit. Separately, we evaluated whether the shortcut-suppressing augmentation policy of Wang et al.[26] attenuates the attack. In our implementation the policy did not reduce demographic decodability at all: a race detector retrained under it reached AUROC 0.983 against 0.977 without it, attack success was not reduced at any tested rate ($\mathrm{ASR}_{\mathrm{rel}}$ augmented against default: 0.147 versus 0.135 at pr = 0.50, $p = 0.88$; 0.221 versus 0.221 at 0.75, $p = 0.996$; 0.282 versus 0.361 at 1.0, $p = 0.15$). Because the intervention did not achieve its own precondition, this arm tests neither the defense nor the threat model, and we report it as an unsuccessful implementation, not as a negative result about augmentation.

S3 Supplementary Note: calibration of the predicted-race subgroup proxy

On NIH-CXR14 and VinDr-CXR, recorded race is unavailable and subgroups are approximated by top and bottom terciles of a held-out race detector's predicted $P(\text{Black})$. Because a tercile is a fixed third of a cohort whereas the true target group is not, this proxy is calibrated wherever true labels and predictions coexist (Table S12).

Rank agreement is excellent: on the MIMIC test split the proxy recovers the true-label transfer effect with Spearman $\rho = 0.99$ over 27 paired estimates, and top-tercile purity is 98.6% where the true target share is 44.9%. Magnitude agreement is not. The regression of the proxy estimate on the true-label estimate has slope 1.45 where the target group is 44.9% of the cohort, 0.66 where it is 21.2%, and 0.53 on CheXpert where it is 8.5%. The direction of the bias therefore depends on whether the target group exceeds one third of the cohort: above that share the top tercile is nearly pure and the contrast is sharper than the true one; below it the tercile is diluted and the contrast is blunted. Three prevalence points are a described pattern, not an estimated relationship.

The proxy accordingly licenses the existence, direction, dose–response and cohort ordering of the external transfer effect, but not its magnitude. The CheXpert analysis with recorded self-reported race is therefore the primary external evidence in the main text, and the NIH and VinDr transfer magnitudes (Table S16) are reported as directional and, given the target shares involved, most likely understated.

S4 Supplementary Note: fairness-aware retraining as a defense

Three fairness-aware retraining methods were trained on the corrupted data and re-scored against the installed MIMIC race attack at the reference strength (pr = 0.75, DenseNet-121, three seeds). Methods are given in the main manuscript. These three interventions were run before the operating-point analysis and are reported at a fixed threshold of 0.5; by the argument of Sec. 2.3 of the main text the magnitudes below are inflated relative to a clinically defensible operating point, though the ordering of the three is not affected. Full results are in Table S9.

Two of the three remove the failure. Inverse-prevalence reweighting and Group distributionally robust optimization both neutralize it while preserving overall AUROC, each overshooting into a reversed disparity (ASR_{rel} -0.279 and -0.367 against $+0.333$ undefended). Overshooting is worth stating plainly: these interventions do not restore parity, they invert it, and a service adopting either should measure subgroup sensitivity afterwards rather than assume the problem is solved.

Adversarial debiasing does not, and the diagnostic shows why. Adversarial debiasing through a gradient-reversal demographic head fails at every adversarial weight tested. The best result across λ from 0.01 to 10 was 0.279 ± 0.065, and at both low and high λ the failure was *stronger* than undefended, reaching 0.539 at $\lambda = 10$. A freshly trained linear probe still reads race from the penultimate features at AUROC 0.935 on average, rising to 0.958 at $\lambda = 10$, while the adversarial head's own validation accuracy sits at the majority-class rate of 0.784 for every $\lambda \geq 0.3$ and every seed. The adversary collapsed to predicting the majority demographic and supplied no useful gradient, so raising λ regularized the encoder without removing the signal it was meant to remove. This is a negative result about a standard gradient-reversal configuration and not evidence that representation-level debiasing cannot work. It does indicate that any such defense must be validated by probing the representation directly, because the adversary's own training loss reports success while the demographic signal is still fully present.

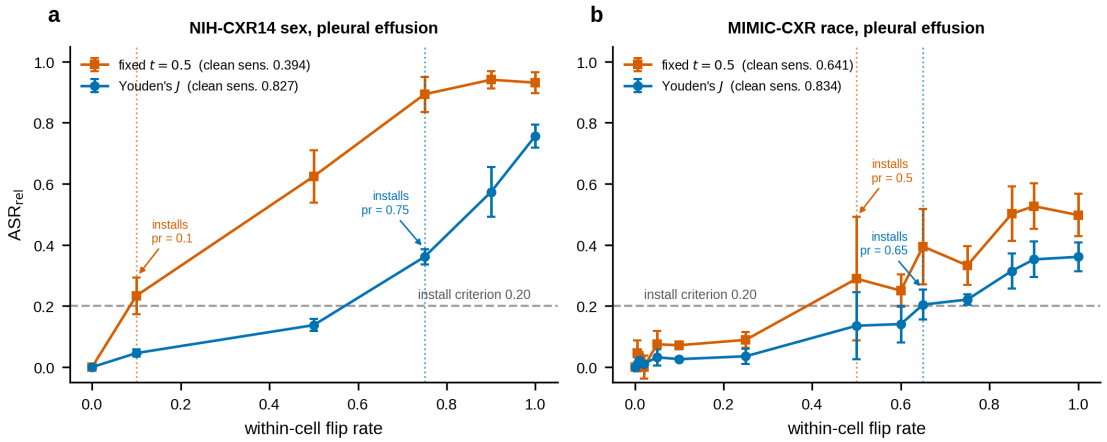

Figure S1: The apparent gap between the NIH sex and MIMIC race axes closes once evaluation is anchored to a defensible operating point. **a**, Relative attack success rate against within-cell flip rate for the NIH-CXR14 sex attack, evaluated at a fixed threshold of 0.5 (where the clean model reaches 0.394 sensitivity) and at the Youden's-J operating point (clean sensitivity 0.827). The dashed horizontal line marks the install criterion. At the fixed threshold the attack appears to clear the criterion at a flip rate of 0.10; at Youden's J it does not do so until 0.75. **b**, The same comparison for the MIMIC race attack, which moves by two steps of the grid, from 0.50 to 0.65. Points are means over three seeds and error bars one standard deviation.

Table S1: Head-to-head comparison: Kulkarni et al.[10] versus this work.

Dimension	Kulkarni et al. (MIDL 2024)	This work
Poisoning primitive	demographic-conditional label flip (positive → "no finding")	identical
Conceptual frame	adversarial bias and underdiagnosis, within the fairness-robustness literature	the same primitive as a *semantic backdoor*[1] and *subpopulation poisoning*[5] instance, within the backdoor attack-and-defense literature
Dose–response model	**logistic** vulnerability index $\nu = \beta$; crossover points discussed in appendix	operational installation point under three criteria fixed before analysis, on a 14-rate grid; four-parameter logistic reported as shape evidence
Operating point	Youden's J; AUROC reported as potentially masking the bias	four operating points compared; the installation point and the apparent cross-axis ordering both shown to depend on this choice
Audit metric	anticipated: thresholded error rates preferred over AUROC	quantified detection rate and false-positive rate for both audits, at matched calibration, over 768 evaluations
Role of prevalence correlation	not examined	matched-cohort control: correlation is not required; removing it raises visibility cost fivefold
Demographic axis	sex, age, intersections; race deferred to future work	race (primary) and sex; the intersectional and age-stratified footprint of the race-conditioned attack characterized (Table S18), though intersections are not themselves targeted
Target findings	pneumonia	pleural effusion (MIMIC and NIH) and pneumothorax (NIH) attacked; cardiomegaly co-trained but never targeted, and used as an untouched-label control (Table S18b)
Modalities	chest radiography	chest radiography, histopathology, dermoscopy, electrocardiography
Architectures	DenseNet-121 (ResNet-50, Inception-v3 in appendix)	six, including two transformers
Foundation models	none	RAD-DINO, BiomedCLIP, MedSigLIP (probe and fine-tune)
Cross-cohort	RSNA → CheXpert, MIMIC	MIMIC → CheXpert (recorded race), NIH, VinDr
Backdoor defenses	named as future work; robust statistics identified as promising	five detectors at matched false-positive rate, including SPECTRE
Fairness defenses	none	reweighting, Group DRO, adversarial debiasing over six λ with a representation probe
Clinical translation	none	missed findings per 1,000, number needed to harm, ratio to baseline disparity

Table S2: MIMIC-CXR race attack: full dose–response at the Youden's-J operating point (DenseNet-121, pleural effusion, unmatched cohort, three seeds, mean ± SD). Install criteria are $\mathrm{ASR}_{\mathrm{rel}} \geq 0.20$, attacked-minus-control gap ≥ 0.05, and $|\Delta\mathrm{AUROC}| \leq 0.03$. The attack has no detectable effect through pr = 0.25 and first meets all three criteria at pr = 0.65.

Poison rate	n flipped	$\mathrm{ASR}_{\mathrm{rel}}$ attacked	$\mathrm{ASR}_{\mathrm{rel}}$ control	ΔAUROC overall	Criteria met
0.000	0	0.000 ± 0.000	0.000	0.0000	
0.005	23	0.010 ± 0.023	0.008	+0.0006	
0.010	47	0.022 ± 0.017	0.012	+0.0004	
0.020	94	0.011 ± 0.002	0.005	+0.0003	
0.050	237	0.032 ± 0.026	0.015	−0.0001	
0.100	474	0.026 ± 0.006	0.014	−0.0009	
0.250	1,185	0.035 ± 0.025	−0.004	−0.0001	
0.500	2,371	0.135 ± 0.110	0.028	−0.0018	
0.600	2,845	0.141 ± 0.061	0.011	−0.0033	
0.650	**3,082**	$\mathbf{0.205 \pm 0.049}$	0.045	−0.0041	**all three**
0.750	3,556	0.221 ± 0.017	0.008	−0.0052	all three
0.850	4,030	0.315 ± 0.057	0.043	−0.0075	all three
0.900	4,267	0.353 ± 0.058	0.047	−0.0080	all three
1.000	4,742	0.361 ± 0.048	0.043	−0.0138	all three

A quadratic term over the full grid is significant ($F_{1,39} = 14.32$, $P = 5.2 \times 10^{-4}$) and a four-parameter logistic is preferred over a linear fit by 8.9 AIC. The logistic inflection is not identified on this cohort because $\mathrm{ASR}_{\mathrm{rel}}$ is still rising at pr = 1.0 (Table S7).

Table S3: Installation point by cohort and operating point. Every operating point other than the fixed threshold of 0.5 is derived on the clean seed-matched model's validation split and applied unchanged to the attacked model; the clean model's validation sensitivity at each is given so that relative-metric inflation can be assessed. Entries marked n/a denote settings that never meet all three criteria at that operating point; the binding criterion is given in the footnote. Entries marked ‡ meet all three criteria but without demographic selectivity, and are not counted as installed attacks. Setting those aside, every setting that installs at a clinically defensible operating point does so between pr = 0.50 and 0.75.

	Fixed $t = 0.5$		Youden's J		Sensitivity 0.80		Specificity 0.90	
Cohort	clean sens	install	clean sens	install	clean sens	install	clean sens	install
MIMIC race, unmatched	0.641	0.50	0.834	**0.65**	0.800	0.65	0.687	0.50
MIMIC race, matched	0.531	0.50	0.835	**0.75**	0.800	0.75	0.701	0.50
NIH sex, effusion	0.394	0.10	0.827	**0.75**	0.800	0.75	0.635	0.50
NIH sex, pneumothorax	0.165	0.50	0.803	**0.75**	0.801	0.75	0.703	0.75
PCam site	0.769	0.50	0.813	**0.75**	0.800	0.75	0.851	n/a
ISIC source	0.757	0.50	0.837	n/a	0.799	0.50	0.754	0.50
PTB-XL sex	0.869	0.75‡	0.950	n/a	0.799	0.50‡	0.969	n/a

Installation points are for DenseNet-121, the architecture common to every cohort, except PTB-XL, which uses the one-dimensional ResNet. Binding criteria for the n/a entries: PCam at specificity 0.90 and ISIC at Youden's J fail the *visibility* criterion, not the effect criterion, because at pr = 1.0 PCam reaches $\text{ASR}_{\text{rel}} = 1.000$ with ΔAUROC -0.315, and ISIC reaches 0.977 with -0.211 (Table S6); PTB-XL at Youden's J and at specificity 0.90 likewise fails only the visibility criterion, at the single rate (pr = 1.0) that clears the other two. ‡PTB-XL meets all three criteria at these two operating points but does so *without demographic selectivity*: at the sensitivity-matched point and pr = 0.50 the control cell reaches $\text{ASR}_{\text{rel}} = 0.817$ against the attacked cell's 0.935, so what the criteria register is a global collapse of the operating point rather than a subgroup-targeted failure. We therefore do not count PTB-XL as an installed attack anywhere, and note that an absolute attacked-minus-control gap criterion is not by itself sufficient to establish selectivity once both cells approach complete failure (Sec. 2.7 of the main text). Excluding those two cells, every setting that installs at a clinically defensible operating point does so between pr = 0.50 and 0.75. Sensitivity of every installation point to install criteria of 0.10, 0.15 and 0.30 is given in the source-data file accompanying this table.

Table S4: Matched versus unmatched MIMIC race cohort at the Youden's-J operating point (DenseNet-121, pleural effusion). The matched cohort equalizes effusion prevalence across race by subsampling, removing the demographic–label correlation while leaving race decodability unchanged. The attack installs in both, later and at greater visibility cost in the matched cohort. Unmatched $n = 3$ seeds; matched $n = 5$ on the low-rate grid and $n = 3$ above.

	Unmatched			Matched		
Poison rate	ASR_{rel} att.	ctrl	ΔAUROC	ASR_{rel} att.	ctrl	ΔAUROC
0.000	0.000	0.000	0.0000	0.000	0.000	0.0000
0.005	0.010	0.008	+0.0006	-0.000 ± 0.055	0.025	−0.0005
0.010	0.022	0.012	+0.0004	-0.003 ± 0.059	0.012	+0.0001
0.020	0.011	0.005	+0.0003	-0.007 ± 0.072	0.008	+0.0003
0.050	0.032	0.015	−0.0001	-0.009 ± 0.042	−0.001	−0.0001
0.100	0.026	0.014	−0.0009	0.004 ± 0.050	−0.002	−0.0008
0.500	0.135	0.028	−0.0018	0.096 ± 0.111	0.002	−0.0058
0.650	**0.205**	0.045	**−0.0041**	n/a	n/a	n/a
0.750	0.221	0.008	−0.0052	$\mathbf{0.284 \pm 0.040}$	0.030	**−0.0202**
1.000	0.361	0.043	−0.0138	0.355 ± 0.180	0.012	−0.0465

Installation point pr = 0.65 unmatched, 0.75 matched. Visibility cost at the respective installation points is −0.004 against −0.020, a roughly fivefold increase; the matched attack leaves the ±0.03 visibility band entirely at pr = 1.0.

Table S5: Architecture sweep at the Youden's-J operating point (MIMIC unmatched race, pleural effusion, three seeds, mean ± SD, except EfficientNet-B4 at pr = 0.50, where one of the three runs returned a non-finite validation loss and is excluded, $n = 2$). Bold entries meet all three install criteria. Three of the six architectures install at pr ≤ 0.75; ResNet-50 and EfficientNet-B4 require the entire target cell, and ConvNeXt-T does not meet the effect criterion at any tested rate. The final column gives the same quantity at a fixed threshold of 0.5, where all six install at pr ≤ 0.75: the architecture-dependence is an operating-point effect.

		ASR_{rel} at Youden's J			Installation point	
Architecture	Family	pr = 0.50	pr = 0.75	pr = 1.00	Youden's J	$t = 0.5$
ResNet-50	convolutional	0.062 ± 0.009	0.110 ± 0.007	$\mathbf{0.244 \pm 0.024}$	1.00	0.75
ConvNeXt-T	convolutional	0.063 ± 0.079	0.140 ± 0.045	0.195 ± 0.041	n/a	0.75
EfficientNet-B4	convolutional	0.108 ± 0.000	0.170 ± 0.028	$\mathbf{0.252 \pm 0.083}$	1.00	0.50
DenseNet-121	convolutional	0.135 ± 0.110	$\mathbf{0.221 \pm 0.017}$	$\mathbf{0.361 \pm 0.048}$	**0.65**	0.50
ViT-B/16	transformer	0.146 ± 0.053	$\mathbf{0.249 \pm 0.062}$	$\mathbf{0.202 \pm 0.032}$	**0.75**	0.50
Swin-T	transformer	0.137 ± 0.037	$\mathbf{0.250 \pm 0.110}$	$\mathbf{0.247 \pm 0.090}$	**0.75**	0.50

At pr = 0.75 the convolutional mean is 0.160 ($n = 4$ architectures) and the transformer mean 0.250 ($n = 2$); Welch $t = -3.77$, $P = 0.033$, Shapiro–Wilk $P = 0.92$ on the convolutional means ($P = 0.41$ over all six). Six architectures are too few to establish this contrast, and it does not hold along the whole grid: at pr = 1.0 the ordering reverses (convolutional 0.263, transformer 0.225). DenseNet-121's installation point of 0.65 is read off the full 14-rate grid (Table S2); the remaining architectures were run on the three-rate subset shown, so their installation points are resolved only to that grid.

Table S6: Non-radiology modalities at the Youden's-J operating point (three seeds, mean over seeds). Site and acquisition source are dataset provenance attributes, not patient characteristics. On PatchCamelyon and ISIC the attack reaches near-complete success but leaves the ±0.03 visibility band, so the binding constraint at high poison rates is detectability rather than efficacy; on PTB-XL the control cell rises with the attacked cell, so no rate produces a demographically selective failure (Table S3).

		pr = 0.50		pr = 0.75		pr = 1.00	
Cohort (shortcut)	Model	ASR_{rel}	ΔAUROC	ASR_{rel}	ΔAUROC	ASR_{rel}	ΔAUROC
PCam (site)	DenseNet-121	0.076	−0.008	0.718	−0.027	1.000	**−0.315**
PCam (site)	ViT-B/16	0.170	−0.010	0.337	−0.020	1.000	**−0.394**
ISIC (acquisition source)	DenseNet-121	0.134	−0.017	0.675	**−0.031**	0.977	**−0.211**
ISIC (acquisition source)	ViT-B/16	0.362	−0.004	0.832	**−0.033**	0.926	**−0.187**
PTB-XL (sex)	ResNet-1D	0.078	−0.009	0.113	−0.017	0.347	−0.038

Bold ΔAUROC values exceed the ±0.03 visibility criterion. PTB-XL control-cell ASR_{rel} is 0.070, 0.072 and 0.175 at the three rates, tracking the attacked cell closely.

Table S7: Four-parameter logistic fits to the dose–response, $\text{ASR}(\text{pr}) = d+(a-d)/(1+(\text{pr}/c)^b)$ with a fixed at 0, at the Youden's-J operating point. The inflection c is reported with a bias-corrected and accelerated bootstrap interval over seeds. Where ASR_{rel} has not saturated within the swept range the fitted inflection runs to the boundary of its support and is reported as unidentified; for those settings the criterion-based installation point in Table S3 is the reported quantity.

Cohort	rates	inflection c	BCa 95% CI	ΔAIC vs linear	identified
MIMIC race, unmatched	14	→ bound	0.60–0.94	−8.9	no
MIMIC race, matched	9	0.61	0.04–0.78	−2.8	yes
NIH sex, effusion	6	0.82	0.81–0.84	−23.1	yes
NIH sex, pneumothorax	5	0.93	0.72–0.93	−3.4	yes
PCam site	4	0.67	0.53–0.71	−5.6	yes
ISIC source	4	0.67	0.60–0.75	−10.8	yes
PTB-XL sex	4	→ bound	0.95–1.00	−1.4	no

Negative ΔAIC favours the logistic. We additionally regressed the fitted inflection on each setting's shortcut-detector AUROC to test whether decodability predicts the installation point, and found no support (slope −2.60, 95% CI −7.79 to 2.59, $R^2 = 0.25$, $P = 0.25$, Spearman $\rho = -0.51$, $P = 0.24$, $n = 7$). Four of seven settings lie above AUROC 0.99, so the comparison has little resolving power and we draw no conclusion in either direction.

Table S8: Audit detection over 768 evaluations spanning seven cohorts, seven architectures, four operating points and thirteen poison rates, with 164 clean models providing the null. The conventional absolute rule uses the flag thresholds an auditor would ordinarily apply; matched calibration sets each audit's threshold at the 95th percentile of its own statistic over clean models within each cohort and operating point. Detection rates are k/n with Wilson 95% intervals. False-positive rates under matched calibration are leave-one-out, so they are not read in-sample.

Flag rule	Subset	Subgroup-AUROC audit	Subgroup-FNR audit	McNemar	P
Conventional absolute	installed attacks	18/287 (6.3%)	268/287 (93.4%)	n/a	n/a
Conventional absolute	clean models (FPR)	0/164 (0.0%)	33/164 (20.1%)	n/a	n/a
Matched 5% FPR	installed attacks	210/287 (73.2%, CI 67.8–78.0)	275/287 (95.8%, CI 92.8–97.6)	$\chi^2 = 63.0$	2.1×10^{-15}
Matched 5% FPR	all attacked	404/768 (52.6%)	537/768 (69.9%)	$\chi^2 = 109.6$	1.2×10^{-25}
Matched 5% FPR	clean models (leave-one-out FPR)	24/164 (14.6%)	17/164 (10.4%)	n/a	n/a

Among installed attacks at matched calibration all 65 discordant cases favour the FNR audit. Effect magnitude on MIMIC race at the installation point (pr = 0.65), Youden's-J operating point: the subgroup-AUROC gap shifts by 0.0055, which is 4.4 clean standard deviations but 0.11 of its conventional flag threshold; the subgroup-FNR gap shifts by 0.121, which is 13.1 clean standard deviations and 1.21 of its threshold. The FNR figure is operating-point specific, since the same shift read at a fixed threshold of 0.5 is 16.5 standard deviations and 1.52 of threshold, which is the caution of Sec. 2.3 applied to our own audit statistic; the AUROC figure, being rank-based, is identical at all four operating points. No clean model was trained on NIH sex with pneumothorax as the target, so neither audit could be calibrated there and neither fires (0/60 each); those 60 evaluations, 11 of which are installed attacks, are retained in the denominators above for completeness but carry no information about either audit. Excluding them, the matched-calibration rates among installed attacks are 275/276 (99.6%, CI 98.0–99.9) for the FNR audit and 210/276 (76.1%, CI 70.7–80.7) for the rank audit; the discordant pairs and hence the McNemar statistic are unchanged. Both audits perform poorly on the matched cohort, the subgroup-AUROC audit detecting 36/136 and the subgroup-FNR audit 40/136, where five clean models are available. The two audits do not reach the same leave-one-out false-positive rate under matched calibration (14.6% and 10.4% respectively), so the comparison is not strictly matched; the FNR audit is nonetheless both more sensitive and more specific under this rule.

Table S9: Fairness-aware retraining against the installed MIMIC race attack (DenseNet-121, pr = 0.75, three seeds, evaluated at a fixed threshold of 0.5). Reweighting and Group DRO neutralize the attack, both over-correcting; adversarial debiasing does not at any adversarial weight. By the argument of Sec. 2.3 of the main text the ASR_{rel} magnitudes here are inflated relative to a defensible operating point, though the ordering of the three interventions is not affected. The probe AUROC is that of a freshly trained linear classifier reading race from the penultimate features, and is the diagnostic distinguishing failure to defend from failure to remove the demographic signal.

Intervention	ASR_{rel}	Overall AUROC	Adversary acc.	Probe AUROC
None (undefended)	0.333	n/a	n/a	n/a
Reweighting	−0.279	0.881	n/a	n/a
Group DRO	−0.367	0.887	n/a	n/a
Adversarial debiasing				
$\lambda = 0.01$	0.392 ± 0.118	0.887	0.794	0.922
$\lambda = 0.1$	0.399 ± 0.044	0.888	0.784	0.918
$\lambda = 0.3$	0.322 ± 0.036	0.890	0.784	0.923
$\lambda = 1$	0.279 ± 0.065	0.890	0.784	0.938
$\lambda = 3$	0.431 ± 0.010	0.888	0.784	0.950
$\lambda = 10$	0.539 ± 0.059	0.886	0.784	0.958

Reweighting uses inverse-prevalence sample weights; Group DRO minimizes worst-group loss. The adversary's validation accuracy is identical to four decimal places for every $\lambda \geq 0.3$ and every seed, and equals the majority-class rate: the gradient-reversal head collapsed to predicting the majority demographic and supplied no useful gradient. Probe AUROC rises monotonically with λ from $\lambda = 0.1$ upward, so increasing the adversarial weight made race *more* linearly decodable from the representation, not less. Chance probe AUROC is 0.500. The undefended row is the attacked model before any intervention; overall AUROC for the clean seed-matched model on this three-seed set is 0.890.

Table S10: Clinical consequence of the installed attack at pr = 0.75, evaluated at an operating point where the clean model reaches 0.80 sensitivity (MIMIC race, DenseNet-121, three seeds, 10,000-draw cluster bootstrap over patients). Prevalence and subgroup composition are those of the held-out MIMIC test split. No downstream outcome is modelled.

Quantity	Value	95% CI
Additional missed effusions per 1,000 imaged target-subgroup patients	30.6 ± 5.3	24.0–37.7
Number needed to harm (imaged patients)	33	27–43
Increase in target-subgroup false-negative rate	0.165	0.133–0.200
Target:control sensitivity ratio, clean model	0.882	n/a
Target:control sensitivity ratio, attacked model	0.681	0.616–0.742

The clean model's own disparity (0.882) is the baseline against which the attacked disparity (0.681) should be read: the attack adds roughly 1.7 times the gap already present without any attack, widening the sensitivity-ratio gap from 0.118 to 0.319, which is what makes it plausible as documented bias. As an arithmetic illustration only, a service performing 100,000 frontal chest radiographs annually with this cohort's 21.2% target-subgroup composition would accrue approximately 650 additional missed findings per year.

Table S11: External transfer of the installed MIMIC race attack to CheXpert, using recorded self-reported race. Inference only, using the attacked and seed-matched clean MIMIC checkpoints without modification; three seeds. Cohort: 118,645 frontal studies, 10,144 `BLACK_OR_AA` and 108,501 `WHITE`, effusion prevalence 44.6% and 45.8% respectively, with no overlap with any model's training data. The transfer effect is the attacked-minus-clean difference of the between-subgroup false-negative-rate gap.

Operating point	Transfer effect, recorded race	Predicted-tercile estimate
Fixed $t = 0.5$	$+0.087 \pm 0.006$	$+0.051$
Youden's J	$+0.080 \pm 0.011$	$+0.042$
Sensitivity 0.80	$+0.086 \pm 0.003$	$+0.047$
Specificity 0.90	$+0.086 \pm 0.002$	$+0.052$

The MIMIC race detector transfers to this cohort at AUROC 0.952, indicating that the trigger is decodable across institutions. Unlike the installation point, the transfer magnitude is stable across operating points. The predicted-tercile column is shown to calibrate the proxy used on cohorts without recorded race (Supplementary Note S3); at this cohort's 8.5% target share it understates the effect by roughly half.

Table S12: Calibration of the predicted-race tercile proxy against recorded race, wherever both coexist. Rank agreement is excellent throughout; magnitude agreement depends on the target subgroup's share of the cohort, because a tercile is a fixed third whereas the true group is not.

Cohort	Target share	Top-tercile purity	Proxy-on-true slope	Spearman ρ
MIMIC test, leakage-free subset	44.9%	0.986	1.45 (1.37–1.52)	0.99
MIMIC test, full split	21.2%	0.605	0.66 (0.55–0.77)	0.91
CheXpert	8.5%	n/a	0.53 (0.33–0.73)	0.50

The MIMIC leakage-free subset excludes subjects present in the race detector's training data. Three prevalence points are a described pattern, not an estimated relationship. The proxy licenses existence, direction, dose–response and ordering of the transfer effect; it does not license magnitude.

Table S13: Poison rate versus absolute flipped-label count. The eligible cell was downsampled to three sizes while the training-set size was held exactly constant at 116,598 by backfilling with held-out target-subgroup negatives, so that rate and count could be varied independently. Two equal-count diagonals are realised by construction. Youden's-J operating point, three seeds per cell.

Cell scale	Cell size	pr = 0.50	pr = 0.75	pr = 1.00
0.25	1,186	0.080 (n=593)	0.126 (n=889)	0.174 (n=1,186)
0.50	2,371	0.091 (n=1,185)	0.108 (n=1,778)	0.096 (n=2,371)
1.00	4,742	0.112 (n=2,371)	0.172 (n=3,556)	0.285 (n=4,742)

Equal-count contrasts: at $n \approx 1{,}185$ flipped labels, 0.091 versus 0.174 ($P = 0.46$); at $n = 2{,}371$, 0.096 versus 0.112 ($P = 0.81$). Neither of the two equal-count contrasts, nor any of the nine equal-rate contrasts, differs at $P < 0.05$. All four candidate models relating $\mathrm{ASR}_{\mathrm{rel}}$ to log(count), rate, both or neither lie within 2.7 AIC of one another, with the nominal best leading by 0.4, so the design does not discriminate between them at this resolution; partial R^2 is 0.059 for log(count) and 0.043 for rate. This is reported as a bound on what three seeds and three cell sizes can resolve, not as a dissociation. One consequence of the design is inherent and should be noted: downsampling the positive cell while holding the subgroup size constant necessarily lowers within-group effusion prevalence, from 0.221 to 0.111 to 0.055 across the three arms.

Table S14: Foundation-model supply chain (MIMIC race, pleural effusion). Mode A is a frozen linear probe (three seeds); Mode B a full fine-tune (two seeds). The final column gives the lowest poison rate meeting all three install criteria, or "none" where no tested rate does. These runs are evaluated at a fixed threshold of 0.5; by the argument of the main text their installation points should be read as lower bounds.

Encoder	Mode	pr = 0.25	pr = 0.50	pr = 1.00	Install
RAD-DINO	A: frozen probe	0.164	0.328	0.641	0.50
	B: fine-tune	0.182	0.339	0.788†	0.50
BiomedCLIP	A: frozen probe	0.134	0.219	0.431	0.50
	B: fine-tune	0.148	0.235	0.640	0.50
MedSigLIP	A: frozen probe	0.126	0.242	0.486	0.50
	B: fine-tune	0.059	0.179	0.624†	none

†Visibility criterion exceeded (ΔAUROC -0.039 and -0.043 respectively). Fine-tuning does not lower the installation point relative to the frozen probe for any encoder.

Table S15: Clean baseline subgroup performance, establishing the pre-existing disparity the attack compounds and the sensitivity available to it at each operating point. DenseNet-121, mean over five seeds on the held-out test split. These are the baselines of the original attack runs; the clean validation sensitivities in Table S3 come from the three-seed rescoring and are not the same seed set, so the two differ by up to 0.12 in subgroup sensitivity at a fixed threshold of 0.5.

Cohort	Finding	Operating point	Sensitivity att./ctrl	Subgroup AUROC att./ctrl	Overall AUROC
MIMIC (race)	pleural effusion	fixed $t = 0.5$	0.454 / 0.533	0.880 / 0.888	0.884
MIMIC (race)	pleural effusion	Youden's J	0.798 / 0.858	0.880 / 0.888	0.884
MIMIC (race)	pneumothorax	fixed $t = 0.5$	0.054 / 0.060	0.834 / 0.806	0.821
MIMIC (race)	cardiomegaly	fixed $t = 0.5$	0.335 / 0.209	0.799 / 0.791	0.797
NIH (sex)	pleural effusion	fixed $t = 0.5$	0.401 / 0.409	0.883 / 0.877	0.880
NIH (sex)	pleural effusion	Youden's J	0.821 / 0.831	0.883 / 0.877	0.880
NIH (sex)	pneumothorax	fixed $t = 0.5$	0.074 / 0.125	0.858 / 0.879	0.869

The very low sensitivities at a fixed threshold of 0.5, particularly for pneumothorax, are the reason relative subgroup-error metrics computed there are inflated: $\mathrm{ASR}_{\mathrm{rel}}$ is normalized by $1 - \mathrm{FNR}_{\mathrm{clean}}$, so a clean sensitivity of 0.394 inflates it roughly 2.5-fold relative to a perfectly sensitive model, and roughly 2.1-fold relative to this cohort's Youden's-J operating point, where the clean sensitivity is 0.827.

Table S16: Transfer of the attacked MIMIC model to cohorts without recorded race, stratified into top and bottom terciles of predicted $P(\text{Black})$ by a held-out race detector (three seeds, fixed threshold 0.5). These strata are model-defined rather than label-defined, and given the target shares involved the magnitudes are most likely understated (Supplementary Note S3); the CheXpert analysis in Table S11 is the primary external evidence.

Cohort	Poison rate	$\mathrm{FNR}_{\mathrm{high}}$	$\mathrm{FNR}_{\mathrm{low}}$	Gap	Transfer effect
NIH	0.00	0.387	0.281	0.106	n/a
NIH	0.75	0.563	0.319	0.244	0.138 ± 0.049
NIH	1.00	0.732	0.371	0.361	0.255 ± 0.033
VinDr	0.00	0.400	0.333	0.067	n/a
VinDr	0.75	0.483	0.382	0.102	0.035 ± 0.164
VinDr	1.00	0.700	0.527	0.173	0.106 ± 0.110

VinDr is a Vietnamese cohort in which the `BLACK_OR_AA` and `WHITE` contrast carries limited meaning; it is retained as a distribution-shift stress test, not as evidence of demographic transfer.

Table S17: NIH-CXR14 sex attack at the Youden's-J operating point (DenseNet-121, female attacked and male control, three seeds, mean over seeds). Install criteria are $\text{ASR}_{\text{rel}} \geq 0.20$, attacked-minus-control gap ≥ 0.05, and $|\Delta\text{AUROC}| \leq 0.03$. At this operating point the attack is sub-threshold at pr = 0.10 and first meets all three criteria at 0.75, inside the 0.50–0.75 band the race axis also falls in; at higher rates it meets the effect and gap criteria but exceeds the visibility criterion. At a fixed threshold of 0.5, where the clean model reaches only 0.394 sensitivity on its validation split, the same runs appear to install at pr = 0.10 (Fig. S1).

Poison rate	ASR_{rel} F	ASR_{rel} M	ΔAUROC	Criteria met
0.00	0.000	0.000	0.000	
0.10	0.046	0.024	−0.001	
0.50	0.137	0.013	−0.006	
0.75	**0.361**	0.028	−0.020	**all three**
0.90	0.573	0.016	−0.038	visibility exceeded
1.00	0.757	0.039	−0.114	visibility exceeded

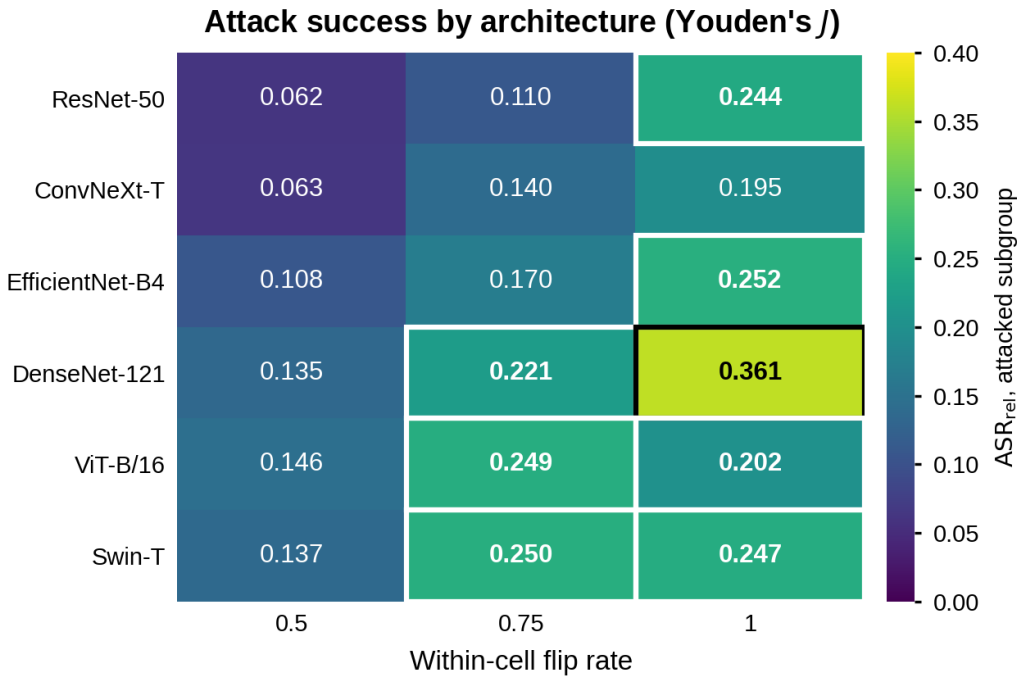

Figure S2: Architecture by poison-rate attack success. Relative attack success rate in the attacked subgroup for six backbones across three within-cell flip rates at the Youden's-J operating point (MIMIC unmatched race, pleural effusion, mean of three seeds). Outlined cells in bold type meet all three install criteria. Three of the six architectures install at a flip rate of 0.75 or below; ResNet-50 and EfficientNet-B4 do so only at 1.0, and ConvNeXt-T does not meet the effect criterion at any tested rate. DenseNet-121's installation point of 0.65 lies between the columns shown and is read off the full 14-rate grid (Table S2). The two transformer backbones show higher mean susceptibility than the four convolutional backbones at a flip rate of 0.75, though not at 1.0. Colour encodes attack success on a perceptually uniform sequential scale, with the numerical value printed in each cell. Drawn from the same rescored runs as Table S5.

Table S18: Specificity of the installed backdoor. (a) The failure induced in the attacked subgroup is distributed uniformly across sex and age within it. **(b)** It is confined to the finding whose labels were flipped; the two co-trained findings are unaffected. MIMIC-CXR unmatched cohort, DenseNet-121, three seeds, each finding evaluated at its own Youden's-J threshold taken from the seed-matched clean model's validation split; values are mean ± s.d. over seeds. Age tertiles of the test split are < 57, 57–71 and > 71 years. Clean sensitivity is shown in (a) because $\mathrm{ASR}_{\mathrm{rel}}$ is normalized by it, so cells with different baselines remain comparable; the differences within the attacked subgroup are within seed variation in every case. The uniformity is unfavourable for auditing: the attack leaves no concentrated intersectional cell for a subgroup dashboard to isolate. In (b) the only movement outside the attacked finding appears at pr = 1.00 on pneumothorax and is larger in the control subgroup than in the attacked one, which identifies it as generic degradation and not spillover of the backdoor. Intersections are characterized here as a footprint; they are not themselves targeted.

(a) Attack footprint within and across demographic strata, pleural effusion

Subgroup	Stratum	n pos.	Sens. clean	$\mathrm{ASR}_{\mathrm{rel}}$ pr=0.65	$\mathrm{ASR}_{\mathrm{rel}}$ pr=0.75
Attacked (`BLACK_OR_AA`)	female	750	0.760	0.197 ± 0.047	0.209 ± 0.034
	male	545	0.757	0.216 ± 0.055	0.237 ± 0.010
	age < 57	399	0.743	0.193 ± 0.050	0.217 ± 0.022
	age 57–71	412	0.714	0.241 ± 0.078	0.247 ± 0.017
	age > 71	484	0.810	0.186 ± 0.031	0.204 ± 0.029
Control (`WHITE`)	female	3,643	0.862	0.044 ± 0.024	0.008 ± 0.019
	male	4,209	0.819	0.045 ± 0.030	0.008 ± 0.027
	age < 57	1,670	0.788	0.051 ± 0.018	0.008 ± 0.032
	age 57–71	2,850	0.838	0.043 ± 0.029	0.006 ± 0.028
	age > 71	3,332	0.866	0.043 ± 0.031	0.011 ± 0.014

(b) Effect on the two co-trained, untargeted findings

	Labels	$\mathrm{ASR}_{\mathrm{rel}}$ attacked / control			ΔAUROC
Finding	flipped	pr=0.65	pr=0.75	pr=1.00	pr=0.75
pleural effusion	yes	**0.205** / 0.045	**0.221** / 0.008	**0.361** / 0.043	−0.005
pneumothorax	no	−0.052 / 0.010	0.013 / 0.048	0.073 / 0.117	−0.005
cardiomegaly	no	−0.039 / −0.041	−0.010 / −0.011	0.028 / 0.002	−0.001

Clean overall AUROC is 0.890 for pleural effusion, 0.835 for pneumothorax and 0.785 for cardiomegaly. Only pleural-effusion labels are ever flipped; the model carries a three-label sigmoid head throughout, so the two untargeted findings are trained on the same images under the same optimization.

Table S19: Reversing the direction of the attack destroys its selectivity. The published attack poisons `BLACK_OR_AA`×pleural effusion and leaves `WHITE` untouched; the reverse arm poisons `WHITE` and leaves `BLACK_OR_AA` untouched. Cohort, architecture, seeds, seed-matched clean baselines and Youden's-J thresholds are identical, so the only difference is which subgroup's labels are flipped and, in consequence, how many labels that is. Values are mean ± s.d. over three seeds. The final column is the ratio of control to target attack success: a subgroup-selective attack keeps it near zero, and it stays between 0.61 and 0.86 for every rate of the reverse arm. The install criteria were specified to detect a targeted failure and do not test for selectivity, so they score the reverse arm as installing at pr = 0.50; that verdict should not be read as a successful backdoor, because at that rate the untouched control subgroup has already lost 28.5% of its remaining sensitivity. The reverse arm is better described as global degradation with a subgroup gradient, the same signature that appears on the most separable non-radiology shortcuts (Sec. 2.7 of the main text). At pr = 1.0 it destroys the model (ΔAUROC −0.227).

Poisoned subgroup	pr	Labels flipped	% of train labels	ASR_{rel} target	ASR_{rel} control	Control / target
`BLACK_OR_AA` (published) cell = 4,742	0.25	1,185	0.99%	0.035 ± 0.025	-0.004 ± 0.017	−0.11
	0.50	2,371	1.97%	0.173 ± 0.125	0.043 ± 0.056	0.25
	0.65	3,082	2.57%	$\mathbf{0.205 \pm 0.049}$	0.045 ± 0.027	0.22
	0.75	3,556	2.96%	0.221 ± 0.017	0.008 ± 0.022	0.04
	1.00	4,742	3.95%	0.361 ± 0.048	0.043 ± 0.039	0.12
`WHITE` (reverse) cell = 28,695	0.25	7,173	5.97%	0.118 ± 0.062	0.102 ± 0.075	0.86
	0.50	14,347	11.94%	0.431 ± 0.165	0.285 ± 0.133	0.66
	0.65	18,651	15.52%	0.675 ± 0.122	0.412 ± 0.133	0.61
	0.75	21,521	17.91%	0.908 ± 0.042	0.562 ± 0.041	0.62
	1.00	28,695	23.88%	0.996 ± 0.003	0.849 ± 0.069	0.85

Visibility cost, ΔAUROC against the seed-matched clean model, for the published arm: −0.000, −0.002, −0.004, −0.005, −0.014 across the five rates. For the reverse arm: −0.003, −0.011, −0.021, −0.034, −0.227. The reverse arm leaves the ±0.03 visibility band at pr = 0.75; the published arm never does.

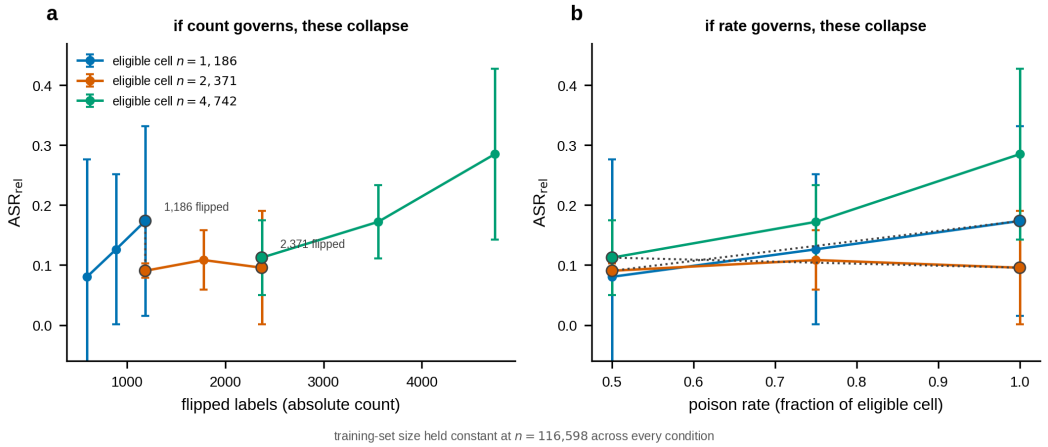

Figure S3: Poison rate and absolute flipped-label count are not separated at this resolution. **a**, Relative attack success rate against the number of flipped labels, with points coloured by the size of the eligible cell; if the attack were governed by absolute count these series would collapse onto one curve. **b**, The same values against poison rate; if the attack were governed by rate these series would collapse instead. Neither collapse is clean. Filled markers connected by a dotted line mark the two equal-count diagonals, at approximately 1,185 and 2,371 flipped labels, where the same number of labels is flipped at different rates. Points are means over three seeds and error bars one standard deviation; the training-set size is held exactly constant across all conditions.

Table S20: The backdoor reaches patients whose race is not recorded, in proportion to what the model reads from the image. The published cohorts retain only `WHITE` and `BLACK_OR_AA` patients; this is their complement, every frontal MIMIC-CXR study whose subject falls outside those two categories (73,274 images, 24,716 subjects, 16,698 pleural-effusion positives, 30.1% of frontal MIMIC), scored by inference only and subject-disjoint from the attack cohort by construction. **(a)** Deciles of $P(\text{Black} \mid \text{image})$ from a held-out MIMIC race detector. **(b)** The same patients grouped by the race category actually recorded in MIMIC-IV. Attacked model at pr $= 0.75$ against its seed-matched clean model at the same Youden's-J threshold, mean $\pm$ s.d. over three seeds. Attack success rises monotonically with the predicted score and reaches 0.222 ± 0.017 in the top decile, indistinguishable from the 0.221 ± 0.017 measured on recorded `BLACK_OR_AA` patients inside the cohort (Table S2). The recorded categories order themselves by their mean predicted score, not by anything in the label. These patients cannot appear in a subgroup audit keyed on recorded race, and the category-level averages in (b) understate by up to sixfold what is happening to individuals within them.

(a) By predicted $P(\text{Black} \mid \text{image})$ decile

Decile	n pos.	Mean P(Black)	FNR clean	FNR attacked	ASR_{rel}
1 (lowest)	2,111	0.000	0.141	0.139	-0.003 ± 0.025
2	1,956	0.000	0.145	0.146	0.002 ± 0.027
3	1,883	0.000	0.170	0.169	-0.002 ± 0.029
4	1,846	0.000	0.167	0.176	0.011 ± 0.026
5	1,822	0.000	0.168	0.180	0.014 ± 0.022
6	1,767	0.001	0.164	0.180	0.019 ± 0.029
7	1,626	0.004	0.168	0.196	0.034 ± 0.031
8	1,536	0.026	0.186	0.235	0.060 ± 0.030
9	1,320	0.260	0.192	0.274	0.101 ± 0.018
10 (highest)	831	0.912	0.233	0.403	$\mathbf{0.222 \pm 0.017}$

(b) By recorded MIMIC-IV race category

Recorded category	Images	n pos.	Mean P(Black)	FNR clean $\rightarrow$ attacked	ASR_{rel}
OTHER	6,989	1,824	0.170	$0.175 \rightarrow 0.221$	0.056 ± 0.030
Hispanic/Latino	13,394	2,512	0.157	$0.194 \rightarrow 0.239$	0.055 ± 0.023
No race recorded	25,012	3,443	0.165	$0.179 \rightarrow 0.208$	0.034 ± 0.018
Asian	7,262	2,164	0.090	$0.157 \rightarrow 0.184$	0.032 ± 0.025
American Indian, Pacific Islander, multiple	975	322	0.092	$0.155 \rightarrow 0.171$	0.018 ± 0.017
UNKNOWN	8,859	3,167	0.056	$0.147 \rightarrow 0.162$	0.017 ± 0.036
White, non-US subcategory	9,752	2,883	0.011	$0.166 \rightarrow 0.175$	0.011 ± 0.022
Declined / unable to obtain	1,031	383	0.049	$0.138 \rightarrow 0.145$	0.007 ± 0.045

Across the whole cohort the attack adds 0.026 ± 0.021 to the false-negative rate at pr $= 0.75$ and 0.054 ± 0.024 at pr $= 0.65$. Strata with fewer than 50 positives are not reported.

Table S21: Conditioning the attack on a different demographic axis, and on a smaller intersectional cell. MIMIC-CXR unmatched cohort, DenseNet-121, pleural effusion, Youden's-J operating point, mean ± s.d. over three seeds, against the same seed-matched clean models used throughout. **(a)** Age reproduces the failure on a third demographic axis; the intersectional cell does not meet the effect criterion at any rate tested, so narrowing the target group does not produce the same failure for fewer flipped labels. **(b)** The within-cell poison rate is held fixed at 0.65 and only the target cell changes, giving four conditions on one cohort that span tenfold in absolute flipped-label count. Attack success is not constant across them, which rules out an account of the installation threshold in terms of the within-cell rate alone; it is close to log-linear in the count ($r = 0.994$). We do not draw the converse conclusion, because the four conditions also differ in the attribute being conditioned on. The final column shows that demographic selectivity falls as the cell grows: the largest effects are the least backdoor-like. The age arm was extended down to pr = 0.05 so that its installation point is bracketed rather than merely bounded: it fails the effect criterion at 0.05, 0.10 and 0.25 and meets all three from 0.50 upward, placing the age installation point in (0.25, 0.50] against 0.65 for race on the same cohort, architecture and seeds. Age therefore requires the smaller fraction of the two, consistent with its larger eligible cell, and the dose–response has the same flat-then-steep shape on both.

(a) Dose–response on two further conditionings

Target cell	pr	Labels flipped	ASR_{rel} target	ASR_{rel} control	ΔAUROC	Criteria met
Age < 65 cell = 13,621	0.05	681	0.026 ± 0.026	0.015 ± 0.022	-0.001 ± 0.001	
	0.10	1,362	0.035 ± 0.025	0.016 ± 0.022	-0.000 ± 0.001	
	0.25	3,405	0.115 ± 0.036	0.057 ± 0.024	-0.002 ± 0.000	
	0.50	6,810	$\mathbf{0.320 \pm 0.085}$	0.139 ± 0.051	-0.008 ± 0.003	**all three**
	0.65	8,853	0.433 ± 0.091	0.145 ± 0.055	-0.017 ± 0.001	all three
	0.75	10,215	0.496 ± 0.034	0.170 ± 0.030	-0.021 ± 0.001	all three
BLACK_OR_AA×female cell = 2,763	0.50	1,381	0.061 ± 0.026	-0.003 ± 0.013	-0.001 ± 0.001	
	0.65	1,795	0.097 ± 0.035	0.018 ± 0.038	-0.001 ± 0.001	
	0.75	2,072	0.104 ± 0.021	0.014 ± 0.027	-0.002 ± 0.000	

(b) Fixed within-cell rate (pr = 0.65), varying cell size

Target cell	Cell size	Labels flipped	% of train labels	ASR_{rel} target	Control / target
BLACK_OR_AA×female	2,763	1,795	1.49%	0.097 ± 0.035	0.18
BLACK_OR_AA (published)	4,742	3,082	2.57%	0.205 ± 0.049	0.22
Age < 65	13,621	8,853	7.37%	0.433 ± 0.091	0.33
WHITE	28,695	18,651	15.52%	0.675 ± 0.122	0.61

Pearson correlation between ASR_{rel} and the base-10 logarithm of the flipped-label count across the four conditions is 0.994. The same four conditions at a fixed *count* are not available, because the cells are of fixed size; the cell-size factorial that holds the training-set size constant is Table S13.

Table S22: Relabelling whole patients instead of individual images does not lower the cost of installing the backdoor. MIMIC-CXR unmatched cohort, DenseNet-121, pleural effusion, Youden's-J operating point, mean ± s.d. over three seeds, against the same seed-matched clean models used throughout. The two arms flip an identical number of labels at each rate and differ only in how those flips are distributed over patients: the published image-level arm draws eligible image rows at random, leaving 74% of target patients who hold more than one eligible image with mutually contradictory pleural-effusion labels, whereas the subject-level arm flips every eligible image of a randomly drawn patient until the same budget is exhausted, leaving none. The subject-level arm is the more natural unit for a relabelling vendor acting on the patient record. Both arms first meet all three install criteria at the same within-cell rate, pr = 0.65, and neither differs from the other beyond seed-level variation. Within-patient label consistency is therefore not what limits the attack, and the installation points reported throughout this study are not an artefact of the image-level design.

Poisoning unit	pr	Labels flipped	Patients touched	ASR_{rel} target	ASR_{rel} control	ΔAUROC	Criteria met
Image (published)	0.50	2,371	898	0.135 ± 0.110	0.028 ± 0.047	-0.002 ± 0.001	
Patient	0.50	2,371	629	0.130 ± 0.033	0.009 ± 0.031	-0.003 ± 0.001	
Image (published)	**0.65**	3,082	1,025	$\mathbf{0.205 \pm 0.049}$	0.045 ± 0.027	-0.004 ± 0.002	**all three**
Patient	**0.65**	3,082	826	$\mathbf{0.216 \pm 0.092}$	0.043 ± 0.049	-0.004 ± 0.001	**all three**

The eligible cell holds 4,742 target-subgroup pleural-effusion positives from 1,232 patients. "Patients touched" is the mean number of distinct patients holding at least one flipped label, over three seeds; the patient-level arm concentrates the same budget on fewer patients by construction and exhausts it exactly, with no shortfall at either rate. Image-level comparators at pr = 0.65 and at pr = 0.50 for two of three seeds come from the dose–response sweep of Table S2; all six pairs share the clean model and the operating point of their seed.

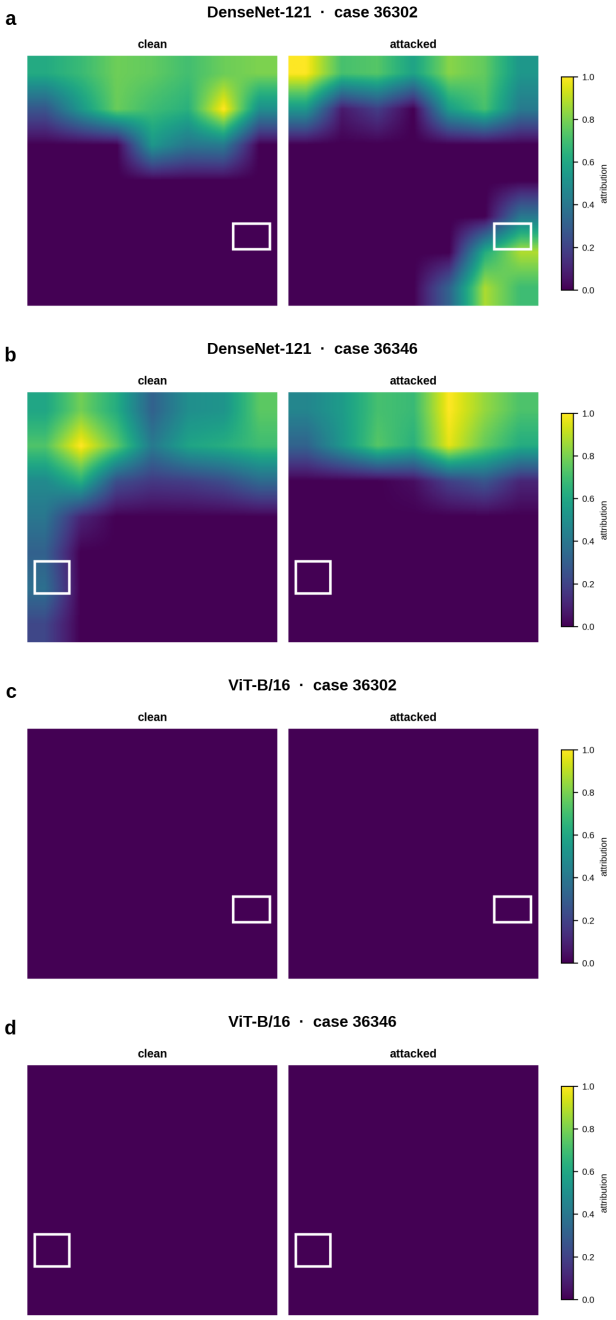

Figure S4: Spatial attribution for clean and attacked models. GradCAM maps for DenseNet-121 (**a**, **b**) and attention maps for ViT-B/16 (**c**, **d**) on two representative ChestX-Det10 cases, with the clean model on the left and the attacked model on the right of each pair. Ground-truth pleural-effusion bounding boxes are outlined in white. Attribution intensity uses a perceptually uniform sequential colour scale on a fixed 0–1 range shared by every panel, so panels are directly comparable and a flat panel indicates a flat attribution map rather than a rendering failure. The ViT-B/16 attention maps are near-uniform on both the clean and the attacked model, which is why panels **c** and **d** carry no visible structure. Localization intersection-over-union with the ground-truth boxes drifts from 0.143 to 0.135 on average over the two architectures and is unchanged on ViT-B/16 (0.117 in both conditions), so spatial attribution is not a reliable stand-alone detector.

References

[1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR vol. 108*, pages 2938–2948, 2020. arXiv:1807.00459.

[2] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security)*, pages 1505–1521, 2021. arXiv:2005.03823.

[3] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. doi: 10.1016/S2589-7500(22)00063-2.

[4] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

[5] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3104–3122, 2021. doi: 10.1145/3460120.3485368.

[6] David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020), LNCS vol. 12457*, pages 162–177. Springer, 2021. doi: 10.1007/978-3-030-67658-2_10.

[7] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8930–8938, 2021. doi: 10.1609/aaai.v35i10.17080.

[8] Tianci Liu, Haoyu Wang, Feijie Wu, Hengtong Zhang, Pan Li, Lu Su, and Jing Gao. Towards poisoning fair representations. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16487.

[9] Eunice Chan and Hanghang Tong. Adversarial bias: data poisoning attacks on fairness. In *IEEE International Conference on Big Data (IEEE BigData)*, 2025. arXiv:2511.08331.

[10] Pranav Kulkarni, Andrew Chan, Nithya Navarathna, Skylar Chan, Paul H. Yi, and Vishwa S. Parekh. Hidden in plain sight: undetectable adversarial bias attacks on vulnerable patient populations. In *Medical Imaging with Deep Learning (MIDL), PMLR vol. 250*, pages 793–821, 2024. arXiv:2402.05713.

[11] Jason Uwaeze, Pranav Kulkarni, Vladimir Braverman, Michael A. Jacobs, and Vishwa S. Parekh. Generative counterfactual augmentation for bias mitigation. In *Proc. IEEE/CVF Intl. Conf. on Computer Vision Workshops (ICCVW), CVAMD*, pages 1164–1171, 2025.

[12] Guangyao Zheng, Michael A. Jacobs, Vladimir Braverman, and Vishwa S. Parekh. Towards fair medical AI: adversarial debiasing of 3D CT foundation embeddings, 2025. arXiv:2502.04386.

[13] Jiaqi Xue, Qian Lou, and Mengxin Zheng. BadFair: backdoored fairness attacks with group-conditioned triggers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8257–8270, 2024. doi: 10.18653/v1/2024.findings-emnlp.484. arXiv:2410.17492.

[14] Jiaqi Xue, Mengxin Zheng, Yi Sheng, Lei Yang, Qian Lou, and Lei Jiang. TrojFair: trojan fairness attacks. In *ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS)*, pages 47–56, 2024. doi: 10.1145/3689217.3690620. arXiv:2312.10508.

[15] Nicholas Furth, Abdallah Khreishah, Guanxiong Liu, NhatHai Phan, and Yaser Jararweh. Un-fair trojan: targeted backdoor attacks against model fairness. In *IEEE 9th Intl. Conf. on Software Defined Systems (SDS)*, pages 1–9, 2022. doi: 10.1109/SDS57574.2022.10062890.

[16] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint*, 2019. arXiv:1912.02771.

[17] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11957–11965, 2020. doi: 10.1609/aaai.v34i07.6871. arXiv:1910.00033.

[18] Dazhong Rong, Guoyao Yu, Shuheng Shen, et al. Clean-image backdoor attacks. In *Artificial Neural Networks and Machine Learning – ICANN 2024, LNCS vol. 15025*, pages 187–202. Springer, 2024. doi: 10.1007/978-3-031-72359-9_14. arXiv:2403.15010.

[19] Binyan Xu, Fan Yang, Di Tang, Xilin Dai, and Kehuan Zhang. Breaking the stealth-potency trade-off in clean-image backdoors with generative trigger optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026. arXiv:2511.07210.

[20] Yangxu Yin, Honglong Chen, Yudong Gao, et al. FFCBA: feature-based full-target clean-label backdoor attacks. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, 2025. doi: 10.1145/3746027.3755227. arXiv:2504.21054.

[21] Ruinan Jin, Chun-Yin Huang, Chenyu You, and Xiaoxiao Li. Backdoor attack on unpaired medical image-text foundation models: a pilot study on MedCLIP. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 272–285, 2024. doi: 10.1109/SaTML59370.2024.00020. arXiv:2401.01911.

[22] Asif Hanif, Fahad Shamshad, Muhammad Awais, et al. BAPLe: backdoor attacks on medical foundational models using prompt learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024. doi: 10.1007/978-3-031-72390-2_42. arXiv:2408.07440.

[23] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 2043–2059, 2022. doi: 10.1109/SP46214.2022.9833644. arXiv:2108.00352.

[24] Farhad Abtahi, Fernando Seoane, Ivan Pau, and Mario Vega-Barbas. Data poisoning vulnerabilities across health care artificial intelligence architectures: analytical security framework and defense strategies. *Journal of Medical Internet Research*, 28:e87969, 2026. doi: 10.2196/87969.

[25] Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14(1):4314, 2023. doi: 10.1038/s41467-023-39902-7.

[26] Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth P. Seastedt, Judy W. Gichoya, and Leo A. Celi. Drop the shortcuts: image augmentation improves fairness and decreases AI detection of race and other demographics from medical images. *eBioMedicine*, 102:105047, 2024. doi: 10.1016/j.ebiom.2024.105047.

[27] William Lotter. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15(1):7465, 2024. doi: 10.1038/s41467-024-52003-3.

[28] Eike Petersen, Enzo Ferrante, Melanie Ganz, and Aasa Feragen. Are demographically invariant models and representations in medical imaging fair? *arXiv preprint*, 2023. arXiv:2305.01397.

[29] Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, 2025. doi: 10.1136/bmj-2024-081554.

[30] Joseph E. Alderman, Joanne Palmer, Elinor Laws, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, 2025. doi: 10.1016/S2589-7500(24)00224-3.

[31] Michael D. Abràmoff, Michelle E. Tarver, Nilsa Loyo-Berrios, et al. Considerations for addressing bias in artificial intelligence for health equity. *npj Digital Medicine*, 6(1):170, 2023. doi: 10.1038/s41746-023-00913-9.

[32] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 3315–3323, 2016.

[33] Ha-Hieu Pham, Hai-Dang Nguyen, Dang P. M. Cao, Thanh-Huy Nguyen, Min Xu, Trung-Nghia Le, Ulas Bagci, and Huy-Hieu Pham. Who gets missed in the tail? Thresholded subgroup underdiagnosis in long-tailed chest X-ray classification, 2026. arXiv:2607.07717.

[34] Brandon Dominique, Prudence Lam, Nicholas Kurtansky, Jochen Weber, Kivanc Kose, Veronica Rotemberg, and Jennifer Dy. On the role of calibration in benchmarking algorithmic fairness for skin cancer detection. *Machine Learning for Biomedical Imaging (MELBA)*, 3: 618–636, 2025. doi: 10.59275/j.melba.2025-ae66. arXiv:2511.07700.

[35] Han-Jay Shu, Shun-Ting Chang, Rodrigo Rosa Gameiro, Judy Wawira Gichoya, Leo Anthony Celi, and Po-Chih Kuo. Impact of exposure parameters on deep learning models in chest radiography and implications for deployment. *Radiology: Artificial Intelligence*, 8(4):e250731, 2026. doi: 10.1148/ryai.250731.

[36] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning (CHIL), PMLR vol. 174*, pages 204–233, 2022. arXiv:2203.12609.

[37] Nicholas J. Jackson, Chao Yan, and Bradley A. Malin. Enhancement of fairness in AI for chest x-ray classification. In *AMIA Annual Symposium Proceedings*, volume 2024, pages 551–560, 2024. PMID 40417510; electronically published 22 May 2025.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.