

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

Saptarshi Purkayastha

saptpurk@iu.edu

Indiana University Indianapolis

Parvati Naliyatthalizchayil

Indiana University Indianapolis

Judy W. Gichoya

Emory University

Article

Keywords:

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10919324/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

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

Saptarshi Purkayastha^{1,*} Parvati Naliyatthaliyazchayil¹ Judy Wawira Gichoya²

¹Department of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis, IN, USA

²Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA, USA

*Corresponding author: saptpurk@iu.edu

Abstract

Medical imaging models read race and sex from pixels alone, making demographic identity a usable backdoor trigger. Flipping positive labels to negative within one group, with no image altered, silently suppresses that group’s disease detection. Across seven cohorts, three foundation-model encoders and four modalities, the backdoor installed once two thirds of that group’s positive labels for one finding were flipped, 2.6% of the training set. Overall AUROC moved 0.004 and the comparison group stayed untouched. It cost 30.6 missed pleural effusions per 1,000 imaged patients of the targeted group and transferred to another health system. Harm reached patients whose race was never recorded, through the pixel shortcut rather than the chart entry, so an audit on recorded demographics is blind to 30% of studies. Five backdoor detectors failed. What works is auditing subgroup false-negative rate at the deployed threshold, which flagged 96% of installed failures compared to 73% for a rank-based audit.

1 Introduction

Imaging models cleared for clinical use do not perform equally for all patients. Chest-radiograph classifiers systematically underdiagnose historically under-served groups^{1,2}, and bias analyses of chest-radiograph foundation models report that detection of common findings degrades for Black patients while overall metrics stay strong^{3,4}. Benchmarked against board-certified radiologists across five globally sourced datasets, expert-level vision-language models underdiagnose female and Black patients across a wide range of pathologies⁵. Because such errors resemble health-care disparities that already exist, they are easy to blame on the data and hard to notice after deployment^{6,7}.

This study asks a practical question about that resemblance. Suppose the training labels for one demographic group are systematically wrong for a single finding. How wrong must they be before the deployed model quietly stops detecting that finding in that group, and would any check a clinical service actually runs catch it?

The question matters because there are two routes to the same corrupted training set, and only one of them involves an attacker. The everyday route is ordinary label failure. A report-extraction pipeline that performs worse on one population’s reports, an annotation vendor whose readers under-call one group, or a contributing site whose radiologists document a finding differently will all under-report that group’s positives. The adversarial route is a party who wants a competitor’s device to fail after deployment and who holds labels for part of the training data. Labels are routinely outsourced to annotation vendors, extracted automatically from radiology reports, or aggregated across contributing sites, so that access is unremarkable. Both routes yield the same label distribution and the same deployed model, and for a service running that model the harm and the remedy are identical. We describe the deliberate version because it lets us control the amount of corruption and measure the dose, but nothing in the result depends on intent.

What makes the failure hard to see is the feature the model latches onto. Convolutional and transformer models predict self-reported race from chest radiographs with high accuracy, from images in which no obvious anatomical correlate reveals it, and the signal is spread across anatomical regions and the frequency spectrum rather than confined to one patch⁸. The mechanism is debated: confounders mediate part of demographic prediction in some modalities⁹, and acquisition and processing parameters contribute measurably in chest radiography, to the point that correcting for them reduces underdiagnosis bias¹⁰. Exposure settings recorded in DICOM metadata behave as a shortcut in their own right, and may be one route by which demographic information reaches the pixels¹¹. The fact this work depends on is not disputed under either account. A demographic attribute is recoverable from the pixels at high accuracy, is present in every image of the target group, and is hard to remove without degrading the diagnostic task^{6,12}, which are the defining conditions of a learned shortcut^{13–15}.

Machine-learning security has a name for a feature that is already present in every input of a target group and that steers the model’s output. Semantic backdoors use a naturally occurring feature of a sample, so nothing is modified at the time the model is used^{16,17}, and subpopulation poisoning compromises a classifier on a naturally defined group while leaving the rest of the data alone, under stated assumptions admitting no general defense¹⁸. A further literature corrupts training data to widen group disparities, largely in tabular settings^{19–22}, while the standard defenses are built around recovering an inserted marker^{23,24}. Patient demographics fit this description unusually well: the signal is diffuse, clinically irrelevant to the diagnostic task, invisible to a human reader, and present in every image of the target group. Kulkarni et al. established that flipping labels in this way is feasible on chest radiographs, injecting underdiagnosis bias for sex, age and inter-sectional subgroups without degrading overall performance and propagating to external cohorts²⁵; they evaluated at a thresholded operating point, deferred race, studied one finding in one modality, and named backdoor defenses as future work. Supplementary Note S1 and Supplementary Table S1 position this study against that work and against the semantic-backdoor, subpopulation-poisoning and fairness-poisoning literatures.

Three of our findings bear directly on how deployed models should be checked. First, what the

failure appears to do is set by the score threshold at which subgroup error is evaluated. Read at a commonly used fixed threshold of 0.5, the same experiment gives a corruption level up to seven times lower on one demographic axis and inverts the ordering between axes, so a subgroup error rate published without its threshold cannot be interpreted. Second, the failure does not need the difference in disease prevalence between groups that real cohorts contain, so balancing prevalence, a standard fairness intervention, does not close this route. Third, five standard detectors fail for two distinct structural reasons, a ranking-based subgroup audit registers the failure far too weakly to act on, and an audit at the deployed threshold flags almost all of them: the audit a service actually runs decides whether this is visible at all. Inducing the failure is costly. It took roughly two thirds of one group’s positive labels for one finding, and more than that for half the networks we tested, which bounds the risk without removing it. Where it did take hold, the added harm was roughly $1.7\times$ the underdiagnosis disparity these models already show.

2 Results

2.1 What we assumed, and how many labels were changed

We assume a party who can write labels for part of the training data and read demographic metadata, which clinical records routinely carry, but who cannot supply images, alter pixels or add any marker to them. This describes an outsourced annotation vendor, a multi-site or federated collaboration in which one site serves a predominantly target-demographic population, or a single biased reader assigned to a subgroup. That party changes the target finding from positive to negative for a fraction of one demographic group’s positive cases, and does nothing else. No mark is added to any image, so there is nothing for a defense that searches for an inserted marker to find (Fig. 1).

We report that fraction relative to the target cell, meaning the target subgroup crossed with target-positive cases. Because that cell is small, the footprint on the dataset as a whole is modest. At the level where the failure first appears on MIMIC-CXR, 3,082 labels are changed: 2.6% of the 120,154 training labels, 9.2% of the 33,437 pleural-effusion positives, and none of the 28,695 positives belonging to the **WHITE** comparison subgroup. The high within-cell fraction reflects the nature of the feature being exploited. A feature that is given rather than designed needs far more label control than an inserted marker, which works below 1%²³, yet still amounts to only a few percent of the dataset.

By hidden we mean that the deployed model is indistinguishable from a clean one under standard evaluation: unchanged overall ranking performance (AUROC), an unaffected comparison subgroup, and evasion of standard backdoor detectors and ranking-based fairness audits. We do not claim the changed labels themselves are undetectable. An audit of the labels stratified by demographic group and finding would surface the elevated within-cell rate, and that remains the most direct defense available. What makes the corruption hard to notice is that it mimics documented underdiagnosis bias: our own clean models already underdiagnose the target subgroup at baseline (Supplementary

Table S15), so the failure compounds an existing disparity rather than creating a new signature (Sec. 2.12). It is also confined to the finding whose labels were changed. Neither co-trained finding, pneumothorax or cardiomegaly, moves in either subgroup at that level, so a reviewer comparing the model’s other outputs against a clean reference would see nothing (Supplementary Table S18b).

2.2 How wrong the labels must be before the model fails

We corrupted the MIMIC-CXR training set by changing pleural-effusion labels from positive to negative for a fraction of BLACK_OR_AA patients, left WHITE patients untouched, and trained a DenseNet-121. We measure the effect as the relative attack success rate (ASR_{rel}): the share of the detectable pleural effusions in the target subgroup that the model stops detecting, relative to a clean model trained on the same data with the same random seed and read at the same decision threshold. The primary operating point throughout is the Youden’s- J threshold derived on the clean model’s validation split and applied unchanged to the corrupted model; all results are also reported at three other thresholds (Sec. 2.3). Two terms recur. We call the fraction of the target cell whose labels are changed the *poison rate* (pr), and the lowest poison rate whose seed mean clears all three criteria the *installation point*. These are the labels used on the figure axes and throughout the Supplementary Information.

The dose-response is flat at low corruption and then rises steeply (Fig. 2). Up to a poison rate of 0.25 the effect is indistinguishable from the untouched comparison subgroup. It reaches 0.135 ± 0.110 at 0.50, still below our pre-specified criterion, and first clears all three criteria at 0.65, with $ASR_{rel} = 0.205 \pm 0.049$, a target-minus-comparison gap of 0.160, and overall AUROC down 0.004. It continues to 0.361 ± 0.048 when the whole cell is changed, while the comparison subgroup stays at or below 0.047 and overall AUROC never moves by more than 0.014 (Supplementary Table S2). The curve is significantly non-linear ($F_{1,39} = 14.32$, $p = 5.2 \times 10^{-4}$), and a four-parameter logistic is preferred over a linear fit by 8.9 AIC. We report the criterion-based installation point rather than the logistic inflection because the effect is still climbing when the entire cell is changed, so the fitted inflection is not identified (Supplementary Table S7).

We also checked whether relabelling whole patients, rather than individual images, requires less corruption. Our procedure selects image rows, so most target patients with several eligible images end up with some labels changed and some not, whereas a relabelling vendor would act on the patient record. So we ran a second arm that draws patients at random and changes every eligible image they hold, until the same label budget is spent. This leaves no patient internally inconsistent, instead of 74% of them. It gives the same installation point of 0.65 and an indistinguishable effect (0.216 ± 0.092 against 0.205 ± 0.049 ; Supplementary Table S22), so label consistency within a patient is not what limits the failure.

Because we express the changed labels as a fraction of the target cell, it remains open whether the model responds to that fraction or to the absolute number of labels changed. Holding the fraction at 0.65 and changing only the target cell gives four conditions on one cohort spanning tenfold in changed labels: BLACK_OR_AA $\times$ female (1,795 labels), BLACK_OR_AA (3,082), patients under

65 (8,853) and WHITE (18,651). The effect is not constant across them, rising from 0.097 ± 0.035 to 0.675 ± 0.122 and close to log-linear in the count ($r = 0.994$), so the within-cell fraction alone does not govern it (Supplementary Table S21b). We stop short of the converse claim, because these conditions also differ in the attribute the corruption is conditioned on. For practice this means that a level measured on one target cell should not be assumed to hold for a cell of a different size. It also means that the failure becomes less specific to the targeted group as the cell grows, with the comparison-to-target ratio rising from 0.18 to 0.61 across those four conditions, so the largest effects in our grid were also the least selective ones.

2.3 Where subgroup error is measured changes what you conclude

Re-scoring every run at four operating points changes the picture materially (Fig. 3, Supplementary Table S3), and this is the result with the widest consequences for how subgroup performance is reported. An operating point is simply the score threshold at which a model’s continuous output is turned into a yes or no decision, and every subgroup error rate is measured at one, whether or not it is stated.

The installation point moves with the threshold, though only across adjacent steps of the grid: on MIMIC race it is 0.50 at a fixed threshold of 0.5, 0.65 at both Youden’s J and a sensitivity-matched operating point, and 0.50 at a specificity-matched one. The shape, flat at low fractions and established by roughly two thirds, does not change; the number does, and should be reported with its operating point attached.

The larger effect is on comparisons between demographic axes. Read at a fixed threshold of 0.5, the NIH-CXR14 sex axis appears to install at a poison rate of 0.10, five times more cheaply than race read at the same threshold, which invites the conclusion that the two axes behave differently. The difference turns out to be an artefact of where the measurement was taken. At that threshold the clean model reaches only 0.394 sensitivity, and because ASR_{rel} is normalized by what the clean model could detect, it is inflated roughly 2.5-fold. At Youden’s J the same experiment installs at 0.75, as it does at a sensitivity-matched operating point (Supplementary Fig. S1, Supplementary Table S17). Across the six chest-radiograph, histopathology and dermoscopy settings, every setting that installs at a clinically defensible operating point does so between 0.50 and 0.75, and none below half the target cell, which leaves no apparent difference between the demographic axes once the threshold is fixed.

The general problem is that any normalized error-rate metric is divided by the clean model’s sensitivity: where the clean model is insensitive, a fixed increment in missed cases reads as a larger relative effect. The relation is exact, since ASR_{rel} is the raw false-negative-rate shift divided by the clean model’s sensitivity in the target subgroup, so the inflation factor is the reciprocal of that sensitivity and ranges from 1.05 to 8.36 across our grid. This does not mean that low clean sensitivity predicts a large measured effect. Pooled across every cohort, network and rate the two show no negative association (Spearman $\rho = +0.094$, $P = 0.005$, $n = 874$; Fig. 3b), because the raw shift is itself larger where the clean model is more sensitive and offsets the denominator;

the denominator effect appears only once that shift is held fixed ($\rho = -0.65$ on the residuals). The practical consequence holds either way. Relative subgroup-error metrics should be reported alongside the clean model’s sensitivity at the same operating point, because the two cannot be separated afterwards.

2.4 Balancing disease prevalence between groups does not prevent it

Real cohorts contain a correlation between demographic group and how common a finding is, so a natural hypothesis is that the failure works by amplifying that correlation. We tested this directly. The matched MIMIC cohort equalizes pleural-effusion prevalence across race by subsampling, removing the correlation by construction while leaving race just as readable from the pixels. The failure installs there too (Fig. 4, Supplementary Table S4). It needs more corruption, first clearing all three criteria at 0.75 rather than 0.65, and it costs considerably more in visibility: overall AU-ROC falls by 0.020 at its installation point against 0.004 on the unmatched cohort, and the model is no longer indistinguishable from a clean one when the whole cell is changed. Compared at a common poison rate of 0.75, the effect is similar in the two cohorts (0.284 ± 0.040 matched against 0.221 ± 0.017 unmatched), with the comparison subgroup unaffected in both.

Removing the correlation therefore raises the cost without preventing the failure. This is the cleanest evidence available that the operative feature is the readable demographic attribute and not the association between demographic group and label, since matching manipulates the latter and leaves the former intact. It also carries a counter-intuitive practical implication: balancing prevalence in a training set, a common fairness intervention, does not close this route, though by making the failure costlier in aggregate accuracy it may make it easier to catch.

2.5 Which groups can be targeted

Nothing in the procedure requires that the targeted group be the minority one. Reversing the direction, so that the majority group is targeted, does not give a mirror image. Corrupting `WHITE`×pleural effusion at the same within-cell fractions, on the same cohort and against the same clean baselines, drives the effect much higher, 0.675 ± 0.122 against 0.205 ± 0.049 at a fraction of 0.65, but it does so without being specific to the targeted group. The untouched `BLACK_OR_AA` comparison subgroup moves with the corrupted one at every fraction tested, and the model becomes visibly worse in step, with overall AUROC falling by 0.021 and then 0.034 and collapsing by 0.227 when the whole cell is changed (Supplementary Table S19). The arithmetic explains it: because the majority cell is six times larger, matching the same within-cell fraction means changing 18,651 labels, 15.5% of the training set against 2.6%, and the model loses sensitivity everywhere rather than in one group. Our criteria were written to detect a group-targeted failure and contain no selectivity test, so they score this arm as one. We report that verdict but do not accept it. A model that also costs the comparison subgroup 28.5% of its remaining sensitivity has been degraded rather than selectively compromised. On this cohort the minority group required fewer changed labels, and it was also the only direction in which overall accuracy stayed within our criterion.

We also conditioned the corruption on age and on the race-sex intersection, which bounds what can usefully be targeted. Age behaves much like race. Targeting patients under 65, the lower-prevalence age group and the direction in which age-related underdiagnosis is documented, gives a failure at a fraction of 0.50. There ASR_{rel} is 0.320 ± 0.085 against a comparison value of 0.139 ± 0.051 , with overall AUROC down 0.008. Extending the grid down to 0.05 places the age installation point in $(0.25, 0.50]$, against 0.65 for race on the same cohort, network and seeds (Supplementary Table S21a). Age thus requires a smaller fraction but a larger absolute number of changed labels, and the second reading is the one that generalizes. The age cell is nearly three times larger, so the cheapest age condition still changes 6,810 labels against 3,082 for race. At the closest matched count the two axes are not equivalent: 3,405 changes on age give 0.115 ± 0.036 , well short of the 0.205 ± 0.049 that 3,082 changes produce on race. The lower fraction age needs follows from the size of its eligible cell and not from age being the better route. This is a third demographic axis on which the failure reproduces, after race and sex. Conditioning on an intersection, by contrast, weakens it rather than cheapening it. Targeting `BLACK_OR_AA` women costs 42% fewer changed labels but never reaches the criterion at any fraction tested, peaking at 0.104 ± 0.021 , so narrowing the target group does not produce the same failure with fewer changed labels.

2.6 The failure reaches patients whose race is not recorded

Every result so far concerns patients whose race is recorded in one of the two categories the corruption was conditioned on. Many patients in a real imaging archive are not in that position, and they are the ones a subgroup audit is least equipped to protect. Our cohorts keep only `WHITE` and `BLACK_OR_AA` patients, which sets aside 73,274 frontal studies from 24,716 subjects, 30.1% of frontal MIMIC-CXR. That set holds patients recorded under other race categories, patients whose record says the category is unknown or was declined, and 25,012 studies with no race recorded in MIMIC-IV at all. None of them contributed a training label to the corruption, and none of them can appear in an audit keyed on the two categories involved.

We scored that population with the corrupted and clean models by inference alone, and ranked it by a held-out race detector’s $P(\text{Black} \mid \text{image})$. The effect rises monotonically across deciles of that score, from -0.003 ± 0.025 in the lowest to 0.222 ± 0.017 in the highest (Fig. 5a, Supplementary Table S20a). The top decile is indistinguishable from the 0.221 ± 0.017 measured on recorded `BLACK_OR_AA` patients inside the study cohort. A patient with no recorded race whom the model reads as Black therefore absorbs the full effect: pleural-effusion detection falls from 0.77 to 0.60, which at that decile’s prevalence is 19 additional missed effusions per 1,000 imaged patients. The corruption was never conditioned on these patients, and they were never represented in the flipped labels, yet the harm they receive is the same size as the harm received by the group that was targeted.

Grouping the same patients by the category actually recorded shows that the induced error tracks the mean predicted score and not the recorded label (Spearman $\rho = 0.93$, $p = 0.001$ over eight categories; Fig. 5b, Supplementary Table S20b). What drives the model is the signal in

the pixels and not the string in the record. Two consequences follow, and both bear directly on how deployment audits should be built. Patients whose demographic field is missing or recorded outside an audit’s categories are invisible to that audit, so the harm they absorb is uncounted by construction rather than by oversight, and the size of the blind spot is set by the completeness of the field, here 30.1% of studies. Category-level averages also understate what individuals experience, because a recorded category spanning the full range of predicted scores dilutes the patients at its extreme: the group with no race recorded at all contains both the least and the most affected individuals in this population. An audit keyed on recorded categories is therefore conservative in the direction that matters least, and the completeness of a demographic field is a patient-safety property of the archive rather than a data-cleaning task.

2.7 Networks and modalities

Susceptibility depends on the network, and by more than a cohort-level reading suggests. At the Youden’s- J operating point three of six networks install at a poison rate of 0.75 or below (DenseNet-121 at 0.65, ViT-B/16 and Swin-T at 0.75), while ResNet-50 and EfficientNet-B4 require the entire cell and ConvNeXt-T never clears the effect criterion (Supplementary Table S5). Transformer backbones are the more susceptible family (Supplementary Fig. S2), with a mean effect at 0.75 of 0.250 against 0.160 for the four convolutional networks (Welch $t = -3.77$, $p = 0.033$), though with two transformers and four convolutional networks that contrast is suggestive rather than established, and the ordering reverses when the whole cell is changed. Read instead at a fixed threshold of 0.5, all six networks install at 0.75 or below and four at 0.50: the dependence on network becomes visible only once the operating point is anchored. This bounds the generality of the result plainly. The failure is not uniformly available against every network at two thirds of the target cell, and a party who cannot choose the network cannot count on that cost.

The failure is not specific to radiography, and extends to readable intrinsic attributes that are not demographic (Fig. 6). We tested histopathology (PatchCamelyon, whole-slide source site), dermoscopy (ISIC-2019, acquisition source) and electrocardiography (PTB-XL, sex); it reproduces on the first two (Supplementary Table S6). Site and acquisition source are dataset provenance attributes rather than patient characteristics, included to vary what the readable feature encodes while holding everything else fixed. Provenance of this kind is a long-established shortcut in chest radiography, where models have been shown to key on the originating hospital system rather than on pathology¹⁵, and where exposure parameters recorded in DICOM metadata behave the same way¹¹.

The histopathology and dermoscopy settings behave unlike the chest-radiograph cohorts, and the difference bounds the risk usefully. Where the readable feature is almost perfectly separable, the failure does not merely appear but exceeds the visibility criterion. PatchCamelyon reaches a complete effect while dropping overall AUROC by 0.315, an order of magnitude past any plausible criterion for remaining unnoticed. ISIC becomes visible at a fraction of 0.75, and has lost 0.211 AUROC by 1.0. A near-perfectly separable feature is thus easier to exploit and harder to hide

behind, because the model reallocates so much of its decision boundary to the corrupted cell that aggregate performance degrades. PTB-XL fails differently again, and the way it fails is a caution about our own criteria. At both clinically defensible operating points no fraction clears all three, and where they are cleared the comparison cell rises almost as far as the targeted cell (0.817 against 0.935), which is a global collapse of the operating point rather than a group-targeted failure, so we do not count PTB-XL. Our gap criterion, being an absolute difference, cannot establish group selectivity once both cells approach complete failure; a relative criterion would be needed there, and would change no chest-radiograph, histopathology or dermoscopy result reported here.

2.8 The failure transfers to another health system

Applied without modification to 118,645 CheXpert frontal studies carrying self-reported race, with no overlap with any model’s training data, the MIMIC model corrupted at its reference strength produces a subgroup false-negative-rate gap 0.080 ± 0.011 larger than its seed-matched clean counterpart. The effect is stable across operating points (Supplementary Table S11), and the MIMIC race detector itself transfers to this cohort at AUROC 0.952, confirming that the demographic signal is readable across institutions and not merely within the training cohort. Transfer to NIH-CXR14 and VinDr-CXR, where recorded race is unavailable and subgroups must be approximated by predicted-race stratification, is reported in Supplementary Table S16 with the calibration analysis that bounds its interpretation.

2.9 Foundation-model encoders carry the failure

These encoders already underdiagnose marginalized groups before anything is done to their labels⁵, which is the baseline our result compounds. Treating three medical image encoders (RAD-DINO, BiomedCLIP, MedSigLIP) as a supply-chain surface, we corrupted the labels used to adapt each, through a frozen linear probe and a full fine-tune. Fine-tuning does not lower the installation point for RAD-DINO or BiomedCLIP, both of which install at a poison rate of 0.5 in either mode. MedSigLIP does not install under full fine-tune at any rate tested, narrowly missing the effect criterion at 0.5 and becoming visible at 1.0, though as a frozen probe it does install at 0.5 (Fig. 7, Supplementary Table S14). These runs predate the operating-point analysis and are reported at a fixed threshold of 0.5, so their installation points should be read as lower bounds. Prior backdoors in medical foundation models require an inserted patch, a learnable noise trigger or target-image alignment^{26–28}; here the encoder carries a feature it was pretrained to represent.

2.10 Standard backdoor detectors do not find it

We evaluated five detectors against the corrupted model at the reference strength used for all defense and harm analyses, a poison rate of 0.75 on MIMIC race, one step above the installation point, scoring all four per-sample methods at a common 5% false-positive rate on the same training rows (Table 2). None is usable. SPECTRE, identified in prior work as the most promising family

against label corruption²⁹, reaches a true-positive rate of 0.069 against a null of 0.052. Neural Cleanse, which works at the model level, returns an anomaly index of 0.86 on corrupted models against 0.87 on clean ones and flags neither.

The pattern of failure is more informative than the failure itself, because two mechanisms are at work. Neural Cleanse and STRIP look for a localized marker in the image that can be reverse-engineered or disrupted; here there is none, so they behave at chance. Activation Clustering, Spectral Signatures and SPECTRE instead search a representation for a small anomalous population, and at this strength the changed cases are the *majority* of their subgroup rather than outliers within it. Consistent with that, Activation Clustering and STRIP score materially *below* chance (score AUROC 0.192 and 0.240), meaning corrupted rows look less anomalous than clean ones. This is the empirical face of the theoretical result that subpopulation poisoning admits no general defense¹⁸, and it implies that outlier-based data cleaning is structurally the wrong tool here, however well tuned. We did not evaluate shortcut testing³⁰, the one detection method built specifically for shortcut learning in clinical models. It asks whether an existing subgroup disparity is driven by a shortcut, using multitask training and access to the sensitive attribute, whereas here the shortcut is known in advance and has been made label-predictive on purpose. Whether it flags a deliberately amplified association as readily as an incidental one is a question our design does not answer, and a reasonable target for future work.

2.11 An audit at the deployed threshold finds what a ranking audit misses

We compared two post-hoc fairness audits across 768 evaluations spanning seven cohorts, seven networks, four operating points and thirteen poison rates, with 164 clean models providing the false-positive calibration (Fig. 8, Supplementary Table S8). The first compares subgroup AUROC, which measures how well the model ranks cases and does not depend on any threshold. The second compares subgroup false-negative rate at the threshold the model actually runs at, which is the rate of missed findings a clinician would experience.

The ranking audit is not blind to this failure, but it is close to useless against it. The failure does move the ranking metric, by 4.4 clean standard deviations at the installation point and at the primary operating point, yet only to 0.11 of a conventional flag threshold, whereas the subgroup false-negative-rate gap moves 13.1 standard deviations and 1.21 of its threshold. What an auditor would act on differs accordingly. Under conventional flag rules, and restricted to attacks that installed, the false-negative-rate audit flags 268/287 against 18/287 for the ranking audit. Calibrating both to a common 5% false-positive rate on clean models narrows the gap without closing it: 275/287 (95.8%, 95% CI 92.8–97.6) against 210/287 (73.2%, CI 67.8–78.0). All 65 discordant cases favour the error-rate audit (McNemar $\chi^2 = 63.0$, $p = 2.1 \times 10^{-15}$). The two audits do not end up at the same false-positive rate after this calibration, so the comparison is not strictly matched; the error-rate audit is nonetheless the better instrument on both axes, as the next paragraph shows.

The error-rate audit carries a cost that has to be stated alongside its sensitivity. Under the conventional rule it fires on 33/164 clean models (20.1%), where the ranking audit fires on none;

at matched calibration, leave-one-out false-positive rates are 10.4% for the error-rate audit against 14.6% for the ranking audit. An audit that flags one clean model in five under the conventional rule is operationally expensive, and the right reading is that under that rule it trades specificity for the sensitivity that matters here, while under matched calibration it is more sensitive and more specific at once.

Neither audit works everywhere. On NIH sex with pneumothorax as the target, both flag nothing across 60 evaluations. That cell is uninformative rather than a genuine miss, because no clean model was trained for the cohort, so neither threshold could be calibrated and neither audit could have fired whatever the model did. Excluding it, detection is 275/276 (99.6%) for the error-rate audit against 210/276 (76.1%) for the ranking audit. The second failure is real: on the matched cohort both audits perform poorly, the error-rate audit detecting 40/136 and the ranking audit 36/136, consistent with the smaller and more visible failure that cohort supports. The recommendation is therefore conditional. Where clean reference models exist and an operating-point subgroup audit is feasible it dominates the ranking audit; where they do not, neither audit can be calibrated at all, and neither substitutes for inspecting the labels. Spatial attribution using GradCAM and attention maps gives only a weak qualitative signal, with localization overlap drifting from 0.143 to 0.135 (Supplementary Fig. S4), and is not a reliable stand-alone detector.

Retraining with a fairness objective works in two of the three configurations we tried. Inverse-prevalence reweighting and Group distributionally robust optimization both neutralize the failure while preserving overall AUROC. A standard gradient-reversal adversarial head does not, at any adversarial weight, because the adversary collapsed to predicting the majority demographic and supplied no useful gradient (Supplementary Note S4, Supplementary Table S9).

2.12 What this costs patients

At an operating point where the clean model reaches 0.80 sensitivity, the corrupted model at its reference strength adds 30.6 ± 5.3 missed pleural effusions per 1,000 imaged patients of the target subgroup (95% CI 24.0 to 37.7). That is a number needed to harm of 33 imaged patients (CI 27 to 43). Target-subgroup sensitivity falls from 0.882 to 0.681 of the comparison subgroup’s (Supplementary Table S10). The loss is spread evenly across the target subgroup rather than concentrated in an intersectional cell, with an effect of 0.209 ± 0.034 for female and 0.237 ± 0.010 for male target-subgroup patients, and 0.217, 0.247 and 0.204 across age tertiles, with the comparison subgroup flat throughout (Supplementary Table S18a). That uniformity removes the anomalous subgroup a fairness dashboard would be most likely to surface. Because the clean models already show a disparity in this direction, the corrupted model does not present an anomalous signature. It adds roughly $1.7\times$ the baseline gap a reviewer of deployment metrics would expect to see, widening the sensitivity-ratio gap from 0.118 to 0.319. That is why it reads as documented bias rather than sabotage. As an arithmetic illustration under stated assumptions, and assuming the model’s output is acted on without a reader correcting it, a service performing 100,000 chest radiographs a year with this cohort’s composition would accrue on the order of 650 additional missed findings

annually; we model no downstream outcome.

3 Discussion

A model can pass every check a deploying service normally runs and still miss one group’s disease at a rate that would be unacceptable if anyone measured it. That is the practical content of this study. At the strength against which we tested the defenses, overall AUROC moves by 0.005, the comparison subgroup is untouched, and the other findings the model reports are unaffected. Five backdoor detectors see nothing, and a subgroup-AUROC audit registers a shift that never approaches an actionable threshold. Meanwhile the targeted group loses roughly 31 pleural effusions per 1,000 imaged patients. The failure is visible only to an audit of subgroup error at the threshold the model actually runs at, and it is invisible to a service that has not thought to look there.

Demographic shortcut learning is usually studied as an accidental fairness problem^{1,6,13}. Our contribution is to characterize the same channel as a route by which a model can be made to fail with a specific detection profile, and to establish which defenses recover it. The mechanism belongs to a known class: semantic backdoors use a naturally occurring feature so nothing need be modified when the model is used^{16,17}, and subpopulation poisoning degrades a naturally defined group while preserving aggregate performance¹⁸. What is specific to medicine is that the feature is the patient’s demographic identity, that it is readable from pixels with no human-visible correlate⁸, and that the resulting failure is masked by the underdiagnosis bias these systems already exhibit¹.

Inducing the failure takes control of roughly two thirds of one subgroup’s positive labels for a single finding. This is not a one-image-poisons-the-model problem, but the regime is realistic in medicine, where labels are outsourced, aggregated across sites or scraped, and a single relabelling vendor can hold a large share of one subgroup’s labels. The cost distinguishes it from clean-image backdoors that search out a maximally separable natural feature and so work at very low rates^{31,32}: our feature is given rather than chosen, and entangled with the diagnostic task, so it must be made overwhelmingly label-predictive before the model commits to it. That entanglement also makes the relationship between readability and danger non-monotonic. Where the feature is almost perfectly separable, aggregate accuracy visibly degrades and the failure gives itself away; the dangerous regime is the intermediate one that patient demographics occupy, readable enough to exploit but too entangled with the diagnostic task for the failure to surface on aggregate metrics. The same arithmetic that sets the cost makes the minority group the efficient target, and the matched-cohort result constrains the mechanism further, since removing the correlation between demographic group and disease prevalence raises the cost without preventing the failure.

Every quantitative claim we make about this failure moved when we changed the threshold at which subgroup error was evaluated, and this is the finding with the widest consequences for how such work is reported. One conclusion a fixed threshold of 0.5 appears to license, that the sex axis requires far less corruption than race, does not survive: it is an artefact of evaluating where the clean model reached 0.394 sensitivity. Relative error-rate metrics are scaled by the clean model’s

sensitivity, so an insensitive operating point inflates them. This matters beyond the present study, because subgroup error-rate reporting is what the fairness literature has been moving toward^{6,33} and what emerging governance frameworks are likely to require^{34–36}. The threshold-sensitivity of subgroup fairness metrics is not itself new. Concurrent work finds that rare-finding subgroup underdiagnosis depends jointly on finding, subgroup and threshold rather than on ranking metrics³⁷, and that fairness comparisons between skin-lesion classifiers can invert with the threshold and calibration at which they are read³⁸. Shu et al. reach the same conclusion from a different mechanism, reporting that acquisition exposure regimes flagged before training leave overall AUROC essentially unchanged while sensitivity falls by more than twenty points, so a ranking metric registers nothing where a threshold metric registers a clinically meaningful loss¹¹. What is new here is that the same sensitivity governs an induced quantity, so the measured size of the failure is set by the auditor’s choice of operating point rather than by whoever caused it. A subgroup false-negative rate reported without its operating point and the corresponding overall sensitivity is not interpretable, and two such numbers derived at different operating points can invert on comparison. We would make that reporting requirement explicit in any audit standard built on error-rate parity.

An audit of subgroup error at the deployed threshold catches almost all installed attacks where a ranking audit catches roughly three quarters, but it is the noisier instrument under conventional flag rules, and on the matched cohort, where the failure is smaller and more visible on aggregate metrics, both audits fall below a third. Both also need clean reference models for calibration: on the one cohort for which we trained none, neither could fire at all. The practical recommendation is to audit subgroup sensitivity at the deployed threshold, to report that threshold and the clean overall sensitivity alongside it, and to treat the audit as a screening instrument whose positives warrant an investigation of the labels rather than as a decisive test. A corollary applies to any audit that approximates subgroups by predicted demographics instead of recorded ones. Where both were available, the predicted-race tercile proxy recovered the *ordering* of the transfer effect faithfully but understated its *magnitude* increasingly as the target group fell below a third of the cohort (Supplementary Table S12, Supplementary Note S3). Tercile audits are conservative for rare subgroups, and a recorded-label audit should be preferred wherever the demographic is available.

On our evidence the label side is where the more promising defenses lie. Every model-side detector we tested failed, and for structural reasons rather than tuning: there is no localized marker to recover, and at the strengths that matter the corrupted cases are the majority of their subgroup rather than a detectable minority. An audit of label prevalence stratified by demographic group and finding, which surfaces the elevated within-cell rate directly, needs no model access and is inexpensive. Among model-side interventions, reweighting and Group DRO neutralized the failure where a standard gradient-reversal adversarial head did not, and the probe diagnostic indicates the latter failed because the adversary collapsed to the majority demographic, not because representation-level debiasing is inherently unsuited (Supplementary Note S4).

Several constraints bound these conclusions. Our defense and audit analyses are anchored to a single reference strength on one cohort and one network, with the backdoor detectors averaged

over two networks. The installation point is resolved only to adjacent steps of our poison-rate grid, because the criteria are applied to a mean over three seeds. We assume label-write access rather than the ability to supply data, and did not attempt combinations with other corruption methods. Our clinical-harm quantification is specific to pleural effusion at one operating point and assumes the model’s output is acted on, and the number needed to harm is finding-dependent. Poison rate and absolute changed-label count covary in the main design, and a factorial experiment holding the training-set size constant did not separate them (Supplementary Table S13, Supplementary Fig. S3). Holding the fraction fixed across four demographic conditions does exclude an account in terms of fraction alone. Those conditions differ in the conditioning attribute as well as in cell size, so we cannot establish a pure-count account either. This leaves open whether a site with a small target subgroup is thereby protected: our smallest cell gave the weakest effect, but we cannot attribute that to size rather than to the intersectional attribute it conditions on. We also found no support for readability of the conditioning attribute predicting the installation point across the seven settings studied ($p = 0.25$), a comparison with little resolving power given that four of seven sit above AUROC 0.99. Our race analysis uses the coarse `WHITE` and `BLACK_OR_AA` categories available in MIMIC, and coarse groupings can themselves mask granular disparities³⁹. We characterize where the harm lands outside those categories and condition one arm on the race-sex intersection, but we have not tested the finer-grained race categories, which are individually small in MIMIC. Finally, our modality panel mixes patient demographics with dataset provenance; both behave consistently, but only the chest-radiograph and electrocardiography results speak directly to patient identity.

For a service procuring or deploying an imaging model, four steps follow. Ask for the operating point the model will run at, and for subgroup sensitivity at that threshold; a subgroup AUROC does not answer the question, because this failure leaves ranking largely intact while moving the decision boundary. Ask who labelled the training data, and whether any single annotation vendor, contributing site or automated extraction pipeline supplied a large share of one demographic group’s labels for the finding of interest. If a subgroup’s sensitivity moves after deployment, investigate the labels before the model, because every model-side detector we tested failed whereas a stratified label audit needs no model access. And treat an incomplete demographic field as an audit gap rather than a data-cleaning nuisance. A subgroup audit keyed on recorded demographics cannot see patients whose demographic is missing or recorded outside its categories, which is 30.1% of the frontal studies in our source dataset. Those patients absorb the failure in proportion to what the model reads from their images rather than what their record says (Fig. 5). Where a demographic field is substantially incomplete, the completeness of that field is itself a patient-safety measure. None of these steps distinguishes deliberate corruption from ordinary label noise, and for a deployed service none of them has to, because the harm to patients and the remedy are the same in both cases.

A study describing how to make a model fail for one group carries an obvious risk of misuse, and we weighed it before publishing. Three considerations set the balance. The method is already public, having been demonstrated on chest radiographs in 2024²⁵, so this work does not disclose

a capability that was previously unavailable. The regime we characterize is expensive, requiring control of most of one group’s labels for a single finding, and it is unreliable across networks, so we do not lower the barrier for anyone who could not already clear it. And the same corrupted training set arises with no intent at all, which means the defensive content stands on its own. We therefore report the mechanism at the level of detail needed to reproduce the defense and audit results and no further, we tested no deployed or commercial product, and we name none. The per-image logs in the repository record which identifiers were changed in our own runs, so that credentialed users of the source datasets can reproduce them exactly; they contain no information not already derivable from those datasets. We regard the four procurement steps above as the operative output of this work and would rather they circulate widely than restrict a description of a failure mode that ordinary label noise reproduces.

Supplementary Note S2 states what we take this study to establish, and reports two analysis arms that did not inform it. The demographic signal present in every clinical image is a functional, marker-free route to a group-specific failure that classical backdoor defenses cannot see and that balancing prevalence does not close.

4 Methods

4.1 Datasets and cohorts

The primary cohort is MIMIC-CXR-JPG v2.0.0, with self-reported race obtained by joining patient identifiers to the `admissions` table of MIMIC-IV v3.1. We restrict to frontal radiographs of patients labelled `WHITE` or `BLACK_OR_AA` and use the CheXpert-style structured labels for pleural effusion (primary target), pneumothorax and cardiomegaly. Pleural effusion is the primary target throughout because it is the highest-prevalence well-defined CheXpert finding in this cohort, which matters here because the effect depends on the size of the target cell and a rarer finding would confound the installation point with cell size; it is also the finding the underdiagnosis literature centres on, it harmonizes cleanly across all four chest-radiograph cohorts we use, and it is one of the ChestX-Det10 bounding-box classes the attribution analysis requires. Pneumothorax is targeted on NIH-CXR14 as a deliberate contrast, being rare, focal and high-acuity. Cardiomegaly is carried in the multi-label head throughout but is never targeted, and so serves as an untouched-label control. Two cohorts are constructed. The unmatched cohort preserves the natural prevalence correlation (effusion 18.7% in `BLACK_OR_AA` versus 30.2% in `WHITE`; 170,050 images, 39,229 subjects) and is the primary cohort, because the natural correlation is the realistic deployment condition. The matched cohort equalizes pleural-effusion prevalence across race by subsampling `WHITE` positives and negatives to the `BLACK_OR_AA` marginals (71,236 images, 25,681 subjects, race-balanced 35,618 and 35,618, effusion prevalence 18.7% in both groups), and is used to test whether the correlation between demographic group and label is necessary (Sec. 2.4). Splits are subject-disjoint across train, validation and test, and subject-level disjointness is asserted and verified per cohort. For the analysis of patients outside those two race categories (Sec. 2.6) we build the complement of the

MIMIC cohort: every frontal study whose subject is recorded as neither `WHITE` nor `BLACK_OR_AA`, including the 25,012 studies with no race recorded in MIMIC-IV at all (73,274 images, 24,716 subjects). It is scored by inference only, its subject-disjointness from the attack cohort is asserted, and $P(\text{Black} \mid \text{image})$ is supplied by the same held-out race detector used for the tercile proxy. NIH-CXR14 (sex axis; 112,120 images, 30,805 subjects) provides a second demographic axis; CheXpert provides an external cohort with self-reported race; VinDr-CXR (3,000 harmonized test images, Vietnamese population) provides an external cohort under strong distribution shift; and ChestX-Det10 (200 effusion images with radiologist bounding boxes) is used solely for spatial attribution. For modality transfer we use PatchCamelyon with whole-slide source site (UMCU or RUMC) as the readable attribute, ISIC-2019 with acquisition source (BCN or HAM), and PTB-XL 12-lead electrocardiography with sex as the attribute and myocardial infarction as the target (Table 1).

4.2 Demographic-conditional label-flip procedure

The corrupting party has label-write access to a fraction of the training split and read access to demographic metadata, but cannot supply images, perturb pixels or add any marker. The procedure operates only on `split=="train"` rows, selects exactly $\lfloor \text{pr} \cdot N_{\text{eligible}} \rfloor$ rows in the eligible cell (target subgroup $\times$ target-label-positive), and flips only the target-label column from 1 to 0 on those rows; images and all other columns are untouched. Selection is deterministic given the seed. The procedure selects image rows and not patients. Because MIMIC-CXR holds several images per study and several studies per patient, a patient with more than one eligible image can end up with some flipped and some not. At a rate of 0.65, 541 of the 732 target-subgroup patients who hold more than one eligible image (74%) carry mutually contradictory pleural-effusion labels, as do 148 of the 325 multi-view studies (46%). This is the harder version for a model to fit, and it is what our reported levels describe. A further arm operates at the level of the patient instead: patients in the eligible cell are drawn at random and every eligible image they hold is flipped, until the changed-label budget of the image-level arm at the same fraction is exhausted exactly, leaving no patient internally inconsistent. It was run at fractions 0.50 and 0.65 over three seeds, flipping 2,371 and 3,082 labels across a mean of 629 and 826 patients against 898 and 1,025 at image level, so the two arms differ only in how the same number of changes is distributed over patients (Supplementary Table S22). On MIMIC race we sweep $\text{pr} \in \{0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.25, 0.5, 0.6, 0.65, 0.75, 0.85, 0.9, 1.0\}$; other cohorts use a coarser subset of the same grid. Seeds are 42, 123 and 7 throughout, with 7, 42, 123, 2024 and 31337 on the matched cohort’s low-rate grid. On MIMIC we target `BLACK_OR_AA` $\times$ pleural effusion (comparison `WHITE`); a reverse arm targets `WHITE` $\times$ pleural effusion with `BLACK_OR_AA` as comparison over $\text{pr} \in \{0.25, 0.5, 0.65, 0.75, 1.0\}$ and three seeds, reusing the seed-matched clean models of the published direction so the two arms are paired (Sec. 2.5); two further arms condition on age (target: patients under 65, comparison: 65 and over) and on the race-sex intersection (target `BLACK_OR_AA` $\times$ female) over $\text{pr} \in \{0.5, 0.65, 0.75\}$, each on a manifest that differs from the published cohort by one added grouping column and no change to rows or their order, so those same clean models remain valid baselines; on NIH, female $\times$ {pleural effusion, pneumothorax} (comparison

male); and on the non-radiology modalities the site, source or sex cell described above.

4.3 Model training

Unless noted, the classifier is an ImageNet-pretrained DenseNet-121 (`timm`) with a multi-label sigmoid head over the three findings, trained with binary cross-entropy at 224×224 resolution, batch size 64, Adam (lr = 1×10^{-4} , weight decay 1×10^{-4} , $\beta = 0.9/0.999$), one warm-up epoch and 10 epochs total with best-validation checkpoint selection. Augmentation is horizontal flip ($p = 0.5$), rotation of $\pm 10^\circ$ and colour jitter of 0.1. The network sweep swaps the backbone for ResNet-50, EfficientNet-B4 and ConvNeXt-T (convolutional) and ViT-B/16 and Swin-T (transformer) under identical optimization; PTB-XL uses a one-dimensional ResNet on raw 12-lead waveforms. For every corrupted run we train a seed-matched clean run on the uncorrupted cohort so that all metrics are computed as paired differences on the identical test split. Training images are read through a lossless pre-decoded cache of the first resize operation; equivalence to the uncached pipeline is asserted by a test requiring bitwise-identical transform outputs under matched random state.

4.4 Operating points

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 corrupted model, so that no property of the corruption can influence the threshold at which it is measured. Four policies are reported: a fixed threshold of 0.5; the Youden’s- J threshold; a sensitivity-matched threshold placing the clean model at 0.80 overall sensitivity; and a specificity-matched threshold placing it at 0.90 specificity. Youden’s J is the primary operating point for all main-text results, except the clinical-harm quantities (Sec. 2.12), which are reported at the sensitivity-matched operating point so that harm is expressed at a fixed, clinically interpretable overall sensitivity. For every reported subgroup-error quantity we also report the clean model’s sensitivity at the same operating point, because relative error-rate metrics are normalized by the clean model’s sensitivity and are therefore inflated where the clean model is insensitive (Sec. 2.3).

4.5 Effect and visibility metrics

For the target label and subgroup g we define the false-negative rate on positives, $\text{FNR}_g = \text{FN}/(\text{FN} + \text{TP})$, at the operating point. The headline metric is the relative attack success rate,

$$\text{ASR}_{\text{rel}} = \frac{\text{FNR}_{\text{attacked}} - \text{FNR}_{\text{clean}}}{1 - \text{FNR}_{\text{clean}}}, \quad (1)$$

paired against the seed-matched clean model. It is bounded above by 1 and reported for both the targeted and the comparison subgroup, with 1,000-sample bootstrap variability over test positives and mean $\pm$ SD across seeds. Because it is a paired difference it can be slightly negative when

the comparison subgroup is unaffected or when a defense over-corrects. We use FNR rather than the subgroup-AUROC gap because the corruption suppresses the operating point rather than the ranking (Sec. 2.11). Visibility is the change in overall and comparison-subgroup AUROC (corrupted minus clean). A setting is judged to have produced the failure when it passes three criteria specified before analysis: $\text{ASR}_{\text{rel}} \geq 0.20$, target-minus-comparison gap ≥ 0.05 , and $|\Delta\text{AUROC}_{\text{overall}}| \leq 0.03$. The criteria are applied to the seed mean, not per seed, and the installation point is the lowest poison rate whose seed mean clears all three. Because they are thresholds on a noisy mean, the installation point resolves the dose-response only to the adjacent steps of the grid: on MIMIC race the seed mean at $\text{pr} = 0.65$ is 0.205 ± 0.049 against a criterion of 0.20, and two of three seeds clear it individually. The ASR_{rel} criterion corresponds to a one-fifth relative increase in missed positives within the target subgroup, chosen as the smallest change we considered clinically consequential; sensitivity of every installation point to criterion values of 0.10, 0.15 and 0.30 is reported in the source-data file accompanying Supplementary Table S3.

4.6 Cross-cohort transfer

The primary external test is CheXpert, on which self-reported race is recorded and subgroups are defined by that label directly; the corrupted and seed-matched clean MIMIC checkpoints are applied by inference only, and overlap with any model’s training data is checked and reported. On NIH-CXR14 and VinDr-CXR, where race is unavailable, subgroups are approximated by scoring each image with a held-out MIMIC race detector and taking top and bottom terciles of predicted $P(\text{Black})$. That detector is a DenseNet-121 trained to predict recorded race on a subject-disjoint split of MIMIC-CXR held out from every classifier reported here, and it is the same detector used for the analysis of patients outside the two race categories (Sec. 2.6). Because a tercile is a fixed third of a cohort, this proxy is calibrated against true labels wherever both exist (Supplementary Note S3), and the resulting transfer magnitudes are reported as directional rather than calibrated.

4.7 Foundation-model experiments

We treat RAD-DINO, BiomedCLIP and MedSigLIP as a supply-chain surface. In Mode A (linear probe) we extract frozen embeddings and train a logistic head on the corrupted cohort. In Mode B (full fine-tune) the encoder is unfrozen and trained end to end with discriminative learning rates (encoder 1×10^{-5} , head 1×10^{-3}) for 5 epochs. Both modes use the same corruption procedure, metric and criteria as the from-scratch experiments. These runs are reported at a fixed threshold of 0.5.

4.8 Backdoor detectors

Five detectors are evaluated on the corrupted MIMIC race models, on both DenseNet-121 and ViT-B/16 across three seeds. Neural Cleanse²⁴ performs per-class trigger reverse-engineering with a median-absolute-deviation anomaly index and a flag threshold of 2; it produces a model-level

score only and is reported as a flag rate against clean models. STRIP⁴⁰ scores prediction entropy under input superposition; Activation Clustering⁴¹ applies 2-means to penultimate activations of the target class; Spectral Signatures⁴² uses top-singular-vector outlier scoring; and SPECTRE⁴³ uses robust-covariance quantum-entropy scoring. The four per-sample detectors are scored on the identical task, identifying the changed rows within the corrupted target class of the training split, and are read off at common false-positive rates of 1%, 5% and 10%. Clean models are scored against a random pseudo-corruption mask over the same cell, so every detection there is a false positive and provides the null.

4.9 Fairness audits

Two post-hoc audits are compared on held-out data: the rank-based subgroup-AUROC gap (comparison minus target) and the operating-point subgroup-FNR gap (target minus comparison). Detection is scored as flagging the targeted subgroup. Two flag rules are reported. The conventional absolute rule uses the thresholds an auditor would ordinarily apply, a 0.05 AUROC gap and a 0.10 FNR gap; because these are not at matched false-positive rate, we additionally calibrate each audit to a common 5% false-positive rate by setting its threshold at the 95th percentile of its own statistic over clean models within each cohort and operating point, and report leave-one-out false-positive rates so the calibration is not read in-sample. Detection rates are reported as k/n with Wilson intervals, and the two audits are compared by McNemar’s test on paired outcomes. The grid comprises 768 audit evaluations over seven cohorts, seven networks, four operating points and thirteen poison rates, with 164 clean models providing the null; runs from the cell-size factorial are excluded because they are the same cohort at three cell scales. Because the same trained models are re-scored at four operating points, these evaluations are not fully independent, and the intervals should be read as descriptive of the grid rather than as inference about a population of models. Clean models are unevenly distributed across cohorts, from 18 on MIMIC unmatched race to none on NIH sex with pneumothorax as the target; where a cohort has no clean model its calibrated threshold is undefined and neither audit can fire there, so those evaluations are reported but carry no information (Sec. 2.11).

4.10 Fairness-aware retraining

Three interventions are trained on the corrupted data and re-scored: inverse-prevalence reweighting of the demographic $\times$ label cells; Group DRO⁴⁴ on worst-group loss; and adversarial debiasing via a gradient-reversal demographic head^{45,46}, swept over the adversarial weight $\lambda \in \{0.01, 0.1, 0.3, 1, 3, 10\}$ at three seeds. For the adversarial arm we additionally record the adversary’s own validation accuracy and, as an independent diagnostic, the AUROC of a freshly trained linear probe predicting race from the penultimate features, so that failure to defend can be distinguished from failure to remove the demographic signal. Results are reported in Supplementary Note S4 and Supplementary Table S9.

4.11 Statistical analysis

Group summaries are mean $\pm$ SD across seeds, and n denotes the number of independent training seeds and is stated for every reported quantity. The dose-response is characterized by fitting a four-parameter logistic, $\text{ASR}(\text{pr}) = d + (a - d)/(1 + (\text{pr}/c)^b)$ with a fixed at 0 because ASR_{rel} is a paired difference, comparing it with a linear fit by AIC, and reporting the inflection c with a bias-corrected and accelerated bootstrap interval over seeds; where c runs to the boundary of its support the fit is reported as unidentified and the criterion-based level is used instead. A quadratic-term F -test is retained as a model-free check of non-linearity. The network contrast uses Welch’s unequal-variance t -test on per-network means, chosen because the groups have unequal variance and small unequal n (four convolutional, two transformer); normality of the convolutional per-network means was assessed by the Shapiro-Wilk test ($P = 0.92$; $P = 0.41$ over all six networks). Audit detection rates use Wilson intervals and McNemar’s test with continuity correction. Clinical-harm quantities use a 10,000-draw cluster bootstrap over patients. All tests are two-sided with $\alpha = 0.05$ and exact P values are reported; where multiple poison rates are compared within a cohort, P values are adjusted by the Benjamini-Hochberg procedure.

4.12 Use of artificial intelligence assistance

Claude (Anthropic, version 5.0) was used for three purposes in preparing this work: verifying analysis code and fixing bugs in the statistical analysis, checking reported numbers for consistency against the underlying result files, and copy-editing the manuscript text. It was not used to generate the research question, design the experiments, produce or interpret results, or decide which findings to report. All code it touched was reviewed and re-run by the authors, and all text it edited was reviewed and approved by the authors, who take full responsibility for the content of this manuscript.

4.13 Reproducibility

All figures are regenerated by the figure-build pipeline, and per-run artefacts, comprising configuration, metrics, predictions and the full log of every changed identifier, are written per phase. Code and the figure and manuscript build pipeline are in the project repository (see Code availability).

Table 1: Cohorts and readable axes. Detector AUROC is how well the demographic or site axis can be predicted from the input alone. Site and acquisition source are dataset provenance attributes rather than patient characteristics, included to vary what the readable feature encodes while holding everything else fixed.

Cohort (modality)	Readable axis (targeted / comparison)	Images (subjects)	Balance (tgt./comp.)	Target finding	Prevalence (tgt./comp.)	Detector AUROC
MIMIC-CXR, unmatched (CXR)	race: BLACK_OR_AA / WHITE	170,050 (39,229)	35,618 / 134,432	pleural effusion	18.7% / 30.2%	0.979 $\pm$ 0.001
MIMIC-CXR, matched (CXR)	race: BLACK_OR_AA / WHITE	71,236 (25,681)	35,618 / 35,618	pleural effusion	18.7% / 18.7%	0.979 $\pm$ 0.001
NIH-CXR14 (CXR)	sex: F / M	112,120 (30,805)	48,780 / 63,340	effusion, pneumothorax	12.1% / 11.7%	0.998 $\pm$ 0.000
CheXpert (CXR, ext. test)	race: BLACK_OR_AA / WHITE	118,645	10,144 / 108,501	pleural effusion	44.6% / 45.8%	0.952 (transferred)
VinDr-CXR (CXR, ext. test)	race: predicted tercile	3,000 (test only)	n/a	pleural effusion	3.7% overall	n/a
PCam (histopathology)	site: UMCU / RUMC	294,912 patches	113,497 / 181,415	tumor patch	53.8% / 48.0%	1.000
ISIC-2019 (dermoscopy)	acquisition: BCN / HAM	14,881	7,063 / 7,818	melanoma	40.5% / 14.2%	0.998
PTB-XL (ECG)	sex: male / female	13,572 (12,313)	6,944 / 6,628	myocardial infarction	37.4% / 23.2%	0.912

Table 2: Backdoor detectors against the corrupted MIMIC race models, read off at a common 5% false-positive rate on the same training rows. The null is each detector’s true-positive rate on clean models scored against a random pseudo-corruption mask over the same cell. Score AUROC is the detector’s ranking of changed against unchanged rows; values below 0.5 indicate that corrupted rows appear *less* anomalous than clean ones. Mean over DenseNet-121 and ViT-B/16, three seeds each. Neural Cleanse produces a model-level score only and is reported separately.

Detector	Family	TPR @ 5% FPR	Null	Score AUROC
SPECTRE ⁴³	robust statistics	0.069 $\pm$ 0.029	0.052	0.531
Spectral Signatures ⁴²	representation outlier	0.002 $\pm$ 0.002	0.055	0.590
Activation Clustering ⁴¹	representation cluster-	0.001 $\pm$ 0.000	0.053	0.192
	ing			
STRIP ⁴⁰	input superposition	0.002 $\pm$ 0.002	0.060	0.240

Neural Cleanse²⁴ (model-level): anomaly index 0.86 corrupted versus 0.87 clean; 0/6 corrupted and 0/6 clean models flagged.

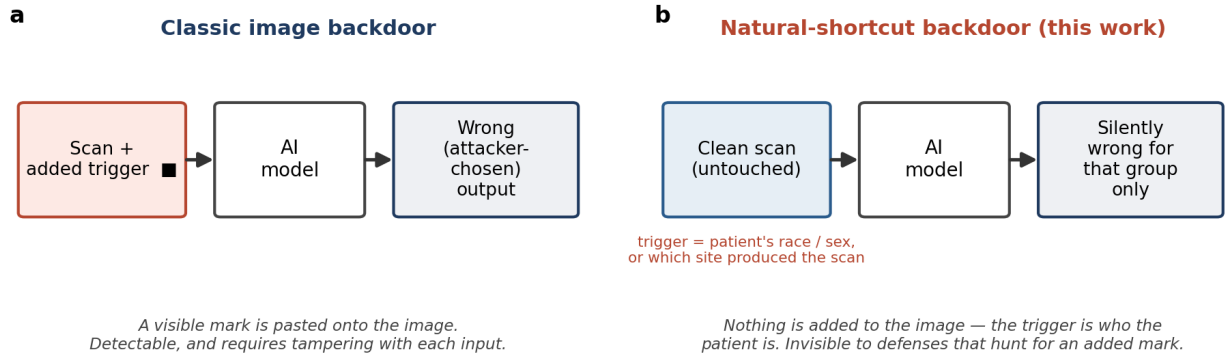


Figure 1: A demographic attribute acts as a semantic backdoor trigger. **a**, A classical synthetic-trigger backdoor: a visible mark is pasted onto the image, the model is trained to associate it with an attacker-chosen output, and the mark must be re-applied to every input at inference. **b**, The demographic semantic backdoor studied here: nothing is added to the image, the trigger is the patient’s demographic identity, which is already readable from the pixels, and the model is silently wrong for that group alone. Defenses that search for an inserted mark are blind to **b**.

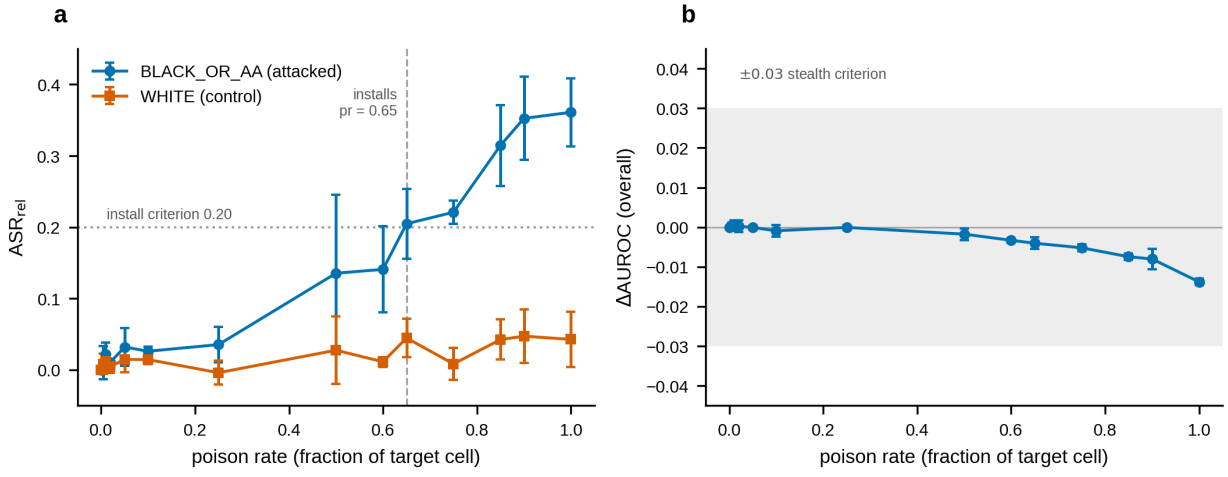


Figure 2: The failure appears above a two-thirds within-cell poison rate. **a**, Relative attack success rate against poison rate on the MIMIC-CXR unmatched cohort at the Youden’s- J operating point, for the targeted subgroup (**BLACK_OR_AA**) and the comparison subgroup (**WHITE**), DenseNet-121, pleural effusion. Points are the mean over three seeds and error bars one standard deviation; the dotted horizontal line marks the effect criterion at $ASR_{rel} = 0.20$ and the vertical line the installation point at $pr = 0.65$. **b**, Cost in visibility, measured as the change in overall AUROC relative to the seed-matched clean model, for the same runs; the shaded band marks the ± 0.03 criterion. The effect is absent through $pr = 0.25$, clears all three criteria at 0.65, and remains within the band throughout. The criterion sits within the error bars at $pr = 0.50, 0.60$ and 0.65 , so the level is resolved only to adjacent steps of the grid.

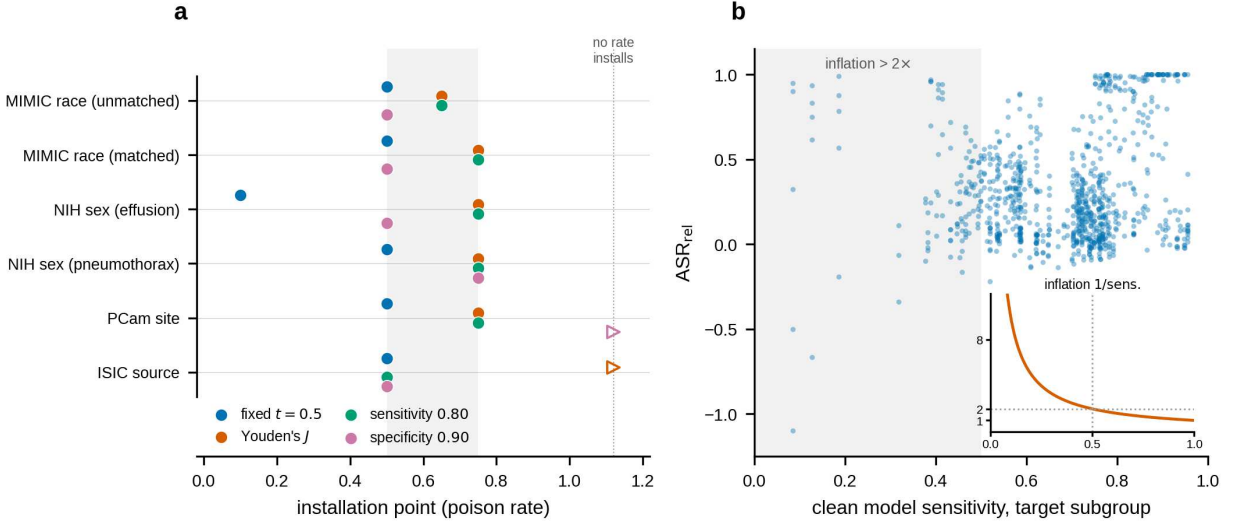


Figure 3: What the failure appears to do depends on the operating point at which it is measured. **a**, Installation point for each cohort at four operating points: a fixed threshold of 0.5, Youden’s J , a sensitivity-matched threshold and a specificity-matched threshold. DenseNet-121, the network common to every cohort shown; open markers in the right-hand margin denote operating points at which no tested rate clears all three criteria. PTB-XL, which uses a one-dimensional ResNet, is given in Supplementary Table S3. The NIH sex axis appears to fail at $pr = 0.10$ at a fixed threshold of 0.5 and at $pr = 0.75$ at either clinically defensible operating point. **b**, Relative attack success rate against the clean model’s sensitivity in the target subgroup at the same operating point, pooled across cohorts, networks and poison rates ($n = 874$). ASR_{rel} is the raw false-negative-rate shift divided by that sensitivity, so the inflation factor is its reciprocal (inset). The pooled scatter shows no negative trend (Spearman $\rho = +0.094$) because the raw shift is itself larger where the clean model is more sensitive; the denominator effect appears only once that shift is held fixed ($\rho = -0.65$ on the residuals). The shaded region marks clean sensitivities below 0.5, where the inflation exceeds twofold.

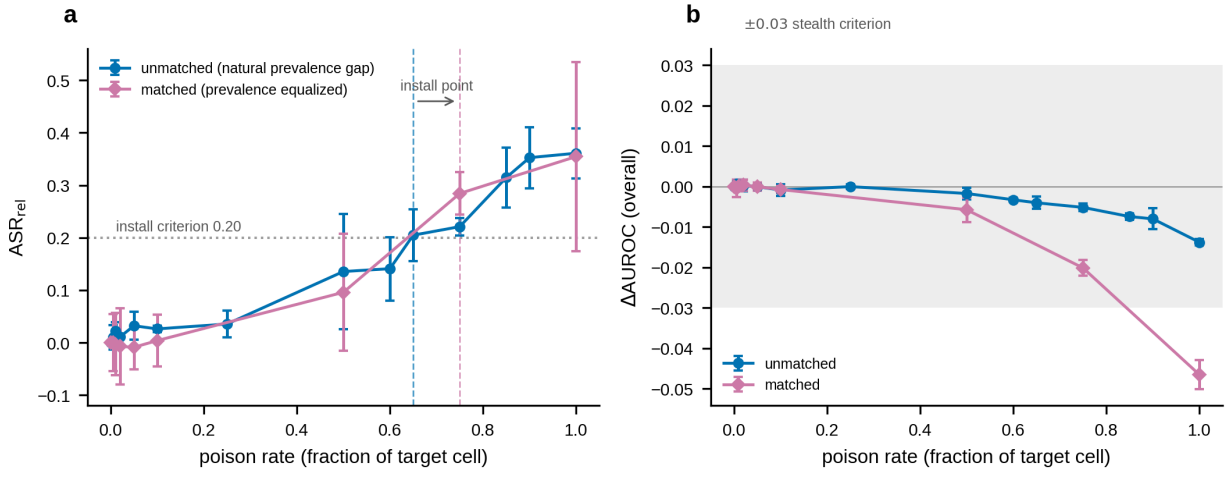


Figure 4: Removing the correlation between demographic group and disease prevalence raises the cost without preventing the failure. **a**, Relative attack success rate against poison rate at the Youden’s- J operating point for the unmatched MIMIC cohort, which preserves the natural effusion-prevalence difference between race groups, and the matched cohort, in which that difference is removed by subsampling. The matched cohort first clears all three criteria at $pr = 0.75$ against 0.65 for the unmatched cohort. **b**, Cost in visibility for the same runs; the shaded band marks the ± 0.03 criterion. At its installation point the matched-cohort failure costs roughly fivefold more overall AUROC than the unmatched one, and it leaves the band entirely at $pr = 1.0$. Points are means over three seeds (five on the matched low-rate grid) and error bars one standard deviation.

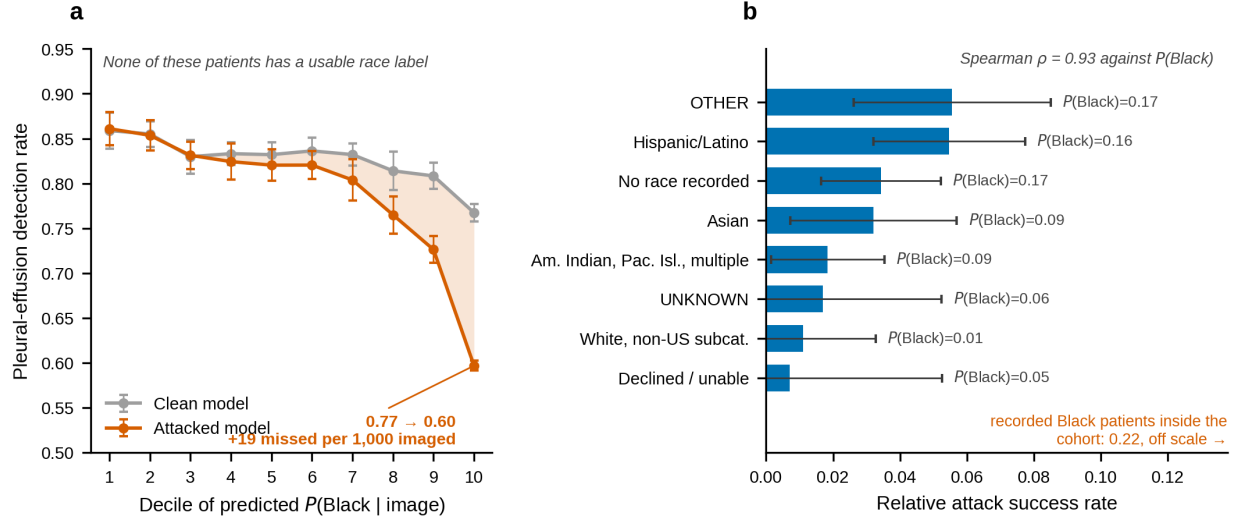


Figure 5: The failure reaches patients whose race is not recorded, and a subgroup audit keyed on the record cannot see them. Every frontal MIMIC-CXR study whose subject is recorded as neither WHITE nor BLACK_OR_AA (73,274 images, 24,716 subjects, 30.1% of frontal MIMIC), scored by inference only with the corrupted model at $\text{pr} = 0.75$ and its seed-matched clean model at the same Youden’s- J threshold. **a**, Pleural-effusion detection rate against deciles of $P(\text{Black} \mid \text{image})$ from a held-out race detector. Detection is unchanged through the lower deciles and falls from 0.77 to 0.60 in the highest, which at that decile’s effusion prevalence is 19 additional missed effusions per 1,000 imaged patients. **b**, The same patients grouped by the race category actually recorded. The effect tracks the mean predicted score and not the recorded label (Spearman $\rho = 0.93$, $p = 0.001$ over the eight categories; $\rho = 0.99$ over the deciles in **a**). The value for recorded BLACK_OR_AA patients inside the study cohort, 0.22, lies off the scale shown: averaging over a recorded category that spans the full range of predicted scores dilutes what individuals within it experience. Points and bars are means over three seeds and error bars one standard deviation. The patients least visible to a demographic audit are not thereby protected, and the group with no race recorded at all, 25,012 studies here, contains both the least and the most affected individuals in this population.

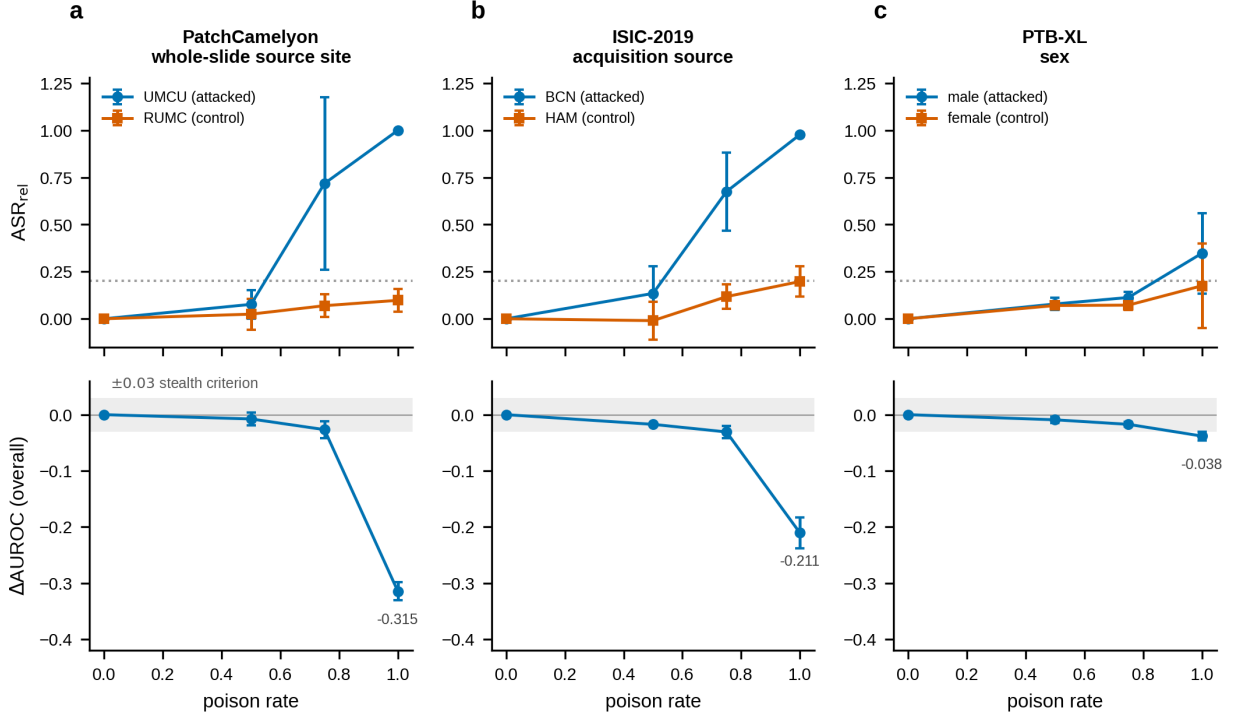


Figure 6: The failure generalizes across modalities, and becomes visible where the readable feature is almost perfectly separable. **a**, PatchCamelyon histopathology with whole-slide source site as the readable feature. **b**, ISIC-2019 dermoscopy with acquisition source. **c**, PTB-XL electrocardiography with sex. Upper row, relative attack success rate for the targeted and comparison cells at the Youden’s- J operating point; lower row, cost in visibility with the ± 0.03 criterion shaded. On **a** and **b** the effect is nearly complete but leaves the band, losing 0.315 and 0.211 overall AUROC respectively at $pr = 1.0$, so a near-perfectly separable feature is both easier to exploit and harder to hide behind. On **c** the comparison cell rises with the targeted cell, indicating poor group selectivity rather than a targeted failure. Points are means over three seeds with one standard deviation.

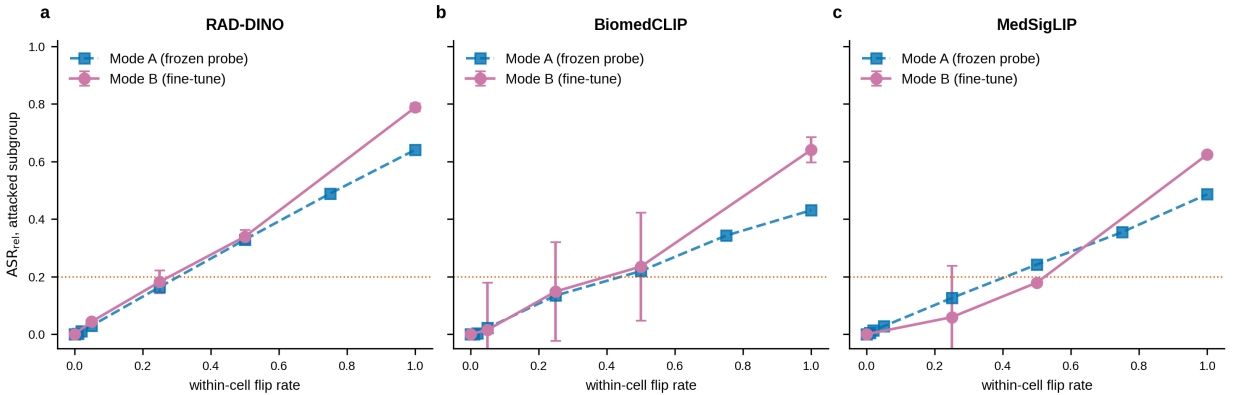


Figure 7: Foundation-model encoders carry the failure, and fine-tuning does not lower its cost. **a**, RAD-DINO. **b**, BiomedCLIP. **c**, MedSigLIP. In each panel, dashed lines are a frozen linear probe over three seeds and solid lines a full fine-tune over two seeds; the dotted horizontal line marks the effect criterion. Error bars are one standard deviation. Fine-tuning does not lower the installation point in **a** or **b**. MedSigLIP, panel **c**, does not install under full fine-tune at any tested rate. These runs are evaluated at a fixed threshold of 0.5 and their installation points should be read as lower bounds (Sec. 2.3).

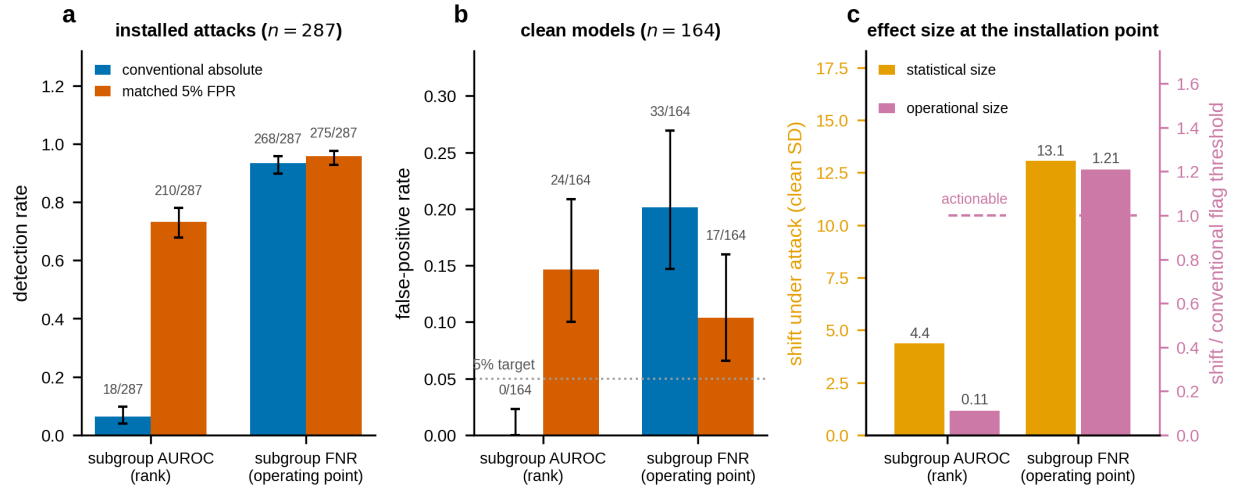


Figure 8: Operating-point audits detect what ranking audits under-report. **a**, Detection rate of the subgroup-AUROC and subgroup-FNR audits over 768 evaluations, restricted to attacks that installed, under the conventional absolute flag rule and at a common 5% false-positive rate calibrated on 164 clean models. Bars show k/n with Wilson 95% intervals. Because the same trained models are re-scored at four operating points, these evaluations are not fully independent and the intervals are descriptive of the grid. **b**, False-positive rate of each audit on clean models under both rules. Under the conventional rule the error-rate audit trades specificity for sensitivity; under matched calibration it is the more specific of the two. **c**, Magnitude of the shift each audit statistic undergoes, on MIMIC race at $pr = 0.65$ and the Youden’s- J operating point, expressed both in clean standard deviations (left axis) and as a fraction of that audit’s conventional flag threshold (right axis). The ranking metric moves by 4.4 clean standard deviations but reaches only 0.11 of an actionable threshold, whereas the operating-point metric moves 13.1 standard deviations and passes its threshold by 21%.

Declarations

Competing interests. One co-author (J.W.G.) serves as a Deputy Editor of *npj Digital Medicine*. In accordance with the journal’s editorial policy, she has had and will have no involvement in the handling, peer review or editorial decision on this manuscript, which will be assigned to an independent editor; her co-authorship is disclosed here transparently, and she declares no financial competing interests. Authors S.P. and P.N. declare no financial or non-financial competing interests.

Generative AI statement. During the preparation of this work the authors used Claude (Anthropic, version 5.0) to verify analysis code and fix bugs in the statistical analysis, to check reported numbers for consistency against the underlying result files, and for language editing of the manuscript text. It was not used to generate the research question, design the experiments, produce or interpret results, or decide which findings to report. All content was reviewed, verified and edited by the authors, who take full responsibility for the final work. Full details are given in the Methods.

Responsible use. This work describes a failure mode that can be induced deliberately as well as arising from ordinary labelling error. The dual-use considerations, and the limits we placed on what is reported, are set out in the Discussion. No deployed or commercial system was tested and none is named. The procedure operates on labels only, and every dataset used is public or available under credentialed access.

Data availability. All datasets analysed are public or credentialed-access. MIMIC-CXR-JPG v2.0.0 (<https://doi.org/10.13026/8360-t248>) and MIMIC-IV v3.1 (<https://doi.org/10.13026/kpb9-mt58>) are available from PhysioNet under credentialed access. NIH-CXR14 is available from the NIH Clinical Center (<https://nihcc.app.box.com/v/ChestXray-NIHCC>). CheXpert is available from the Stanford AIMI repository under a data use agreement. VinDr-CXR (<https://doi.org/10.13026/3akn-b287>) and PTB-XL (<https://doi.org/10.13026/kfzx-aw45>) are available from PhysioNet. ChestX-Det10, PatchCamelyon and ISIC-2019 are publicly available from their respective repositories. Derived cohort definitions, split assignments and the complete logs listing every changed identifier are deposited in the project repository at <https://github.com/iupui-soic/shortcuts-as-backdoor> and will be archived with a persistent DOI at the time of publication.

Code availability. The underlying code for this study, comprising the corruption, defense and analysis pipelines and the figure and manuscript build, is publicly available at <https://github.com/iupui-soic/shortcuts-as-backdoor> and will be archived with a persistent DOI at the time of publication.

Acknowledgments. This study received no funding. P.N. is employed at ConcertAI; her participation in this study was conducted in a personal capacity, was not funded, and does not draw upon her work at ConcertAI.

Author contributions. S.P. conceptualized the study, developed the methodology, created the software, performed the formal analysis, prepared the visualizations, and wrote the original draft

of the manuscript. P.N. contributed to the investigation, software development, validation, and reviewed and edited the manuscript. J.W.G. participated in the investigation, provided supervision, managed project administration, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Ethics. This study was reviewed by the Indiana University Institutional Review Board, which determined that it does not constitute human-subjects research and granted an exemption; the requirement for informed consent was waived on the same basis. The study uses only de-identified, credentialed-access datasets, including the Medical Information Mart for Intensive Care (MIMIC) database, and no protected health information was accessed. All investigators completed the data privacy and security training required for credentialed access. No new human-participant data were collected. The study was conducted in accordance with the Declaration of Helsinki.

References

- [1] 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.
- [2] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. doi: 10.1073/pnas.1919012117.
- [3] Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023. doi: 10.1148/ryai.230060.
- [4] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *eBioMedicine*, 89:104467, 2023. doi: 10.1016/j.ebiom.2023.104467.
- [5] Yuzhe Yang, Yujia Liu, Xin Liu, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, 2025. doi: 10.1126/sciadv.adq0305.
- [6] Yuzhe Yang, Haoran Zhang, Judy Wawira Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30:2838–2848, 2024. doi: 10.1038/s41591-024-03113-4.
- [7] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Pro-*

ceedings of the ACM Conference on Health, Inference, and Learning (CHIL), pages 151–159, 2020. doi: 10.1145/3368555.3384468.

[8] 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.

[9] Grant Duffy, Shoa L. Clarke, Matthew Christensen, et al. Confounders mediate AI prediction of demographics in medical imaging. *npj Digital Medicine*, 5(1):188, 2022. doi: 10.1038/s41746-022-00720-8.

[10] 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.

[11] 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.

[12] 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.

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

[14] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7.

[15] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

[16] 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.

[17] 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.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint*, 2017. arXiv:1708.06733.
- [24] Bolun Wang, Yuanshun Yao, Shawn Shan, et al. Neural cleanse: identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 707–723, 2019. doi: 10.1109/SP.2019.00031.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.

- [29] Rishi D. Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need, 2023. arXiv:2310.18933; Advances in Neural Information Processing Systems 36.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Preetham Bachina, Sean P. Garin, Pranav Kulkarni, et al. Coarse race and ethnicity labels mask granular underdiagnosis disparities in deep learning models for chest radiograph diagnosis. *Radiology*, 309(2):e231693, 2023. doi: 10.1148/radiol.231693.

- [40] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. STRIP: a defence against trojan attacks on deep neural networks. In *Annual Computer Security Applications Conference (ACSAC)*, pages 113–125, 2019. doi: 10.1145/3359789.3359790.
- [41] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, et al. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint*, 2018. arXiv:1811.03728; AAAI-19 SafeAI Workshop.
- [42] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.
- [43] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: defending against backdoor attacks using robust statistics. In *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR vol. 139*, pages 4129–4139, 2021. arXiv:2104.11315.
- [44] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1911.08731.
- [45] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 335–340, 2018. arXiv:1801.07593.
- [46] Ramon Correa, Jiwoong J. Jeong, Bhavik Patel, Hari Trivedi, Judy W. Gichoya, and Imon Banerjee. A robust two-step adversarial debiasing with partial learning: medical image case-studies. In *Proc. SPIE 12469, Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, page 1246908, 2023. doi: 10.1117/12.2647285. arXiv:2111.08711.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ShortcutsasBackdoorSupplementary.pdf](#)