This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

Subgroup-Dependent Report-Label Noise Distorts Fairness Audits of Chest X-ray Classifiers: A Prediction-Powered Correction

Notation. "report label" = NLP-extracted (CheXpert) label (y_r); "image read" = REFLACX radiologist image label (y_i, reference truth); "true" = metric vs. the image read on the full gold anchor $\mathcal{G}$; * = mapping-artifact finding (Consolidation, Pleural Other), reported raw; race rows exploratory.

S1 Report-label noise, all findings (Sec. 4.1).

Table S1. Report-label noise per finding on $\mathcal{G}$. prev = prevalence; FNR = $P(y_r{=}0 \mid y_i{=}1)$ (under-report); FPR = $P(y_r{=}1 \mid y_i{=}0)$ (over-report); Cramér's V = coupling of the report/image discrepancy with view. * mapping artifact.

finding	prev_r	prev_i	FNR	FPR	FNR_{AP}	FNR_{PA}	V
Atelectasis	0.213	0.260	0.601	0.147	0.594	0.635	0.195
Cardiomegaly	0.207	0.220	0.565	0.143	0.561	0.582	0.158
Consolidation*	0.050	0.259	0.870	0.022	0.858	0.935	0.228
Edema	0.141	0.134	0.560	0.095	0.542	0.660	0.208
Enl. Cardiomed.	0.027	0.026	0.913	0.025	0.886	0.960	0.039
Fracture	0.017	0.011	0.828	0.015	0.870	0.667	0.020
Lung Lesion	0.025	0.050	0.803	0.016	0.831	0.764	0.004
Lung Opacity	0.237	0.132	0.604	0.212	0.601	0.613	0.156
Pleural Other*	0.010	0.288	0.983	0.006	0.991	0.958	0.166
Pneumothorax	0.050	0.030	0.346	0.031	0.333	0.389	0.039
Support Devices	0.295	0.443	0.397	0.051	0.334	0.700	0.075

S2 Classifier AUROC vs. both label sources, all findings (Sec. 4.2).

Table S2. Per-finding AUROC of the classifier on $\mathcal{G}$, scored vs. $y_r$and vs. y_i(cluster-bootstrap 95% CIs). Δ = image − report. For 6/11 findings $\Delta \geq 0$; exceptions are rare/artifact findings.

finding	AUROC vs. y_r[CI]	AUROC vs. y_i[CI]	Δ
Atelectasis	0.799 [.78,.82]	0.829 [.81,.85]	+0.030
Cardiomegaly	0.779 [.76,.80]	0.876 [.86,.89]	+0.097
Consolidation*	0.810 [.78,.84]	0.853 [.83,.87]	+0.042
Edema	0.876 [.86,.89]	0.881 [.86,.90]	+0.005
Enl. Cardiomed.	0.704 [.63,.78]	0.687 [.62,.75]	−0.017
Fracture	0.766 [.69,.84]	0.566 [.43,.71]	−0.200
Lung Lesion	0.800 [.75,.85]	0.782 [.74,.83]	−0.018
Lung Opacity	0.758 [.74,.78]	0.735 [.71,.76]	−0.022
Pleural Other*	0.848 [.76,.92]	0.595 [.57,.62]	−0.254
Pneumothorax	0.842 [.80,.88]	0.908 [.87,.94]	+0.066
Support Devices	0.904 [.89,.92]	0.926 [.91,.94]	+0.022

S3 Full fairness-gap table (Table 2).

Table S3. All fairness gaps (sensitivity, $A-B$): naive | PPI++ [95% CI] | true. SIG = significance flip, SIGN = sign flip. Of the 43 conclusions tested (35 finding-level, 8 macro) 14 flip; the 35 finding-level rows are listed here and the 8 per-axis macro rows are omitted, so 12 of the 14 flips appear below and all 14 are in Table 3. "—" in the naive column = the report essentially never labels that finding in that cell. Mean|error| vs. true over the 34 gaps with a defined naive estimate: 0.071 naive, 0.042 PPI++. * mapping artifact.

Pair	Finding	naive	PPI++ [CI]	true	flip
sex M-F	Atelectasis	+0.014	+0.010 [−0.06, 0.09]	+0.054	—
sex M-F	Cardiomegaly	−0.042	+0.027 [−0.04, 0.10]	+0.010	SIGN
sex M-F	Consolidation*	−0.033	+0.008 [−0.08, 0.10]	+0.049	SIGN
sex M-F	Edema	+0.008	+0.027 [−0.07, 0.13]	+0.047	—
sex M-F	Enl. Cardiomed.	+0.201	+0.162 [−0.15, 0.47]	+0.111	—
sex M-F	Lung Lesion	−0.021	−0.095 [−0.33, 0.14]	−0.099	—
sex M-F	Lung Opacity	+0.066	+0.142 [0.00, 0.27]	+0.102	SIG
sex M-F	Pleural Other*	+0.382	+0.100 [−0.01, 0.20]	+0.107	—
sex M-F	Pneumothorax	−0.114	−0.274 [−0.56, −0.02]	−0.192	SIG
sex M-F	Support Devices	+0.035	+0.029 [−0.04, 0.10]	+0.075	—
age <40-80+	Atelectasis	−0.055	−0.046 [−0.23, 0.12]	−0.030	—
age <40-80+	Cardiomegaly	−0.260	−0.007 [−0.21, 0.13]	−0.057	SIG
age <40-80+	Consolidation*	+0.060	+0.113 [−0.06, 0.26]	+0.132	—
age <40-80+	Lung Opacity	−0.140	−0.316 [−0.60, −0.05]	−0.221	—
age <40-80+	Pleural Other*	—	−0.110 [−0.28, 0.11]	−0.097	—
age <40-80+	Support Devices	−0.011	−0.019 [−0.13, 0.08]	+0.017	—
race W-Black	Cardiomegaly	+0.030	+0.030 [−0.06, 0.14]	+0.003	—
race W-Black	Support Devices	+0.079	+0.008 [−0.09, 0.11]	+0.073	SIG
race W-Asian	Cardiomegaly	−0.018	−0.082 [−0.12, −0.05]	−0.083	SIG
race W-Asian	Support Devices	+0.002	−0.090 [−0.19, 0.02]	−0.007	SIGN
race W-Hispanic	Cardiomegaly	+0.118	+0.238 [−0.01, 0.57]	+0.134	—
race W-Hispanic	Support Devices	+0.030	+0.039 [−0.16, 0.24]	+0.102	—
race W-Other	Cardiomegaly	+0.043	+0.048 [−0.11, 0.31]	+0.060	—
race W-Other	Support Devices	+0.004	−0.094 [−0.21, 0.09]	−0.104	SIGN
view AP-PA	Atelectasis	+0.144	+0.147 [0.03, 0.30]	+0.198	—
view AP-PA	Cardiomegaly	+0.157	+0.108 [0.01, 0.22]	+0.095	—
view AP-PA	Consolidation*	+0.365	+0.158 [0.03, 0.29]	+0.224	—
view AP-PA	Edema	+0.146	+0.287 [0.12, 0.49]	+0.224	—
view AP-PA	Enl. Cardiomed.	+0.232	+0.476 [0.19, 0.78]	+0.350	SIG
view AP-PA	Lung Lesion	−0.037	−0.201 [−0.42, 0.01]	−0.223	—
view AP-PA	Lung Opacity	+0.232	+0.252 [0.07, 0.44]	+0.223	—
view AP-PA	Pleural Other*	−0.179	−0.255 [−0.36, −0.15]	−0.269	SIG
view AP-PA	Support Devices	+0.235	+0.348 [0.24, 0.45]	+0.355	—
race W-Minority	Cardiomegaly	+0.037	+0.057 [−0.03, 0.15]	+0.017	—
race W-Minority	Support Devices	+0.048	−0.013 [−0.09, 0.07]	+0.039	SIGN

S4 Report-label noise by subgroup, all findings (Sec. 3.4, Sec. 4.4).

Table S4: Report-label under-report rate FNR = $P(y_r{=}0 \mid y_i{=}1)$ and over-report rate FPR = $P(y_r{=}1 \mid y_i{=}0)$ for every finding at every subgroup level, with Wilson 95% CIs and image-positive (n^+) / image-negative (n^-) counts. Model-independent (label noise only). † cell below the ≥ 20 image-positive inclusion criterion. Race rows exploratory. * mapping artifact.

finding	subgroup	n^+	n^-	FNR [95% CI]	FPR [95% CI]
Atelectasis	pooled	680	1936	0.601 [0.564,0.638]	0.147 [0.132,0.164]
	sex M	337	882	0.611 [0.558,0.662]	0.162 [0.139,0.188]
	sex F	310	947	0.577 [0.522,0.631]	0.130 [0.110,0.153]
	age <40	48	233	0.562 [0.423,0.693]	0.073 [0.046,0.114]
	age 40-59	153	548	0.562 [0.483,0.638]	0.128 [0.102,0.158]
	age 60-79	314	766	0.611 [0.557,0.664]	0.158 [0.134,0.185]

continued on next page

(continued)

finding	subgroup	n^+	n^-	FNR [95% CI]	FPR [95% CI]
	age 80+	132	282	0.606 [0.521,0.685]	0.206 [0.163,0.257]
	race White	451	1140	0.592 [0.546,0.636]	0.171 [0.150,0.194]
	race Black	93	307	0.645 [0.544,0.735]	0.098 [0.069,0.136]
	race Asian	28	71	0.643 [0.458,0.793]	0.085 [0.039,0.172]
	race Hispanic	16	101	0.625 [0.386,0.815]†	0.089 [0.048,0.161]†
	race Other	24	60	0.583 [0.388,0.755]	0.183 [0.106,0.299]
	view AP	554	1082	0.594 [0.552,0.634]	0.198 [0.175,0.223]
	view PA	126	854	0.635 [0.548,0.714]	0.083 [0.066,0.104]
Cardiomegaly	pooled	575	2041	0.565 [0.524,0.605]	0.143 [0.129,0.159]
	sex M	279	940	0.591 [0.533,0.647]	0.135 [0.115,0.158]
	sex F	260	997	0.546 [0.485,0.606]	0.148 [0.128,0.172]
	age <40	23	258	0.435 [0.256,0.632]	0.093 [0.063,0.135]
	age 40-59	106	595	0.575 [0.480,0.665]	0.121 [0.097,0.150]
	age 60-79	274	806	0.577 [0.517,0.634]	0.158 [0.134,0.184]
	age 80+	136	278	0.574 [0.490,0.654]	0.187 [0.146,0.237]
	race White	349	1242	0.596 [0.544,0.646]	0.141 [0.123,0.161]
	race Black	105	295	0.552 [0.457,0.644]	0.183 [0.143,0.231]
	race Asian	21	78	0.524 [0.324,0.717]	0.128 [0.071,0.220]
	race Hispanic	23	94	0.609 [0.408,0.778]	0.128 [0.075,0.210]
	race Other	21	63	0.381 [0.208,0.591]	0.095 [0.044,0.193]
	view AP	465	1171	0.561 [0.516,0.606]	0.179 [0.158,0.202]
	view PA	110	870	0.582 [0.488,0.670]	0.094 [0.077,0.115]
Consolidation*	pooled	677	1939	0.870 [0.843,0.893]	0.022 [0.017,0.030]
	sex M	348	871	0.885 [0.847,0.914]	0.030 [0.020,0.043]
	sex F	295	962	0.861 [0.817,0.896]	0.015 [0.009,0.024]
	age <40	49	232	0.857 [0.733,0.929]	0.013 [0.004,0.037]
	age 40-59	169	532	0.882 [0.824,0.922]	0.028 [0.017,0.046]
	age 60-79	294	786	0.867 [0.824,0.901]	0.015 [0.009,0.026]
	age 80+	131	283	0.885 [0.820,0.929]	0.035 [0.019,0.064]
	race White	442	1149	0.876 [0.842,0.903]	0.025 [0.018,0.036]
	race Black	82	318	0.890 [0.804,0.941]	0.009 [0.003,0.027]
	race Asian	28	71	0.893 [0.728,0.963]	0.014 [0.002,0.076]
	race Hispanic	16	101	0.938 [0.717,0.989]†	0.030 [0.010,0.084]†
	race Other	24	60	0.833 [0.641,0.933]	0.050 [0.017,0.137]
	view AP	570	1066	0.858 [0.827,0.884]	0.028 [0.020,0.040]
	view PA	107	873	0.935 [0.871,0.968]	0.015 [0.009,0.025]
Edema	pooled	350	2266	0.560 [0.508,0.611]	0.095 [0.083,0.108]
	sex M	177	1042	0.531 [0.458,0.603]	0.092 [0.076,0.111]
	sex F	152	1105	0.586 [0.506,0.661]	0.094 [0.078,0.113]
	age <40	16	265	0.688 [0.444,0.858]†	0.030 [0.015,0.058]†
	age 40-59	72	629	0.528 [0.414,0.639]	0.068 [0.051,0.091]
	age 60-79	158	922	0.557 [0.479,0.632]	0.099 [0.081,0.120]
	age 80+	83	331	0.554 [0.447,0.656]	0.175 [0.138,0.220]
	race White	223	1368	0.556 [0.490,0.620]	0.094 [0.079,0.110]
	race Black	47	353	0.574 [0.433,0.705]	0.099 [0.072,0.135]
	race Asian	10	89	0.600 [0.313,0.832]†	0.124 [0.070,0.208]†
	race Hispanic	18	99	0.611 [0.386,0.797]†	0.030 [0.010,0.085]†
	race Other	14	70	0.357 [0.163,0.612]†	0.157 [0.090,0.260]†
	view AP	297	1339	0.542 [0.485,0.598]	0.143 [0.126,0.163]
	view PA	53	927	0.660 [0.526,0.773]	0.025 [0.017,0.037]
Enlarged Cardiomediastinum	pooled	69	2547	0.913 [0.823,0.960]	0.025 [0.020,0.032]
	sex M	29	1190	0.862 [0.694,0.945]	0.024 [0.016,0.034]
	sex F	38	1219	0.947 [0.827,0.985]	0.024 [0.017,0.034]
	age <40	1	280	1.000 [0.207,1.000]†	0.025 [0.012,0.051]†
	age 40-59	17	684	0.882 [0.657,0.967]†	0.020 [0.012,0.034]†
	age 60-79	29	1051	0.966 [0.828,0.994]	0.023 [0.015,0.034]
	age 80+	20	394	0.850 [0.640,0.948]	0.030 [0.018,0.052]
	race White	40	1551	0.875 [0.739,0.945]	0.021 [0.015,0.030]
	race Black	13	387	0.923 [0.667,0.986]†	0.018 [0.009,0.037]†
	race Asian	2	97	1.000 [0.342,1.000]†	0.052 [0.022,0.115]†

continued on next page

(continued)

finding	subgroup	n^+	n^-	FNR [95% CI]	FPR [95% CI]
	race Hispanic	4	113	1.000 [0.510,1.000]†	0.009 [0.002,0.048]†
	race Other	4	80	1.000 [0.510,1.000]†	0.013 [0.002,0.067]†
	view AP	44	1592	0.886 [0.760,0.950]	0.032 [0.024,0.042]
	view PA	25	955	0.960 [0.805,0.993]	0.014 [0.008,0.023]
Fracture	pooled	29	2587	0.828 [0.655,0.924]	0.015 [0.011,0.021]
	sex M	16	1203	0.812 [0.570,0.934]†	0.019 [0.013,0.029]†
	sex F	11	1246	0.909 [0.623,0.984]†	0.011 [0.007,0.019]†
	age <40	1	280	1.000 [0.207,1.000]†	0.014 [0.006,0.036]†
	age 40-59	11	690	0.818 [0.523,0.949]†	0.012 [0.006,0.023]†
	age 60-79	10	1070	0.800 [0.490,0.943]†	0.015 [0.009,0.024]†
	age 80+	5	409	1.000 [0.566,1.000]†	0.022 [0.012,0.041]†
	race White	17	1574	0.882 [0.657,0.967]†	0.020 [0.014,0.028]†
	race Black	2	398	1.000 [0.342,1.000]†	0.003 [0.000,0.014]†
	race Asian	1	98	1.000 [0.207,1.000]†	0.000 [0.000,0.038]†
	race Hispanic	0	117	–	0.017 [0.005,0.060]†
	race Other	1	83	0.000 [0.000,0.793]†	0.000 [0.000,0.044]†
	view AP	23	1613	0.870 [0.679,0.955]	0.015 [0.010,0.022]
	view PA	6	974	0.667 [0.300,0.903]†	0.016 [0.010,0.027]†
Lung Lesion	pooled	132	2484	0.803 [0.727,0.862]	0.016 [0.012,0.022]
	sex M	59	1160	0.746 [0.622,0.839]	0.017 [0.011,0.026]
	sex F	65	1192	0.862 [0.757,0.925]	0.016 [0.010,0.025]
	age <40	9	272	0.778 [0.453,0.937]†	0.004 [0.001,0.021]†
	age 40-59	30	671	0.800 [0.627,0.905]	0.015 [0.008,0.027]
	age 60-79	60	1020	0.867 [0.758,0.931]	0.019 [0.012,0.029]
	age 80+	25	389	0.680 [0.484,0.828]	0.023 [0.012,0.043]
	race White	93	1498	0.753 [0.656,0.829]	0.021 [0.015,0.030]
	race Black	12	388	1.000 [0.758,1.000]†	0.003 [0.000,0.014]†
	race Asian	6	93	0.833 [0.436,0.970]†	0.011 [0.002,0.058]†
	race Hispanic	3	114	1.000 [0.439,1.000]†	0.009 [0.002,0.048]†
	race Other	1	83	1.000 [0.207,1.000]†	0.012 [0.002,0.065]†
	view AP	77	1559	0.831 [0.732,0.899]	0.017 [0.011,0.024]
	view PA	55	925	0.764 [0.637,0.856]	0.015 [0.009,0.025]
Lung Opacity	pooled	346	2270	0.604 [0.552,0.654]	0.212 [0.196,0.230]
	sex M	186	1033	0.597 [0.525,0.665]	0.245 [0.220,0.272]
	sex F	149	1108	0.617 [0.537,0.692]	0.183 [0.162,0.207]
	age <40	25	256	0.680 [0.484,0.828]	0.137 [0.100,0.184]
	age 40-59	81	620	0.667 [0.559,0.760]	0.192 [0.163,0.225]
	age 60-79	162	918	0.580 [0.503,0.654]	0.220 [0.194,0.248]
	age 80+	67	347	0.567 [0.448,0.679]	0.288 [0.243,0.338]
	race White	231	1360	0.628 [0.564,0.687]	0.240 [0.218,0.264]
	race Black	53	347	0.604 [0.469,0.724]	0.159 [0.124,0.201]
	race Asian	8	91	0.625 [0.306,0.863]†	0.209 [0.138,0.303]†
	race Hispanic	13	104	0.462 [0.232,0.709]†	0.125 [0.075,0.202]†
	race Other	9	75	0.667 [0.354,0.879]†	0.173 [0.104,0.274]†
	view AP	271	1365	0.601 [0.542,0.658]	0.261 [0.238,0.285]
	view PA	75	905	0.613 [0.500,0.715]	0.139 [0.118,0.163]
Pleural Other*	pooled	754	1862	0.983 [0.971,0.990]	0.006 [0.004,0.011]
	sex M	375	844	0.973 [0.952,0.985]	0.008 [0.004,0.017]
	sex F	332	925	0.991 [0.974,0.997]	0.005 [0.002,0.013]
	age <40	33	248	1.000 [0.896,1.000]	0.000 [0.000,0.015]
	age 40-59	176	525	0.977 [0.943,0.991]	0.004 [0.001,0.014]
	age 60-79	351	729	0.980 [0.959,0.990]	0.008 [0.004,0.018]
	age 80+	147	267	0.986 [0.952,0.996]	0.015 [0.006,0.038]
	race White	501	1090	0.980 [0.964,0.989]	0.007 [0.004,0.014]
	race Black	89	311	0.978 [0.922,0.994]	0.006 [0.002,0.023]
	race Asian	45	54	0.978 [0.884,0.996]	0.000 [0.000,0.066]
	race Hispanic	18	99	1.000 [0.824,1.000]†	0.000 [0.000,0.037]†
	race Other	25	59	1.000 [0.867,1.000]	0.000 [0.000,0.061]
	view AP	564	1072	0.991 [0.979,0.996]	0.007 [0.003,0.013]

continued on next page

(continued)

finding	subgroup	n^+	n^-	FNR [95% CI]	FPR [95% CI]
	view PA	190	790	0.958 [0.919,0.979]	0.006 [0.003,0.015]
Pneumothorax	pooled	78	2538	0.346 [0.250,0.457]	0.031 [0.025,0.039]
	sex M	41	1178	0.390 [0.257,0.543]	0.038 [0.029,0.051]
	sex F	36	1221	0.306 [0.180,0.469]	0.025 [0.018,0.036]
	age <40	7	274	0.000 [0.000,0.354]†	0.026 [0.012,0.052]†
	age 40-59	21	680	0.238 [0.106,0.451]	0.031 [0.020,0.047]
	age 60-79	36	1044	0.500 [0.345,0.655]	0.037 [0.027,0.051]
	age 80+	13	401	0.308 [0.127,0.576]†	0.022 [0.012,0.042]†
	race White	58	1533	0.293 [0.192,0.420]	0.033 [0.025,0.043]
	race Black	9	391	0.556 [0.267,0.811]†	0.018 [0.009,0.036]†
	race Asian	3	96	0.333 [0.061,0.792]†	0.052 [0.022,0.116]†
	race Hispanic	1	116	0.000 [0.000,0.793]†	0.017 [0.005,0.061]†
	race Other	3	81	0.667 [0.208,0.939]†	0.037 [0.013,0.103]†
	view AP	60	1576	0.333 [0.227,0.459]	0.036 [0.027,0.046]
	view PA	18	962	0.389 [0.203,0.614]†	0.024 [0.016,0.036]†
Support Devices	pooled	1160	1456	0.397 [0.370,0.426]	0.051 [0.041,0.063]
	sex M	611	608	0.380 [0.342,0.419]	0.058 [0.042,0.079]
	sex F	493	764	0.418 [0.375,0.462]	0.041 [0.029,0.057]
	age <40	89	192	0.303 [0.218,0.405]	0.042 [0.021,0.080]
	age 40-59	292	409	0.414 [0.359,0.472]	0.044 [0.028,0.068]
	age 60-79	538	542	0.416 [0.375,0.458]	0.046 [0.031,0.067]
	age 80+	185	229	0.357 [0.291,0.428]	0.066 [0.040,0.105]
	race White	774	817	0.416 [0.382,0.451]	0.060 [0.046,0.078]
	race Black	140	260	0.379 [0.302,0.461]	0.027 [0.013,0.055]
	race Asian	45	54	0.444 [0.309,0.588]	0.056 [0.019,0.151]
	race Hispanic	34	83	0.500 [0.341,0.659]	0.036 [0.012,0.101]
	race Other	34	50	0.265 [0.146,0.431]	0.040 [0.011,0.135]
	view AP	960	676	0.334 [0.305,0.365]	0.077 [0.059,0.099]
	view PA	200	780	0.700 [0.633,0.759]	0.028 [0.019,0.042]

S5 Reader agreement on the five-reader subset (Sec. 4.1).

Table S5. REFLACX phases 1–2 (109 images, each read by all five radiologists). κ = Fleiss' κ across the five readers. ΔFNR = range of the report-label under-report rate when the reference is taken to be each single reading in turn (max − min over the five). Pairwise percent agreement across findings is 0.69–0.96; scoring each held-out reader against the majority of the other four gives a single-reader ceiling of 0.72–0.87 sensitivity and 0.78–0.94 specificity on the estimable findings. Enlarged Cardiomediastinum, Fracture and Lung Lesion have 1, 3 and 3 multi-reader image-positives and are not estimable. REFLACX does not release reader identity, so the swap is computed over reading slots and may under-state between-radiologist variance. Among the findings with an estimable ceiling, only Support Devices (0.129) and Atelectasis (0.143) exceed the 0.122 race subgroup spread. ‡ rests on 7 (Edema) and 9 (Lung Opacity) multi-reader image-positives, too few for a stable spread. * mapping artifact.

finding	κ	ΔFNR
Atelectasis	0.32	0.143
Cardiomegaly	0.56	0.111
Consolidation*	0.43	0.070
Edema	0.17	0.276‡
Lung Opacity	0.19	0.267‡
Pleural Other*	0.67	0.067
Pneumothorax	0.58	0.033
Support Devices	0.64	0.129
macro	—	0.069

S6 Sensitivity to the decision threshold (Sec. 4.4).

Table S6. The full analysis repeated at five decision thresholds. flips = number of gap conclusions changing sign or significance between naive and PPI++ (manufactured / masked / sign). naive bias = range of the naive sensitivity bias across subgroups. Coverage is the mean over cells; in the thinnest exploratory cells (rare finding × small subgroup) per-cell coverage falls to 0.46–0.73 across the sweep (0.51 at τ=0.5), which is why the coverage claim in Sec. 4.4 is stated for the adequately-powered cells. The flip count is smallest at the default τ=0.5. § averaged over all 35 finding-level gaps; the main text reports 0.042, averaged over the 34 with a defined naive estimate.

τ	flips (man./mask./sign)	naive bias (range)	max |PPI++ bias| pooled / any cell	mean cov.	mean |gap err| naive → PPI++
0.3	24 (6/7/11)	[−0.037, +0.073]	0.005 / 0.019	0.98	0.046 → 0.034
0.4	16 (5/4/7)	[−0.050, +0.086]	0.004 / 0.019	0.99	0.058 → 0.040
0.5	**14** (2/6/6)	[−0.047, +0.094]	0.009 / 0.019	0.99	0.071 → 0.041§
0.6	17 (4/5/8)	[−0.046, +0.108]	0.011 / 0.027	0.99	0.076 → 0.046
0.7	19 (8/2/9)	[−0.028, +0.104]	0.011 / 0.034	0.99	0.093 → 0.043

S7 Anchor vs. deployment covariates (Sec. 5).

Table S7. Observable covariate distributions of the REFLACX anchor vs. the held-out deployment set (∼68k images). The transfer assumption behind the deployment estimate is untestable within the anchor; the populations already differ, most on view (the anchor is frontal-only) and report-label prevalence. Descriptive only; does not validate transfer.

covariate	anchor	held-out test	$\lvert\Delta\rvert$
images	2616	68372	65756
patients	2199	12636	10437
% male	0.466	0.493	0.027
% female	0.481	0.460	0.020
age mean	62.1	60.3	1.9
age <40	0.107	0.132	0.024
age 40–59	0.268	0.297	0.029
age 60–79	0.413	0.374	0.039
age 80+	0.158	0.150	0.008
race WHITE	0.608	0.570	0.038
race BLACK	0.153	0.158	0.005
race ASIAN	0.038	0.033	0.005
race HISPANIC	0.045	0.055	0.011
race OTHER	0.032	0.037	0.005
race missing	0.124	0.147	0.022
view AP	0.625	0.368	0.258
view PA	0.375	0.266	0.108
y_r prev. Atelectasis	0.213	0.163	0.049
y_r prev. Cardiomegaly	0.207	0.163	0.045
y_r prev. Consolidation	0.050	0.035	0.015
y_r prev. Edema	0.141	0.087	0.054
y_r prev. Enlarged Cardiomediastinum	0.027	0.026	0.000
y_r prev. Fracture	0.017	0.020	0.003
y_r prev. Lung Lesion	0.025	0.032	0.007
y_r prev. Lung Opacity	0.237	0.194	0.042
y_r prev. Pleural Other	0.010	0.009	0.001
y_r prev. Pneumothorax	0.050	0.033	0.017
y_r prev. Support Devices	0.295	0.207	0.089

S8 Reproducibility.

Data. Gold anchor (REFLACX): 2616 images / 2199 patients, both labels, held out from the classifier, split 50/50 by patient for validation. Training pool: 271,325 non-REFLACX images (report label only). Held-out test: 68,372 (report label only), unused in training. The 11 findings have adequate image-read coverage (Pleural Effusion excluded).

Classifier. Per-finding ℓ_2-regularized logistic regression on frozen RAD-DINO embeddings; balanced class weights; threshold 0.5; trained once against $y_r$and reused unchanged. **PPI++.** $\hat{\theta}_{k\ell}(\lambda) = \frac{1}{N}\sum_{\text{large}} \lambda g_{k\ell} + \frac{1}{n}\sum_{\mathcal{G}}(\mathbf{1}\{\hat{y}{=}k, y_i{=}\ell\} - \lambda g_{k\ell})$ with proxy $g_{k\ell} = \mathbf{1}\{\hat{y}{=}k, y_r{=}\ell\}$ and λ minimizing asymptotic variance; 95% CIs from a 1000-replicate patient-level cluster bootstrap. A flip is a change

of sign or significance between naive and PPI++. Seed 42; patient-grouped splits throughout.