Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias

Supplementary Information

Supplementary Table 1: Performance of the racial identity prediction AI models.

Supplementary Table 2: Counts of view types by patient race in the CXP and MXR datasets.

Supplementary Table 3: Sensitivity and specificity of the AI diagnostic models.

Supplementary Table 4: Comparison of original and resampled test sets.

Supplementary Figure 1: Confounder analysis of image processing effects on AI-based racial identity prediction.

Supplementary Figure 2: Confounder analysis of view position effects on AI-based racial identity prediction.

Supplementary Figure 3: Confounder analysis of underdiagnosis bias in MXR.

Supplementary Figure 4: Distribution of AI model scores by view for the diagnostic task.

Supplementary Figure 5: Calibration curves by threshold approach.

Supplementary Figure 6: Average image by patient race.

Supplementary Table 1: Performance of the racial identity prediction AI models. The AI models demonstrate high classification performance in the original CXP and MXR test sets. This performance is largely maintained when controlling for confounders using various approaches: Tst-Res.: test set resampling based on age, sex, and disease prevalence; Tr/Tst-Res.: training set and test set resampling based on age, sex, and disease prevalence; DICOM: testing based on the original DICOM images; BMI: training and test set resampling based on BMI. We note that BMI is only available for 39% of the images in MXR, resulting in lower amounts of training data for the MXR-BMI model. AUROC: area under the receiver operating characteristic curve. Parentheses correspond to 95% confidence interval.

Dataset	Version	**AUROC for Racial Identity Prediction**		
		Asian Patients	Black Patients	White Patients
CXP	Original	0.926 (0.922, 0.930)	0.914 (0.907, 0.920)	0.913 (0.909, 0.917)
MXR	Original	0.931 (0.925, 0.936)	0.955 (0.953, 0.957)	0.947 (0.945, 0.949)
CXP	Tst-Res.	0.903 (0.901, 0.905)	0.899 (0.896, 0.901)	0.907 (0.905, 0.909)
CXP	Tr/Tst-Res.	0.897 (0.895, 0.899)	0.887 (0.885, 0.890)	0.899 (0.897, 0.901)
MXR	Tst-Res.	0.928 (0.927, 0.929)	0.935 (0.934, 0.936)	0.930 (0.928, 0.931
MXR	Tr/Tst-Res.	0.920 (0.919, 0.922)	0.918 (0.917, 0.919)	0.929 (0.928, 0.930)
MXR	DICOM	0.915 (0.909, 0.921)	0.948 (0.946, 0.950)	0.940 (0.937, 0.942)
MXR	BMI	0.734 (0.731, 0.736)	0.867 (0.866, 0.869)	0.821 (0.819, 0.823)

Supplementary Table 2: Counts of view types by patient race in the CXP and MXR datasets. Raw counts and percent by patient race are shown. The number of portable views and a breakdown of standard vs. portable for AP views is also shown for MXR, which is not available for CXP. We note that nearly all portable views have an AP position (97.5% AP, 1.8% missing, <1% other views).

	CXP				**MXR**			
	Self-Reported Race				Self-Reported Race			
View	All	Asian	Black	White	All	Asian	Black	White
All	223414	23272	11961	125491	377110	11121	55614	218049
AP	161590 (72.3%)	16292 (70.0%)	8387 (70.1%)	91254 (72.7%)	147173 (39.0%)	4487 (40.3%)	18993 (34.2%)	93398 (42.8%)
PA	29420 (13.2)	3344 (14.4)	1664 (13.9)	16234 (12.9)	96161 (25.5)	2816 (25.3)	15597 (28.0)	50418 (23.1)
Lateral	32403 (14.5)	3636 (15.6)	1910 (16.0)	18002 (14.3)	117986 (31.2)	3423 (30.8)	20071 (36.1)	63897 (29.3)
Other	1 (<1)	0 (0)	0 (0)	1 (<1)	21 (<1)	0 (0)	1 (<1)	9 (<1)
Unknown	0 (0)	0 (0)	0 (0)	0 (0)	15769 (4.2)	395 (3.6)	952 (1.7)	10327 (4.7)
Portable	-	-	-	-	127314 (33.8)	4029 (36.2)	15142 (27.2)	81261 (37.3)
AP type	All	Asian	Black	White	All	Asian	Black	White
Standard	-	-	-	-	23080 (15.7)	600 (13.4)	4281 (22.5)	14172 (15.2)
Portable	-	-	-	-	124093 (84.3)	3887 (86.7)	14712 (77.5)	79226 (84.8)

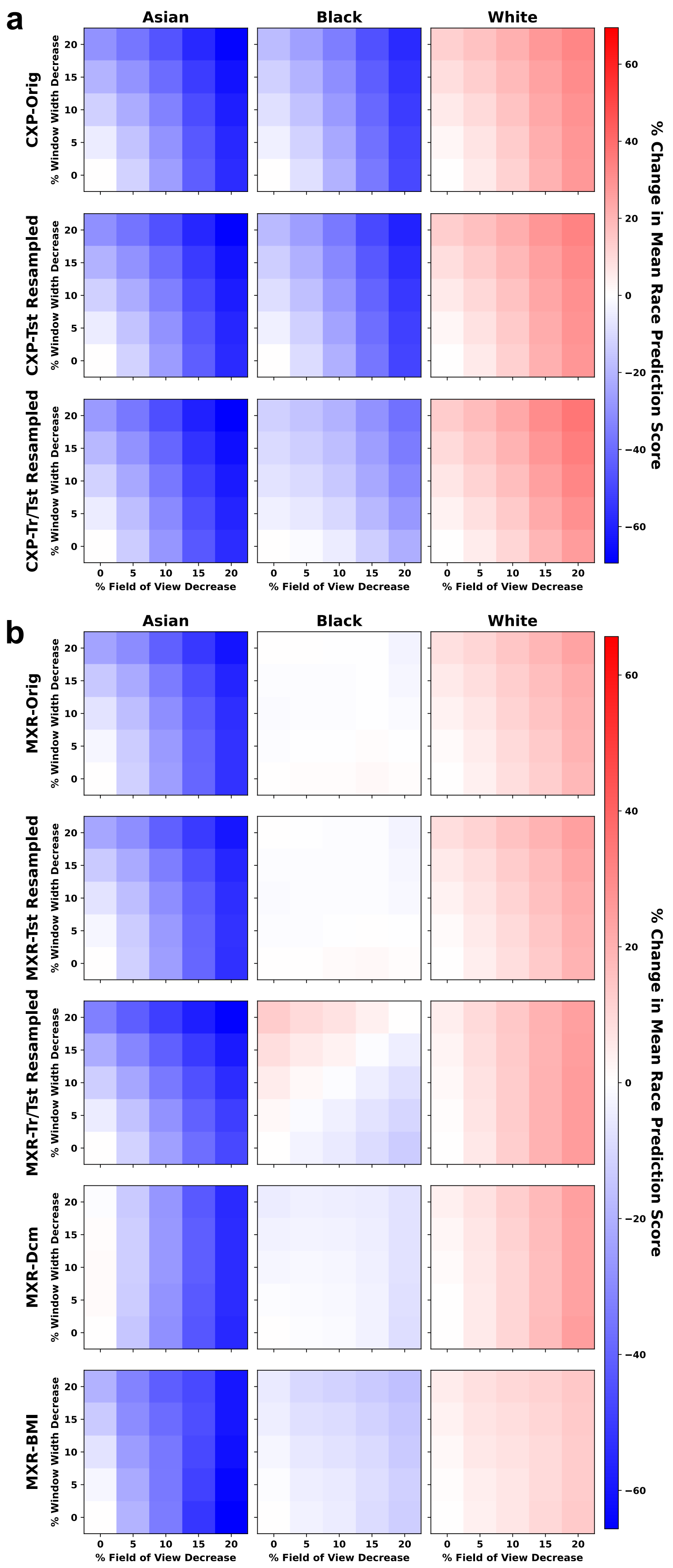
Supplementary Table 3: Sensitivity and specificity of the AI diagnostic models. The sensitivity and specificity of the models for the “No Findings” vs. “Findings Present” task is shown per patient race, dataset, and AI approach. Data Aug: data augmentation approach; Per View: per-view threshold approach. The ± values correspond to standard deviation computed via bootstrapping.

		Sensitivity			**Specificity**		
Datasets	Approach	Asian	Black	White	Asian	Black	White
Train: CXP Test: CXP	Baseline	82.1 ± 0.8	81.6 ± 1.1	83.7 ± 0.5	85.8 ± 1.7	85.0 ± 2.3	78.9 ± 1.0
	Data Aug	82.5 ± 0.9	79.5 ± 1.1	83.6 ± 0.6	85.6 ± 1.8	87.4 ± 2.2	80.3 ± 1.0
	Per View	82.4 ± 0.8	82.8 ± 1.1	84.0 ± 0.5	81.4 ± 2.2	77.5 ± 3.1	74.7 ± 1.3
Train: MXR Test: CXP	Baseline	81.0 ± 0.9	80.0 ± 1.2	81.5 ± 0.6	84.9 ± 1.8	83.0 ± 2.4	80.5 ± 1.0
	Data Aug	81.9 ± 0.9	79.4 ± 1.3	82.6 ± 0.5	84.3 ± 1.9	86.6 ± 2.3	80.4 ± 1.1
	Per View	80.7 ± 0.8	80.6 ± 1.1	81.5 ± 0.6	82.8 ± 1.9	77.9 ± 3.0	77.1 ± 1.2
Train: MXR Test: MXR	Baseline	80.5 ± 1.4	75.8 ± 0.8	83.5 ± 0.4	80.5 ± 1.8	83.5 ± 0.7	74.3 ± 0.6
	Data Aug	80.3 ± 1.4	74.3 ± 0.8	84.0 ± 0.4	79.6 ± 1.8	83.9 ± 0.7	74.0 ± 0.6
	Per View	82.3 ± 1.3	79.1 ± 0.7	83.3 ± 0.4	73.6 ± 2.0	75.6 ± 0.9	67.7 ± 0.8
Train: CXP Test: MXR	Baseline	77.8 ± 1.4	74.1 ± 0.8	81.6 ± 0.4	78.8 ± 1.8	80.0 ± 0.8	69.1 ± 0.6
	Data Aug	77.4 ± 1.4	73.2 ± 0.8	82.7 ± 0.4	80.6 ± 1.7	80.3 ± 0.7	69.9 ± 0.6
	Per View	78.9 ± 1.3	76.2 ± 0.8	81.6 ± 0.4	74.6 ± 2.0	74.2 ± 0.9	64.1 ± 0.7

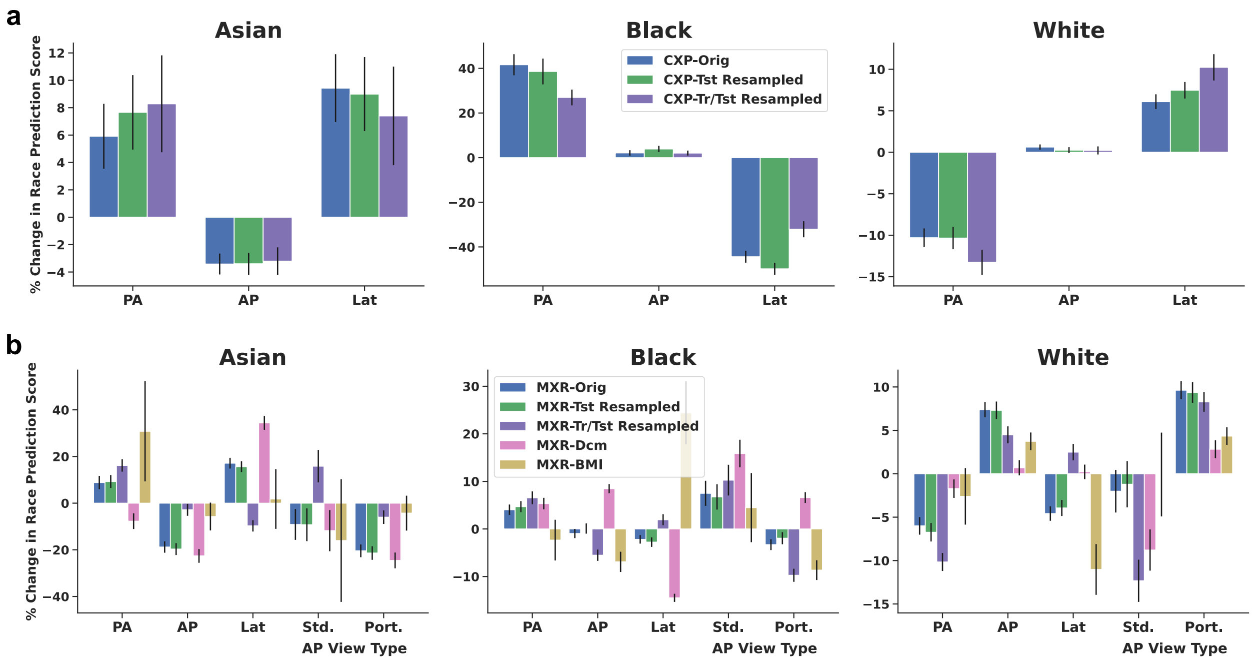
Supplementary Table 4: Comparison of original and resampled test sets. The resampled test sets were created to control for potential confounding factors by sampling the original test sets with replacement to achieve a balance across patient race and approximately equal distributions of age, sex, and disease labels within each patient race. A separate resampled set was also created for MXR based on BMI, which is not available for CXP. As illustrated below, the resampling process results in reduced differences in age, sex, findings presence, and BMI across patient race, while some differences in view proportions remain. Values are calculated at the image level.

		CXP			**MXR**		
Factor	Test Set	Asian	Black	White	Asian	Black	White
Total (%)	Original	13.7	7.3	79.0	3.9	19.8	76.3
	Resampled	33.3	33.1	33.5	33.2	33.5	33.3
Age (Mean)	Original	60.9	56.4	63.2	65.1	60.9	66.6
	Resampled	62.1	61.8	62.3	64.6	64.7	64.9
Male (%)	Original	57.6	50.5	60.6	54.4	40.4	55.1
	Resampled	59.9	59.3	59.6	52.2	52.3	52.2
BMI (Mean)	Original	-	-	-	23.7	29.1	28.4
	Resampled	-	-	-	26.3	28.8	28.2
No Findings (%)	Original	10.6	11.0	9.6	34.6	42.7	32.5
	Resampled	10.0	9.6	9.8	35.9	38.1	33.5
AP View (%)	Original	70.2	69.4	73.0	38.7	34.0	43.9
	Resampled	70.9	70.6	73.2	36.3	36.4	43.2
PA View (%)	Original	14.3	14.1	12.8	25.9	28.3	22.4
	Resampled	13.5	13.5	12.9	27.2	26.8	23.0
Lateral View (%)	Original	15.4	16.5	14.2	31.8	36.0	29.0
	Resampled	15.6	15.9	14.0	32.9	35.1	29.1
Portable (%)	Original	-	-	-	33.7	27.4	37.9
	Resampled	-	-	-	31.8	29.2	37.5

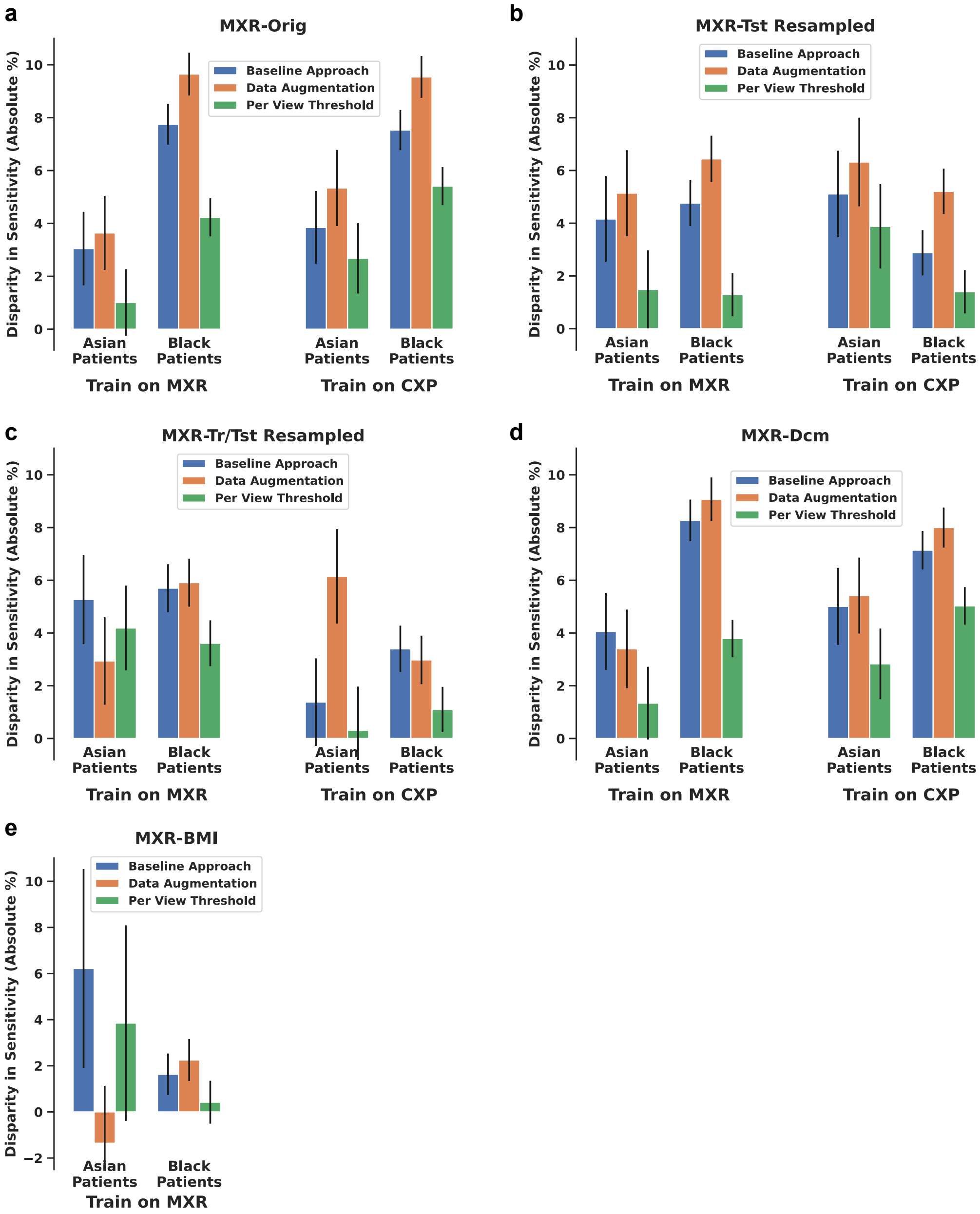
Supplementary Figure 1: Confounder analysis of image processing effects on AI-based racial identity prediction. a CXP-Orig: original CXP results; CXP-Tst Resampled: test set resampling based on age, sex, and disease prevalence; CXP-Tr/Tst Resampled: training and test set resampling based on age, sex, and disease prevalence. **b** MXR-Orig: original MXR results; MXR-Tst Resampled: test set resampling based on age, sex, and disease prevalence; MXR-Tr/Tst Resampled: training and test set resampling based on age, sex, and disease prevalence; MXR-Dcm: testing on original DICOM images; MXR-BMI: training and test set resampling based on BMI. Despite some variation, the observed patterns are largely consistent within each dataset across all conditions. For both MXR and CXP, the test set resampling approach produces highly consistent results compared to the original test sets. When performing both training set and test set resampling, the overall patterns again remain, for instance where decreases in the window width and field of view parameters lead to lower Asian prediction scores in CXP and MXR, lower Black predictions scores in CXP, and higher White prediction scores in CXP and MXR. For the Black prediction score in MXR, patterns with relatively lower magnitudes emerge, which were not present in the original testing. When testing on the DICOM images in MXR, the effects of the field of view parameter are consistent with the original testing, though there is more variation regarding the window width parameter. For both the original and DICOM test sets, a decrease in this parameter results in a statistically significant increase in the White prediction score (p<0.001), whereas there is a statistical significance decrease in the Asian prediction score (p<0.001) but not the Black prediction score (p=0.2) in the original test set and a statistical significance decrease in the Black prediction score (p<0.001) but not Asian prediction score (p=0.58) in the DICOM test set (*p*-values (two-sided) computed via bootstrapping of Pearson correlation at original field of view). Thus, while there is some variation across the training and testing scenarios, the overarching trends and conclusions still hold, where changes in the window width and field of view parameters affect the behavior of the racial identity prediction models in each scenario. Source data are provided as a Source Data file.


a
Asian
Black
White
CXP-Orig
CXP-Tst Resampled
CXP-Tr/Tst Resampled
% Window Width Decrease
0
5
10
15
20
% Field of View Decrease
% Change in Mean Race Prediction Score
60
40
20
0
−20
−40
−60
b
Asian
Black
White
MXR-Orig
MXR-Tst Resampled
MXR-Tr/Tst Resampled
MXR-Dcm
MXR-BMI
% Window Width Decrease
% Field of View Decrease
% Change in Mean Race Prediction Score

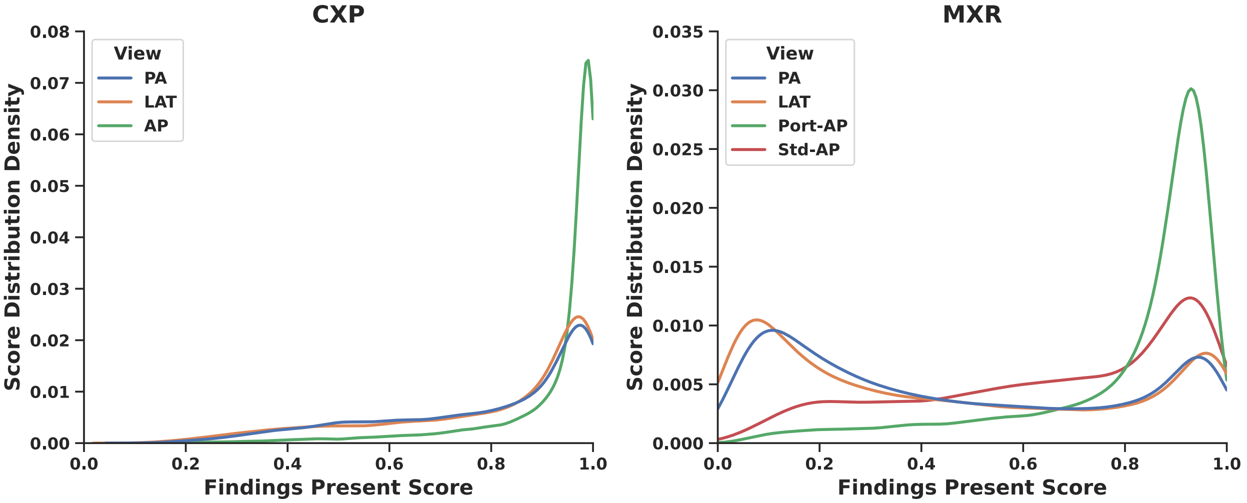
Supplementary Figure 2: Confounder analysis of view position effects on AI-based racial identity prediction. a CXP-Orig: original CXP results; CXP-Tst Resampled: test set resampling based on age, sex, and disease prevalence; CXP-Tr/Tst Resampled: training and test set resampling based on age, sex, and disease prevalence. **b** MXR-Orig: original MXR results; MXR-Tst Resampled: test set resampling based on age, sex, and disease prevalence; MXR-Tr/Tst Resampled: training and test set resampling based on age, sex, and disease prevalence; MXR-Dcm: testing on original DICOM images; MXR-BMI: training and test set resampling based on BMI. For both MXR and CXP, the test set resampling approach produces highly consistent results compared to the original test sets, including trends in the same direction for all 9 (race, view) pairs in CXP and 14 out of 15 in MXR. Similar trends are again observed for all 9 instances in CXP when performing both training and test set resampling, though there is more quantitative and qualitative variation in MXR with 11 of the 15 instances (73%) trending in the same direction as the original results when resampling by age, sex, and disease prevalence, and 13 of 15 (87%) when resampling by BMI. There is also more variation in the DICOM-based evaluation in MXR, though the observed values again trend in the same direction as the original results in 11 (73%) of the instances. Thus, while there is some variation across the conditions, there are also similarities and relatively large differences in race prediction scores by view position in all scenarios. PA: posterior-anterior; AP: anterior-posterior; Lat: lateral; Std: standard; Port: portable. Error bars correspond to standard deviation computed via bootstrapping. Source data are provided as a Source Data file.



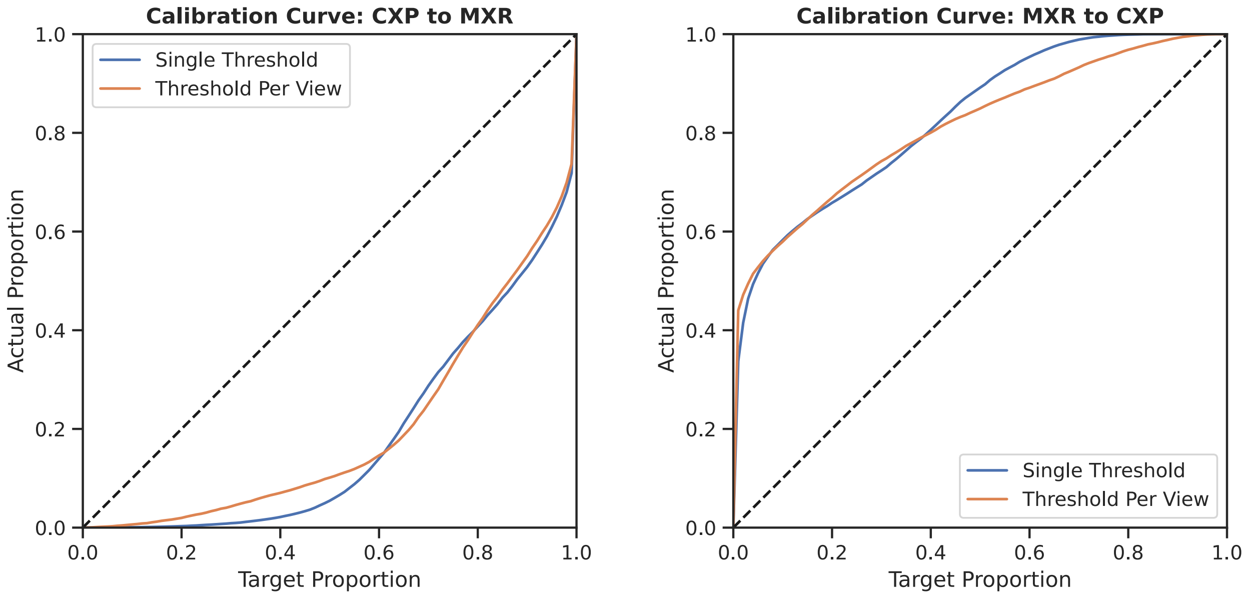
Supplementary Figure 3: Confounder analysis of underdiagnosis bias in MXR. Disparities in sensitivity are observed in the original MXR results (**a** MXR-Orig), when performing test set resampling based on age, sex, and disease prevalence (**b** MXR-Tst Resampled), when performing training and test set resampling based on age, sex, and disease prevalence (**c** MXR-Tr/Tst Resampled), when testing on the original DICOM images (**d** MXR-Dcm), and when performing training and test set resampling based on BMI (**e** MXR-BMI). The resampling approaches tend to reduce the overall disparity magnitude, but the view-specific score thresholds reduce these differences even further. The threshold approach additionally reduces the disparities in the DICOM-based testing, with similar magnitudes observed as the original test set. Error bars correspond to standard deviation computed via bootstrapping. Source data are provided as a Source Data file.



Supplementary Figure 4: Distribution of AI model scores by view for the diagnostic task. A kernel density estimate of the model score is shown for baseline diagnostic models trained and tested on each dataset. The score corresponds to the binary task of "No Findings" vs. "Findings Present" for the diagnostic models. PA: posterior-anterior; AP: anterior-posterior; LAT: lateral; Std: standard; Port: portable. Source data are provided as a Source Data file.



Supplementary Figure 5: Calibration curves by threshold approach. Calibration curves were generated to assess calibration across datasets for the “No Findings” diagnostic task. Left: Calibration from CXP to MXR; Right: Calibration from MXR to CXP. In each scenario, thresholds were calculated using the first dataset and assessed in the second dataset to compare the target proportion of images flagged (i.e., above the threshold) to the observed proportion of images flagged. For instance, the CXP to MXR curve is created based on calculating thresholds in the CXP validation set for a CXP-trained model and assessing these thresholds in the MXR test set, where this process is performed over the full range of potential operating points in increments of 0.01 from 0 to 1. The per-view thresholds were computed based on view position alone (i.e., AP, PA, Lateral) for this experiment, as the “standard” vs. “portable” view indicator is not available in the CXP dataset. Consistent with the general challenge of AI calibration, the models demonstrate miscalibration, though the average calibration error is slightly lower for the per-view threshold approach. For CXP to MXR, the average calibration error is 0.316 for the single threshold and 0.301 for the per-view thresholds, for a difference of 0.015 (95% CI: 0.014, 0.017). For MXR to CXP, the average calibration error is 0.326 for the single threshold and 0.311 for the per-view thresholds, for a difference of 0.016 (95% CI: 0.014, 0.017). Thus, while calibration is a general challenge, we do not find evidence that the per-view threshold strategy is any worse than the standard approach and could potentially help in some scenarios. Source data are provided as a Source Data file.



Supplementary Figure 6: Average image by patient race. Average images per patient race were computed using the default preprocessing for the CXP training split, MXR training split, and the BMI-resampled test split for MXR. An aggregate average image was computed as the average over the per-race average images. Differences between the per-race average images and the aggregate average image within each dataset are also shown with the scale bar indicating the percent difference from the mean image. Red areas indicate regions where the per-race average image has higher pixel intensity than the aggregate average image, whereas blue areas correspond to lower intensity. Overall, differences are quite subtle and could be caused by a number of factors. Yet the differences at the edges of images and in the relative contrast between lung and non-lung regions could potentially relate to the effects observed in the window width and field of view parameter analysis. Source data are provided as a Source Data file.