Supplementary Information

Supplementary Results 1: RoentGen-v2 Checkpoint Selection

To select the best generative model checkpoint we computed several metrics to measure the quality of synthetic images across three distinct axes: (a) alignment with the provided findings text and demographics from the prompt, (b) similarity of generated images to real images, and (c) diversity among synthetic images generated from the same text prompt. Supplementary Table 1 summarizes the synthetic image quality metrics at each model checkpoint.

(a) Text Prompt Alignment

For each text prompt in the validation set we generated synthetic CXRs. To evaluate how well synthetic CXRs depicted radiological findings that were included in the prompt, a pretrained classification model trained on real images (torch XRV [1]) was used to predict disease labels for the synthetic images. Similarly, to evaluate how well synthetic CXRs depicted the demographic information from the prompt, separate XRV models, trained on real data, were used to predict sex, race and age. The disease labels and patient metadata of the original text prompt represented the ground truth labels for the respective tasks. We report area under the receiver-operating characteristic curve (AUROC) for the diseases, accuracy for sex and race, and root mean square error (RMSE) for age predictions. Since the XRV classifiers themselves are imperfect, we compute a reference baseline by evaluating the classifier performance on the real images from the validation set (see Supplementary Table 1, 'real data' row).

The checkpoints at $7.5k$ training steps (equivalent to 21 epochs) and $10k$ training steps (equivalent to 28 epochs) achieved the highest average disease AUROC (0.81 compared to 0.88 for the real data baseline). The original RoentGen [2] model achieved an average disease AUROC of 0.82 (see Supplementary Table 1, 'RoentGen' row), showing we maintain the disease fidelity while introducing metadata variables in the conditional image generation process. For demographic attributes, sex and race accuracy for synthetic images were 100% and 99% respectively. As a reference, the XRV sex and race accuracy in classifying the real images were 97% and 95% respectively. This demonstrates the sex and race attributes were represented accurately by RoentGen-v2 . The more challenging task of age prediction also had accurate results, with only a slightly larger root mean square error (8.9 yrs) compared to the real data (7.1 yrs). We note that training past $15k$ optimization steps (equivalent to 43 epochs) led to decreased classification performance for disease, sex and race, likely due to model overfitting. Overall, we observe near-perfect accuracy for demographics instruction following along with disease instruction following on par with the original RoentGen model.

(b) Real–Synthetic Image Similarity

We used Fréchet Inception Distance (FID) score to compare the distribution of generated images with the distribution of the real ground truth images [3]. Out of all model checkpoints, the $10k$ training steps (equivalent to 28 epochs) checkpoint achieved the lowest FID score (76.8), indicating the most "realistic" image distribution. By comparison, the FID score of images generated by RoentGen [2] was 96.1.

While we aim to minimize the distance between the two image distributions, having individual synthetic images being exact replicas of real samples would be sub-optimal. To check for such model overfitting, we computed the image-level multi-scale structural similarity index (MS-SSIM) between generated images and their real counterparts, based on the same text prompt. We similarly computed the embedding-level similarity by measuring the cosine similarity of image embeddings, extracted using the image encoder of a domain-specific foundation model (BioViL [4]). In the case of MS-SSIM, the highest observed score was 0.41 at the $5k$ steps checkpoint (equivalent to

14 epochs), while the lowest observed score was 0.30 at $30k$ and $40k$ steps (equivalent to 86 and 115 epochs), with all other checkpoints scoring in-between. For the BioViL metric, the highest observed score was 0.52 at the $60k$ steps checkpoint (equivalent to 172 epochs), while the lowest observed score was 0.35 at $5k$ steps (equivalent to 14 epochs), with all other checkpoints scoring in-between. We observed balanced similarity scores (MS-SSIM and BioViL) between synthetic and original images for all model checkpoints, indicating we do not overfit to the real samples.

(c) Intra-prompt Diversity of Generated Images

To measure the diversity of synthetic images generated by RoentGen-v2 , we generated multiple images using different random seeds for each text prompt in the validation set. We computed pairwise similarity scores (MS-SSIM and cosine similarity of BioViL embeddings) between images generated from the same prompt. For both metrics, a score of 1.00 would indicate that the model collapsed and is generating identical samples, thus lower scores are desirable. In the case of MS-SSIM, the highest observed score was 0.46 at the $5k$ steps checkpoint (equivalent to 14 epochs), while the lowest observed score was 0.24 at $30k$ and $40k$ steps (equivalent to 86 and 115 epochs), with all other checkpoints scoring in-between. For the BioViL metric, the highest observed score was 0.72 at the $5k$ steps checkpoint (equivalent to 14 epochs), while the lowest observed score was 0.52 at $20k$ steps (equivalent to 57 epochs), with all other checkpoints scoring in-between. Balanced intra-prompt similarity scores verify the ability of the model to generate distinct and varied images at multiple checkpoints.

Supplementary Tables

Supplementary Table 1 RoentGen-v2 checkpoints: synthetic image quality metrics.

Checkpoint number of training steps (training epochs)	Text Prompt Alignment[1]									Real-Synthetic Similarity[2]			Intra-prompt Diversity[3]	
	Atelectasis	Cardiomegaly	Edema	Effusion	Pneumothorax	Average AUROC ↑	Sex Acc. ↑	Race Acc. ↑	Age RMSE ↓	FID ↓	MS-SSIM	BioViL	MS-SSIM ↓	BioViL ↓
real data	0.75	0.83	0.92	0.81	0.92	0.88	0.97	0.95	7.1	0.0	1.00	1.00	-	-
RoentGen	0.80	0.86	0.81	0.93	0.72	0.82	-	-	-	96.1	0.36	0.59	0.42	0.77
5k (14 ep)	0.72	0.81	0.84	0.73	0.87	0.79	1.00	0.98	9.4	77.2	0.41	0.35	0.46	0.72
7.5k (21 ep)	0.78	0.81	0.84	0.72	0.93	0.81	1.00	0.99	8.7	81.9	0.39	0.36	0.39	0.73
10k (28 ep)	0.76	0.77	0.82	0.78	0.90	**0.81**	1.00	0.98	8.9	**76.8**	0.37	0.41	0.37	0.66
12.5k (35 ep)	0.74	0.83	0.80	0.73	0.89	0.80	1.00	0.98	7.9	86.4	0.37	0.43	0.36	0.62
15k (43 ep)	0.76	0.80	0.79	0.73	0.89	0.80	0.99	0.98	8.4	87.4	0.34	0.44	0.31	0.57
20k (57 ep)	0.75	0.77	0.72	0.70	0.89	0.76	0.97	0.95	8.5	103.5	0.31	0.47	0.26	0.52
30k (86 ep)	0.70	0.78	0.71	0.66	0.89	0.75	0.98	0.96	8.7	101.3	0.30	0.48	0.24	0.54
40k (115 ep)	0.69	0.80	0.72	0.64	0.89	0.75	0.97	0.94	8.8	99.2	0.30	0.49	0.24	0.56
50k (143 ep)	0.72	0.81	0.73	0.65	0.90	0.76	0.98	0.95	8.9	97.9	0.31	0.51	0.26	0.60
60k (172 ep)	0.72	0.81	0.74	0.65	0.91	0.77	0.98	0.96	8.8	90.9	0.32	0.52	0.28	0.60

Abbreviations: AUROC: area under the receiver-operating curve; FID: Frechet Inception Distance, calculated using embeddings from an ImageNet-pretrained InceptionV3 network; MS-SSIM: multi-scale structural similarity index; BioViL: indicates cosine similarity between image embeddings obtained from BioViL image encoder.

[1]Text Prompt Alignment Metrics. For individual disease labels, showing binary AUROC. For sex and race, showing classification accuracy. For age, showing root mean square error (RMSE).

[2]Real-Synthetic Similarity Metrics. For FID, lower scores are better. For MS-SSIM and BioViL, higher scores indicate more resemblance to real images; however, a score of 1.00 would indicate model overfitting, i.e. synthetic samples identical to their real counterparts.

[3]Intra-prompt Diversity Metrics. Lower scores indicate more diversity; a score of 1.00 would indicate model collapse, i.e. always identical synthetic samples.

Supplementary Table 2 Description of chest radiography datasets.

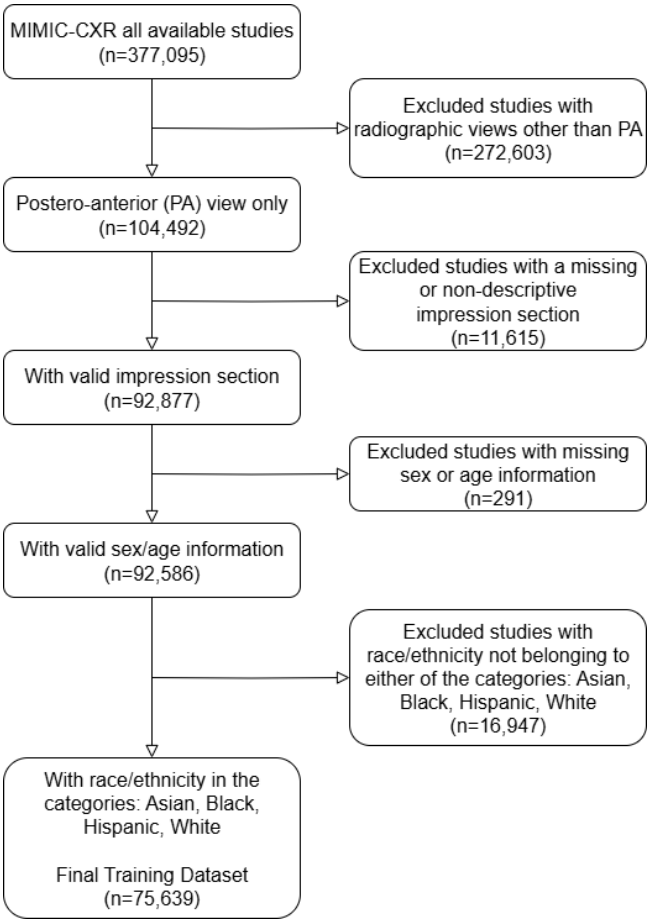
Subgroup	**Attribute**	MIMIC-CXR			CheXpert	NIH	PadChest	VinDr
	Split	train	val	test	test	test	test	test
	Views	PA	PA	PA	PA	PA	PA	PA,AP
Total	No.images	66,760	1,295	7,584	20,531	47,075	59,036	3,000
Sex	Female	32,593 (48.8%)	514 (39.7%)	3,670 (48.4%)	7,132 (34.7%)	20,872 (44.3%)	30,695 (52.0%)	-
	Male	34,167 (51.2%)	781 (60.3%)	3,914 (51.6%)	13,399 (65.3%)	26,203 (55.7%)	28,341 (48.0%)	-
Age	0-18	-	-	-	-	1,922 (4.1%)	1,789 (3.0%)	-
	18-40	9,599 (14.4%)	118 (9.1%)	1,186 (15.6%)	3,622 (17.6%)	12,684 (26.9%)	5,614 (9.5%)	-
	40-60	24,167 (36.2%)	459 (35.4%)	2,606 (34.4%)	7,179 (35.0%)	20,810 (44.2%)	16,666 (28.3%)	-
	60-80	25,941 (38.8%)	574 (44.3%)	3,061 (40.4%)	7,722 (37.6%)	11,244 (23.9%)	22,682 (38.4%)	-
	80+	7,053 (10.6%)	144 (11.2%)	731 (9.6%)	2,008 (9.8%)	415 (0.9%)	12,293 (20.8%)	-
Race/ Ethnicity [1]	Asian	2,452 (3.6%)	14 (1.1%)	299 (3.9%)	2,333 (11.4%)	-	-	-
	Black	13,685 (20.5%)	218 (16.8%)	1,500 (19.8%)	1,174 (5.7%)	-	-	-
	Hispanic	5,521 (8.3%)	82 (6.3%)	700 (9.2%)	2,401 (11.7%)	-	-	-
	White	45,102 (67.6%)	981 (75.8%)	5,085 (67.1%)	11,395 (55.5%)	-	-	-
Target Diseases [2]	No Finding	34,028 (51.0%)	472 (36.4%)	3,783 (49.9%)	3,671 (17.9%)	27,379 (58.2%)	24,180 (41.0%)	2,051 (68.4%)
	Atelectasis	8,795 (13.2%)	210 (16.2%)	1,050 (13.8%)	4,531 (22.1%)	4,029 (8.6%)	2,477 (4.2%)	86 (2.9%)
	Cardiomegaly	7,074 (10.6%)	173 (13.4%)	831 (11.0%)	2,709 (13.2%)	1,122 (2.4%)	5,422 (9.2%)	309 (10.3%)
	Consolidation	2,209 (3.3%)	83 (6.4%)	209 (2.8%)	2,546 (12.4%)	1,059 (2.2%)	643 (1.1%)	96 (3.2%)
	Edema	4,332 (6.5%)	142 (11.0%)	504 (6.6%)	1,704 (8.3%)	181 (0.4%)	147 (0.3%)	0 (0.0%)
	Effusion	9,560 (14.3%)	267 (20.6%)	1,258 (16.6%)	6,542 (31.9%)	4,678 (9.9%)	2,032 (3.4%)	111 (3.7%)
	Pneumonia	8,806 (13.2%)	247 (19.1%)	976 (12.9%)	2,305 (11.2%)	438 (0.9%)	2,281 (3.9%)	246 (8.2%)
	Pneumothorax	1,696 (2.5%)	31 (2.4%)	197 (2.6%)	1,266 (6.2%)	2,386 (5.1%)	121 (0.2%)	18 (0.6%)

Supplementary Table 3 Description of demographic subgroup composition of chest radiography datasets.

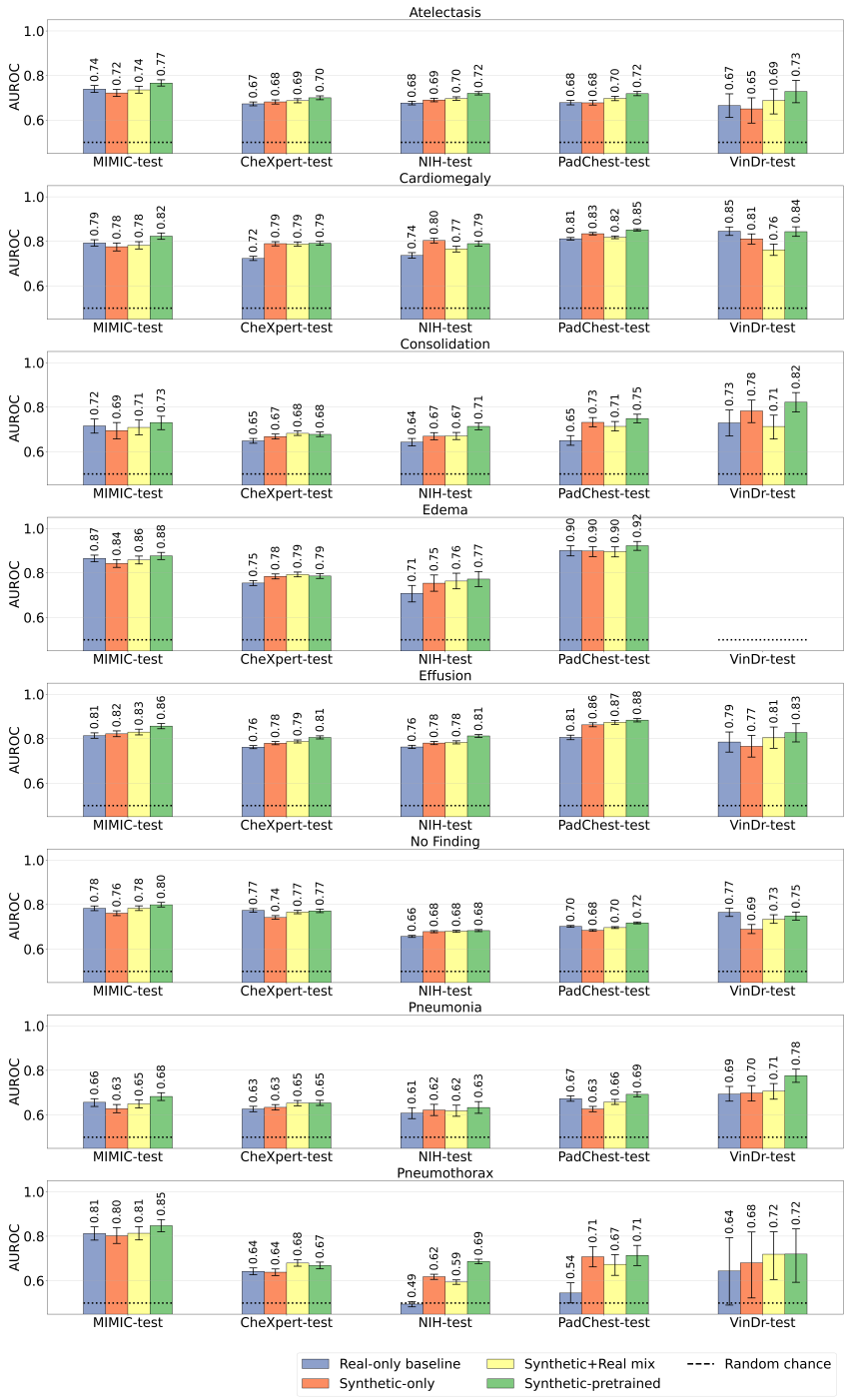
Intersectional Subgroup	MIMIC-CXR			CheXpert	NIH	PadChest
	train	val	test	test	test	test
Total	66,760	1,295	7,584	20,531	47,075	59,036
White Male	25,141	612	2,880	7,572	-	-
White Female	19,961	369	2,205	3,550	-	-
Black Male	5,366	105	576	653	-	-
Black Female	8,319	113	924	487	-	-
Hispanic Male	2,429	62	325	1,350	-	-
Hispanic Female	3,092	20	375	992	-	-
Asian Male	1,231	2	133	1,394	-	-
Asian Female	1,221	12	166	886	-	-
Male aged under 18	0	0	0	0	1,071	933
Female aged under 18	0	0	0	0	851	856
Male aged $18-40$	4,058	51	469	2,212	6,834	2,720
Female aged $18-40$	5,541	67	717	1,330	5,850	2,894
Male aged $40-60$	12,774	286	1,370	4,558	11,166	7,447
Female aged $40-60$	11,393	173	1,236	2,445	9,644	9,216
Male aged $60-80$	13,893	371	1,681	5,087	6,880	11,343
Female aged $60-80$	12,048	203	1,380	2,428	4,364	11,334
Male aged over 80	3,442	73	394	1,230	252	5,898
Female aged over 80	3,611	71	337	738	163	6,395

[1]Note: NIH and PadChest datasets do not have race/ethnicity information available.

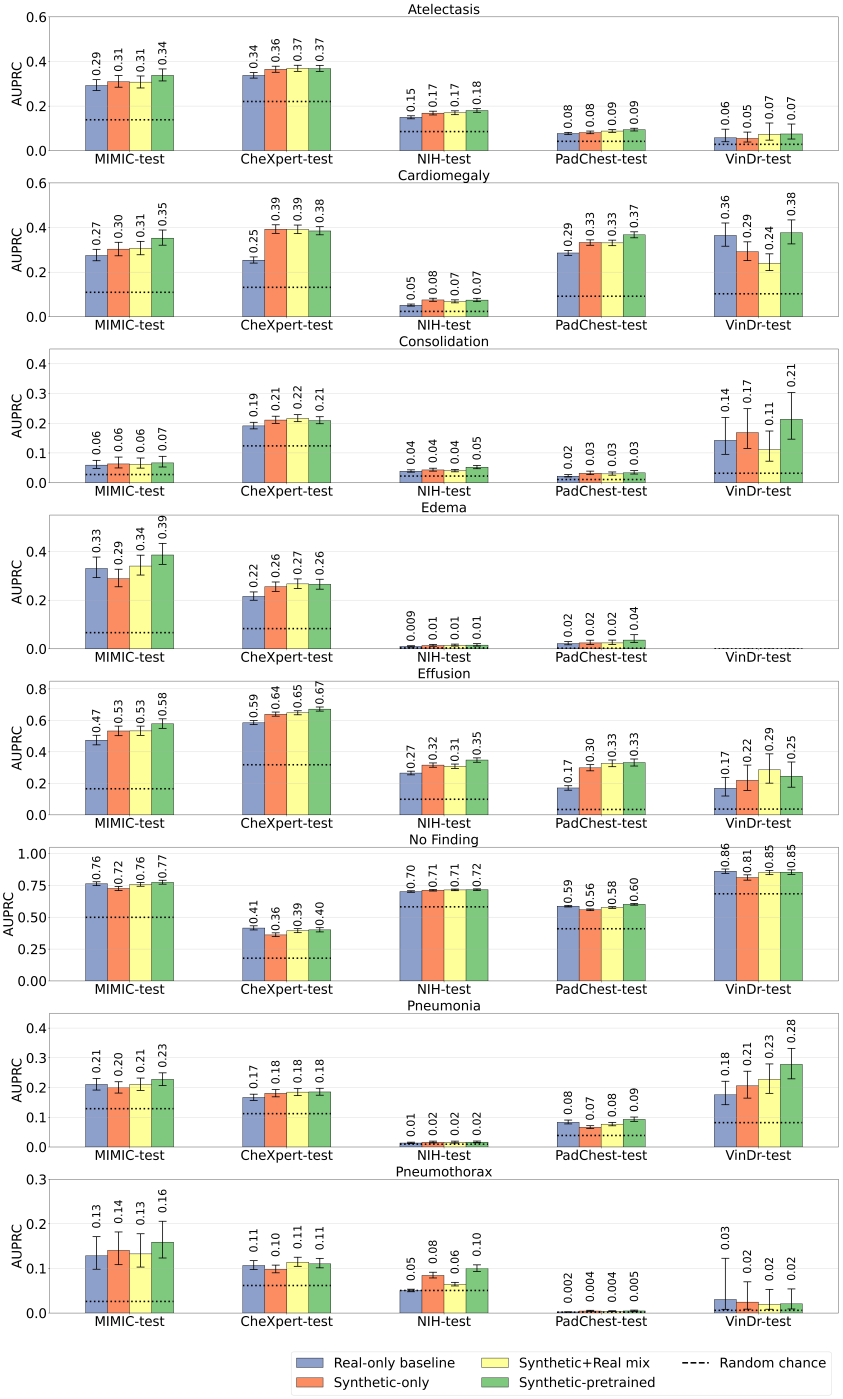
Supplementary Figures



Supplementary Figure 1 Selection criteria for the training dataset of the generative model.



Supplementary Figure 2 Classification performance of models trained on all available real and/or synthetic data according to four strategies: (i) Real-only (baseline) model trained on $66k$ real CXRs, (ii) Synthetic-only model trained on $565k$ synthetic CXRs, (iii) Synthetic+Real mix model trained on combined $66k$ real and $565k$ synthetic CXRs, and (iv) Synthetic-pretrained model, which underwent supervised pretraining on $565k$ synthetic CXRs followed by fine-tuning on $66k$ real CXRs. In scenarios (i)–(iii) models were initialized using ImageNet pretrained weights; in scenario (iv) the model was trained from scratch. Y-axis shows binary area under the receiver-operating curve (AUROC), with each panel focusing on one label. A random chance classifier would score 0.50 AUROC. For 'Edema' label, there are no patients with positive label in VinDr-test. Each model is evaluated in-distribution on MIMIC-test and out-of-distribution on four external datasets.



Supplementary Figure 3 Classification performance of models trained on all available real and/or synthetic data according to four strategies: (i) Real-only (baseline) model trained on 66k real CXRs, (ii) Synthetic-only model trained on 565k synthetic CXRs, (iii) Synthetic+Real mix model trained on combined 66k real and 565k synthetic CXRs, and (iv) Synthetic-pretrained model, which underwent supervised pretraining on 565k synthetic CXRs followed by fine-tuning on 66k real CXRs. In scenarios (i)–(iii) models were initialized using ImageNet pretrained weights; in scenario (iv) the model was trained from scratch. Y-axis shows binary area under the precision-recall curve (AUPRC), with each panel focusing on one label. A random chance classifier would score AUPRC equal to the label prevalence in the test dataset, hence the large differences in scores between labels and datasets. Each model is evaluated in-distribution on MIMIC-test and out-of-distribution on four external datasets. For 'Edema' label, there are no patients with positive label in VinDr-test; the disease prevalence is 0.004 in NIH-test, and 0.002 in PadChest-test. For 'Pneumonia' label, the disease prevalence in NIH-test is 0.009. For 'Pneumothorax' label, the disease prevalence is 0.002 in PadChest-test, and 0.006 in VinDr-test.

Supplementary Note 1: Disease-specific Classification Metrics

Supplementary Figure 2 shows the disease-specific AUROC metrics across all models and datasets. The trend observed in the macro-averaged multi-label AUROC was reflected in the binary classification AUROCs of individual imaging findings. Generally, the synthetic-pretrained model achieved the highest AUROC scores across datasets and diseases, with only a few exceptions. For Cardiomegaly on VinDr, there was no significant difference between the real-only baseline the synthetic-pretrained model (p-value= 0.80). For Consolidation on MIMIC, there was no significant difference between the real-only baseline the synthetic-pretrained model (p-value= 0.18). Lastly, for the No Finding label on CheXpert, there was no significant difference between the real-only baseline the synthetic-pretrained model (p-value= 0.43), and on VinDr the real-only model achieved higher AUROC than the synthetic-pretrained model (p-value= 0.02).

Supplementary Figure 3 shows the disease-specific AUPRC metrics across all models and datasets. Across the disease labels of Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Pneumonia and Pneumothorax, the synthetic-pretrained model achieved the best binary AUPRC. For the 'No Finding' label, there was no significant difference between the proposed synthetic pretraining strategy and the baseline model.

Supplementary Note 2: Details of Pretrained CXR Classifiers

- Torch X-ray Vision library [1]: XRV version 1.3.5
- Disease classification model (XRV):
  `xrv.models.DenseNet(weights="densenet121-res224-all")`
- Race classification model (XRV):
  `xrv.baseline_models.emory_hiti.RaceModel()`
- Age prediction model (XRV):
  `xrv.baseline_models.riken.AgeModel()`
- Sex classification model from [5]:
  `SexModelResNet.load_from_checkpoint("sex_model_chexpert_resnet_all.ckpt")` with weights from https://github.com/biomedia-mira/chexploration/tree/main/prediction
- Downstream disease classification models initialized with natural image (ImageNet) pretrained weights used the PyTorch weights
  `DenseNet121_Weights.IMAGENET1K_V1`.

Supplementary References

[1] Cohen, J. P. *et al.* *TorchXRayVision: A library of chest X-ray datasets and models* (2022). URL https://github.com/mlmed/torchxrayvision.

[2] Bluethgen, C. *et al.* A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering* (2024). URL http://dx.doi.org/10.1038/s41551-024-01246-y.

[3] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. *Gans trained by a two time-scale update rule converge to a local nash equilibrium*, NIPS'17, 6629–6640 (Curran Associates Inc., Red Hook, NY, USA, 2017).

[4] Boecking, B. *et al.* *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, 1–21 (Springer Nature Switzerland, 2022). URL http://dx.doi.org/10.1007/978-3-031-20059-5_1.

[5] Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *eBioMedicine* **89**, 104467 (2023). URL http://dx.doi.org/10.1016/j.ebiom.2023.104467.