# Source Matters: Source Dataset Impact on Model Robustness in Medical Imaging

Dovile Juodelyte[1], Yucheng Lu[1], Amelia Jiménez-Sánchez[1], Sabrina Bottazzi[2], Enzo Ferrante[3], and Veronika Cheplygina[1]

[1] IT University of Copenhagen, Denmark
{doju,yucl,amji,vech}@itu.dk
[2] Universidad Nacional de San Martín, Argentina
sbottazzi@estudiantes.unsam.edu.ar
[3] CONICET - Universidad Nacional del Litoral, Argentina
eferrante@sinc.unl.edu.ar

**Abstract.** Transfer learning has become an essential part of medical imaging classification algorithms, often leveraging ImageNet weights. However, the domain shift from natural to medical images has prompted alternatives such as RadImageNet, often demonstrating comparable classification performance. However, it remains unclear whether the performance gains from transfer learning stem from improved generalization or shortcut learning. To address this, we investigate potential confounders – whether synthetic or sampled from the data – across two publicly available chest X-ray and CT datasets. We show that ImageNet and RadImageNet achieve comparable classification performance, yet ImageNet is much more prone to overfitting to confounders. We recommend that researchers using ImageNet-pretrained models reexamine their model robustness by conducting similar experiments. Our code and experiments are available at https://github.com/DovileDo/source-matters.

**Keywords:** Transfer Learning · Classification · Domain Shift · Shortcuts

## 1 Introduction

Machine learning models hold immense promise for revolutionizing healthcare. However, their deployment in real-world clinical settings is hindered by various challenges, with one of the most critical being their hidden reliance on spurious features [24]. Recent research has highlighted the detrimental effects of this reliance, including bias against demographic subgroups [2], limited generalization across hospitals [25], and the risk of clinical errors that may harm patients [17].

Despite transfer learning becoming a cornerstone in medical imaging, its impact on model generalization remains largely unexplored. Pre-training on ImageNet has become a standard practice due to its success in 2D image classification. While some studies have explored alternative medical source datasets for pre-training [3,15,23,26,14], ImageNet continues to serve as a strong baseline.

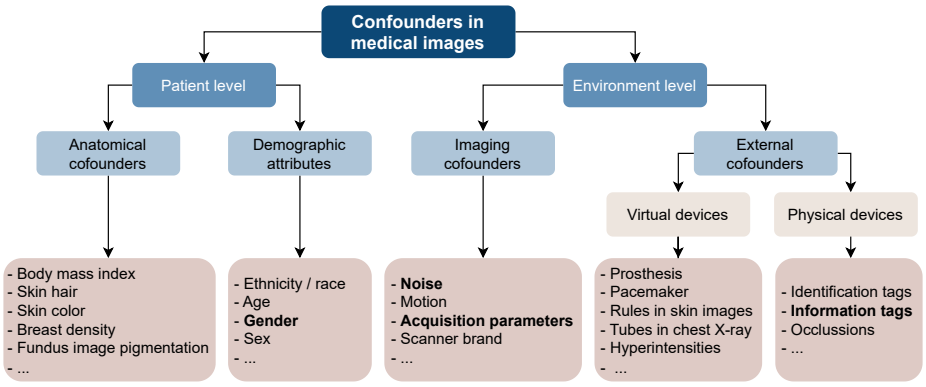



Fig. 1: **MICCAT**: Medical Imaging Contextualized Confounder Taxonomy. Instances of confounders investigated in this paper are highlighted in bold.

Recent literature suggests that the size of the source dataset may matter more than its domain or composition [19,8,6]. However, [13] demonstrated performance improvements through source dataset pruning. In this context, we argue that cross-domain transfer can be problematic, especially when source dataset selection is solely based on classification performance, as it may inadvertently lead to shortcut learning rather than genuine improvements in generalization.

In this paper, we investigate how the domain of the source dataset affects model generalization. First, we conceptualize confounding factors in medical images and systematically assess model robustness by generating synthetic or sampling real-world confounders commonly found in chest X-rays and CT scans. Second, we compare models pre-trained on natural (ImageNet) and medical (RadImageNet) datasets across X-ray and CT tasks and show substantial differences in robustness to shortcut learning despite comparable predictive performance. While transfer learning has been observed to enhance model robustness [11,12], our results suggest that it may not hold true when transferring across domains, cautioning against using ImageNet pre-trained models in medical contexts due to their susceptibility to shortcut learning. Furthermore, our findings highlight the limitations of conventional performance metrics based on i.i.d. datasets, which fail to discern between genuine improvements in generalization and shortcut learning. Thus, we advocate for a more nuanced evaluation of transfer learning effectiveness to ensure the reliability and safety of machine learning applications in clinical settings.

## 2 Method

### 2.1 MICCAT: towards a standardized taxonomy for medical imaging confounders

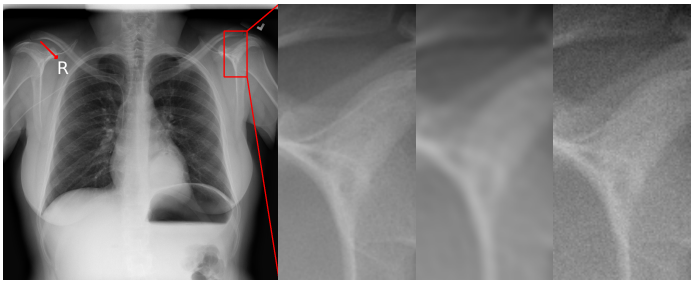
To the best of our knowledge, there is no standardized taxonomy for classifying potential confounders in medical images. Thus, to better structure our robustness analysis, we propose a new taxonomy: Medical Imaging Contextualized Confounder Taxonomy (MICCAT).

Previous work has shown that standard demographic attributes such as sex, age, or ethnicity may act as confounders, leading to shortcut learning and potentially disadvantaging historically underserved subgroups [2]. However, solely focusing on standard protected demographic attributes may overlook other specific factors related to clusters of patients for which the systems tend to fail [7]. In MICCAT, we identify these as 'contextualized confounders', as they are often domain or context-specific, associated with particular image modalities, organs, hospitalization conditions, or diseases.

First, MICCAT differentiates between *patient level* and *environment level* confounders. At the *patient level*, we make a distinction between standard *demographic attributes* (e.g., sex, age, race) and contextualized *anatomical confounders*, which arise from inherent anatomical properties of the organs and human body or disease variations in images. This distinction is crucial as standard demographic attributes often serve as proxies for underlying causes of learned shortcuts. For instance, ethnicity may proxy skin color in dermatoscopic images. Identifying the true shortcut cause allows for more targeted interventions to mitigate biases. We define the concept of *environment level* confounders, which stem from contextualized *external* or *imaging confounders*. The former encompass physical or virtual elements in images due to external factors like hospitalization devices or image tags, while the latter include characteristics related to the imaging modality itself, such as noise, motion blur, or differences in intensities due to equipment or acquisition parameters. Figure 1 illustrates this taxonomy with examples for each category. While we have found over 20 papers with various shortcut examples, space constraints prevent listing them all. We recommend [2,20] for detailed discussions on specific examples and [18] for an interesting analysis of common pitfalls that humans encounter when analyzing medical images, which can lead to shortcut learning.

**Confounders studied in this paper.** We cover the confounder taxonomy by investigating four examples of confounders:

- An external confounder (*a tag*) placed in the upper left corner of the image, representing confounding features introduced by various imaging devices across or within hospitals (Fig. 5a).
- Two typical imaging confounders: *denoising* (Fig. 2c), widely used by various vendors to reduce noise for enhanced readability [9], and *Poisson noise* (Fig. 2d), originating from quantum statistics of photons, which cannot be




(a) (b) (c) (d)

Fig. 2: **Synthetic artifacts**: (a) *A tag* with a red arrow for reference, (b) a zoomed-in view of the original image, (c) *Denoising* by low-pass filter with cutoff frequency (see Eq. 1) of $D_0 = 200$px, and (d) *Poisson noise* with $N_0 = 2 \times 10^6$ (see Eq. 2). The parameters used here are to emphasize subtle local variations such as the smoothing effect of the low-pass filter and the graininess introduced by the Poisson noise. For our experiments, we use $D_0 = 500$px and $N_0 = 2 \times 10^7$ which are imperceptible.

mitigated through hardware engineering, unlike noise introduced by circuit-related artifacts [22].
– A patient-level confounder where we leverage *patient gender*, which is easily accessible in metadata, as a proxy for a broader spectrum of anatomical confounders. We use the same label for this variable as in the original dataset.

### 2.2 Experimental Design

We investigate the impact of source dataset domain on model generalization by comparing ImageNet [5] and RadImageNet [15] models, which are fine-tuned using binary prediction tasks for findings in open-access chest X-ray (NIH CXR14 [21]) and CT (LIDC-IDRI [1]) datasets curated to include systematically controlled confounders.

**Confounder generation.** *A tag* is placed further away from the edges (starting at $200 \times 200$px in the original image of $1024 \times 1024$px), to ensure it remains intact during training despite augmentations applied (Fig. 5a).

The simplest method for *Denoising* is applying low-pass filtering which entails converting the input image from the spatial to the frequency domain using Discrete Fourier Transform (DFT), followed by element-wise multiplication with the low-pass filter $H_{LPF}(u, v)$ to generate the filtered image:

$$H_{LPF}(u, v) = \begin{cases} 1, D(u, v) \leq D_0 \\ 0, \text{otherwise} \end{cases} \quad (1)$$

where $D(u, v)$ represents the distance from the origin in the frequency domain, and $D_0$ is the specified cutoff frequency. In our experiments, we set $D_0 = 500$px. Subsequently, the high-frequency suppressed image is reconstructed

Table 1: Target datasets used for fine-tuning. T: *tag*, D: *denoising*, N: *noise*.

| Task | Confounder | # images in test/dev | % split train/val | % class split pos/neg | Image size | Batch size |
|---|---|---|---|---|---|---|
| Lung mass (NIH CXR14 [21]) | T, D, N | 83/248 | 90/10 | 30/70 | $512 \times 512$ | 32 |
| Lung mass (LIDC-IDRI [1]) | T, D, N | 1710/500 | 80/20 | 50/50 | $362 \times 362$ | 32 |
| Atelectasis (NIH CXR14 [21]) | Gender | 400/400 | 85/15 | 50/50 | $256 \times 256$ | 64 |

in the spatial domain via the Inverse Discrete Fourier Transform (IDFT), resulting in a smoothing effect (see Fig. 2c).

*Poisson noise* originating from quantum statistics of photons is formulated as a Poisson random process:

$$(p_r + N_p) = \mathcal{P}(p_r) \tag{2}$$

where $N_p$ represents Poisson noise, which notably affects image quality under low-dose conditions (e.g., low-dose CT and X-ray screenings), while the linear recording $p_r = \exp(-p_a) N_0$ is obtained via the reversed conversion from attenuation $p_a$ given the prior information of the source intensity $N_0$. To simulate low-dose screening, we add Poisson noise to the image (Fig. 2d), adjusting the $N_0$ parameter to control noise levels. We aim for minimal noise, setting $N_0 = 2 \times 10^7$ after visually examining the noise to ensure it remains barely imperceptible.

*Patient gender* is sampled to correlate 'Female' with the label.

**Evaluation.** To investigate shortcut learning systematically, we construct development datasets for fine-tuning, focusing on a binary classification task. We introduce previously mentioned confounders (e.g., 'Female') into the positive class with a controlled probability $p_{\text{art}} \in \{0, 0.1, 0.2, 0.5, 0.8, 1\}$ to deliberately influence the learning process, replicating scenarios where real-world data may contain confounders. To assess the presence of shortcut learning, we evaluate the fine-tuned models using a dedicated out-of-distribution (o.o.d.) test set. In this test set, we introduce the same artifact used during fine-tuning to the negative class with $p_{\text{art}} = 1$, such that the models are tested on instances where artifacts appear in the opposite class compared to what they encountered during training. Fine-tuned model classification performance is evaluated using the AUC (area under the receiver operating characteristic curve).

**Medical targets.** We create separate binary classification tasks for lung mass detection using subsets of images sourced from two datasets: the chest X-ray NIH CXR14 [21] subset annotated by clinicians [16], and the chest CT dataset LIDC-IDRI [1] annotated by four radiologists. From the latter, we sample paired positive and negative 2D slices from the original 3D scans using nodule ROI annotations, representing any kind of lesions and their nearby slices without remarkable findings. We include synthetic artifacts (*a tag, denoising,* and *Poisson noise*) in both tasks. For the case where patient gender serves as the confounding feature, we sample posterior to anterior (PA) images from the NIH CXR14 dataset [21] to construct a binary classification task for atelectasis. We deliberately limit the size of our development datasets, encompassing both balanced and unbalanced class distributions to cover a spectrum of clinical scenarios.

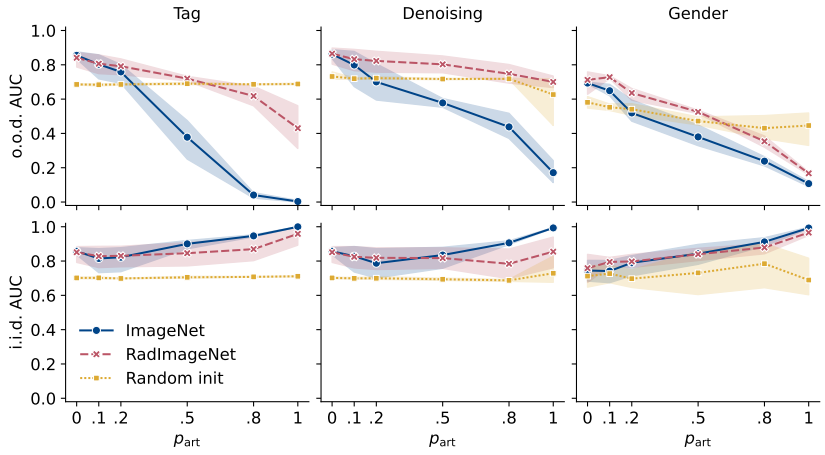



Fig. 3: Mean AUC across five-fold cross-validation with 95% confidence interval for lung mass (left and middle) and atelectasis (right) prediction in chest X-rays. Increasing correlation between artifact (*a tag* (left), *denoising* (middle), or *gender* (right)) and the label leads to decreased o.o.d. performance (on o.o.d. test set as described in Sec. 2.2) (top row), while i.i.d. performance increases (bottom row). RadImageNet's pre-trained models exhibit lesser degradation in o.o.d. performance compared to ImageNet's pre-trained models, suggesting that ImageNet may over-rely on spurious correlations in the target dataset.

Data splits for training, validation, and testing preserve class distribution and are stratified by patient. Further details are available in Table 1.

**Fine-tuning details.** We use ResNet50 [10] as the backbone with average pooling and a dropout layer (0.5 probability). The models are trained using cross-entropy loss with Adam optimizer (learning rate: $1 \times 10^{-5}$) for a maximum of 200 epochs with early stopping after 30 epochs of no improvement in validation loss (AUC for the balanced tasks). This configuration, established during early tuning, proved flexible enough to accommodate different initializations and target datasets. During training, we apply image augmentations including random rotation (up to 10 degrees), width and height shifts, shear, and zoom, all set to 0.1, with a fill mode set to 'nearest'. Models were implemented using Keras [4] library and fine-tuned on an NVIDIA Tesla A100 GPU card.

## 3 Results and Discussion

**RadImageNet is robust to shortcut learning.** Figure 3 shows that both ImageNet and RadImageNet pre-trained models exhibit comparable performance on independent and identically distributed (i.i.d.) test sets, however, when subjected to o.o.d. test sets, notable differences emerge. Specifically, ImageNet's o.o.d. performance on X-rays, confounded by *tag, denoising,* and *patient gender*,

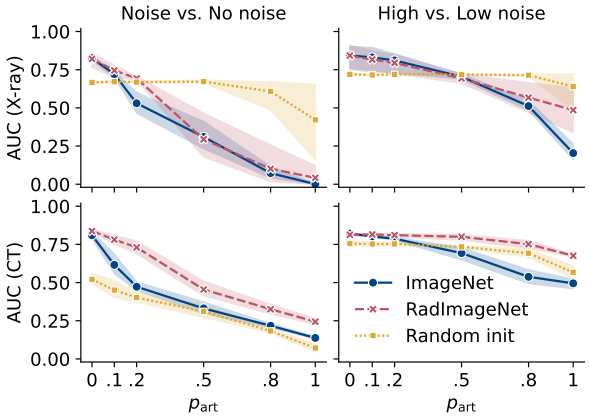



Fig. 4: Mean o.o.d. AUC across five-fold cross-validation with a 95% confidence interval for lung mass prediction in chest X-rays and CT scans. In X-rays (top row), both ImageNet and RadImageNet pre-trained models show similar reliance on Poisson noise. However, RadImageNet is more robust in CT scans (bottom row). When the confounding feature is high vs low noise level, both ImageNet and RadImageNet are less sensitive (right column), compared to noise vs no noise (left column).

drops more compared to RadImageNet, indicating ImageNet's higher reliance on spurious correlations. This could be because certain features, for instance, *a tag* (letters), may serve as a discriminative feature in ImageNet, e.g., for the computer keyboard class. However, RadImageNet is invariant to such features as they are not consistently associated with specific labels across different classes, and this invariance transfers to the target task. Similar trends were observed in the CT dataset, with the o.o.d. AUC decreasing from 0.84 to 0.02 for ImageNet, and to 0.22 for RadImageNet (for *tag*); and from 0.7 to 0.01 for ImageNet, and from 0.83 only to 0.6 for RadImageNet (for *denoising*).

Although *tag* and *denoising* are designed to replicate real-world artifacts, they lack the diversity found in real-world scenarios. *Patient gender* presents a more realistic confounder. Here, the performance gap between ImageNet and RadImageNet is smaller (by 0.12 on average for $p_{\text{art}} \geq 0.1$) yet remains statistically significant (permutation test, $0.008 < p\text{-value} < 0.032$, for $p_{\text{art}} \geq 0.1$, detailed in Table 2 in the supplementary material). This suggests that RadImageNet's resilience to shortcuts extends to more realistic variations in confounders, further emphasizing its robustness in medical image classification.

Random initialization appears to be robust to shortcut learning, with consistent o.o.d. performance as $p_{\text{art}}$ increases. However, this is mainly due to the unbalanced class distribution in the lung mass prediction task within the NIH CXR14 dataset, where randomly initialized models tend to predict the overrepresented negative class. Conversely, in the case of a balanced class distribution

in the CT target dataset, the o.o.d. performance of randomly initialized models deteriorates to a similar degree as that of ImageNet-initialized models.

**Shortcuts come in all shapes and sizes.** ImageNet and RadImageNet both heavily rely on Poisson noise in X-rays (Fig. 4, upper left) but RadImageNet shows greater robustness to noise in CT scans compared to ImageNet (Fig. 4, lower left). It is important to note that Poisson noise manifests differently in X-rays and CT scans. In X-rays, Poisson noise introduces graininess characterized by random and pixel-wise independent variations, while in CT scans, it appears as streak artifacts structurally correlated to projections and thus not random in the image domain (see Fig. 5 in the supplementary material).

To understand the impact of this difference, we directly introduce Poisson noise $N_0 = 2 \times 10^7$ in the image domain for CT scans, mimicking the pixel-wise independence seen in X-rays. However, since CT scans inherently contain noise, this introduces a confounding feature of high versus low levels of noise, as opposed to the original confounder of noise versus no noise.

To simulate a corresponding scenario in X-rays, we generate two levels of Poisson noise: $N_0 = 2 \times 10^7$ for the positive images and $N_0 = 1 \times 10^7$ for the negative images (reversed for the o.o.d. test set). Both models exhibit a smaller drop in o.o.d. performance across modalities, indicating a reduced reliance on the noise shortcut (Fig. 4, right column). This suggests that discerning between high and low noise levels presents a more challenging task than simply detecting the presence of noise.

RadImageNet maintains its robustness in CT scans, while in X-rays, RadImageNet relies on noise to a similar extent as ImageNet. This may be explained by the absence of X-ray images in RadImageNet's pre-training data, leading to a lack of robust X-ray representations that would resist pixel-wise independent noise – a phenomenon less common in CT, MR, and ultrasound, modalities encompassed in RadImageNet. This highlights that even transferring from a medical source of a different modality may lead to overfitting on confounders.

In our exploration of the confounder taxonomy, we found that RadImageNet models are generally more robust to shortcut learning. However, there is some variability within the category of *imaging confounders*, and the importance of the source dataset domain in *anatomical confounders* seems to be lower. Expanding the scope to include more examples of confounders would offer a more comprehensive understanding of the taxonomy landscape and provide insights into the nuances within each category, facilitating better-informed source dataset selection and evaluation strategies.

## 4 Conclusion

Our study sheds light on the critical role that the domain of the source dataset plays in model generalization in medical imaging tasks. By systematically investigating confounders typically found in X-rays and CT scans, we have uncovered substantial differences in robustness to shortcut learning between models pre-trained on natural and medical image datasets. Our findings caution against the

blind application of transfer learning across domains. We advocate for a more nuanced evaluation to improve the reliability and safety of machine learning applications in clinical settings.

**Data use declaration.** This study uses publicly available pre-trained ImageNet [5] (from Keras [4]) and RadImageNet [15] weights as well as publicly available medical imaging datasets NIH CXR14 [21] (available at https://nihcc.app.box.com/v/ChestXray-NIHCC), and LIDC-IDRI [1].

**Acknowledgments.** This study was funded by the Novo Nordisk Foundation (grant number NNF21OC0068816). The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health as well as NIH Clinical Center, and their critical role in the creation of the free publicly available LIDC/IDRI Database and NIH CXR14 dataset used in this study.

## References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P.: Data From LIDC-IDRI [Data set]. The Cancer Imaging Archive (2015)
2. Banerjee, I., Bhattacharjee, K., Burns, J.L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B.N., Shiradkar, R., Gichoya, J.: "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. Journal of the American College of Radiology **20**(9), 842–851 (Sep 2023)
3. Cheplygina, V.: Cats or cat scans: Transfer learning from natural or medical image source data sets? Current Opinion in Biomedical Engineering **9**, 21–27 (2019)
4. Chollet, F., et al.: Keras (2015)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
6. Entezari, R., Wortsman, M., Saukh, O., Shariatnia, M.M., Sedghi, H., Schmidt, L.: The Role of Pre-training Data in Transfer Learning (Mar 2023), arXiv:2302.13602 [cs]
7. von Euler-Chelpin, M., Lillholm, M., Vejborg, I., Nielsen, M., Lynge, E.: Sensitivity of screening mammography by density and texture: a cohort study from a population-based screening program in Denmark. Breast cancer research: BCR **21**(1), 111 (Oct 2019)
8. Gavrikov, P., Keuper, J.: Does Medical Imaging learn different Convolution Filters? (Oct 2022), arXiv:2210.13799 [cs, eess]
9. Hasegawa, A., Ishihara, T., Thomas, M.A., Pan, T.: Noise reduction profile: A new method for evaluation of noise reduction techniques in ct. Medical physics **49**(1), 186–200 (2022)

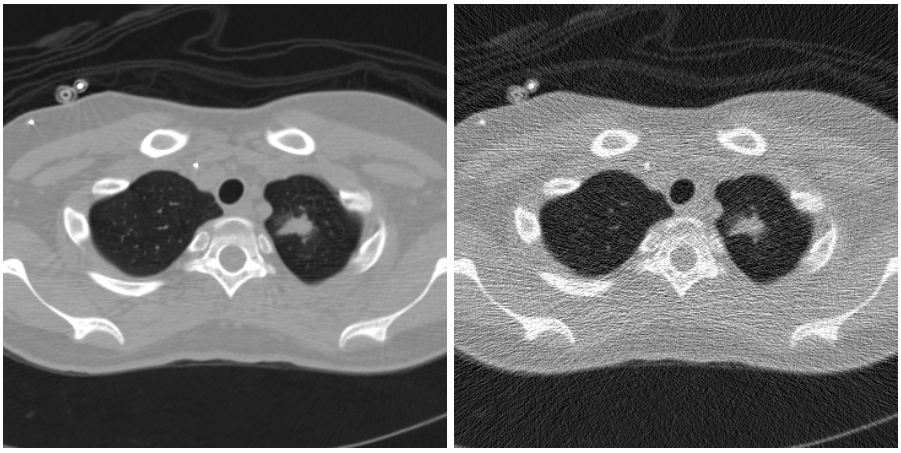
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (Jun 2016)
11. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2712–2721. PMLR (09–15 Jun 2019)
12. Islam, M., Li, Z., Glocker, B.: Robustness Stress Testing in Medical Image Classification. In: Sudre, C.H., Baumgartner, C.F., Dalca, A., Mehta, R., Qin, C., Wells, W.M. (eds.) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 167–176. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023)
13. Jain, S., Salman, H., Khaddaj, A., Wong, E., Park, S.M., Mądry, A.: A data-based perspective on transfer learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3613–3622. IEEE, Vancouver, BC, Canada (Jun 2023)
14. Juodelyte, D., Jiménez-Sánchez, A., Cheplygina, V.: Revisiting hidden representations in transfer learning for medical imaging. Transactions on Machine Learning Research (2023)
15. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y.: RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. Radiology: Artificial Intelligence **4**(5), e210315 (Sep 2022), publisher: Radiological Society of North America
16. Nabulsi, Z., Sellergren, A., Jamshy, S., Lau, C., Santos, E., Kiraly, A.P., Ye, W., Yang, J., Pilgrim, R., Kazemzadeh, S., Yu, J., Kalidindi, S.R., Etemadi, M., Garcia-Vicente, F., Melnick, D., Corrado, G.S., Peng, L., Eswaran, K., Tse, D., Beladia, N., Liu, Y., Chen, P.H.C., Shetty, S.: Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. Scientific Reports **11**(1), 15523 (Sep 2021)
17. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: ACM Conference on Health, Inference, and Learning. pp. 151–159 (2020)
18. Peh, W.C.: Pitfalls in diagnostic radiology. Springer (2014)
19. Ramanujan, V., Nguyen, T., Oh, S., Farhadi, A., Schmidt, L.: On the connection between pre-training data diversity and fine-tuning robustness. Advances in Neural Information Processing Systems **36** (2023)
20. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434. Springer (2023)
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3462–3471. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017)
22. Wei, K., Fu, Y., Zheng, Y., Yang, J.: Physics-based noise modeling for extreme low-light photography. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8520–8537 (2021)
23. Wen, Y., Chen, L., Deng, Y., Zhou, C.: Rethinking pre-training on medical imaging. Journal of Visual Communication and Image Representation **78**, 103145 (Jul 2021)

24. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P.N., Thadaney-Israni, S., Goldenberg, A.: Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine **25**(9), 1337–1340 (Sep 2019)
25. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11), e1002683 (Nov 2018), publisher: Public Library of Science
26. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models Genesis. Medical Image Analysis **67**, 101840 (Jan 2021)

## Supplementary material

Table 2: Permutation t-test for the hypothesis that the o.o.d. AUC of ImageNet is lower than that of RadImageNet for atelectasis prediction in chest X-rays confounded by patient gender.

| $p_{\mathbf{art}}$ | $t$-**statistic** | $p$-**value** |
|---|---|---|
| 0 | -0.5 | 0.817 |
| 0.1 | -4.1 | 0.008 |
| 0.2 | -2.8 | 0.032 |
| 0.5 | -4.2 | 0.008 |
| 0.8 | -4.8 | 0.008 |
| 1 | -5.9 | 0.008 |



(a) Original CT scan (b) Poisson noise

Fig. 5: Here $N_0 = 4 \times 10^3$ to emphasize streak artifacts introduced by the Poisson noise. For our experiments, we use $N_0 = 4 \times 10^{35}$ which is imperceptible.