Look What the Probes Dragged In! Real-World Chest X-ray Shortcuts in MedCLIP

Nikolette Pedersen[1,*], Regitze Sydendal[1,*], Veronika Cheplygina[1], and Théo Sourget[1]

Pattern Recognition Revisited Lab (PURRlab)
IT University of Copenhagen, Denmark
{nizp,resy,vech,tsou}@itu.dk

Abstract. Vision-language models, such as contrastive language-image pre-training (CLIP)-based approaches, have reached state-of-the-art (SOTA) results in medical artificial intelligence. However, recent work reveals that CLIP-based models remain vulnerable to shortcuts. We investigate how real-world shortcuts manifest across different layers of the medical CLIP-based model, MedCLIP, and its vision encoder, a frozen ResNet-50. We attach 17 linear classification probes to the intermediate layers of the ResNet-50 and train them on three different dataset configurations and targets: NIH-CXR14 (pneumothorax) and PadChest (cardiomegaly and pneumothorax). This setup allows us to observe model behaviour during evaluation using subgroup-based calibration and layer-wise confidence curves. We find that the final linear probes achieve a high AUROC but poor calibration in the models. The layer-wise confidence analyses suggest that shortcuts emerge at different depths. Patterns consistent with localised shortcuts, such as drains, appear at later layers, while patterns consistent with diffuse shortcuts, such as scanner-specific noise patterns, emerge earlier, aligning with previous work. Finally, we conduct a manual analysis of the images, which reveals data quality issues in both NIH-CXR14 and PadChest. Our findings underscore that even SOTA models remain vulnerable to shortcuts, and the need for high-quality and well-annotated datasets to draw solid conclusions. Code can be found on our GitHub: `https://github.com/nikodice4/MedCLIP_shortcuts`.

1 Introduction

Vision-language models (VLMs), which leverage both images and medical reports, show high performance in medical AI, yet previous work raises questions about their fairness [9], and their tendency to rely on shortcuts [4] such as chest drains in pneumothorax classification [11].

Recent studies [14,2] investigate these biases and how shortcuts arise during training, yielding results worth further investigation. For VLMs, Sourget et al. [14] show that six contrastive language-image pre-training (CLIP)-based models perform better on chest X-rays with drains than without, indicating

* Equal contribution

shortcut reliance. Three of the models, CXR-CLIP [17], CheXzero [20], and MedCLIP [19], produce predicted probabilities clustered around 0.5 in zero-shot classification, despite CheXzero and MedCLIP achieving area under the receiver operating characteristic curve (AUROC) scores that are within typical ranges for these tasks. For convolutional neural networks (CNNs), Boland et al. [2] investigate layer-wise confidence using linear probes trained on synthetically biased data, and show that models trained on biased data tend to be overconfident [2]. This makes MedCLIP worth exploring, since if a model relies on shortcuts, we would expect overconfidence, rather than near-random probabilities [14]. We explore MedCLIP in a linear probing setting, rather than a zero-shot classification setting.

Given these insights, we investigate MedCLIP's reliance on shortcuts across layers. We attach 17 linear classification probes to the intermediate building blocks in MedCLIP's vision encoder, a frozen ResNet-50, training them on three dataset configurations: pneumothorax classification on NIH-CXR14, and cardiomegaly and pneumothorax classification on PadChest. We then visualise the calibration and layer-wise confidence curves for each probe. Our contributions are as follows: i) implementing linear classification probes in MedCLIP and training the probes on real-world medical data, rather than curated biased data, ii) investigating MedCLIP's vision encoder (ResNet-50), revealing poor calibration and shortcut learning at different layers and iii) a manual analysis of the model predictions, highlighting the importance of data quality.

2 Related Work

Despite deep learning models reaching strong performance on chest X-ray classification [5,13,17], multiple works show that high-performance metrics alone do not reveal whether a model is learning clinically relevant features and could be impacted by shortcut learning [4,18]. A typical shortcut in pneumothorax classification is the presence of a chest drain, which correlates with positive cases of pneumothorax. Oakden-Rayner et al. [11] find that a DenseNet-121 trained on NIH-CXR14 achieves an AUROC of 0.940 on images with drains versus 0.770 without, and Jiménez-Sánchez et al. [7] find the same pattern on CheXpert and NIH-CXR14. Sourget et al. [14] conduct a similar analysis for several CLIP-based models in a zero-shot setting and obtain similar findings, but also show the poor calibration of all models.

Boland et al. [2] study how different kinds of shortcuts emerge at different layer depths by attaching linear probes to intermediate layers of CNNs trained on synthetically biased data. A diffuse shortcut is uniform noise spread across the image, and emerges in the earlier layers, while localised shortcuts, such as a red square placed at a fixed or random location in the images, appear in the later layers. They also introduce bias mitigation strategies targeting the intermediate layers using knowledge distillation from an unbiased teacher model.

Other works describe the quality and challenges of publicly available datasets and how these affect AI models. Rafferty and Rajan [12] identify four major

dataset challenges: label quality and semantic noise from automated annotations, demographic and institutional biases, technical artefacts promoting shortcut learning, and limitations of evaluation practices. Testing across MIMIC-CXR, CheXpert, NIH-CXR14 and PadChest, they find that no architecture proves consistently robust to domain shift, establishing that dataset characteristics are the primary drivers of failure in chest radiography, rather than network design. Jiménez-Sánchez et al. [6] also discuss the current limitations of literature reviews focusing mostly on methods and less on datasets. They propose a living review to track datasets and their research artifacts, such as shortcuts, and discuss important considerations for dataset curation.

Building on the work of Boland et al. [2] and Sourget et al. [14], we aim to identify the depths at which real-world shortcuts emerge in MedCLIP and study its calibration with linear probes. We also perform a manual analysis highlighting data quality issues in large publicly available medical datasets.

3 Investigating MedCLIP's Shortcuts with Linear Probes

3.1 Datasets

NIH-CXR14 This dataset was released in collaboration with the National Institutes of Health (NIH), containing 112,120 chest X-ray images from 30,805 unique patients, collected between 1992 and 2015, with 14 thoracic disease labels mined from reports using natural language processing techniques. We focus on pneumothorax, also known as collapsed lung, which occurs when air leaks into the space between the lung and the chest wall [10].

To analyse drain-related shortcuts, we combine two supplementary annotation datasets. NEATX (Non-Expert Annotations of Tubes in X-rays) by [3] provides manual chest drain annotations for 3,709 positive samples of pneumothorax from NIH-CXR14. Since NEATX only contains positive cases of pneumothorax, we supplement it with automatically generated drain labels from [14], who train a DenseNet-121 on the NEATX labels and apply it to the negative-pneumothorax cases in NIH-CXR14.

PadChest The second dataset is PadChest (Pathology Detection in Chest radiographs), one of the largest open-source datasets containing 160,861 chest X-ray images from 67,625 patients collected at the San Juan Hospital of Alicante in Spain between 2009 and 2017, with multi-labelled annotated reports in Spanish. It covers five different projections: PA, L, AP, AP-horizontal, and COSTAL.

We use two target labels: cardiomegaly and pneumothorax. Cardiomegaly (enlarged heart) [1] was chosen for its higher class prevalence and sex balance across patients. Pneumothorax is heavily underrepresented at 0.4%, but allows for a comparison of the same disease across two datasets.

PadChest also includes three registered patient sexes: "female", "male" and "other". The 14 patients registered as "other" are included in the training set, but not represented in the sex subgroup plots.

Table 1: Complete overview of all datasets, stratified by their splits (training, validation, test) and diagnostic label, with further stratification by the subgroups (sex, scanner, drains) we investigate. Scanner types are abbreviated as IDC (ImagingDynamicsCompanyLtd) and PMS (PhilipsMedicalSystems) and positive and negative are abbreviated as Pos. and Neg.

		Female Patients		**Male Patients**		**Other Patients**		**IDC**		**PMS**		**Drain**		**No Drain**		**Total**	**Pos. Ratio**
Dataset	**Split**	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	N	(%)
NIH-CXR14	Train	29,148	1,111	38,314	1,052	-	-	-	-	-	-	-	-	-	-	69,625	3.1
(Pneumothorax)	Val	7,564	243	8,861	231	-	-	-	-	-	-	-	-	-	-	16,899	2.8
	Test	9,483	1,231	13,448	1,434	-	-	-	-	-	-	10,494	1,692	12,437	973	25,596	10.4
PadChest	Train	30,862	3,518	32,030	2,705	6	1	31,649	3,034	31,249	3,190	-	-	-	-	69,122	9.0
(Cardiomegaly)	Val	7,620	881	7,999	689	4	0	7,917	774	7,707	796	-	-	-	-	17,194	9.1
	Test	9,716	1,044	10,183	822	3	0	10,061	909	9,843	957	-	-	-	-	21,770	8.6
PadChest	Train	34,287	93	34,580	155	7	0	34,654	29	34,220	219	-	-	-	-	69,122	0.4
(Pneumothorax)	Val	8,483	18	8,653	35	4	0	8,685	6	8,456	47	-	-	-	-	17,194	0.3
	Test	10,734	26	10,949	56	3	0	10,966	4	10,722	78	-	-	-	-	21,770	0.4

Preprocessing the Datasets For PadChest, we follow the preprocessing methodology of [16], and remove all vertical X-ray projections, keeping only the following projections: PA, AP, AP-horizontal. We also exclude rows where the disease label is "suboptimal study", "exclude" or "Unchanged". The NIH-CXR14 dataset already provides a train and test split. For PadChest, we split the dataset into a train (80% of the data) and a test split (20% of the data). Both train sets are then split again to use 20% as validation data. All splits were done by patient identifier to avoid any data leakage. All information about the sizes of the datasets and subgroups can be found in Table 1.

3.2 Linear Probe Setting in MedCLIP and Evaluation

MedCLIP is a pre-trained VLM combining vision and language encoders, pretrained on CheXpert, MIMIC-CXR, COVID, and RSNA Pneumonia [19]. The vision and text encoders map inputs to a shared embedding space, where the predicted cosine similarity between projected embeddings is aligned to a Unified Medical Language System (UMLS)-derived similarity matrix. Standard CLIP-based models require paired image-text data, which is scarce in medical settings, and treat semantically similar but unpaired samples as negatives. MedCLIP addresses both issues by decoupling image-text pairs using medical knowledge extracted via the UMLS, which enables training on unpaired data and dramatically increases the dataset size for training.

To investigate how the classification confidence of MedCLIP's vision encoder develops throughout the layers, we follow the methodology of Boland et al. [2]. Each linear probe consists of an average pooling layer, followed by a single-linear fully connected neural network. They attach the probes to the intermediate layers of their CNN architectures.

We attach linear probes to the output of each building block of our frozen ResNet-50 architecture, and one final probe is placed at the last average pooling layer, providing the final predictions for the input. In total, we train 17 probes. To assess the calibration of the model, we use the predictions from the final probe, the main classification task being disease vs. no disease.

We train the attached linear probes three separate times, each time with a different dataset configuration. Only the weights of the linear probes are updated during training, as the frozen ResNet-50 captures representations from MedCLIP's pre-training. The training loss is the sum of the cross-entropy losses across all 17 linear probes (16 are intermediate, plus the final probe on the average-pooled output). Since the backbone is frozen and probes do not share weights, each probe receives gradients only from its own loss term. All probes were trained using the same hyperparameters: epochs: 100, batch size: 32, learning rate: 0.00001. To mitigate overfitting, we monitor the summed validation loss after each epoch with early stopping after 15 epochs without improvement.

3.3 Calibration and Confidence Curves

We plot the mean predicted probabilities from the final probe against the fraction of true positive cases of the given disease in each bin to produce our calibration curve. A perfectly calibrated model is one whose predicted probabilities match the true frequency of a given disease. This means, among all cases assigned a probability of p, a fraction p are actually positive.

To quantify how classification confidence develops throughout the model, we follow the methodology of [2], who define the model's confidence, $C(X)$, as the deviation from the maximum uncertainty of 0.5, where a higher value indicates a greater certainty in the prediction. [2] find that the model becomes more overconfident when training on curated biased data than when training on clean data. As we are not dealing with synthetically biased data, we do not expect to see the exact same results. Following the taxonomy from [8], physical devices like chest drains are categorised as external shortcuts. We therefore hypothesise that we will see spikes in the later layers, as the chest drains act more like a localised shortcut than a diffuse one, since a physical device aligns with [2]'s synthetic red squares rather than the synthetic noise. In contrast, the different X-ray machines are categorised as imaging shortcuts and can produce image noise, which we hypothesise will align with a diffuse shortcut. However, the IDC images also contain an "R" marker in the X-rays, which can resemble a localised shortcut.

4 Results

4.1 Good Overall AUROC Across Dataset Configurations

Figure 1 shows the final probes' AUROC scores across all test set configurations. Pneumothorax on NIH-CXR14 achieves a global AUROC score of 0.839, with a small gap of 0.028 between drains (0.840) and no drains (0.812). There is no notable difference between the female and male patients' AUROC scores. Cardiomegaly on PadChest achieves the highest global AUROC score across all three classification tasks, at 0.905. The subgroup IDC achieves the highest AUROC score across all dataset configurations and subgroups, reaching 0.917,

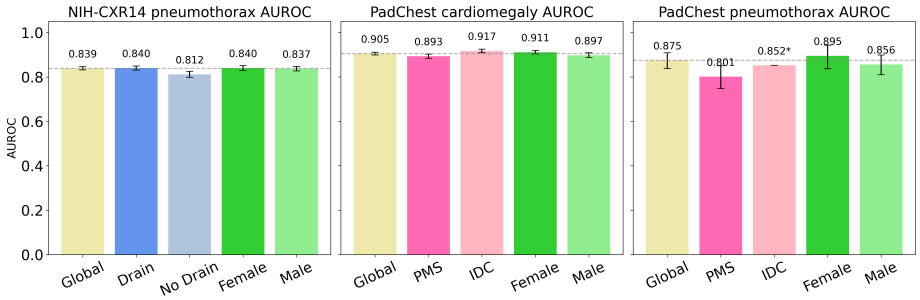


Fig. 1: AUROC scores across all dataset configurations and their given subgroups, with the dashed lines indicating the global AUROC scores. The confidence intervals are displayed as black error bars. "*" means that there is no confidence interval as there are not enough positive samples for the bootstrapping.

while its counterpart, PMS, reaches 0.893. Pneumothorax on PadChest achieves a global AUROC score of 0.875. PMS and IDC have AUROC scores of 0.801 and 0.852, respectively, with PMS having the lowest AUROC score across all subgroups. Further, we see large confidence intervals, which is due to the small number of positive cases in this dataset configuration.

4.2 Generally Poorly Calibrated Models

Figure 2 shows calibration curves and confidence curves, with each column corresponding to one of the three dataset configurations. Pneumothorax on NIH-CXR14, split by whether a patient has a drain, shows the model is poorly calibrated and overconfident, though marginally better calibrated for the drain subgroup. For cardiomegaly on PadChest, split by X-ray machine, we see that this is the best calibrated model, and it is also the dataset configuration with the most positive cases. Still, the model is overconfident, and we see no real difference between the subgroups. Pneumothorax on PadChest is the most miscalibrated model. The IDC subgroup is the worst calibrated. This is notably also the subgroup with the least positive cases. For the PadChest pneumothorax patient sex subgroup, it looks similar to its scanner machine counterpart, and there are no differences between the two subgroups. Overall, every model is miscalibrated to some degree, and the number of positive cases for the dataset configurations visibly influences the curves.

4.3 Confidence Curves Show Signs of Shortcuts

For pneumothorax on NIH-CXR14, split by drain status, all four curves stay low and stable across the first 13 building blocks. The groups diverge from block 13, where the positive cases (drain in blue and no drain in orange) reach high

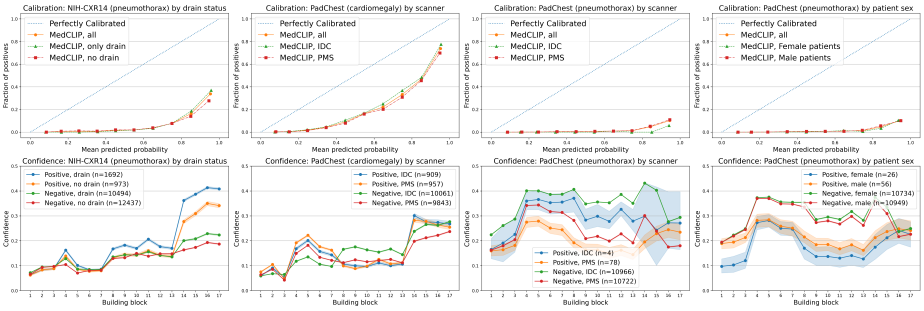
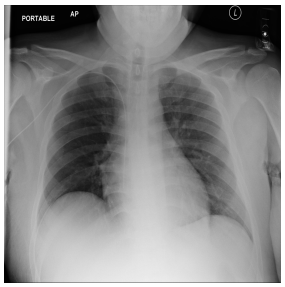


Fig. 2: Calibration (top) and confidence (bottom) curves across all dataset configurations and their given subgroups. In the calibration curves, the x-axis shows the mean predicted probability per bin and the y-axis the fraction of positive cases. The diagonal represents perfect calibration, with deviations above or below the diagonal indicating miscalibration. In the confidence curves, the x-axis represents the individual building blocks with an attached probe, and the y-axis shows the mean confidence of the model's predictions, with 95% confidence intervals obtained via 1,000 bootstrap resamples per probe.

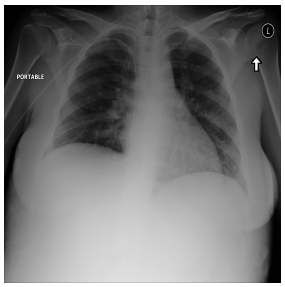
confidence between 0.35 and 0.41, while the negative cases stay below 0.25, however the green curve (drain) remains higher than the red curve (no drain). These findings align with [2], who found localised shortcuts spike in later layers. Cardiomegaly on PadChest, split by X-ray machine, shows all four curves remain low, spiking between building blocks 4 and 6 and again at building block 13. All four lines lie in the same range and yield a relatively low confidence, around 0.23-0.28, therefore these spikes do not resemble the patterns of shortcuts. Pneumothorax on PadChest shows spikes in the earlier layers. With only 4 positive pneumothorax cases, the IDC curve (in blue) is very uncertain. All lines spike at building block 3 and afterwards remain stable in their trajectory. The IDC curve (in green) for the negative cases peaks at 0.42, but all curves end at similar levels. These spikes do resemble the diffuse shortcuts of [2], which further aligns with previous research that also finds how different X-ray machines in an unbalanced setting can lead to shortcuts [15]. However, the IDC X-rays also have an "R" marker in the corner, which on its own can resemble a localised shortcut. Therefore, the plots are not enough evidence to conclude whether the shortcuts are diffuse or localised.

5 Discussion and Conclusions

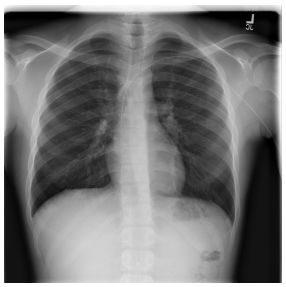
Looking at the confidence throughout the building blocks of the vision encoder, we find signs of shortcut learning in the MedCLIP model, which aligns with previous research [14]. This suggests that, despite their scale and multi-modality, MedCLIP, and other CLIP-based models, are not inherently robust, even though they are state-of-the-art models. Furthermore, with these findings, we show ev-



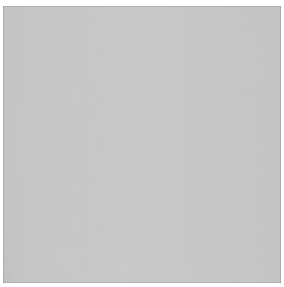
(a) NIH-CXR14 (image ID 00020945_039.png) has no pneumothorax. A small drain can be seen in the top left corner, but it was annotated as having no drain.



(b) NIH-CXR14 (image ID 00013377_010.png) has no pneumothorax. A small drain can be seen in the top left corner, but it was annotated as having no drain.



(c) NIH-CXR14 (image ID 00017318_015.png) has no pneumothorax. A small drain can be seen in the top right corner, but it was annotated as having no drain.



(d) NIH-CXR14 (image ID 00010007_121.png) is a non-uniform grey image.

(e) PadChest (image ID 216840111366964012339356563862009068132653048_00-050-189.png) is not a chest X-ray, but an X-ray of a skull, annotated with lung diseases.

Fig. 3: Selected standout images identified during the manual analysis.

idence that the synthetic shortcuts presented by [2] behave in a similar way as real-world shortcuts in medical datasets. We also find poor calibration in all models, aligning with the results of [14]. Our manual and exploratory data analysis uncovered several errors: for NIH-CXR14, we find cases of incorrectly entered patient ages and a greyed-out image. For PadChest, patients have conflicting sexes, duplicate images, duplicate metadata, as well as an X-ray of a skull. Further, we found the automatically annotated drains of the negative cases of pneumothorax in NIH-CXR14 were not completely reliable. From visually inspecting the images, we found cases where the images had been labelled with "no drain" but the images do actually contain a drain. This was the case for 3/5 of the images we visually examine. These standout images that we found can be seen in Figure 3.

Due to the data quality issues, the results should be interpreted carefully, as the conclusions are drawn from datasets that contain label and metadata errors. Further, our results also highlight the limitations of methods such as calibration and confidence curves, which can be severely influenced by the number of positive cases available, such as in PadChest for pneumothorax classification. Moreover, as shortcuts are dataset-specific, experiments on other datasets may yield different findings. Similarly, we conducted our analysis on a single CLIP-based model and vision encoder. Conducting our experiments on other architectures and datasets would therefore help to improve the generalisation of our results. Finally, it would be beneficial to extend this analysis beyond CNN-based vision encoders to more recent architectures such as Transformers.

To conclude, we applied the methodology of linear probes in the intermediate layers from [2] to a new model architecture, namely MedCLIP, to study how real-world shortcuts emerge in the vision encoder across three different dataset configurations: NIH-CXR14 for pneumothorax, and PadChest for cardiomegaly and pneumothorax. All dataset configurations showed miscalibrated models. Some confidence curves revealed that the confidence did not develop in a stable manner across the network, but aligned with [2], who state that different shortcuts manifest at different depths of the network. We revealed unreliable drain annotations for the negative-pneumothorax cases, and errors in the publicly available datasets NIH-CXR14 and PadChest. These findings further highlight that scale, architecture and high AUROC scores alone do not make SOTA models robust to shortcuts. The errors found underscore the need for high data quality to make reliable conclusions.

Acknowledgments. VC and TS were supported by Novo Nordisk Foundation grant NNF24OC00926.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Amin, H., Siddiqui, W.: Cardiomegaly. Stat-Pearls [Internet] Treasure Island (FL): StatPearls Publishing (2022)
2. Boland, C., Tsaftaris, S.A., Dahdouh, S.: Preventing shortcut learning in medical image analysis through intermediate layer knowledge distillation from specialist teachers. Machine Learning for Biomedical Imaging **3**, 447–476 (2025). https://doi.org/https://doi.org/10.59275/j.melba.2025-8888, https://melba-journal.org/2025:020
3. Cheplygina, V., Damgaard, C., Eriksen, T.N., Juodelyte, D., Jiménez-Sánchez, A.: Augmenting chest x-ray datasets with non-expert annotations. In: Ali, S., Hogg, D.C., Peckham, M. (eds.) Medical Image Understanding and Analysis. pp. 133–144. Springer Nature Switzerland, Cham (2026)
4. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z, https://doi.org/10.1038/s42256-020-00257-z
5. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nature Reviews Cancer **18**(8), 500–510 (2018)
6. Jiménez-Sánchez, A., Avlona, N.R., de Boer, S., Campello, V.M., Feragen, A., Ferrante, E., Ganz, M., Gichoya, J.W., Gonzalez, C., Groefsema, S., Hering, A., Hulman, A., Joskowicz, L., Juodelyte, D., Kandemir, M., Kooi, T., Lérida, J.d.P., Li, L.Y., Pacheco, A., Rädsch, T., Reyes, M., Sourget, T., van Ginneken, B., Wen, D., Weng, N., Xu, J.J., Zajaç, H.D., Zuluaga, M.A., Cheplygina, V.: In the Picture: Medical Imaging Datasets, Artifacts, and their Living Review. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. pp. 511–531. ACM, Athens Greece (Jun 2025). https://doi.org/10.1145/3715275.3732035, https://dl.acm.org/doi/10.1145/3715275.3732035
7. Jiménez-Sánchez, A., Juodelyte, D., Chamberlain, B., Cheplygina, V.: Detecting shortcuts in medical images - a case study in chest x-rays. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023). https://doi.org/10.1109/ISBI53787.2023.10230572
8. Juodelyte, D., Lu, Y., Jiménez-Sánchez, A., Bottazzi, S., Ferrante, E., Cheplygina, V.: Source matters: Source dataset impact on model robustness in medical imaging. In: Wu, S., Shabestari, B., Xing, L. (eds.) Applications of Medical Artificial Intelligence. pp. 105–115. Springer Nature Switzerland, Cham (2025)
9. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., Fang, Y., Wang, M.: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12289–12301 (2024)
10. McNamara, C., Fhlatharta, M.N., Eaton, D.A.: Pneumothorax and chest drain insertion. Surgery (Oxford) **44**(3), 197–202 (2026). https://doi.org/https://doi.org/10.1016/j.mpsur.2026.02.003, https://www.sciencedirect.com/science/article/pii/S0263931926000128, cardiothoracic Surgery II
11. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 151–159. Association for Computing Machinery (Apr 2020). https://doi.org/10.1145/3368555.3384468, https://doi.org/10.1145/3368555.3384468

12. Rafferty, A., Rajan, A.: Limitations of public chest radiography datasets for artificial intelligence: Label quality, domain shift, bias and evaluation challenges (2026), https://arxiv.org/abs/2509.15107
13. Shen, J., Zhang, C.J., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P.H., Ming, W.K., et al.: Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR medical informatics **7**(3), e10010 (2019). https://doi.org/10.2196/10010
14. Sourget, T., Restrepo, D., Hudelot, C., Ferrante, E., Christodoulidis, S., Vakalopoulou, M.: Fairness and robustness of clip-based models for chest x-rays. In: Puyol-Antón, E., Ferrante, E., Feragen, A., King, A., Cheplygina, V., Ganz-Benjaminsen, M., Glocker, B., Petersen, E., Lee, H. (eds.) Fairness of AI in Medical Imaging. pp. 11–21. Springer Nature Switzerland, Cham (2026)
15. Sourget, T., Claßen, N., Xu, J.J., van der Goot, R., Cheplygina, V.: Dataset diversity metrics and impact on classification models (2026), https://arxiv.org/abs/2603.15276
16. Sourget, T., Hestbek-Møller, M., Jiménez-Sánchez, A., Junchi Xu, J., Cheplygina, V.: Mask of truth: Model sensitivity to unexpected regions of medical images. Journal of Imaging Informatics in Medicine (2025). https://doi.org/10.1007/s10278-025-01531-5, https://doi.org/10.1007/s10278-025-01531-5
17. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (Sep 2022). https://doi.org/10.1038/s41551-022-00936-9, https://doi.org/10.1038/s41551-022-00936-9
18. Vásquez-Venegas, C., Wu, C., Sundar, S., Prôa, R., Beloy, F.J., Medina, J.R., McNichol, M., Parvataneni, K., Kurtzman, N., Mirshawka, F., et al.: Detecting and mitigating the clever hans effect in medical imaging: A scoping review. Journal of Imaging Informatics in Medicine pp. 1–17 (2024). https://doi.org/10.1007/s10278-024-01335-z
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.256, https://aclanthology.org/2022.emnlp-main.256/
20. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)