The effect of dataset confounding on predictions of deep neural networks for medical imaging

Beatriz Garcia Santa Cruz*[1,2], Andreas Husch[2], and Frank Hertel[1,2]

[1]National Department of Neurosurgery, Centre Hospitalier de Luxembourg, Luxembourg
[2]Interventional Neuroscience group, Luxembourg Center for Systems Biomedicine, University of Luxembourg, Luxembourg

Abstract

The use of Convolutional Neural Networks (CNN) in medical imaging has often outperformed previous solutions and even specialists, becoming a promising technology for Computer-aided-Diagnosis (CAD) systems. However, recent works suggested that CNN may have poor generalisation on new data, for instance, generated in different hospitals. Uncontrolled confounders have been proposed as a common reason. In this paper, we experimentally demonstrate the impact of confounding data in unknown scenarios. We assessed the effect of four confounding configurations: total, strong, light and balanced. We found the confounding effect is especially prominent in total confounder scenarios, while the effect on light and strong confounding scenarios may depend on the dataset robustness. Our findings indicate that the confounding effect is independent of the architecture employed. These findings might explain why models can report good metrics during the development stage but fail to translate to real-world settings. We highlight the need for thorough consideration of these commonly unattended aspects, to develop safer CNN-based CAD systems.

1 Introduction

The use of Machine Learning (ML) and Deep Learning (DL) in medicine is very promising to improve patient care. Such solutions are applied to many medical areas like oncology and neurology. One of the most promising use cases includes assisting the radiologists in the diagnosis process. DL is expected to provide more accurate, faster and objective (in that it reports quantitative analysis) diagnosis [12]. However, these systems might fail to translate into real-world scenarios, presenting multiple challenges for safe applications [19]. It has been reported that ML-based health systems produce systematic errors on patient subgroup classification, consequently generating wrong predictions and flawed risk estimations [21]. Such systematic errors could originate at any stage of developing pipeline, from dataset generation, model development, model evaluation until its final deployment [21]. In a recent example, a systematic review that analysed prediction models for the diagnosis and prognosis of COVID-19 pneumonia reported that almost all of the published models for prediction were poorly documented. Consequently, such models have a high risk of associated bias and their performance was overrated and led to poor generalisation [23]. In the same vein, ref. [4] systematically reviewed the publicly available X–Ray imaging datasets employed to build such models. This work suggested that, without well-documented datasets and/or complementary metadata, models may learn induced bias or uncontrolled confounders as strong features during the model training, which hampers their safe translation into clinical practice. Previous works demonstrate that in case of potential confounding scenarios, shortcuts can have a variable effect on

*Corresponding Author:garciasantacruz.beatriz@gmail.com

the model generalisation [2]. Considering the potential harm, further analysis of the confounding variables with respect to the model generalisation capacity is needed.

1.1 Bias and confounders in DL

The problems of bias and confounders in DL are becoming more prominent due to the harm and long impact effects they may have in high-stakes disciplines, such as health-care, education or justice [16], especially in underrepresented groups [13]. Bias can be defined as a systematic error presented in the data that may result in wrong predictive estimations. In this context, particularly relevant are selection bias and collider bias, where a population subgroup with certain characteristics (e.g. age, gender) is more likely to be selected, having an increased presence in the dataset compared to its presence in the general population. Induced associations between variables may thus affect the sampling likelihood of an individual [5]. Additionally, confounding factors are variables that influence both the predictor and the outcome [20]. The presence of uncontrolled confounders leads to spurious associations, hampering generalizability and transportability [4].

In medical imaging, such difficulties are often augmented by data scarcity, population and prevalence shifts. Common practices to circumvent such issues include mixing datasets from different populations and/or training models with populations that are different from the target population. Nonetheless, these practices can lead to learn spurious correlations from the confounding factors originated from differences in the dataset generative process (e.g. data acquisition devices, population characteristics) [1]. Considering the impact of such errors, induced systematic bias needs further analysis for applications to medical imaging. In this paper, we study the impact of potential unknown confounders caused by dataset composition. To this end, we focus on pneumonia datasets as a case study.

1.2 Pneumonia X-ray manifestations

Pneumonia is an infection mainly caused by bacteria or viruses, that manifests in inflammation of the lungs *alveoli*, which fill with fluid or pus that cause painful breathing. It is especially dangerous in children, older adults and immunosuppressed patients, causing over 15% of deaths in children under 5 years old worldwide [15]. The differential diagnoses of pneumonia includes examination of a chest radiograph (CXR) by trained specialists, often accompanied with the corresponding clinical history and laboratory tests. Pneumonia generally manifests in CXR as an area or areas with increased opacity [3]. An example of CXR is presented in Figure 1. Pneumonia diagnosis on CXR can be hampered by several conditions, including pulmonary *oedema*, *atelectasis*, or lung cancer as well as other characteristics such as patient position, or depth of inspiration [9].

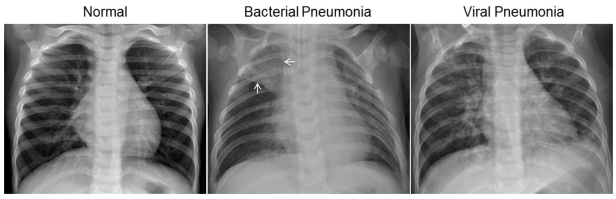

Figure 1: Example of CXRs, the left side panel present a normal case, middle and left panels, pneumonia cases with visible opacities, from: [11]

2 Material and Methods

2.1 Dataset 1: Guangzhou Women and Children's dataset

This CXR dataset contains anterior-posterior images selected from a retrospective study of a paediatric cohort of patients from one to five years old from the Guangzhou Women and Children's Medical Centre at Guangzhou, China. It contains categorical labels for Pneumonia and normal. The labelling involved human expert grading conducted by two specialists [10, 11].

2.2 Dataset 2: RSNA Pneumonia Detection Challenge

The RSNA Pneumonia Detection Challenge dataset is a subset of the NIH CXR14 dataset, which comes from the NIH Clinical Center, United States of America [22]. The labelling involved six human board-certified specialists. Labels consisted of binary classification in positive and negative based on previous findings [14]. The additional annotations from the positive class are not employed in this work.

2.3 Experimental design

To study the effect of potential unknown confounders in a CNN classification network for medical imaging, we simulated different scenarios with different degrees of confounding. For the sake of simplicity we considered only two classes as target labels control and disease (pneumonia) and one confounding factor: the age. Age is divided into two groups: children (from dataset 1) and adults (from dataset 2). Note that, despite we focus on age as the crucial differential factor of the samples coming from these two datasets, other sources of variations such as scanner, protocol acquisition and image preprocessing may also affect the results but are not addressed in this conceptual study.

2.3.1 Dataset preprocesing

All samples were combined into a new dataset with four classes: child control [**C0**] (n= 4273), child disease [**C1**] (n= 1584), adult control [**A0**] (n= 20672) and adult disease [**A1**] (n= 2614). Next, the majority classes of the dataset were randomly under-sampled to match the size of the minority class (n= 1584), creating a balanced dataset. Then, 10% of the dataset was separated to create another balanced dataset for external test. The remaining 90% was used to create 7 different combinations, with different degrees of confounding.

2.3.2 Confounding datasets

Table 1 summarises the composition of each confounding dataset combination.

Table 1: Confounding dataset combinations. Ratio of samples of each class derived from each dataset.

	Children dataset		Adult (RSNA) dataset	
	Disease (C1)	Control (C0)	Disease (A1)	Control (A0)
Total 1	0.501	0	0	0.499
Total 2	0	0.499	0.501	0
Strong 1	0.474	0.024	0.024	0.474
Strong 2	0.024	0.474	0.475	0.024
Light 1	0.425	0.074	0.074	0.425
Light 2	0.074	0.423	0.425	0.074
Balanced	0.25	0.25	0.25	0.25

- **Total confounding:** All the disease samples derive from one age group class while all the control samples derive from the other age group.
- **Strong confounding:** Most samples (95%) from one class (control or disease) derive from one age group (children or adult) and vice-versa.
- **Light confounding:** As described for strong confounder but with an 85%.
- **Balanced:**. The categories are class balanced.

2.4 Training and evaluation

Each experiment proceeds as follows: a model is trained on one combination, then evaluated against the validation set (internal test) and the test set (external test). Each model employs 80% of the dataset for training and 20% for validation. Networks were training using the already per-train models from torch-vision during 15 epoch using the Leslie Smith's 1 cycle policy [18]. Each experiment was conducted 5 times (each time, the validation set was a different random combination). To study the effect of the architecture, the whole study was conducted with three different standard architectures: Resnet50 [6], Densenet121 [7] and squeezenet1.1 [8]. After the internal and external evaluation, next each metrics per class children and control was analysed. In the remainder of the manuscript, we use AUC (area under the ROC curve), disease and control recall at 50% as representative evaluation metrics.

3 Experimental results

This section is structured as follows; In the first place, we presented the results of the seven combinations evaluated with the internal dataset to assess the network accuracy with respect to a dataset with the same confounding scenario, additionally in order to explore the generalisation capacity of the network, an external dataset that represents a balanced scenario was employed. Further, we present the result with more granularity to understand the different behaviour based on the dataset (children vs adult) and the class (disease vs control). Finally, we presented the variation of AUC, disease and control recall with respect to the balanced dataset. Since no significant differences were found across different architectures, we present the detailed results from Resnet50.

In Figure 2, general AUC (on the left), we can observe an overall good performance across confounding combinations, with a general tendency to score higher in strong and total confounding levels in the internal dataset. Nonetheless, the values drop, especially on these combinations, when the model is evaluated using an external dataset. Slight differences in performance can be observed in both combinations of light and strong confounding. These stand out further in the next two charts (middle, and right), showcasing recall for control class and disease class, respectively.

Such variations are explained by analysing the children and adult classes separately. In Table 2, results from the children and the adult classes are presented. The dark colour indicates the majority class for each imbalanced combination and light for the minority. Here, we can observe that samples from the children datasets generally score higher than the adults', not only when they represent the majority class but also in minority conditions. Hence, it seems that the adults' class is more affected by the confounding degree in light and strong confounding situations.

In the two cases of total confounding, the disease and control cases proceeded from distinct age groups (i.e. datasets), respectively. In one combination, the disease cases originate from the children dataset and the control cases from the RSNA dataset. Conversely, the second total confounding combination employed disease cases from the RSNA dataset, and control cases from the children dataset. In both cases, the external test evaluation shows that the network fails to predict examples that were not present during the training. Therefore, the model was not able to generalise beyond the provided examples. This behaviour is further discussed in Section 4.

Next, to assess the variation between disease and control, and between each level of confounding, the variation with respect to the balanced network was analysed for each class as depicted in Figure 3. Negative values represent a drop in performance with respect to balanced. Such drops are more common for the adults´ class in light and strong scenarios, but similar in total confounding scenarios for both classes. Positive values represent an increase in performance, but at the cost of a reduced performance with respect to the other category.

Since many ML research papers commonly approach the performance optimisation on exploring the different architectures rather than datasets. We aim to understand if the behaviour of confounding is affected by the different architectures. Figure 4 depicts the AUC for Validation (Internal test) and Test (External test) as well as the difference between the first two charts. We can observe almost identical performance independently of the architecture employed to build the model.

4 Discussion

This section discusses the confounding effects produced by the seven different combinations of datasets in our experiments.

Confounding effect in internal evaluation. Models reported acceptable scores when evaluated against the validation set (Fig. 2, dark blue), or another dataset with the same confounding characteristics. This is expected, but does not warrant similar performance in external datasets.

Confounding effect in external evaluation. When evaluated against a external balanced dataset, the model's scores drop (Fig. 2, light blue), showcasing a general tendency to reduce the accuracy with

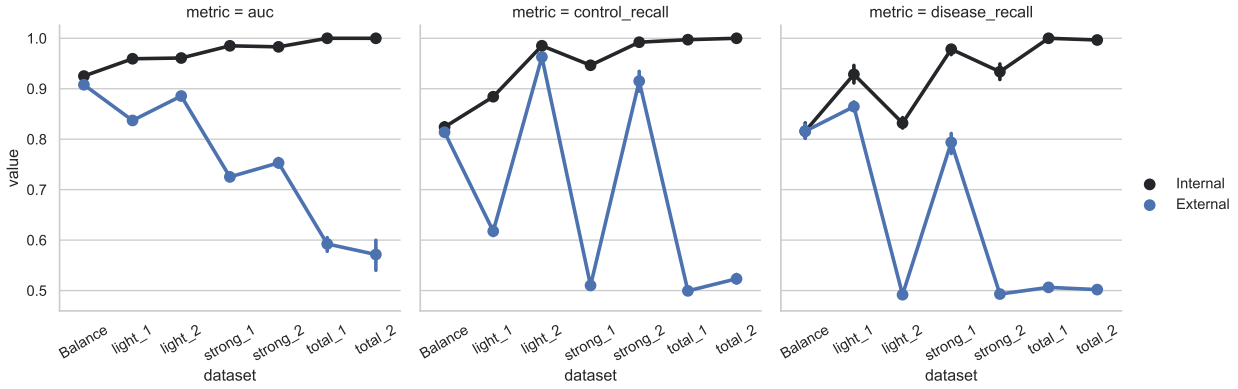

Figure 2: AUC (left panel), control recall (middle), and disease recall (right) metrics from the internal and external evaluation for each confounding dataset combination. Points represent the mean value and the line the standard deviation.

Table 2: External test metrics per children and adult class. Light confounding in blue, strong in green and total in yellow. Dark colour represents the majority class in imbalanced combinations.

		Balanced	**Light_1**	**Light_2**	**Strong_1**	**Strong_2**	**Total_1**	**Total_2**
Children class	AUC	0.9888 ± 0.002	0.9892 ± 0.003	0.9896 ± 0.002	0.9826 ± 0.003	0.9801 ± 0.003	0.8122 ± 0.003	0.7654 ± 0.138
	control recall	0.9686 ± 0.0072	1 ± 0.0072	0.9434 ± 0.0158	1 ± 0.0158	0.8553 ± 0.0457	1.0000 ± 0.0056	0.0377 ± 0.0224
	disease recall	0.8797 ± 0.011	0.7722 ± 0.0259	0.9430 ± 0.0072	0.5949 ± 0.0546	0.9684 ± 0.0137	0.0063 ± 0.0148	1.000 ± 0.000
Adult class	AUC	0.7700 ± 0.0079	0.7484 ± 0.0178	0.7323 ± 0.0078	0.6841 ± 0.0235	0.6956 ± 0.0152	0.5554 ± 0.0421	0.5351 ± 0.0445
	control recall	0.6604 ± 0.0056	0.2327 ± 0.0242	0.9874 ± 0.0082	0.0189 ± 0.0113	0.9937 ± 0.0053	0.0000 ± 0.0028	1.0000 ± 0.0034
	disease recall	0.7532 ± 0.0325	0.9620 ± 0.0045	0.0380 ± 0.0161	1.0000 ± 0.0028	0.0127 ± 0.0035	1.0000 ± 0.0000	0.0063 ± 0.0035

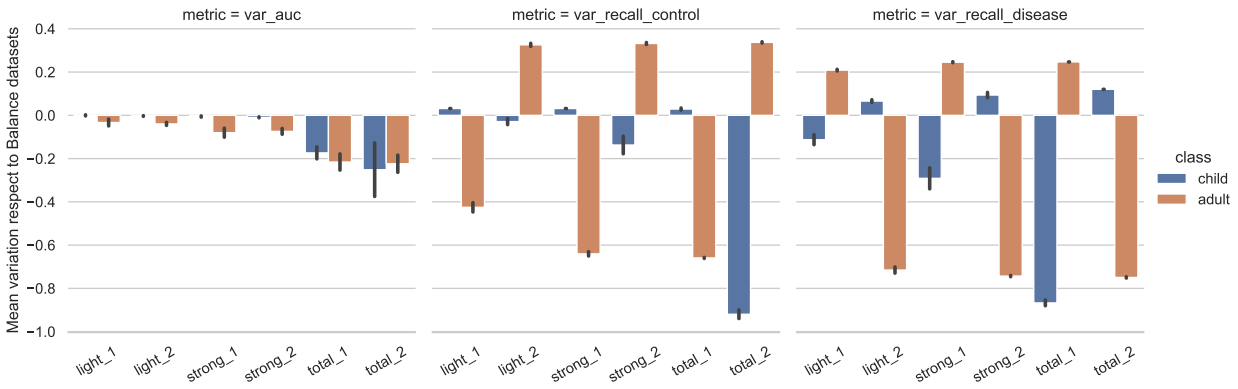

Figure 3: Mean of the variation with respect to the balanced dataset for the external test evaluation. Error bars represent the standard deviation.

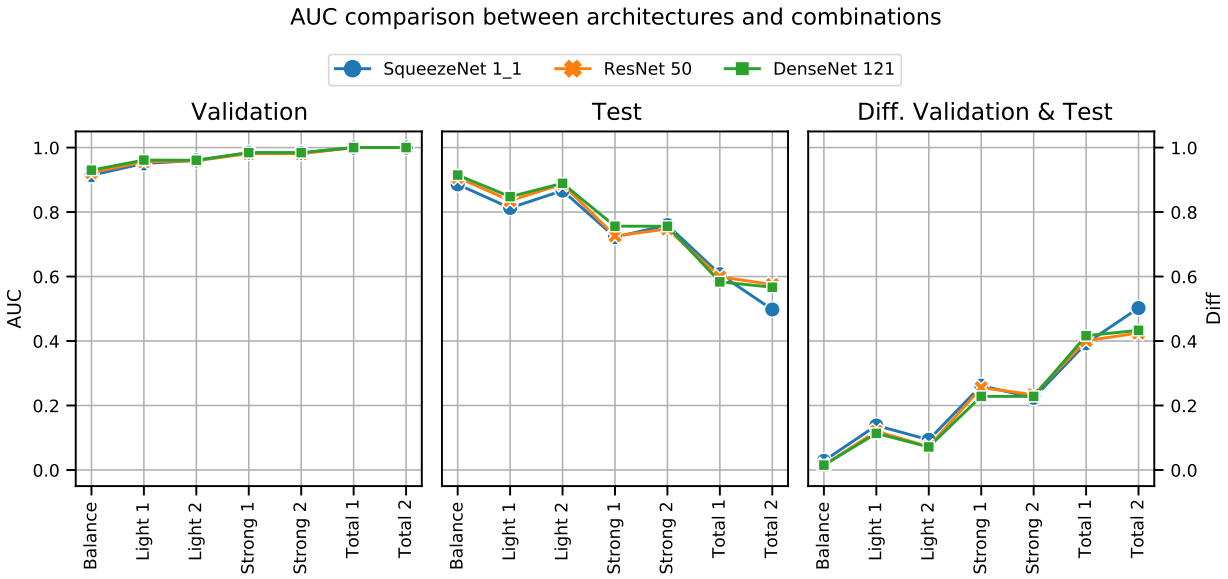

Figure 4: Performance comparison between three different architectures of the seven confounding combinations. Left and centre charts represent the AUC of the network in validation (internal test) and test (external test). The right chart depicts the difference between the previous two charts.

respect to the confounding degree. This may explain why a model can perform well during the development phase but fail during the deployment phase.

Dataset behaviour with respect to the confounding degree. Each dataset has specific characteristics that conferred different robustness against confounding conditions. In our preliminary results, the children class presented a better adaptation to such changes than the adults class. In cases of total confounding (see Table 2) both classes presented similar effects but in the light and strong confounding conditions the more robust dataset (children) was better generalised by the model even with fewer samples. More studies are recommended to better understand this phenomenon. We conjecture that, the more homogeneous and coherent the dataset, the less probably the confounding effect is learned as a strong feature, leading to better generalisation characteristics.

Effect of the confounding degree. In case of **total confounding**, the model fails with a *class recall* of zero for the unseen class and almost 1 for the seen class, suggesting that the models use the age or other dataset-specific characteristics as a learning shortcut (Table 2). This suggests that, in cases of total confounded datasets (for instance when each class stems from a different dataset), the model may have a strong probability of failing to predict the unseen class. This lack of generalisation can result in dangerous translation issues when deployed in real settings. A prominent case of this was the combination of controls from the Guangzhou Women and Children's dataset and adult disease samples (COVID-19 pneumonia), which has been reported to be the most common combination in peer-reviewed publications [4].

For the **strong confounding** scenario, all metrics report lower scores with respect to the balanced models, but the negative effect can differ based of the dataset robustness as discussed before. This affects

the disease class more than the control, which might be explained by the higher variability in this class due to the disease manifestation induced diversity. The adults class effect is close to the total confounding scenario, suggesting that unknown confounders due to an uncontrolled class imbalance can lead to dangerous settings for training clinical models.

In **light confounding** conditions, the metrics report lower scores with respect to the balanced models but higher that in the strong confounder (when comparing against the same group age). In the adults group, the effect on disease recall when the training datasets has a presence of just 15% is similar to the total confounder, suggesting similar conclusions than for the strong confounding.

Architecture impact The employed architectures (Figure 4) were not found to have an impact on the confounding effect. These results are in line with other works which emphasise the need for more data work [16] to improve ensure safe and generalisable models.

5 Conclusion and future work

Model robustness and transferability are key requirements for safe clinical models. This work presents an experimental assessment of the effect of the confounder in CNN-based solution for medical image analysis. Both the confounding degree and the dataset characteristics seem to influence the effect of potential uncontrolled confounders in models. These results demonstrate the hazardous effect of confounders when applied to high-stake domains such as healthcare. While many papers focused on a model-centre approach employing benchmark datasets, it is also crucial to consider other data-centre aspects in other to develop safe solutions.

Additionally, these results could explain why some models perform well even in confounding scenarios when the test employed contains comparable confounding characteristics, but fail to translate to different settings such as a hospital or new sampling strategy where a model is deployed to. Importantly, the external evaluation is also susceptible to present confounding characteristics. A better understanding of the dataset generation process (e.g. through better documentation) could help mitigate these issues.

Overall, these problems highlight the necessity of designing an appropriate strategy for model testing and auditing for future clinical use. The employment of metadata seems to have a relevant role in the control for potential confounders. Metadata can be employed during the design and evaluation phases[17]. For the former, it can be used to ensure a balanced class presence in the datasets. In the latter, it can help conduct a disaggregated evaluation to ensure the model fairness and performance for each subgroup of interest.

To confirm and expand these preliminary results, we plan to extend our study, including more confounding factors, imaging modalities and medical disciplines. Further, we aim to have a better undertaking of the phenomena using some of the existing tools within the framework of explainable AI and model robustness.

6 Acknowledgements

Beatriz Garcia Santa Cruz work is supported by the FNR within the PARK-QC DTU (PRIDE17/12244779/PARK-QC) and Pelican award from the Fondation du Pelican de Mie et Pierre Hippert-Faber. The authors would like to thank Dr. Jan Soelter and Salah Ghamizi, MsC for discussions, Dr. Carlos Vega for technical support as well as editing and to Daniele Provervio, MsC for support on manuscript editing.

References

[1] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[2] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.

[3] T. Franquet. Imaging of community-acquired pneumonia. *Journal of thoracic imaging*, 33(5):282–294, 2018.

[4] B. Garcia Santa Cruz, M. N. Bossa, J. Sölter, and A. D. Husch. Public covid-19 x-ray datasets and their impact on model bias – a systematic review of a significant problem. *Medical Image Analysis*, 74:102225, 2021.

[5] G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. D. Smith, et al. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, 11(1):1–12, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[9] B. Kelly. The chest radiograph. *The Ulster medical journal*, 81(3):143, 2012.

[10] D. Kermany, K. Zhang, M. Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2), 2018.

[11] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[12] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.

[13] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[14] RSNA. Rsna pneumonia detection challenge, 2018.

[15] P. Rui, K. Kang, and M. Albert. National hospital ambulatory medical care survey: 2015 emergency department summary tables. *National center for health statistics*, 2017.

[16] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[17] B. G. Santa Cruz, C. Vega, and F. Hertel. The need of standardised metadata to encode causal relationships: Towards safer data-driven machine learning biological solutions. *Computational Intelligence Methods for Bioinformatics and Biostatistics 2021,10.5281/zenodo.5729350 Nov 16, 2021*, 2021.

[18] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[19] C. Vega. From Hume to Wuhan: An Epistemological Journey on the Problem of Induction in COVID-19 Machine Learning Models and its Impact Upon Medical Research. *IEEE Access*, 9:97243–97250, 2021.

[20] E. Vittinghoff, C. E. McCulloch, D. V. Glidden, and S. C. Shiboski. 5 linear and non-linear regression methods in epidemiology and biostatistics. *Handbook of statistics*, 27:148–186, 2007.

[21] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):1–3, 2021.

[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[23] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.