Removal of Confounders via Invariant Risk Minimization for Medical Diagnosis

Samira Zare[1][0000−0002−7828−0562] and Hien Van Nguyen[1]

University of Houston, Houston, TX, USA
szare836@uh.edu

Abstract. While deep networks have demonstrated state-of-the-art performance in medical image analysis, they suffer from biases caused by undesirable confounding variables (e.g., sex, age, race). Traditional statistical methods for removing confounders are often incompatible with modern deep networks. To address this challenge, we introduce a novel learning framework, named ReConfirm, based on the invariant risk minimization (IRM) theory to eliminate the biases caused by confounding variables and make deep networks more robust. Our approach allows end-to-end model training while capturing causal features responsible for pathological findings instead of spurious correlations. We evaluate our approach on NIH chest X-ray classification tasks where sex and age are confounders.

Keywords: Confounder Removal · Invariant Learning · Chest X-ray.

1 Introduction

Deep neural networks for medical data analysis suffer from undesirable biases created by confounding variables. These models can use the confounders as shortcuts to establish spurious connections between clinically meaningless variables and diagnosis outcomes [15,8,17,25]. For example, when we train a model on a chest radiograph dataset where male patients are more likely to have a disease, the classifier can use breast regions to predict the absence of the disease. This behavior leads to a higher error rate for female patients [12]. Fig. 1 provides empirical evidence to illustrate this problem. Specifically, while breast features are not clinically meaningful for performing chest radiograph diagnosis, deep networks rely heavily on these features as indicated by the visualization (see section 3 for more experimental details). Other factors such as age [25,8], text markers on images [5,8], testing condition [6], institutions where data were collected, or even the pubertal development scale of subjects [25] may all function as confounders and impact the model performance. Relying on these confounders will significantly reduce the model's generalizability and reliability [12,20] because the model does not capture true pathologies underlying the diagnostic outcomes.

Frequent deep learning methods simply assume that if confounders such as gender are not directly included to the training dataset, we can remove their effects. A key issue, however, is that strongly correlated features may act as proxies

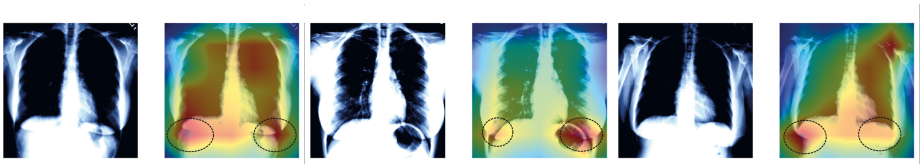

Fig. 1. Activation heat-map visualization for a DenseNet121 trained on a Chest-Xray dataset confounded by sex. As marked in images, the model considers the breasts as abnormalities. This confounding effect culminated in higher false positive rate for females.

for confounders (e.g., inferring gender from chest x-ray images based on anatomical differences) [16]. Traditional statistical methods to mitigate the effects of confounders contain pre-processing, matching or stratification of input variables or post-processing of the outputs [1,14]. However, because of the demand for end-to-end training schemes and large training datasets, these approaches have fallen out of favor with the machine learning community. How to eliminate confounding variables in deep networks remains an important research question. Recent methods use adversarial domain-invariant representation learning to remove the confounder [25]. However, when the class-conditional distributions of input features change between source and target domains, this approach cannot guarantee successful generalization [24].

Invariant Risk Minimization (IRM) [4] is a robust framework for learning invariant features in multi-source domain adaptation. IRM relies on the idea that invariance is a proxy for causality; the causal components of the data remain constant (*invariant*) across environments. Although powerful, IRM requires a discrete set of *environments*, each corresponding to specific data distribution, during the training stage. As a result, one cannot directly apply this method to the confounder removal problem where pre-defined environments are unavailable.

To this end, we propose a novel strategy to optimally split the training dataset into different environments based on the available confounders (e.g., age, sex, test condition). We note that our method does not need the confounders during testing the model. Our experiments show that the generated environments facilitate IRM to learn features highly invariant to confounding variables. In addition, the original IRM formulation enforces the same conditional distribution among all classes, which potentially leads to learning unstable features as discussed in [3]. We propose the first conditional IRM method for medical images to relax this assumption and enable the model to learn more robust features specific to each diagnosis. The main contributions of this paper are as follows:

- We develop a novel confounder removal framework based on the invariant risk minimization theory. We extend this framework to accommodate class-conditional variants, where the invariance learning penalty is conditioned on each class.

- We design a strategy for optimally splitting the dataset into different environments based on the maximum violation of the invariant learning principle.
- We compare the classification performance and visualization of our method to baseline CNN models trained under the traditional empirical risk minimization framework.

Related Works. Many recent studies in the medical domain raise the concern that deep learning models may have biases in their predictions. Larrazabal et al. [12] and Seyyed Kalantari et al. [20] provided experimental evidence that gender, race, age, and socioeconomic status of the patients confound the model predictions. They suggested the models should be trained on a multi-source balanced dataset to minimize the bias. While promising, creating large and diverse medical datasets is time-consuming and expensive. Examples of other techniques for removing the confounding factors are reweighting [21] and stratifying the batch sampling [17]. We note that these methods do not learn the invariant or causal medical pathologies underlying the disease. It can lead to a performance drop in the test environments where the spurious associations are different. Zhao et al. [25] used adversarial domain-invariant representation learning to remove the confounder via a min-max training approach. As mentioned, one major limitation of the adversarial learning method is that we cannot guarantee their successful generalization [24]. It is also possible for the predictor to move in a direction that helps the adversary decrease its loss [23]. Invariant Risk Minimization (IRM) [4] was proposed to overcome the limitations of adversarial domain-invariant representation learning. In this work, we are interested in evaluating the ability of IRM to remove the confounding effects in medical data analysis. Adragna et al. [2] previously utilized IRM to achieve fairness in comment toxicity classification. Our work is different in two ways: i) we introduce a strategy to define training environments when we have a collection of data, ii) we propose the class-conditional invariance learning penalty; our goal is to remove the association between the confounder and model prediction in a more proper way, as we will describe. We also note Rosenfeld et al. [18] argue that IRM does not improve over standard Empirical Risk Minimization in the non-linear setting. However, their theorem only indicates the existence of a non-invariant model that approximately satisfies the IRM criterion. This is similar to showing the existence of neural network parameters that fit the training set but do not generalize to the test set. This is not sufficient to question the efficiency of the method [11].

2 Removal of Confounders via Invariant Risk Minimization (ReConfirm)

In this section, we describe our approach for REmoval of CONFounders via IRM (ReConfirm). We first introduce the IRM approach from the perspective of confounder removal. Then, we will describe our strategy to split a dataset into training environments and our extensions to IRM. Suppose we have a set of training environments $e \in \mathcal{E}_{tr}$ and in each environment, datasets $D_e := \{(x_i^e, y_i^e)\}_{i=1}^{N_e}$

are generated from the same input and label spaces $\mathcal{X} \times \mathcal{Y}$ according to some distribution $P(X^e, Y^e)$. The environments differ in how the labels are spuriously correlated with the confounder variable c. For instance, in environment e_0, older people have a higher probability of being diseased, while in environment e_1, this association is reversed. We aim to find a predictor function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that generalizes well for all confounded environments. Empirical Risk Minimization (ERM), which is the standard method for solving learning problems, minimizes the average loss over all training samples from all environments. On the other hand, IRM looks for features that remain invariant across training environments and ignores the confounding associations specific to each environment. It assumes that f is a compositional model $f := \omega \circ \Phi$, where $\Phi : \mathcal{X} \rightarrow \mathcal{Z}$ is the invariant data representation and $\omega : \mathcal{Z} \rightarrow \mathcal{Y}$ is the classifier on top of Φ. Using this notation, IRM minimizes the total risk of all training environments through a constrained optimization problem:

$$\begin{aligned} &\min_{\Phi,\omega} \sum_{e \in \mathcal{E}_{tr}} R^e(\omega \circ \Phi) \\ \text{subject to} \quad &\omega \in \underset{\bar{\omega}}{\operatorname{argmin}}\ R^e(\bar{\omega} \circ \Phi), \ \text{for all } e \in \mathcal{E}_{tr} \end{aligned} \tag{1}$$

where $R^e := \mathbb{E}_{(X^e,Y^e)\sim D^e}[\mathcal{L}(f(X^e), Y^e)]$ is the risk for environment e and $\mathcal{L}$ is the loss function (e.g. cross entropy). To solve this bi-level optimization problem, the authors in [4] simplified Eq. 1 to:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} R^e(\omega \circ \Phi) + \lambda \|\nabla_{\omega|\omega=1.0} R^e(\omega \circ \Phi)\|^2 \tag{2}$$

where λ is a regularization hyperparameter and ω is a fixed dummy variable [4]. This formulation enables end-to-end network training. While IRM is a compelling approach for domain-generalization problems, we cannot immediately apply it to a medical dataset to remove the confounder variables. IRM requires a set of environments to find features that remain invariant across them. Thus, to take advantage of IRM in discovering the underlying medical pathologies in a confounded dataset, we have to create training environments. Environments should share the same underlying biomarkers that we expect our model to learn. However, they should differ in how the confounders generate spurious correlations. In what follows, we introduce our strategy to split the dataset to ensure learning invariant features while ignoring the confounding effects.

Creating training environments for ReConfirm. To use IRM to remove the spurious effects of confounders, we have to partition our dataset into different environments. Following [4,11], we use only two training environments: e_0 and e_1. The class label and the confounder have strong but spurious correlations in each environment. To construct the environment, we use the *invariant learning principle*. Specifically, ω is simultaneously optimal for all environments due to the constraint in Eq. 1. In addition, for regular loss functions like the cross-entropy and mean squared error, optimal classifiers can be expressed as conditional expectations of the output variable. Therefore, an invariant data

representation function Φ must satisfy the below invariant learning principle:

$$\mathbb{E}[Y|\Phi(x) = h, e_0] = \mathbb{E}[Y|\Phi(x) = h, e_1] \tag{3}$$

where h is in the intersection of the supports of $\Phi(X^e)$. This condition means that IRM learns a set of features such that the conditional distribution of the outcomes given the predictors is invariant across all training environments. Given a collection of data, we seek partitions that maximally violate the invariance principle. This would be equivalent to the worst-case scenario on which we would like to train our model. We note that Creager et al. [7] show that maximizing the invariance principle violation corresponds to maximizing the regularization term in Eq. 2. Motivated by this observation, we maximize the following gap between our environments to find optimal environment splitting:

$$e_0, e_1 = \underset{e_0,e_1}{\operatorname{argmax}}\, g = \underset{e_0,e_1}{\operatorname{argmax}} \left|\mathbb{E}[Y|\Phi(x) = h, e_0] - \mathbb{E}[Y|\Phi(x) = h, e_1]\right| \tag{4}$$

Suppose our input X-ray (MRI, CT scan, etc.) images contain both authentic medical biomarkers and spurious confounding features. The worst-case classifier only relies on the confounder variable c to make predictions, i.e. $\Phi(x) = c$. Then the gap would be $g = \left|\mathbb{E}[Y|c, e_0] - \mathbb{E}[Y|c, e_1]\right|$. Therefore, we are looking for an environment partitioning strategy that maximizes g. One possible solution would be to define the environments based on the agreement between the confounder and the label [7]. This means, for example, in environment e_0, all diseased patients ($y = 1$) are old ($c = 1$) and healthy controls ($y = 0$) are young ($c = 0$), while in environment e_1, all diseased patients ($y = 1$) are young ($c = 0$), and healthy controls ($y = 0$) are old ($c = 1$). In this case, we have the maximum confounding association in each environment, and the gap is $g = 1$, which is its maximum value. The proof can be found in the supplementary materials.

Class-conditional ReConfirm. While the IRM framework promotes invariant learning, it applies one penalty term to all classes (see Eq. 2). This formulation potentially leads to features that are less stable for each class [3]. To address this issue, we propose to use class-conditional penalties. Our formulation can correct confounding effects specific to each class while potentially promoting more diverse features. For instance, age (as a confounder) can impact healthy and diseased patients differently. Rather than learning a global effect as in IRM, our class-conditional method separately models each class's confounding influences. This way, we encourage our model to learn invariant medical pathologies only related to the diseased class (or healthy class as well), thus diversifying the learned features. The mathematical formulation of class-conditional ReConfirm is given as follows:

$$\min_{\Phi} \sum_{e\in\mathcal{E}_{tr}} R^e(\omega \circ \Phi) + \sum_{j=0}^{M-1} \sum_{e\in\mathcal{E}_{tr}} \lambda \|\nabla_{\omega|\omega=1.0} R^{e,j}(\omega \circ \Phi)\|^2 \tag{5}$$

where M is the number of classes and $R^{e,j}(\omega \circ \Phi)$ denotes the risk corresponding to samples from j-th class. A more systematic way to deal with the confounders

is to remove their direct influence on the predictions. For example, we can remove the common effects of age or gender on chest X-ray images or brain MRIs while preserving their actual effects on the development of a disease [25]. We can condition our ReConfirm regularization only on the control class to achieve this behavior.

3 Results and Discussion

Dataset. We particularly evaluate the efficacy of our model on binary classification (normal versus abnormal) of frontal chest radiographs using the NIH Chest-Xray14 (NIH-CXR) dataset [22]. The abnormalities include 14 abnormal findings defined in [22]. This dataset contains age (0–95) and sex (Male and Female) as meta variables. It is shown that the deep learning models have poor performance on female patients and younger patients [12,20]. Given that some anatomical attributes (probably considering the breasts as abnormalities) and age-related alterations [9,13] are reflected in X-ray images, sex and age could be considered as confounders. In order to highlight the ability of our model to mitigate the effect of confounders on predictions, we created a subset of the NIH-CXR dataset where age and gender are confounders for the label. Fig. 2 and 3 show the age and gender distributions within each class.

Implementation Details. We used a densely connected CNN (DenseNet) architecture [10], which has been shown to accomplish state-of-the-art results in X-ray image classification. We used the ImageNet setting to normalize all of the images, and standard online transformations to augment images while training. We also had the Adam optimizer with standard settings, a batch size of 96, a learning rate of 1×10^{-4}, and image size of 224×224. We use the same training settings across all models to ensure fair comparisons. The penalty coefficient λ and the number of training epochs to linearly ramp it up to its full value are among the hyperparameters for the ReConfirm method. Similar to [4,11], we found ramping in λ over several epochs and scaling down the ERM term by the penalty coefficient when it is more than 1 are useful for stable training. All codes will be available at `https://github.com/samzare/ConfounderRemoval` for research purposes.

Sex as Confounder. In our dataset, male patients are more probable to be in the diseased class, as shown in Fig. 2 a. In order to remove the confounding association, we utilize ReConfirm and conditional variants. First, we construct our training environment based on the setting described in Fig. 2 b. In environment e_0 all females are in the control group, while in e_1 this correlation is reversed. We have the ERM model as the baseline where the confounder effects are not removed. We expect that ERM predictions rely more on spurious correlations and consequently have lower performance compared to the variants of ReConfirm methods. We implement ReConfirm, class-conditioned ReConfirm conditioned on all classes (cReConfirm (all)), and only on the control class (cReConfirm (y=0)). Conditioning on all classes would encourage the model to specifically learn features that are invariant (among males and females) in each class. Also,

Table 1. Prediction performance of models trained on sex-confounded dataset.

	Whole Cohort			Male			Female		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ERM	84.23	0.9010	0.7690	83.66	0.9051	0.7521	84.79	0.8971	0.7859
ReConfirm	86.69	**0.9278**	0.7958	85.92	**0.9352**	0.7718	87.46	0.9209	0.8197
cReConfirm (all)	86.41	0.9045	**0.8141**	85.63	0.9042	**0.7972**	87.18	0.9049	**0.8310**
cReConfirm (y=0)	**87.18**	0.9244	0.8099	**85.92**	0.9153	0.7915	**88.45**	**0.9333**	0.8282

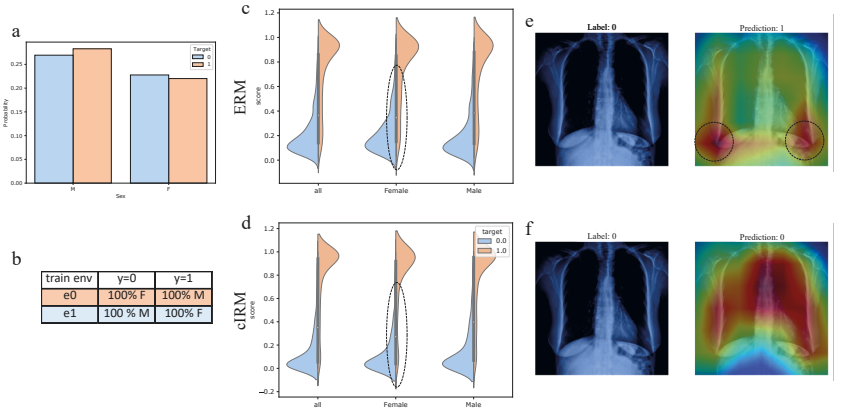

Fig. 2. Difference in sex distribution between normal and abnormal classes resulted in the baseline ERM learning the confounding effects, while cReConfirm removed this effect: **a.** sex distribution for each class, **b.** the ReConfirm environment setting, **c, d.** distribution of prediction scores. **e, f.** qualitative visualization of the learned features.

conditioning only on the control group helps the model learn the "normal" effects of sex that can be observed in the X-ray images (like the breasts). Table 1 lists the perfromance of all models for the balanced whole cohort, female and male patients separately. In the whole cohort setting, we have both male and female patients from both classes. We then restricted the test dataset to one sex to have a closer look at the model performance in each subgroup. We also have the results from worst-case scenarios where sex and class labels are strongly correlated in the supplementary materials. Overall, ReConfirm variants could achieve better performance compared to the ERM method. We could achieve an accuracy of 87.18% on the whole cohort with cIRM conditioned on the control group. To investigate more, we also have the score distributions (Fig. 2 c, d) and GRAD-CAM [19] visualizations (Fig. 2 e, f) for ERM and cReConfirm conditioned on the control group. Note that the score distributions show that the cReConfirm model is more confident in its predictions for all experiments (we used markers to illustrate the difference in Fig. 2). Activations also illustrate how the baseline ERM model is confounded by the sex-related anatomical features in the X-ray. As discussed by Larrazabal et al. [12], female patients are more likely to have false positive predictions. Our visualizations can explain this confounding effect; the ERM model considered the breasts as abnormalities (see Fig. 3 e), while

Table 2. Prediction performance of models trained on age-confounded dataset.

	Whole Cohort			Young			Old		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ERM	82.77	0.884	0.7561	85.67	0.8481	0.7585	80.75	0.9012	0.7551
ReConfirm	86.05	0.8854	**0.8296**	**88.25**	0.8617	**0.8226**	84.52	0.8964	**0.8328**
cReConfirm (all)	85.46	0.9212	0.7771	86.53	**0.9091**	0.7170	84.72	0.9261	0.8041
cReConfirm (y=0)	**86.22**	**0.9237**	0.7911	87.97	0.8986	0.7698	**85.02**	**0.9349**	0.8007

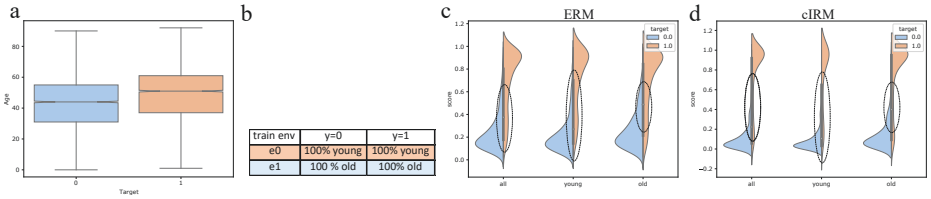

Fig. 3. Difference in age distribution between normal and abnormal classes confounded the baseline ERM while cReConfirm removed this effect: **a.** age distribution for normal and abnormal classes, **b.** the ReConfirm environment setting, **c, d.** distribution of prediction scores for ERM and cReConfirm conditioned on control class.

using ReConfirm, we can learn more meaningful features (Fig. 3 f). While ERM misclassified this case, cReConfirm could find more proper biomarkers and correctly classified the image. More results and visualizations can be found in the supplementary materials.

Age as Confounder. The average age of the diseased class is higher than the healthy controls (48.73 ± 16.77 vs. 42.61 ± 15.65), as shown in Fig. 3 a. We construct our training environments based on the same strategy and the setting we used is described in Fig. 3 b. In order to define the young and old groups, we set a threshold of 40 years old on the age; Seyyed Kalantari et al. [20] have shown that age groups under 40 experienced the highest performance disparities.

Table 2 lists the performance of all models for the balanced whole cohort, younger and older patients separately. The results from worst-case scenarios where age and class labels are strongly correlated are in the supplementary materials. We have ReConfirm, cReConfirm conditioned on all classes, and cReConfirm conditioned on the control group. Conditioning on all classes would preserve the diversity of learned features, while conditioning only on the control group would encourage the model to learn the normal aging effects on the lungs that can be captured in the X-ray images and confound the model. Overall, ReConfirm variants could achieve better performance compared to the ERM method. We have an accuracy of 86.22% on the whole cohort with cReConfirm conditioned on the control group. The score distributions (Fig. 3 c, d) for ERM and cReConfirm conditioned on the control show that our cReConfirm model is more confident in its predictions for all experiments. More results and visualizations can be found in the supplementary materials.

4 Conclusion

In this work, we applied IRM to a binary chest X-ray classification task in order to evaluate its ability in removing the effect of confounder variables. We also proposed two variants of class-conditional ReConfirm; conditioning on all classes results in similar predictive behaviour in each class among different values of confounders, and conditioning only on control class help us to remove direct associations while preserving the indirect ones. Our experiments show that ReConfirm is significantly more robust than ERM against undesirable confounders.

Acknowledgements This research was funded by the National Science Foundation (1910973).

References

1. Adeli, E., Kwon, D., Zhao, Q., Pfefferbaum, A., Zahr, N.M., Sullivan, E.V., Pohl, K.M.: Chained regularization for identifying brain patterns specific to hiv infection. Neuroimage **183**, 425–437 (2018)
2. Adragna, R., Creager, E., Madras, D., Zemel, R.: Fairness and robustness in invariant learning: A case study in toxicity classification. arXiv preprint arXiv:2011.06485 (2020)
3. Ahmed, F., Bengio, Y., van Seijen, H., Courville, A.: Systematic generalisation with group invariant predictions. In: International Conference on Learning Representations (2020)
4. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
5. Badgeley, M.A., Zech, J.R., Oakden-Rayner, L., Glicksberg, B.S., Liu, M., Gale, W., McConnell, M.V., Percha, B., Snyder, T.M., Dudley, J.T.: Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ digital medicine **2**(1), 1–10 (2019)
6. Bustos, A., Payá, A., Torrubia, A., Jover, R., Llor, X., Bessa, X., Castells, A., Carracedo, Á., Alenda, C.: xdeep-msi: Explainable bias-rejecting microsatellite instability deep learning system in colorectal cancer. Biomolecules **11**(12), 1786 (2021)
7. Creager, E., Jacobsen, J.H., Zemel, R.: Environment inference for invariant learning. In: International Conference on Machine Learning. pp. 2189–2200. PMLR (2021)
8. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
9. Gossner, J., Nau, R.: Geriatric chest imaging: when and how to image the elderly lung, age-related changes, and common pathologies. Radiology Research and Practice **2013** (2013)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning. pp. 5815–5826. PMLR (2021)

12. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
13. Mihara, F., Fukuya, T., Nakata, H., Mizuno, S., Russell, W., Hosoda, Y.: Normal age-related alterations on chest radiography: A longitudinal investigation. Acta Radiologica **34**(1), 53–58 (1993)
14. Mohamad Amin, P., Ahmad Reza, B., Mohsen, V.: How to control confounding effects by statistical analysis (2012)
15. Pearl, J.: Causality. Cambridge university press (2009)
16. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568 (2008)
17. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 413–423. Springer (2021)
18. Rosenfeld, E., Ravikumar, P., Risteski, A.: The risks of invariant risk minimization. arXiv preprint arXiv:2010.05761 (2020)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
20. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. In: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium. pp. 232–243. World Scientific (2020)
21. Wang, H., Wu, Z., Xing, E.P.: Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. In: BIOCOMPUTING 2019: Proceedings of the Pacific Symposium. pp. 54–65. World Scientific (2018)
22. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
23. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
24. Zhao, H., Des Combes, R.T., Zhang, K., Gordon, G.: On learning invariant representations for domain adaptation. In: International Conference on Machine Learning. pp. 7523–7532. PMLR (2019)
25. Zhao, Q., Adeli, E., Pohl, K.M.: Training confounder-free deep learning models for medical applications. Nature communications **11**(1), 1–9 (2020)