A Flexible Framework for Simulating and Evaluating Biases in Deep Learning-Based Medical Image Analysis

Emma A. M. Stanley[1,2,3,4(✉)], Matthias Wilms[3,4,5,6], and Nils D. Forkert[1,2,3,4,7]

[1] Department of Biomedical Engineering, University of Calgary, Calgary, Canada
emma.stanley@ucalgary.ca
[2] Department of Radiology, University of Calgary, Calgary, Canada
[3] Hotchkiss Brain Institute, University of Calgary, Calgary, Canada
[4] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada
[5] Department of Pediatrics, University of Calgary, Calgary, Canada
[6] Department of Community Health Sciences, University of Calgary, Calgary, Canada
[7] Department of Clinical Neurosciences, University of Calgary, Calgary, Canada

Abstract. Despite the remarkable advances in deep learning for medical image analysis, it has become evident that biases in datasets used for training such models pose considerable challenges for a clinical deployment, including fairness and domain generalization issues. Although the development of bias mitigation techniques has become ubiquitous, the nature of inherent and unknown biases in real-world medical image data prevents a comprehensive understanding of algorithmic bias when developing deep learning models and bias mitigation methods. To address this challenge, we propose a modular and customizable framework for bias simulation in synthetic but realistic medical imaging data. Our framework provides complete control and flexibility for simulating a range of bias scenarios that can lead to undesired model performance and shortcut learning. In this work, we demonstrate how this framework can be used to simulate morphological biases in neuroimaging data for disease classification with a convolutional neural network as a first feasibility analysis. Using this case example, we show how the proportion of bias in the disease class and proximity between disease and bias regions can affect model performance and explainability results. The proposed framework provides the opportunity to objectively and comprehensively study how biases in medical image data affect deep learning pipelines, which will facilitate a better understanding of how to responsibly develop models and bias mitigation methods for clinical use. Code is available at github.com/estanley16/SimBA.

M. Wilms and N.D. Forkert—Shared last authorship.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_46.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 489–499, 2023.
https://doi.org/10.1007/978-3-031-43895-0_46

1 Introduction

Deep learning for medical image analysis is a key tool to facilitate precision medicine and support clinical decision making. However, it has become increasingly evident that biases in the training data can lead to obstacles for clinical implementation. In this work, we define bias as a property of the data (*e.g.,* class/attribute imbalance, spurious correlations) used for training a model that can lead to shortcut learning and/or failure to adequately represent data subgroups, which may lead to reduced generalizability and/or fairness when applied in real-world scenarios. For instance, such biases have been shown to lead to poor generalization capabilities of models evaluated on cohorts with sociodemographic population statistics different to those that it was trained on [9], which can lead to systematic misdiagnosis of subpopulations [16,18]. Moreover, image acquisition biases can act as spurious correlations to the target (shortcut learning) [8,23].

Due to these problems, a plethora of research has recently gone towards bias mitigation [13,14,25,26] and data harmonization [2,5,23]. However, the utility of real-world medical images to assess and address bias-related challenges is often limited and may not be a comprehensive or sustainable solution. This is because all real-world datasets inherently suffer from biases that can be related to cohort selection, varying scanners and protocols, biases in "ground truth" labels, or any other (un-)known confounding factors associated with the data or the labels. Additionally, many medical imaging datasets do not contain suitable sociodemographic information or representation to adequately investigate the full range of bias scenarios that could be encountered in practice, especially when considering intersectional analyses [22]. Moreover, limitations in the seminal work on underdiagnosis disparities in deep learning models for chest X-ray analysis [18] have been identified since the various sources of bias present in the dataset could not be effectively distinguished from algorithmic bias [3] and known confounding factors were not rigorously accounted for [15]. However, even when known confounders such as disease prevalence between groups are considered, it may not be possible to adequately correct for them and unknown confounders and associated spurious correlations may still exist that go unaccounted for, such as annotation bias in labels used for training. Thus, it is very challenging to understand how biases in medical image data affect deep learning pipelines, especially if it is unknown what biases are present in the dataset, their magnitude and frequency, and how to correct them. As noted by various researchers, "understanding the root cause of bias [...] is a key step towards eliminating that bias" [19]. Therefore, there is a need for a resource that enables researchers to objectively study how biases in medical images affect deep learning models, without the limitations associated with real-world datasets. As a first step towards addressing this need, we propose a flexible framework for generating synthetic neuroimaging data with controlled simulation of realistic biases.

Current methods that have been proposed for fully controlled simulation of features in deep learning datasets, where generating factors can be fully disentangled and are well known in advance, are largely limited to toy problems or

MNIST-like scenarios [4]. On the other hand, a considerable amount of recent research has gone into the supervised and unsupervised disentanglement of generating factors of medical images with generative models that can subsequently be used to synthesize data with specific factor variations [7,11]. However, in such setups, unknown biases could still exist, the true generating factors of real-world data are usually unknown, and it is often impossible to spatially localize an effect. Therefore, we believe that such standard generative models do not offer the flexibility and control of the data generation mechanism that is required to fully analyze the effect of data biases on deep learning models. With our proposed framework, we aim to provide a method for synthesizing realistic image data with a fidelity similar to standard generative models, while still providing a high level of flexibility and control over the type, scale, and proportions of simulated bias features that is usually only available in MNIST-like setups.

In this work, we simulate brain magnetic resonance (MR) images with region-specific morphology variations representing disease and bias effects. We also introduce global morphological variation representing distinct synthetic subjects. This option facilitates bridging the relationship between understanding the impacts of biases alone, and how biases combine with real-world variation when training deep learning models. We utilize neuroimaging data as an initial use case and focus on morphological biases in this work. However, the proposed modular framework is not limited to neuroimaging problems and could be modified to introduce other bias effects, such as gray value effects caused by acquisition parameters or pathologies. Ultimately, this framework can serve as a tool for generating datasets to facilitate analysis of how deep learning models handle various sources of bias. With complete control over the number of samples, types of bias, number of subgroups with different biases, intersections of biased subgroups, and strength and proportion of bias in target classes, datasets generated with this framework can be used as a tool for evaluating how proposed or state-of-the-art models are affected by biases in terms of performance, explainable AI, uncertainty, *etc.*, as well as for benchmarking bias mitigation and data harmonization strategies on a wide range of realistic, controlled scenarios.

The contributions of this paper are summarized as follows: (1) We propose a flexible framework for simulating brain MR datasets, which contain variable morphological disease and bias effects, as a first step towards the controlled and systematic study of how biases in medical imaging data affect deep learning pipelines. (2) We show how this modular framework can be customized to facilitate the investigation of a vast range of data cases that can lead to biased deep learning models. (3) We provide empirical evidence that data generated using this framework can be used to mimic realistic morphological biases in neuroimaging that lead to undesirable performance in a convolutional neural network, and show how these biases can be investigated with explainable AI methods.

2 Methods

The purpose of the proposed framework is to generate a dataset for a multi-class classification problem consisting of synthetic T1-weighted MR images, with N

images I_i and associated labels corresponding to m disease classes. For simplicity, in this description of the methods, we focus on the binary classification task (m=2) with disease labels corresponding to disease (D) and no disease (ND). All images are derived by applying non-linear diffeomorphic transformations to a template image I_T, which represents an average brain morphology. More specifically, we consider three types of transformations: (1) φ_S, a subject morphology, (2) φ_D, a disease (target) effect, and (3) φ_B, a bias effect. φ_S is a global non-linear transformation that deforms I_T into a (simulated) subject morphology. In contrast, φ_D and φ_B are spatially localized deformations that only modify I_T locally to introduce an effect (φ_D) that can be used to differentiate disease classes, and a bias effect (φ_B). In our setup, each synthetic image is generated by sampling the transformations φ_S, φ_D, and optionally φ_B from dedicated generative models (Fig. 1A and Suppl. Mat. Fig. 1). Moreover, we assume that all diffeomorphic transformations are parameterized via stationary velocity fields—*e.g.*, $\varphi_S = \exp(v_S)$, where v_S denotes the velocity field and $\exp(\cdot)$ is the group exponential map from the Log-Euclidean framework, which can be efficiently computed via the scaling-and-squaring algorithm; see [1]. The resulting dataset is defined by the user-specified sample size, number of target disease classes, number of subgroups within the dataset containing bias effects, whether inter-subject variability effects are introduced to the datasets, types and degree of each respective effect, and proportions of each respective class and bias group.

Principal Component Analysis-Based Generative Models for Simulating Effects/Variability. To apply anatomically realistic morphological deformations to our template neuroimaging dataset in this work, we fit a principal component analysis (PCA) to the velocity fields of real T1-weighted MR images of different healthy subjects, which were non-linearly registered to the template image I_T. We treat the resulting low-dimensional affine subspace model as a generative model and sample velocity fields representing a range of real anatomical variation from it. For region-specific effects (φ_D and φ_B), the real T1-weighted MR image velocity fields within the regions defined by a label atlas are masked prior to PCA fitting, whereas the full brain is used in the PCA model for simulating subject morphology (φ_S). Thus, by sampling velocity fields v_D, v_B, and v_S from the latent space of the respective subspace models, we can model disease, bias, and subject morphology as variations within an expected extent of human neuroanatomy.

Disease and Bias Effects. We model disease (φ_D) and bias (φ_B) effects as morphological deformations localized to specific brain regions. We also assume that datasets belonging to each disease class have these localized effects sampled from respective distributions in a bimodal Gaussian mixture model within the PCA subspace of the disease effect model. We assume that bias effects are introduced as an additional morphological deformation in a separate brain region, and that these effects are sampled from a Gaussian distribution within the PCA subspace of the bias effect model. In general, an arbitrary number of target

classes and bias groups can be introduced to the datasets in a similar sampling procedure.

Subject Morphology. To better emulate anatomical variation in clinical data and warrant the use of deep learning models, global morphological variation representing distinct subjects (φ_S) are applied to the entire anatomy within I_T. These are also sampled from a Gaussian distribution within the PCA subspace of the dedicated subject morphology model.

Introducing Effects to the Template Image. The sparsely defined velocity fields for the disease and bias effects, v_D and v_B, are densified using the scattered grid B-spline method [10] to produce a dense velocity field that includes both effects (if present). If inter-subject variability is desired, the conjugate action mechanism [12] is used to transport the deformation field to the "subject" space, where the "subject" is generated using the sampled v_S/φ_S from the subject morphology PCA model.

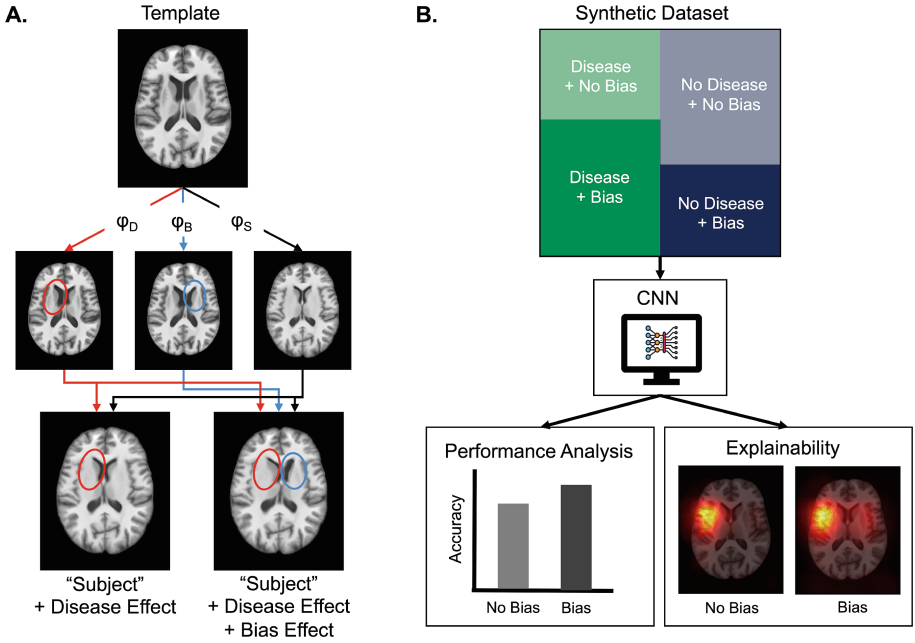

Fig. 1. A) Schematic representing how displacement fields for disease effects (φ_D), bias effects (φ_B), and subject morphology (φ_S) are introduced to a template image I_T to generate custom datasets. B) Synthetic dataset evaluation pipeline used in this paper. A convolutional neural network (CNN) is trained to classify the disease class from a dataset with subgroups containing bias features and evaluated with subgroup performance and explainability.

Framework Customization. For this initial work, we utilize velocity fields from real-world datasets to simulate anatomically realistic effects representing disease features, bias features, and subject morphology via different PCA-based generative models. Although real-world datasets do contain biases, the way in which we propose introducing these effects into the synthetic dataset is highly controlled in such a way where it is known exactly which and how many regions represent either disease or bias effects. Thus, this approach enables a controlled study of bias while benefiting from the utilization of 3D medical images that are representative of real-world clinical data. Moreover, due to the modularity of the proposed framework, such effects can also be introduced through a variety of other methods for generating deformation fields, ranging from highly precise but simple (*e.g.*, single displacement vectors) to more realistic but increasingly complex approaches (*e.g.*, generative models). Furthermore, in this work, we simulate morphological changes in brain images via diffeomorphic transformations as a use case, but the framework can be adapted to other disease or bias effects that would alter the topology (*e.g.*, gray value changes or lesions). Moreover, other imaging modalities or body regions (*e.g.*, cardiac MRI) as well as other generative models (*e.g.*, generative adversarial networks) could be integrated.

3 Experiments and Results

To evaluate our synthetic datasets in a deep learning pipeline, we trained a CNN to predict whether images from biased datasets belong to the disease (D) or no disease (ND) class. More precisely, we evaluated (1) how the proportion of datasets containing bias features within the D class, and (2) how the spatial proximity between the disease region and bias region affect the performance and explainability (XAI) results of a CNN trained to classify D from ND cases (Fig. 1B). All experiments were performed with Keras/Tensorflow v. 2.10.

Simulated Datasets. The SRI24 Normal Adult Brain Anatomy atlas [17] was used as the template image and each PCA model for sampling morphological effects was trained on T1-weighted MRI data from 50 subjects who were part of the IXI database of healthy individuals[1]. Velocity fields for this dataset were estimated by utilizing ITK's VariationalRegistration framework [6,24]. The left insular cortex was selected as the brain region for the disease effect, and the brain regions used to model bias effects were either the left putamen, right putamen, or right postcentral gyrus as defined by the LPB40 atlas [20], depending on the desired spatial proximity. The datasets belonging to the D and ND classes had effects sampled from $\mathcal{N}(0,1)$ and $\mathcal{N}(2,1)$, respectively, along the first principal component of the generative model for the disease region, and the datasets with bias features had effects sampled from $\mathcal{N}(2,1)$ along the first principal components of the models for the respective bias regions. Inter-subject variability effects were sampled from a Gaussian distribution of $\mathcal{N}(0,1)$ along the first 10 principal components of the subject morphology generative model.

[1] https://brain-development.org/ixi-dataset/.

Experiments. To evaluate the effect on model performance and XAI in relation to the proportion of datasets containing bias features within the disease class, the generated datasets had either 60% or 80% of the simulated images from the D class containing the bias feature, with 30% of the simulated images from the ND class containing the bias feature for all experiments. To evaluate the effect of proximity between disease and bias regions, the distances between regions were defined as either near, middle, or far, for the left putamen, right putamen, and right postcentral gyrus, respectively (see Fig. 2B). Each simulated dataset contained a balanced representation of D and ND labels. The proximity experiments were performed under both 60% and 80% conditions defined by the proportion experiments. Model performance with the biased datasets was compared against a baseline experiment in which the datasets do not contain any simulated bias features but only the disease effects.

Model and Training. A CNN was used as a model for predicting whether datasets belonged to the D or ND class. The model consisted of 5 blocks each containing a convolutional layer with (3×3×3) kernel, batch normalization, sigmoid activation, and (2×2×2) max pooling. The convolutional filter sizes were 32, 64, 128, 256, and 512 for each respective block. The sixth block contained average pooling, dropout (rate=0.2), and a dense classification layer with softmax activation. Binary cross entropy loss, Adam optimizer (learning rate = $1e^{-4}$), and batch size 4 with early stopping based on validation loss (patience=30) were used for training. Each experiment simulated and used 500 datasets of voxel dimensions (173×211×155) with a 55%/15%/30% train/validation/test split, stratified by disease and bias labels.

Evaluation. Model performance was evaluated using accuracy, sensitivity, and specificity computed for the aggregate test set, as well as separately for the bias (B) and no bias (NB) groups. Results are reported as the mean ± standard deviation of the models with 5 different weight initialization seeds on the same train/validation/test splits, following [18]. The SmoothGrad (SG) [21] method was used for XAI evaluation. Average SG maps were computed with 25 individual SG maps (5 from each seed) for the datasets in the test set with the bias feature, which were correctly identified as being in the disease class.

Results and Discussion. The results of our evaluation are summarized in Fig. 2, with full quantitative results shown in Tables 1 and 2 in the Supplementary Material. As seen in Fig. 2A, for all conditions with simulated dataset bias, the sensitivity is higher and specificity is lower within the B group, while the opposite was found for the NB group. Due to the higher representation of biased datasets in the D class, it seems reasonable to assume that the model uses the presence of bias features as a shortcut for predicting the disease state, and thus predicts the D class more often for the B group, resulting in fewer true negatives. Within the NB group, the absence of bias features seems to be also

used as a shortcut for predicting the ND class, resulting in a higher number of ND class predictions and consequently fewer true positives. While these shortcuts are apparent in all experiments utilizing biased datasets, there is a stronger relationship between disease–bias region proximity and the degree of the shortcuts (measured by lower specificity in the bias group and lower sensitivity in the NB group) when the dataset has 80% of the D class containing the bias effect compared to 60%. In these 80% conditions, it was observed that the sensitivity in the NB group decreases as a function of region proximity, indicating that the model uses the absence of the bias effect as a shortcut for predicting the ND class more often when regions are further away. Likewise, specificity in the B group increases as a function of region proximity, indicating that the model uses the presence of the bias as a shortcut for predicting the D class label less often when regions are further away. A potential explanation for this may be that the CNN used has a spatially localized receptive field. Thus, when the bias and disease regions are near to each other, the network learns to associate them more closely and predicts the D class label more often for images with the bias effect. When the regions are farther apart from each other, the CNN becomes more tuned to recognize bias effects separately from disease effects. Thus, when the bias effect is not present, the model assumes the image belongs to the ND class.

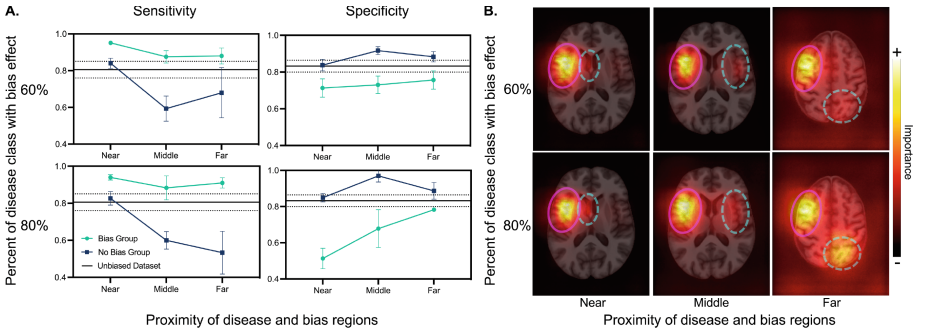

Fig. 2. Results of experiments investigating effect of the proportion of bias effect in the disease class and proximity between disease and bias regions. A) Mean sensitivity (left column) and specificity (right column) for test datasets with bias (green circle) and without bias (blue square). Error bars represent standard deviation over 5 different model initialization seeds with the same train/validation/test split. Black lines represent mean sensitivity and specificity with 95% confidence interval over 5 different model seeds with the same train/validation/test split of dataset containing no bias effects. B) Average SmoothGrad saliency maps with disease region circled in solid/magenta and bias region circled in dashed/cyan. (Color figure online)

Furthermore, as seen in Fig. 2B, with 60% of the D class containing the bias feature, the SG maps show minor activation in the region with the bias effect, particularly in the far proximity condition. However, when 80% of the D class contains the bias feature, the SG maps highlight the bias regions considerably

more intensely for all proximities analyzed. Even though the model still uses prediction shortcuts, which affect performance of the B and NB groups when 60% of the images from the D class exhibit the bias feature, the regions associated with the bias are less clearly identifiable in the group-averaged SG saliency maps, suggesting that XAI may not always be a reliable tool to uncover sources of bias in medical image data.

4 Conclusion

In this work, we presented a flexible and modular framework for simulating bias in medical imaging datasets using realistic morphological effects in neuroimaging as a use case. By sampling brain region-specific morphological variation representing the disease state and bias features from generative models in a controlled manner, we can generate synthetic datasets of arbitrary size and composition, which enables the investigation of a vast range of dataset bias scenarios and corresponding impacts on deep learning pipelines. Directions for future work with this framework are extensive and could include the analysis of more variations of bias proportions and proximities on alternate model architectures (*e.g.*, vision transformers), evaluation of state-of-the-art bias mitigation strategies on various dataset compositions, as well as assessing other potential limitations of explainability methods as a tool for investigating bias. We believe that our work provides a strong foundation for advancing understanding of bias in deep learning for medical image analysis and consequently developing responsible models and methods for clinical use.

Acknowledgements. This work was supported by Alberta Innovates, the Natural Sciences and Engineering Research Council of Canada, the River Fund at Calgary Foundation, Canada Research Chairs Program, and the Canadian Institutes of Health Research.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-Euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006). https://doi.org/10.1007/11866565_113
2. Bashyam, V.M., et al.: The iSTAGING and PHENOM consortia: deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. J. Magn. Reson. Imaging **55**(3), 908–916 (2022)
3. Bernhardt, M., Jones, C., Glocker, B.: Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. Nat. Med. **28**(6), 1157–1158 (2022)
4. Castro, D.C., Tan, J., Kainz, B., Konukoglu, E., Glocker, B.: Morpho-MNIST: quantitative assessment and diagnostics for representation learning. J. Mach. Learn. Res. **20**, 1–29 (2019)

5. Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L.: Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. Neuroimage **228**, 117689 (2021)
6. Ehrhardt, J., Schmidt-Richberg, A., Werner, R., Handels, H.: Variational registration. In: Handels, H., Deserno, T.M., Meinzer, H.-P., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2015. I, pp. 209–214. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46224-9_37
7. Fragemann, J., Ardizzone, L., Egger, J., Kleesiek, J.: Review of disentanglement approaches for medical applications - towards solving the gordian knot of generative models in healthcare (2022). arXiv:2203.11132 [cs]
8. Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects (2019). arXiv:1910.04597 [cs, eess, q-bio]
9. 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. Proc. Natl. Acad. Sci. **117**, 12592–12594 (2020)
10. Lee, S., Wolberg, G., Shin, S.: Scattered data interpolation with multilevel B-splines. IEEE Trans. Visual Comput. Graphics **3**(3), 228–244 (1997)
11. Liu, X., Sanchez, P., Thermos, S., O'Neil, A.Q., Tsaftaris, S.A.: Learning disentangled representations in the imaging domain. Med. Image Anal. **80**, 102516 (2022)
12. Lorenzi, M., Pennec, X.: Geodesics, parallel transport & one-parameter subgroups for diffeomorphic image registration. Int. J. Comput. Vision **105**(2), 111–127 (2013)
13. Luo, L., Xu, D., Chen, H., Wong, T.T., Heng, P.A.: Pseudo bias-balanced learning for debiased chest X-ray classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, pp. 621–631. LNCS, vol. 13438. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_59
14. Marcinkevics, R., Ozkan, E., Vogt, J.E.: Debiasing deep chest X-ray classifiers using intra- and post-processing methods. In: Proceedings of the 7th Machine Learning for Healthcare Conference, pp. 504–536. PMLR (2022)
15. Mukherjee, P., Shen, T.C., Liu, J., Mathai, T., Shafaat, O., Summers, R.M.: Confounding factors need to be accounted for in assessing bias by machine learning algorithms. Nat. Med. **28**(6), 1159–1160 (2022)
16. Puyol-Antón, E., et al.: Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, pp. 413–423. LNCS, vol. 12903. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_39
17. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp. **31**(5), 798–819 (2010)
18. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. **27**(12), 2176–2182 (2021)
19. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Reply to: 'Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms' and 'Confounding factors need to be accounted for in assessing bias by machine learning algorithms'. Nat. Med. **28**(6), 1161–1162 (2022)
20. Shattuck, D.W., et al.: Construction of a 3D probabilistic atlas of human cortical structures. Neuroimage **39**(3), 1064–1080 (2008)

21. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise (2017). arXiv: 1706.03825
22. Stanley, E.A.M., Wilms, M., Forkert, N.D.: Disproportionate subgroup impacts and other challenges of fairness in artificial intelligence for medical image analysis. In: Baxter, J.S.H., et al. (eds.) Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging, pp. 14–25. LNCS, vol. 13755. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-23223-7_2
23. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. **67**, 101879 (2021)
24. Werner, R., Schmidt-Richberg, A., Handels, H., Ehrhardt, J.: Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: a comparison and evaluation study. Phys. Med. Biol. **59**(15), 4247 (2014)
25. Zare, S., Nguyen, H.V.: Removal of confounders via invariant risk minimization for medical diagnosis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 578–587. LNCS, vol. 13438. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_55
26. Zhao, Q., Adeli, E., Pohl, K.M.: Training confounder-free deep learning models for medical applications. Nat. Commun. **11**(11), 6010 (2020)