Bias in Unsupervised Anomaly Detection in Brain MRI

Cosmin I. Bercea[1,2,5(✉)], Esther Puyol-Antón[5,6], Benedikt Wiestler[3], Daniel Rueckert[1,3,4], Julia A. Schnabel[1,2,5], and Andrew P. King[5]

[1] Technical University of Munich, Munich, Germany
cosmin.bercea@tum.de
[2] Helmholtz AI and Helmholtz Center Munich, Munich, Germany
[3] Klinikum Rechts der Isar, Munich, Germany
[4] Imperial College London, London, UK
[5] King's College London, London, UK
[6] HeartFlow Inc, London, UK

Abstract. Unsupervised anomaly detection methods offer a promising and flexible alternative to supervised approaches, holding the potential to revolutionize medical scan analysis and enhance diagnostic performance.

In the current landscape, it is commonly assumed that differences between a test case and the training distribution are attributed solely to pathological conditions, implying that any disparity indicates an anomaly. However, the presence of other potential sources of distributional shift, including scanner, age, sex, or race, is frequently overlooked. These shifts can significantly impact the accuracy of the anomaly detection task. Prominent instances of such failures have sparked concerns regarding the bias, credibility, and fairness of anomaly detection.

This work presents a novel analysis of biases in unsupervised anomaly detection. By examining potential non-pathological distributional shifts between the training and testing distributions, we shed light on the extent of these biases and their influence on anomaly detection results. Moreover, this study examines the algorithmic limitations that arise due to biases, providing valuable insights into the challenges encountered by anomaly detection algorithms in accurately capturing the variability in the normative distribution. Here, we specifically investigate Alzheimer's disease detection from brain MR imaging as a case study, revealing significant biases related to sex, race, and scanner variations that substantially impact the results. These findings align with the broader goal of improving the reliability, fairness, and effectiveness of anomaly detection.

Keywords: Unsupervised Anomaly Detection · Bias · Fairness

1 Introduction

Unsupervised anomaly detection (UAD) methods have gained significant attention in the medical image analysis research literature due to their potential to

S. Wesarg et al. (Eds.): CLIP/FAIMI/EPIMI 2023, LNCS 14242, pp. 122–131, 2023.
https://doi.org/10.1007/978-3-031-45249-9_12

identify anomalies without the need for labeled training data. However, recent literature has shown that UAD methods are vulnerable to non-pathological out-of-distribution (OoD) data [6]. As a result, notable failures in such approaches have raised concerns regarding bias and fairness in their evaluation. For example, Meissen et al. [9] presented cases where polyp detection algorithms achieved excellent performance even when the actual polyps were removed from the error maps. Similarly, Bercea et al. [1] demonstrated nearly perfect OoD detection using popular reconstruction-based methods that solely relied on analyzing background pixels. Moreover, a predominant focus in the recent literature on UAD has been on the detection of hyper-intense lesions in brain MRIs [3,7,11,17]. A recent study demonstrated that many reconstruction-based methods struggled to generalize to other types of anomalies, indicating a bias in the algorithmic performance [2]. In light of these notable failures, it is essential to thoroughly investigate these concerns to enable the development of more robust and reliable models that exhibit fair and unbiased behavior.

There has been relatively little research into bias in UAD techniques. This is in contrast to other medical imaging applications, where in recent years there has been an increasing focus on bias and fairness. For example, Gichoya et al. [4] demonstrated the presence of race-based distributional shifts across several imaging modalities, highlighting the potential for bias when training models with imbalanced data. Additionally, studies such as Larrazabal et al. [8] and Seyyed et al. [13] have examined bias in chest X-ray classification, Guo et al. [5] reviewed work on biases in skin cancer detection algorithms, and Puyol-Anton et al. [12] have identified race bias in cardiac MR segmentation. In recent years, there have been several studies that have examined biases in neuroimaging data [14–16], including the task of Alzheimer's disease (AD) detection [10].

However, these studies have all focused on supervised approaches. In contrast, our study specifically investigates unsupervised models, which are underpinned by the need to learn the normative training distribution and thus could be more susceptible to biases. These studies underscore the importance of addressing bias in medical applications and sand pave the way for further exploration in UAD. The motivation for studying bias in UAD is twofold. First, given that most experimental setups in the research literature have involved distinct data sources for healthy and pathological distributions, it is essential to analyze potential shifts to ensure fair evaluations and prevent correlations that can significantly skew the performance of these methods. Second, as UAD models strive to represent the entire variability of the normative distribution and effectively identify anomalies by isolating pathological shifts, it becomes increasingly more important to identify their algorithmic limitations.

In this work, we investigate both types of biases in anomaly detection, aiming to fill the gap in this important research area. Our focus is on a case study of AD detection from brain MR images. In summary, our main contributions are:

- To the best of our knowledge, this work represents the first comprehensive investigation into the biases present in UAD.

- Through rigorous analysis, we have uncovered evidence of scanner, sex, race, and metrics biases that significantly impact the performance of UAD.
- We examined other factors like age and brain volume but found no additional correlations for the observed performance drops.

Table 1. Datasets. We present the data splits utilized in our experiments. The abbreviations (Abv.) are linked to the experiments in Table 2. We refer to the healthy training distribution as "Control", the healthy cohort from a different distribution as "Healthy". "Alzheimer's (AD)" represent the pathology set. We mark in blue the shifts in distribution compared to the control set.

Abv.	Dataset	#Scans	Group	Race	Sex	Scanner
T	Training (Control)	434	Control	White	Female	Siemens
V	Validation (Control)	54	Control	White	Female	Siemens
C	Test (Control)	122	Control	White	Female	Siemens
AD	Test (Baseline)	131	AD	White	Female	Siemens
H1	Test (Healthy, Scanner Shift 1)	171	Healthy	White	Female	Philips
AD1	Test (Scanner Shift 1)	73	AD	White	Female	Philips
H2	Test (Healthy, Scanner Shift 2)	70	Healthy	White	Female	GE
AD2	Test (Scanner Shift 2)	36	AD	White	Female	GE
H3	Test (Healthy, Sex Shift)	480	Healthy	White	Male	Siemens
AD3	Test (Sex Shift)	188	AD	White	Male	Siemens
H4	Test (Healthy, Race Shift 1)	103	Healthy	Black	Female	Siemens
AD4	Test (Race Shift 1)	16	AD	Black	Female	Siemens
H5	Test (Healthy, Race Shift 2)	18	Healthy	Asian	Female	Siemens
AD5	Test (Race Shift 2)	8	AD	Asian	Female	Siemens

By shedding light on these biases, we strive to enhance the reliability, fairness, and effectiveness of anomaly detection methods in medical imaging, ultimately benefiting both healthcare providers and patients.

2 Materials and Methods

Dataset. Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1]. ADNI offers a rich collection of magnetic resonance imaging (MRI) scans, accompanied by comprehensive metadata including MR scanner information, and demographic factors such as age,

[1] https://adni.loni.usc.edu.

sex, and race. This dataset offers an ideal opportunity to isolate specific factors and evaluate their impact on anomaly detection. In Table 1, we present an overview of the data utilized in this paper, specifically detailing the partitioning of the ADNI dataset into different training, validation and test subsets. As can be seen, our training dataset was acquired from white females using Siemens scanners with a field strength of 3T. This choice was made to maximize the availability of training images and facilitates a more comprehensive assessment of the model's performance.

UAD Method. We utilized a state-of-the-art variational auto-encoder architecture as our UAD method[2]. This recent model incorporates advanced techniques, including perceptual and adversarial loss functions, to enhance the accuracy of image reconstructions, while constraining the latent distribution using the Kullback-Leibler divergence.

Table 2. Bias in UAD. The conventional approach to evaluating UAD methods involves using the control set from the training distribution (denoted as 'C' in Table 1) as the healthy subjects during testing. We present these results as the Naive AD Detection. A relative increase ▲ x% and decrease ▼ x% in performance compared to the baseline (computed as (b-a)/a*100) signifies the presence of bias, while ▶ x% suggests no bias. To focus solely on the methodological bias, we also report the True AD Detection, which involves using both healthy and pathological subjects from the same source at test time, such as H1/AD1. ↑ x% demonstrates improved performance. We show the distributions of the residual errors and visualize the bias shifts in Fig. 2.

Test set	Naive AD Detection Evaluation & Methodological Bias			True AD Detection Methodological Bias		
	Data	AUROC ↑	AUPRC ↑	Data	AUROC ↑	AUPRC ↑
Baseline	C/AD	64.60	64.82	C/AD	64.60	64.82
Scanner (Philips)	C/AD1	50.30 ▼ 28%	41.71 ▼ 55%	H1/AD1	54.72 ▼ 18%	35.23 ▼ 84%
Scanner (GE)	C/AD2	60.22 ▼ 7%	30.73 ▼ 111%	H2/AD2	64.64 ▶ 0%	59.80 ▼ 8%
Sex (Male)	C/AD3	86.68 ▲ 25%	89.77 ▲ 28%	H3/AD3	61.69 ▼ 5%	38.39 ▼ 69%
Race (Black)	C/AD4	55.53 ▼ 16%	12.77 ▼ 408%	H4/AD4	56.43 ▼ 14%	15.07 ▼ 330%
Race (Asian)	C/AD5	65.06 ▶ 1%	9.65 ▼ 572%	H5/AD5	73.61 ↑ 12%	67.59 ↑ 4%

Metrics. We use a range of metrics to evaluate the performance of our method from different perspectives. To assess the reconstruction quality of the methods, we use the mean absolute error (MAE). To assess the anomaly detection ability of our method, we use area under the receiver operator curve (AUROC) and area under the precision-recall curve (AUPRC). Additionally, we include the subjective assessment of a clinician expert to evaluate the quality of reconstructions and the localization of anomalies. Finally, for assessing statistical significance,

[2] https://github.com/Project-MONAI/GenerativeModels.

we used Pearson's correlation to identify the impact of potential confounders on the residual errors, and the Kolomogorov-Smirnov test to identify distributional shifts between the AD sets of the training distribution and AD sets of the target distributions. We considered results with p-values lower than 0.05 significant.

3 Experiments and Results

In Subsect. 3.1, we conduct a comprehensive evaluation of the proposed method under ideal conditions, where no distributional shifts other than the pathological one are expected. This evaluation provides insight into the ability of the model to accurately detect AD. Subsequently, in Subsect. 3.2, we systematically introduce changes in the data distribution by modifying a single factor other than pathology, such as MRI scanner manufacturer, sex, or race and evaluate the impact of biases on the performance. Finally, in Subsect. 3.3 we perform further analysis to uncover potential causes of the performance drops.

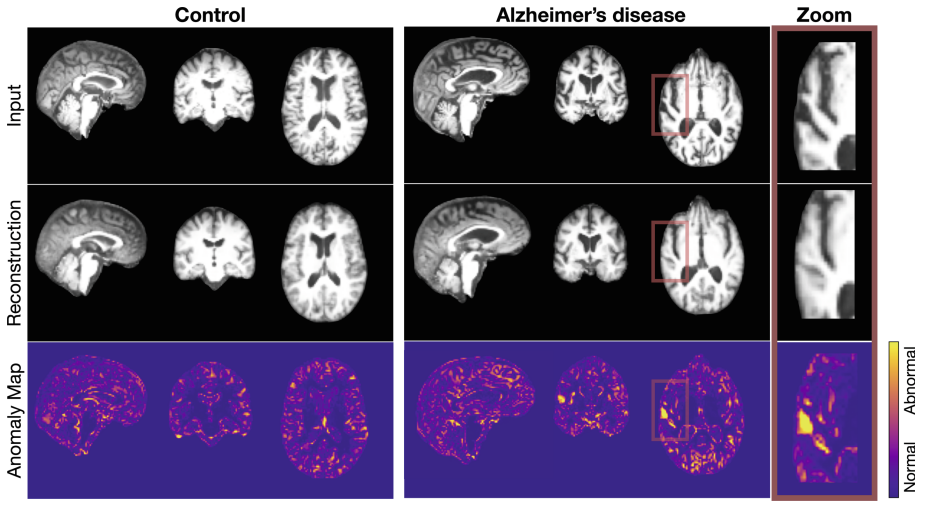

Fig. 1. Qualitative results of the baseline experiment. In the case of AD (right), the atrophy in the Sylvian fissure, which is a typical feature of the disease, is reduced in the reconstruction. This leads to a clear highlight in the anomaly map.

3.1 Baseline Performance

We first evaluated the baseline performance under ideal conditions, where the only factor of change was the presence of AD pathology. See Table 2 for quantitative results and Fig. 1 for a visual example. A clinician assessed that the VAE effectively reverses AD-related pathological changes, such as ventricle dilation

and Sylvian fissure abnormalities. Consequently, such areas are highlighted in the residual anomaly map and thus can be readily interpreted for their plausibility. The distributions plot in Fig. 2 (Baseline) shows increased reconstruction errors for AD compared to the control set. Therefore, the method achieved moderate discriminative performance in detecting Alzheimer's pathology with AUROC and AUPRC scores of approximately 65%.

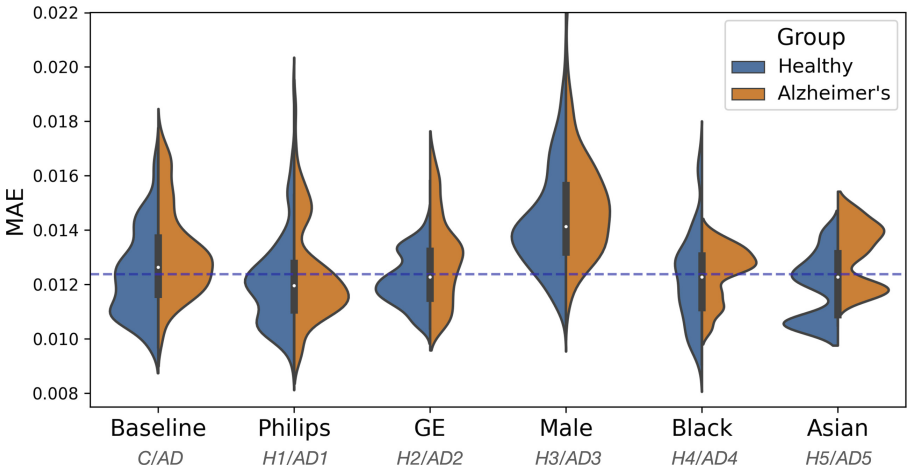

Fig. 2. The distributions of residual errors for different distributional shifts demonstrates the impact of biases. Evaluation bias is characterized by a shift in the overall mean of the residual error, either above or below the mean of the training distribution (shown as a dotted line). Methodological biases are observed when comparing the distributions of healthy and AD groups within a specific shift (violin), revealing a lack of clear distinction between the two distributions. See Table 2 for numerical results.

3.2 Impact of Bias

In this section, we conducted controlled experiments to systematically investigate the impacts of distributional shifts caused by various variables, such as scanner type, sex, and race, aiming to identify potential sources of bias that might impact the performance of UAD. We summarize the results quantitatively in Table 2, and visualize the distributions of the residual errors in Fig. 2.

We considered two distinct scenarios in our analysis. The first scenario, which we call the "naïve" approach, is commonly used in the research literature. Since there is a lack of healthy data within many publicly available datasets, most UAD use alternative sources of healthy data as controls and evaluate their performance on pathology data from a different distribution. However, this approach introduces additional distributional shifts into the evaluation process beyond the presence of pathology itself. Consequently, the evaluation is not clinically realistic and many cases of failure reported in the literature can be attributed to these

confounders. In the second scenario, we aimed to address these confounders and isolate the impact of pathology by using an evaluation set that includes both healthy and pathological subjects from the same source. In doing so, we sought to assess any methodological shortcomings and evaluate the effects of bias on UAD. In both scenarios, we observed indications of significant bias stemming from factors such as scanner type, sex and race within the target groups.

Domain Shifts. First, we observed a significant domain shift in the distribution of residual errors, as depicted in Fig. 2. This shift manifests as a uniform vertical shift of both healthy and AD distributions compared to the baseline distribution. Note that this would lead to artificially high UAD performance when controls from the baseline distribution are employed. It is essential to recognize and address these evaluation biases, as they have the potential to strongly influence and distort the UAD results, as demonstrated in Table 2.

Metrics Bias. Next, we examined the bias introduced by the choice of evaluation parameters. Specifically, we found that the AUROC metric tended to be too optimistic, especially when the healthy and pathological samples were highly imbalanced. In some cases the AUROC failed to recognize the presence of bias, e.g., GE scanner shifts (0% performance difference) and Asian race shifts (1% performance difference) in Table 2. Instead, AUPRC emerged as a more robust measure for assessing performance, demonstrating that it is more suitable for evaluating imbalanced datasets.

Scanner Shifts. When using the conventional approach of using the training distribution control set as a reference (Naive AD Detection), we observed a performance decrease of 55% in AURPC for the Philips scanner (C/AD1) compared to the baseline. Similarly, the GE scanner (C/AD2) showed a substantial 111% decrease in AUPRC. To isolate methodological bias, we used both healthy and pathological distributions from the same source (True AD Detection). Interestingly, the performance on Philips (H1/AD1) still showed a considerable drop of 84%, while for GE (H2/AD2) we only observed a minor 8% performance drop.

Sex Shifts. A notable distinction between Naive and True AD detection performance occurs when examining the presence of sex bias. In the naive approach, AD detection performance for the male group increased significantly to an AUPRC of 89.77. However, a closer examination of the distribution plot shown in Fig. 2 reveals that both the healthy and pathology distributions show higher residual errors. This finding suggests a pronounced evaluation bias associated with the naive approach. In contrast, when considering the True AD scenario, which takes into account both healthy and pathological distributions from the same source, the performance dropped dramatically to only 38.39%.

Race Shifts. The race shift analysis revealed notable biases in anomaly detection performance. Specifically, we observed a considerable decrease in performance of 330% when evaluating samples from the black race. Conversely, there was a slight improvement in the True AD Detection performance for subjects from the Asian race, albeit with a limited number of samples.

Table 3. Sources of Bias. Statistical correlations among age, ventricular volume (VV), Hippocampal volume (HV), and whole brain volume (WBV) are examined to explore potential underlying causes for performance drops in the presence of domain shifts. Significant correlations between the analyzed confounds and increased/decreased residual errors in AD samples are denoted by a checkmark (✓), while no correlation is indicated by a cross (✗). Significant shifts in the AD distributions are highlighted in **bold** and the combination of both (✓ and **bold**) is shown in red.

Shift →	Philips (AD1)		GE (AD2)		Male (AD3)		Black (AD4)		Asian (AD5)	
Age	✗	(**0.224**, **0.015**)	✗	(**0.255**, **0.041**)	✓	(**0.346**, **0.000**)	✗	(0.321, 0.082)	✓	(0.381, 0.172)
VV	✓	(0.216, 0.089)	✓	(0.265, 0.073)	✓	(**0.373**, **0.000**)	✓	(**0.643**, **0.002**)	✗	(0.429, 0.105)
HV	✓	(0.119, 0.728)	✗	(**0.403**, **0.002**)	✓	(0.195, 0.063)	✓	(**0.551**, **0.015**)	✗	(0.277, 0.543)
WBV	✓	(0.230, 0.065)	✓	(0.207, 0.242)	✓	(**0.494**, **0.000**)	✓	(**0.581**, **0.008**)	✓	(0.476, 0.051)

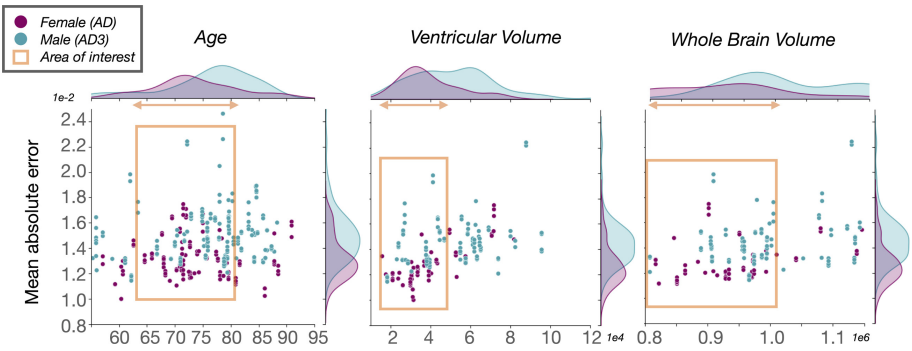

Fig. 3. Visual analysis of the correlations for the sex shift reveals distribution differences in various factors compared to the AD training distribution (distribution plots on top). However, these factors do not seem to be the cause of the performance drop. A closer examination of the target area (highlighted by an orange rectangle) indicates that males have larger overall residual errors than females, suggesting an unaccounted underlying cause for the performance drops. (Color figure online)

3.3 Sources of Bias

In this section, our objective is to explore potential causes for the observed performance drops under different shifts. Table 3 presents the results of our statistical analysis, focusing on distributional shifts resulting from potential confounders, including age, ventricular volume (VV), Hippocampal volume (HV), and whole brain volume (WBV). We mark in red identified significant correlations, where a confounder significantly (according to Pearson's correlation p-values) impacts the residual errors (marked with a checkmark) and there is a significant (according to a Kolmogorov-Smirnov test) distributional shift between the training and target AD distributions (indicated in bold). To summarise, we identified significant correlations for the male and black female distributions. We further inspect the sex shift visually in Fig. 3. The analysis demonstrates elevated residual errors

for males, even within the shared ranges of the analyzed confounding factors between males and females. This suggests the presence of another underlying cause beyond the factors evaluated. Further investigations, exploring additional potential confounders are necessary to uncover potential explanations and causal factors contributing to the observed performance drops in the presence of bias.

4 Conclusion

In conclusion, our study highlights the presence of bias, including bias due to scanner, sex and race, in the performance of UAD algorithms. The results indicate that non-pathological distributional shifts can introduce significant distortions in UAD performance. These biases not only impact the overall error distribution, i.e., evaluation bias, but also affect the ability of the methods to accurately detect AD disease. It is essential to understand and address these biases in order to develop robust and reliable UAD algorithms. Future research should prioritize efforts to mitigate these biases and ensure accurate and precise detection in diverse populations and imaging environments.

Acknowledgements. C.I.B. is in part supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

References

1. Bercea, C.I., Rueckert, D., Schnabel, J.A.: What do we learn? Debunking the myth of unsupervised outlier detection. arXiv preprint arXiv:2206.03698 (2022)
2. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Generalizing unsupervised anomaly detection: towards unbiased pathology screening. In: International Conference on Medical Imaging with Deep Learning (2023)
3. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. **64**, 101713 (2020)
4. Gichoya, J.W., B., et al.: AI recognition of patient race in medical imaging: a modelling study. Lancet. Digit. Health **7500**(22), e406–e414 (2022)
5. Guo, L.N., Lee, M.S., Kassamali, B., Mita, C., Nambudiri, V.E.: Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection - a scoping review. J. Am. Acad. Dermatol. **87**(1), 157–159 (2021)
6. Heer, M., Postels, J., Chen, X., Konukoglu, E., Albarqouni, S.: The OOD blind spot of unsupervised anomaly detection. In: Medical Imaging with Deep Learning (2021). https://openreview.net/forum?id=ZDD2TbZn7X1
7. Kascenas, A., Pugeault, N., O'Neil, A.Q.: Denoising autoencoders for unsupervised anomaly detection in brain MRI. In: International Conference on Medical Imaging with Deep Learning (2022)
8. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. U S A **117**(23), 12592–12594 (2020)
9. Meissen, F., Lagogiannis, I., Kaissis, G., Rueckert, D.: Domain shift as a confounding variable in unsupervised pathology detection. In: Medical Imaging with Deep Learning (2022). https://openreview.net/forum?id=6tsAzh_tnyF

10. Petersen, E., et al.: Feature robustness and sex differences in medical imaging: a case study in MRI-based Alzheimer's disease detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 88–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_9
11. Pinaya, W.H.L., et al.: Unsupervised brain anomaly detection and segmentation with transformers. arXiv preprint arXiv:2102.11650 (2021)
12. Puyol-Antón, E., et al.: Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. Front. Cardiovasc. Med. **9**, 859310 (2022)
13. Seyyed-Kalantari, L., Zhang, H., McDermott, M., et al.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. **27**, 2176–2182 (2021)
14. 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.) EPIMI ML-CDS TDA4BiomedicalImaging 2022. LNCS, vol. 13755, pp. 14–25. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-23223-7_2
15. Stanley, E.A.M., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. J. Med. Imaging **9**(6), 061102 (2022)
16. Wang, R., Chaudhari, P., Davatzikos, C.: Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies. Proc. Natl. Acad. Sci. U S A **120**(6), e2211613120 (2023)
17. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 289–297. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_32