Mitigating Calibration Bias Without Fixed Attribute Grouping for Improved Fairness in Medical Imaging Analysis

Changjian Shui[1,2](✉), Justin Szeto[1,2], Raghav Mehta[1,2], Douglas L. Arnold[3,4], and Tal Arbel[1,2]

[1] Center for Intelligent Machines, McGill University, Montreal, Canada
{maxshui,jszeto,raghav,arbel}@cim.mcgill.ca
[2] MILA, Quebec AI Institute, Montreal, Canada
[3] Department of Neurology and Neurosurgery, McGill University, Montreal, Canada
douglas.arnold@mcgill.ca
[4] NeuroRx Research, Montreal, Canada

Abstract. Trustworthy deployment of deep learning medical imaging models into real-world clinical practice requires that they be calibrated. However, models that are well calibrated overall can still be poorly calibrated for a sub-population, potentially resulting in a clinician unwittingly making poor decisions for this group based on the recommendations of the model. Although methods have been shown to successfully mitigate biases across subgroups in terms of model accuracy, this work focuses on the open problem of mitigating calibration biases in the context of medical image analysis. Our method does not require subgroup attributes during training, permitting the flexibility to mitigate biases for different choices of sensitive attributes without re-training. To this end, we propose a novel two-stage method: Cluster-Focal to first identify poorly calibrated samples, cluster them into groups, and then introduce group-wise focal loss to improve calibration bias. We evaluate our method on skin lesion classification with the public HAM10000 dataset, and on predicting future lesional activity for multiple sclerosis (MS) patients. In addition to considering traditional sensitive attributes (e.g. age, sex) with demographic subgroups, we also consider biases among groups with different image-derived attributes, such as lesion load, which are required in medical image analysis. Our results demonstrate that our method effectively controls calibration error in the worst-performing subgroups while preserving prediction performance, and outperforming recent baselines.

Keywords: Fairness · Bias · Calibration · Uncertainty · Multiple Sclerosis · Skin Lesion · Disease activity prediction

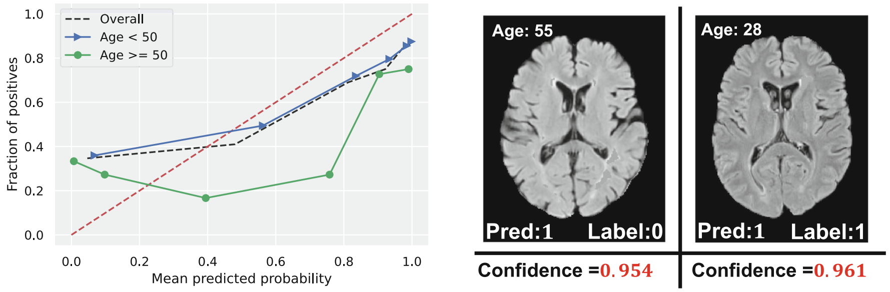
C. Shui and J. Szeto—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_19.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 189–198, 2023.
https://doi.org/10.1007/978-3-031-43898-1_19

1 Introduction

Deep learning models have shown high prediction performance on many medical imaging tasks (e.g., [3,15,21,24]). However, deep learning models can indeed make errors, leading to distrust and hesitation by clinicians to integrate them into their workflows. In particular, models that show a tendency for overconfident incorrect predictions present real risk to patient care if deployed in real clinical practice. One way to improve the trustworthiness of a model is to ensure that it is well-calibrated, in that the predicted probabilities of the outcomes align with the probability of making a correct prediction [8]. While several methods have been shown to successfully improve calibration on the *overall* population [8,16], they cannot guarantee a small calibration error on *sub-populations*. This can lead to a lack of fairness and equity in the resulting diagnostic decisions for a subset of the population. Figure 1(a) illustrates how a deep learning model can achieve good calibration for the overall population and for younger patients, but produces significantly overconfident and incorrect predictions for older patients.



(a) Reliability diagram: ERM

(b) Example Calibration Results: MS

Fig. 1. Illustration of calibration bias for a model that predicts future new lesional activity for multiple sclerosis (MS) patients. (a) Reliability diagram: ERM (training without considering any fairness) exhibits good calibration overall and also for younger patients, whereas it produces significantly overconfident and incorrect predictions for older patients. (b) Two MS patients depicting highly confident predictions, with incorrect results on the older patient and correct results on the younger patient. Poorer calibration for older patients results in older patients being more likely to be incorrect with high confidence.

Although various methods have been shown to successfully mitigate biases by improving prediction performance (e.g. accuracy) in the worst-performing subgroup [1,13,18,27,28], improved prediction performance does not necessarily imply better calibration. As such, this paper focuses on the open problem of mitigating calibration bias in medical image analysis. Moreover, our method does not require subgroup attributes during the training, which permits the flexibility to mitigate biases for different choices of sensitive attributes without re-training.

This paper proposes a novel two-stage method: Cluster-Focal. In the first stage, a model f_{id} is trained to identify poorly calibrated samples. The samples are then clustered according to their calibration gap. In the next stage, a prediction model f_{pred} is trained via group-wise focal loss. Extensive experiments are performed on (a) skin lesion classification, based on the public HAM10000 dataset [3], and (b) on predicting future new lesional activity for multiple sclerosis (MS) patients on a proprietary, federated dataset of MRI acquired during different clinical trials [2,7,26]. At test time, calibration bias mitigation is examined on subgroups based on sensitive demographic attributes (e.g. age, sex). In addition, we consider subgroups with different image-derived attributes, such as lesion load. We further compare Cluster-Focal with recent debiasing methods that do not need subgroup annotations, such as EIIL (Environment Inference for Invariant Learning) [4], ARL (Adversarially Reweighted Learning) [10], and JTT (Just Train Twice) [14]. Results demonstrate that Cluster-Focal can effectively reduce calibration error in the worst-performing subgroup, while preserving good prediction performance, when split into different subgroups based on a variety of attributes (Fig. 2).

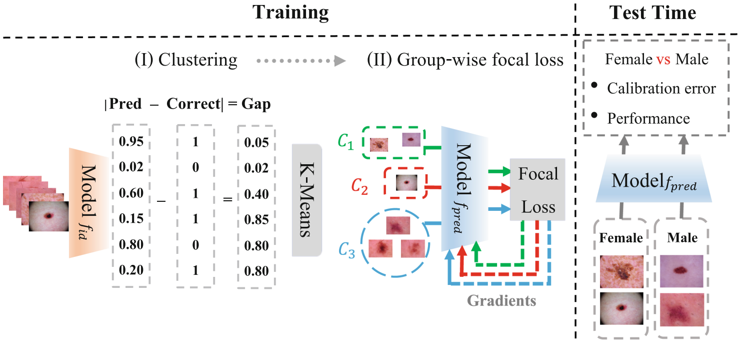


Fig. 2. Cluster-Focal framework. The training procedure is a two-stage method, poorly calibrated sample identifications (clustering) and group-wise focal loss. At test time, the trained model f_{pred} is deployed, then calibration bias and prediction performance are evaluated across various subgroup splittings such as sex or age. (Female/male patients are visualized as an example.)

2 Methodology

We propose a two-stage training strategy, Cluster-Focal. The first stage consists of *identifying different levels of poorly calibrated samples*. In the second stage, we introduce a group-wise focal loss to mitigate the calibration bias. At test time, our model can mitigate biases for a variety of relevant subgroups of interest.

We denote $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ as a dataset, where $\mathbf{x}_i$ represents multi-modal medical images and $y_i \in \{1, 2, \dots\}$ are the corresponding ground-truth class label. A neural network f produces $\hat{p}_{i,y} = f(y|\mathbf{x}_i)$, the predicted probability for a class y given $\mathbf{x}_i$. The predicted class for an $\mathbf{x}_i$ is defined as $\hat{y}_i = \text{argmax}_y \; \hat{p}_{i,y}$, with the corresponding prediction confidence $\hat{p}_i = \hat{p}_{i,\hat{y}_i}$.

2.1 Training Procedure: Two-Stage Method

Stage 1: Identifying Poorly Calibrated Samples (Clustering). In this stage, we first train a model f_{id} via ERM [25], which implies training a model by minimizing the average training cross entropy loss, without any fairness considerations. f_{id} is then used to identify samples that have potentially different calibration properties. Concretely, we compute the gap between prediction confidence $\hat{p}_i$ and correctness via f_{id}:

$$\text{gap}(\mathbf{x}_i) = |\hat{p}_i - \mathbf{1}\{\hat{y}_i = y_i\}|, \tag{1}$$

where $\hat{p}_i$ is the confidence score of the predicted class. Intuitively, if $\text{gap}(\mathbf{x}_i)$ is small, the model made a correct and confident prediction. When $\text{gap}(\mathbf{x}_i)$ is large, the model is poorly calibrated (i.e. incorrect but confident) for this sample. When the model makes a relatively under-confident prediction, $\text{gap}(\mathbf{x}_i)$ is generally in between the two values. We apply *K-means* clustering on the gap values, $\text{gap}(\mathbf{x}_i)$, to identify K clusters $(C_1, \ldots, C_K)$ with different calibration properties.

Stage 2: Group-Wise Focal Loss. We then train a prediction model f_{pred} with a group-wise focal loss on the clusters $C_1, \ldots, C_K$ identified in the first stage. Formally, the following loss is used:

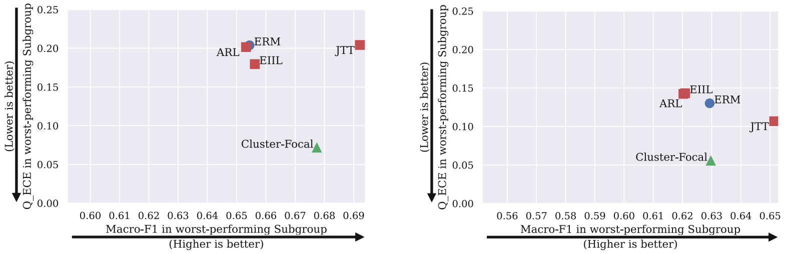
$$\mathcal{L}_{\text{g-focal}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{C_k}(f_{\text{pred}}),$$

where $\mathcal{L}_{C_k}(f_{\text{pred}}) = -\mathbb{E}_{(\mathbf{x}_i, y_i) \sim C_k} \left[(1 - f_{\text{pred}}(y_i|\mathbf{x}_i))^{\gamma} \log(f_{\text{pred}}(y_i|\mathbf{x}_i)) \right]$ with $\gamma > 0$. Intuitively, the focal loss penalizes confident predictions with an exponential term $(1 - f_{\text{pred}}(y_i|\mathbf{x}_i))^{\gamma}$, thereby reducing the chances of poor calibration [16]. Additionally, due to clustering based on $\text{gap}(\mathbf{x}_i)$, poorly calibrated samples will end up in the same cluster. The number of samples in this cluster will be small compared to other clusters for any model with good overall performance. As such, doing focal loss separately on each cluster instead of on all samples will implicitly increase the weight of poorly calibrated samples and help reduce bias.

2.2 Test Time Evaluation on Subgroups of Interest

At test time, we aim to mitigate the calibration error for the **worst-performing subgroup** for various subgroups of interest [6]. For example, if we consider sex (M/F) as the sensitive attribute and denote $\text{ECE}_{A=M}$ as the expected calibration error (ECE) on male patients, then the worst-performing subgroup ECE is denoted as $\max(\text{ECE}_{A=F}, \text{ECE}_{A=M})$. Following the strategy proposed in [17,19], we use Q(uantile)-ECE to estimate the calibration error, an improved estimator for ECE that partitions prediction confidence into discrete bins with an *equal* *number of instances* and computes the average difference between each bin's accuracy and confidence.

In practice, calibration performance cannot be considered in isolation, as there always exists a *shortcut* model that can mitigate calibration bias but have poor prediction performance, e.g., consider a purely random (under-confident) prediction with low accuracy. As such, there is an inherent **trade-off** between calibration bias and prediction error. When measuring the effectiveness of the proposed method, the objective is to ensure that calibration bias is mitigated without a substantial increase in the prediction error.



(a) Age (≤ 60, > 60)

(b) Sex (male, female)

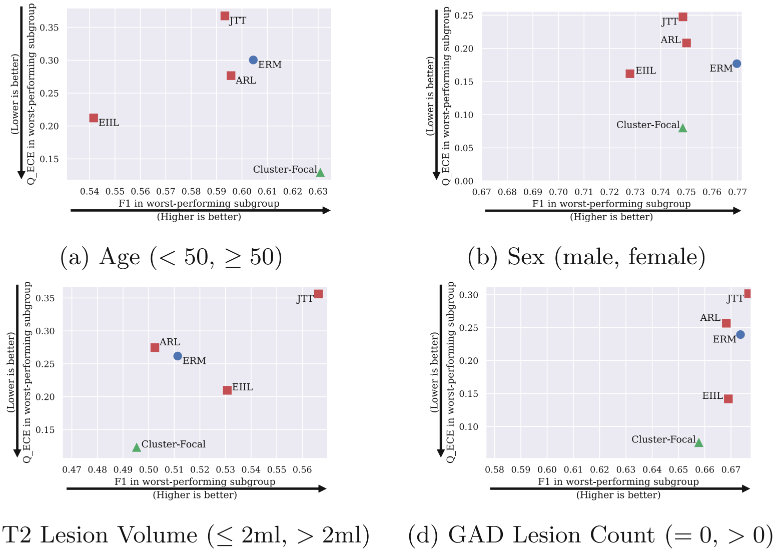
Fig. 3. HAM10000: worst performing subgroup results. Cluster-Focal: Proposed method; ERM: Vanilla model; EIIL, ARL, JTT: Bias mitigation methods. Cluster-Focal demonstrates a better trade-off, significantly improving worst-performing calibration with only a small degradation in prediction performance.

3 Experiments and Results

Experiments are performed on two different medical image analysis tasks. We evaluate the performance of the proposed method against popular debiasing methods. We examine whether these methods can mitigate calibration bias without severely sacrificing performance on the worst-performing subgroups.

Task 1: Skin lesion multi-class ($n = 7$) classification. HAM10000 is a public skin lesion classification dataset containing 10,000 photographic 2D images of skin lesions. We utilize a recent MedFair pipeline [27] to pre-process the dataset into train (80%), validation (10%) and test (10%) sets. Based on the dataset and evaluation protocol in [27], we test two demographic subgroups of interest: age (age ≤ 60, age > 60), and sex (male, female).

Task 2: Future new multiple sclerosis (MS) lesional activity prediction (binary classification). We leverage a large multi-centre, multi-scanner proprietary dataset comprised of MRI scans from 602 RRMS (Relapsing-Remitting MS) patients during clinical trials for new treatments [2,7,26]. The task is to predict the (binary) presence of new or enlarging T2 lesions or Gadolinium-enhancing lesions two years from their current MRI. The dataset was divided



(a) Age (< 50, ≥ 50)

(b) Sex (male, female)

(c) T2 Lesion Volume (≤ 2ml, > 2ml)

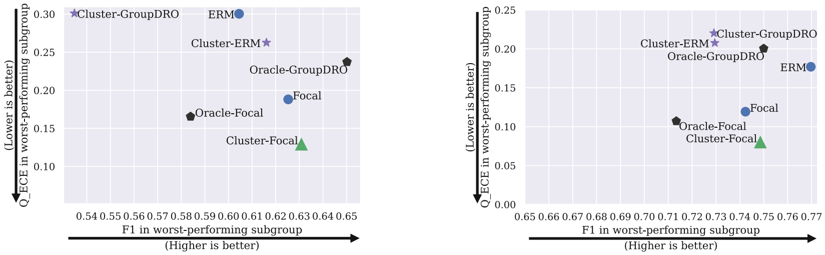
(d) GAD Lesion Count (= 0, > 0)

Fig. 4. MS: worst performing subgroup results. Cluster-Focal: proposed method; ERM: Vanilla Model; EIIL, ARL, JTT: bias mitigation methods.

as follows: training (70%) and test (30%) sets, validation is conducted through 4-fold cross validation in training set. We test model performance on four different subgroups established in the MS literature [5,11,12,22,23]. This includes: age (age < 50, age ≥ 50), sex (male, female), T2 lesion volume (vol ≤ 2.0 ml, vol > 2.0 ml) and Gad lesion count (count $= 0$, count > 0). Age and sex are sensitive demographic attributes that are common for subgroup analysis. The image-derived attributes were chosen because high T2 lesion volume, or the presence of Gad-enhancing lesions, in baseline MRI is generally predictive of the appearance of new and enlarging lesions in future images. However, given the heterogeneity of the population with MS, subgroups *without* these predictive markers can still show future lesional activity. That being said, these patients can form a subgroup with poorer calibration performance.

Implementation Details: We adopt 2D/3D ResNet-18 [9] for Task 1 and Task 2 respectively. All models are trained with Adam optimizer. Stage 1 model f_{id} is trained for 10 (Task 1) and 300 (Task 2) epochs and Stage 2 prediction model f_{pred} for 60 (Task 1) and 600 (Task 2) epochs. We set the number of clusters to 4 and $\gamma = 3$ in group-wise focal loss. Averaged results across 5 runs are reported.

Comparisons and Evaluations: Macro-F1 is used to measure the performance for Task 1 (7 class), and F1-score is used for Task 2 (binary). Q-ECE [16] is used to measure the calibration performance for both tasks. The performance of the proposed method is compared against several recent bias mitigation methods that do not require training with subgroup annotations: ARL [10], which applies a min-max objective to reweigh poorly performing samples; EIIL [4], which pro-



(a) Age ($< 50, \geq 50$) (b) Sex (Male, Female)

Fig. 5. Ablation Experiments for MS. Focal: regular focal loss without stage 1; Cluster-ERM: In stage 2, cross entropy loss is used; Cluster-GroupDRO: In stage 2, GroupDRO loss is used; Oracle-Focal: identified cluster in stage 1 is replaced by the subgroup of interest (oracle); Oracle-GroupDRO: GroupDRO method applied on the subgroups of interest.

poses an adversarial approach to learn invariant representations, and JTT [14], which up-weights challenging samples. Comparisons are also made against ERM, which trains model without any bias mitigation strategy. For all methods, we evaluate the trade-off between the prediction performance and the reduction in Q-ECE error for the **worst-performing subgroups** on both datasets.

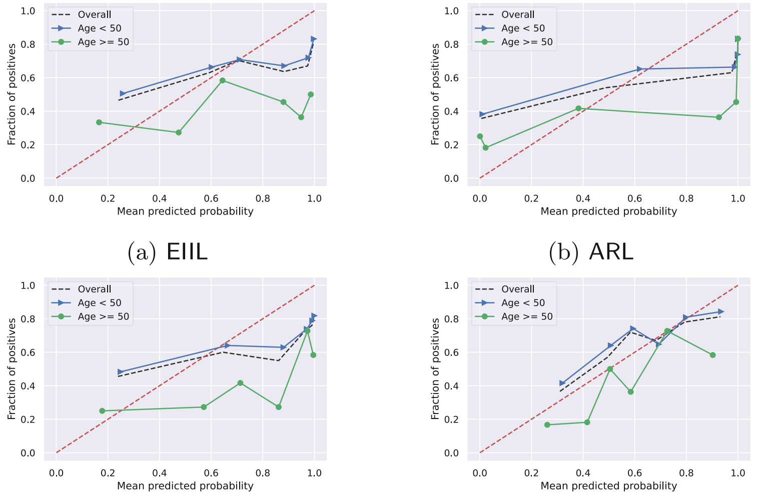
3.1 Results, Ablations, and Analysis

Results: The resulting performance vs. Q-ECE errors tradeoff plots for worst-performing subgroups are shown in Figs. 3 and 4. The proposed method (Cluster-Focal) consistently outperforms the other methods on Q-ECE while having minimal loss in performance, if any. For instance, when testing on sex (male/female) for the MS dataset, (Cluster-Focal) loses around 2% prediction performance relative to (ERM) but has around 8% improvement in calibration error. When testing on sex in the HAM10000 dataset, we only observe a 2% performance degradation with a 4% improvement in Q-ECE.

In addition to subgroups based on sensitive demographic attributes, we investigate how the methods perform on subgroups defined on medical image-derived features. In the context of MS, results based on subgroups, lesion load or Gad-enhancing lesion count are shown in Fig. 4(c–d). The proposed method performs best, with results that are consistent with demographic based subgroups. For Gad-enhancing lesion count, when compared with JTT, Cluster-Focal improves Q-ECE by 20%+ with a reduction in the prediction performance on the worst-performing subgroup of 2%. Detailed numeric values for the results can be found in the Supplemental Materials.

Ablation Experiments: Further experiments are performed to analyze the different components of our method. The following variant methods are considered: (1) Focal: Removing stage 1 and using regular focal loss for the entire training set; (2) Cluster-ERM: Group-wise focal loss in stage 2 is replaced by standard cross entropy; (3) Cluster-GroupDRO: Group-wise focal loss in stage 2 is replaced by GroupDRO [20]; (4) Oracle-Focal: In stage 1, the identified cluster is replaced by the true subgroups evaluated on at test time (oracle); (5) Oracle-GroupDRO: We use GroupDRO with the true subgroups used at test time. Results for MS, shown in Fig. 5, illustrate that each stage of our proposed model is required to ensure improved calibration while avoiding performance degradation for the worst-performing subgroups.

Calibration Curves: Figure 6 shows the reliability diagram for competing methods on Task 2: predicting future new MS lesional activity, with age being the chosen subgroup of interest (also see Fig. 1(a) for ERM results). Results indicate that popular fairness mitigation methods are not able to correct for the calibration bias in older patients (i.e. the worst-performing subgroup). With ARL, for example, most of the predictions were over-confident, resulting in a large calibration error. In contrast, our proposed method (Cluster-Focal) could effectively mitigate the calibration error in the worst-performing subgroup.



(a) EIIL

(b) ARL

(c) JTT

(d) Cluster-Focal (ours)

Fig. 6. MS: Reliability diagram for bias mitigation methods with age-based subgroups: (a) EIIL, (b) ARL, (c) JTT, and (d) Cluster-Focal.

4 Conclusions

In this paper, we present a novel two stage calibration bias mitigation framework (Cluster-Focal) for medical image analysis that (1) successfully controls the trade-off between calibration error and prediction performance, and (2) flexibly overcomes calibration bias at test time without requiring pre-labeled subgroups during training. We further compared our proposed approach against different debiasing methods and under different subgroup splittings such as demographic subgroups and image-derived attributes. Our proposed framework demonstrates smaller calibration error in the worst-performing subgroups without a severe degradation in prediction performance.

Acknowledgements. This paper was supported by the Canada Institute for Advanced Research (CIFAR) AI Chairs program and the Natural Sciences and Engineering Research Council of Canada (NSERC). The MS portion of this paper was supported by the International Progressive Multiple Sclerosis Alliance (PA-1412-02420), the companies who generously provided the MS data: Biogen, BioMS, MedDay, Novartis, Roche/Genentech, and Teva, Multiple Sclerosis Society of Canada, Calcul Quebec, and the Digital Research Alliance of Canada.

References

1. Burlina, P., Joshi, N., Paul, W., Pacheco, K.D., Bressler, N.M.: Addressing artificial intelligence bias in retinal diagnostics. Transl. Vision Sci. Technol. **10**(2), 13–13 (2021)
2. Calabresi, P.A., et al.: Pegylated interferon beta-1a for relapsing-remitting multiple sclerosis (ADVANCE): a randomised, phase 3, double-blind study. Lancet Neurol. **13**(7), 657–665 (2014)
3. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
4. Creager, E., Jacobsen, J.H., Zemel, R.: Environment inference for invariant learning. In: International Conference on Machine Learning, pp. 2189–2200. PMLR (2021)
5. Devonshire, V., et al.: Relapse and disability outcomes in patients with multiple sclerosis treated with fingolimod: subgroup analyses of the double-blind, randomised, placebo-controlled FREEDOMS study. The Lancet Neurology **11**(5), 420–428 (2012)
6. Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A.: Minimax group fairness: algorithms and experiments. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 66–76 (2021)
7. Gold, R., et al.: Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. N. Engl. J. Med. **367**(12), 1098–1107 (2012)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Lahoti, P., et al.: Fairness without demographics through adversarially reweighted learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 728–740 (2020)
11. Lampl, C., You, X., Limmroth, V.: Weekly IM interferon beta-1a in multiple sclerosis patients over 50 years of age. Eur. J. Neurol. **19**(1), 142–148 (2012)

12. Lampl, C., et al.: Efficacy and safety of interferon beta-1b SC in older RRMS patients: a post hoc analysis of the beyond study. J. Neurol. **260**(7), 1838–1845 (2013)
13. 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**(23), 12592–12594 (2020)
14. Liu, E.Z., et al.: Just train twice: improving group robustness without training group information. In: International Conference on Machine Learning, pp. 6781–6792. PMLR (2021)
15. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
16. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Adv. Neural. Inf. Process. Syst. **33**, 15288–15299 (2020)
17. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops, vol. 2 (2019)
18. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nat. Commun. **13**(1), 4581 (2022)
19. Roelofs, R., Cain, N., Shlens, J., Mozer, M.C.: Mitigating bias in calibration error estimation. In: International Conference on Artificial Intelligence and Statistics, pp. 4036–4054. PMLR (2022)
20. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020)
21. Sepahvand, N.M., Hassner, T., Arnold, D.L., Arbel, T.: CNN prediction of future disease activity for multiple sclerosis patients from baseline MRI and lesion labels. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 57–69. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_6
22. Signori, A., Schiavetti, I., Gallo, F., Sormani, M.P.: Subgroups of multiple sclerosis patients with larger treatment benefits: a meta-analysis of randomized trials. Eur. J. Neurol. **22**(6), 960–966 (2015)
23. Simon, J., et al.: Ten-year follow-up of the 'minimal MRI lesion' subgroup from the original CHAMPS Multiple Sclerosis Prevention Trial. Multiple Sclerosis J. **21**(4), 415–422 (2015). Publisher: SAGE Publications Ltd. STM
24. Tousignant, A., Lemaître, P., Precup, D., Arnold, D.L., Arbel, T.: Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. In: Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning (MIDL), vol. 102, pp. 483–492. PMLR, 08–10 July 2019
25. Vapnik, V.: Principles of risk minimization for learning theory. In: Advances in Neural Information Processing Systems, vol. 4 (1991)
26. Vollmer, T.L., et al.: On behalf of the BRAVO study group: a randomized placebo-controlled phase III trial of oral laquinimod for multiple sclerosis. J. Neurol. **261**(4), 773–783 (2014)
27. Zong, Y., Yang, Y., Hospedales, T.: Medfair: benchmarking fairness for medical imaging. In: International Conference on Learning Representations (ICLR) (2023)
28. Zou, J., Schiebinger, L.: AI can be sexist and racist-it's time to make it fair. Nature (2018)