AEquity: A Deep Learning Based Metric for Detecting, Characterizing and Mitigating Dataset Bias

Team Members: Faris F. Gulamali, Ashwin S. Sawant, Jianying Hu, Girish N. Nadkarni

Link to video:

https://youtu.be/SnyV2KNoEuo

Contact Information

Girish N. Nadkarni (girish.nadkarni@mountsinai.org; T - 212-241-1385)

Abstract

Diagnostic and prognostic algorithms can recapitulate and perpetuate systemic biases against underserved populations. Post-hoc technical solutions are often insufficient on its own because they are unable to overcome biased data used to train algorithms. A more data-centric approach may help address bias earlier in the process of development of algorithms. We present a sample-efficient deep-learning based metric (AEq) that measures the learnability of subsets of data representing underserved populations. We then apply a systematic analysis of AEq values across subpopulations in order to identify various manifestations of bias in two healthcare datasets known to be affected. In the first case, we analyzed a computer vision model trained on chest radiographs to investigate its underperformance on chest X-rays from Black individuals for various diagnoses and developed targeted interventions at the dataset level that mitigated bias. We also applied AEq to a cost prediction model to illustrate how the labeling choices lead to bias against Black patients, which can be remedied by predicting comorbidities instead of costs. AEq is a novel and broadly applicable metric which can be applied to advance equity by diagnosing and remediating bias in healthcare datasets.

Github Code

https://github.com/Nadkarni-Lab/AEquity

Methodology Overview

Algorithmic bias is the inability of an algorithm to generalize to a given group, causing selective underperformance (*1*). This is pervasive in healthcare and perpetuates existing disparities. Algorithms developed using standard techniques on diverse datasets can still be significantly biased against individuals based on race/ethnicity or insurance status. For example, the application of standard computer vision models to chest radiographs resulted in selective underdiagnosis in people who are Black, Latino or receive Medicaid (*2*, *3*), and an algorithm used to predict health needs was found to exhibit significant racial bias against Black patients (*4*). This occurs because algorithms developed using data from systems with longstanding discrimination and inequities tend to recapitulate those biases (*5*).

Post-hoc technical solutions can help mitigate bias to a certain extent (*6*). However, these typically involve recalibration which, except in narrow circumstances, implies a tradeoff between sensitivity and specificity in underserved populations, leading to worse performance (*7*). Thus, there is increasing recognition of an urgent need to address bias earlier in the algorithm development pipeline (*5*, *8*).

We developed an approach that measures potential biases by investigating the underlying structure of population-specific subsets of the data (**Fig 1**). This then allows for appropriate augmentation to the existing dataset to better represent that population.

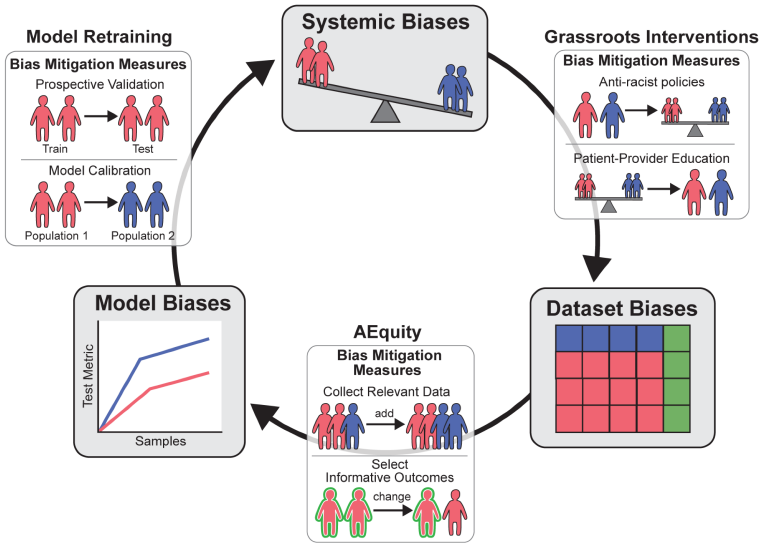

Figure 1. Description of Different Biases and Proposed Role of AEq in Identifying and Mitigating Bias. Systemic biases are captured in datasets. Algorithms trained on these data tend to perpetuate these biases. AEq, the technique described in this paper, examines causes of bias in the dataset and provides actionable feedback to reduce bias early in algorithm development. Prospective validation and calibration of trained models are existing methods of

detecting and mitigating bias. The underlying systemic biases should also be mitigated by policy changes and education.

In this manuscript, we describe a deep learning-based metric that can be used for the characterization and mitigation of social and predictive biases in healthcare datasets. We call this metric AEq because it utilizes an autoencoder to provide actionable, data-driven feedback towards equitable performance of models. We have previously shown that unsupervised learning with an autoencoder can estimate the minimum sample size needed to train a deep neural network on a given dataset - minimum convergence sample estimate (MCSE) (*10*). We define AEq as the MCSE stratified by a group characteristic, such as race or socioeconomic status, and a class such as “pneumonia” or “edema” for Chest X-rays. In general, a higher AEq value for a group within a given diagnostic label suggests that generalization is more difficult and more samples are needed for learning. AEq is reported as log_2(sample size estimate) for interpretability.

Under most constraints, AEq values exhibit dataset-specific properties, and can be used subsequently to guide data-driven, actionable feedback such as selecting informative outcomes, collecting more diverse data or prioritizing collection of data from a specific subgroup (**Fig 2**). In conjunction with model-centric solutions and grassroots interventions, AEq can help mitigate propagation of systemic biases in “black-box” machine learning models. For this tool, we use standard imaging datasets to demonstrate how AEq changes with different types of biases. We then demonstrate its applicability to mitigate under-diagnosis bias in a deep learning model trained on chest radiographs, and an under-allocation bias in a model used to predict healthcare needs. Finally, we provide an easy-to-use, Python library to generate AEq values for tabularized datasets with the demonstration on social and predictive biases.

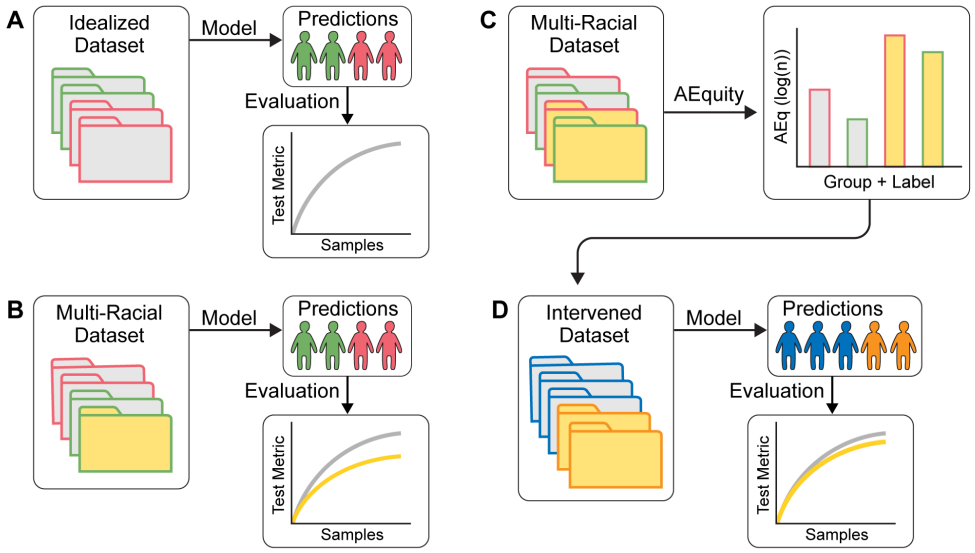

Fig. 2. AEq pipeline in practice. A. The red and green outlines represent the labels. The gray represents a relatively homogenous group. A model is trained on what is thought to be a standard dataset and appears to perform well. (B) The red and green outlines represent

different labels, while the gray represents the over-represented group while yellow represents the under-represented group. When naively trained in a similar manner on a more diverse dataset, the resulting model is biased because it performs worse for a particular group, because of systematic biases reflected in the data. (C) Application of AEq enables disentanglement of dataset-level roots of the resulting bias. AEq is applied directly to the dataset and can generate values for each group and label. (D) Actionable feedback from AEq reduces bias by improving performance for the disadvantaged group. Two types of actionable feedback are depicted – the blue and orange borders represent the selection of more informative labels. Second, the yellow and gray samples added to the dataset highlight a targeted dataset collection.

Value Proposition

We use “groups” to refer to various subpopulations, for example race, gender, or insurance status. We use “class” to refer to diagnostic labels like “pneumonia” and “edema”.

We used standard terminology to separate dataset biases into three distinct categories: sampling bias; complexity bias; and label bias (**Fig 3**) (*1*). Sampling bias arises when there is non-random sampling of patients, and this results in a lower or higher probability of sampling one group over another. Under-representation can negatively impact generalization performance. Sampling bias is primarily driven by data availability – in dermatology datasets, for example, lighter-skinned populations are generally more represented and therefore have better generalizability that darker-skinned populations (*9*). In complexity bias, a group presents more heterogeneously with a diagnosis, and consequently, one group exhibits a class label with greater complexity than another group. Subsequently, generalization performance on that group is worse than other groups for the same class. For example, Hispanic patients with rheumatoid arthritis tend to have a delayed initial presentation to a rheumatologist, and therefore have a greater range of manifestations and severity. As a result, Hispanic individuals present more heterogeneously with rheumatoid arthritis than their counterparts, which makes diagnosis more difficult (*28*). Third, label bias occurs when labels are placed incorrectly at different rates for different groups and can lead to increased misclassification errors for the affected group. For example, Black patients are incorrectly labeled as requiring fewer healthcare resources despite having the same number of comorbidities because of lack of access to care (*4*).

To examine the relationship of AEq and sampling bias, we generated two synthetic groups in which the frequency of class labels are distributed differently (**Fig 3c**). In group 1, the samples are drawn at lower frequencies, and subsequently each label in the group has a higher AEq value. In group 2, the class labels at lower values are drawn at higher frequencies and have lower AEq values. Higher frequency of sampling resulted in easier generalization, and therefore lower AEq values. When sampling bias was the only type of dataset bias, combining groups drove the AEq value closer to the over-sampled data.

Complexity bias was characterized as different groups representing different distributions across a given label. We generated two synthetic datasets with differing complexity across groups by changing the number of informative features (**Fig 3d**) and observed that the AEq of one group was significantly larger than the other group across specific classes. When complexity bias was the only type of dataset bias, combining the groups resulted either in an

increase in the value of AEq or an AEq closer to value that had been higher prior to the combination.

We simulated label bias by generating synthetic data with the same underlying distribution, but with varying percentages of correctly assigned labels (**Fig 3e**). We quantified label bias as the percentage of labels that were considered correct or valid for a given group, and the subsequent discrepancies between those labels. We utilized a standard image dataset but mislabeled class at rates ranging from 10% to 90% and showed that the percentage of mislabeled data did not change the group-specific AEq values, because AEq values for a given group are label-agnostic.

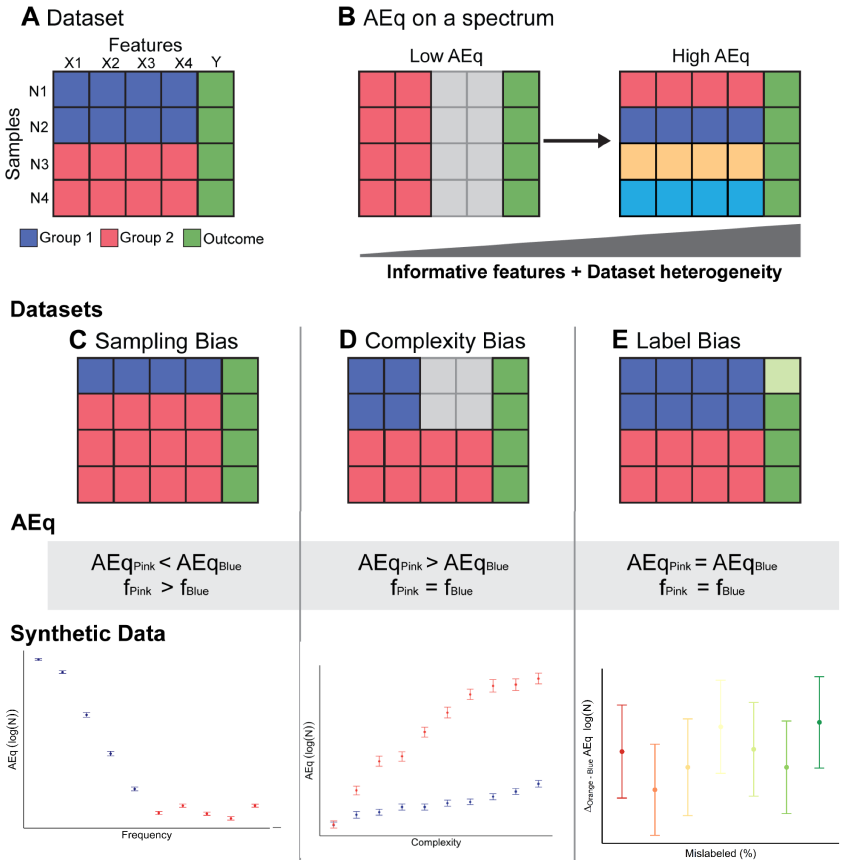

Figure 3. Qualification and Quantification of Bias with AEq. A. A prototypical dataset containing samples from two groups each of which has four informative features. Group 1 is red, group 2 is blue, and the classes are in green. B. The dataset on the left has examples belonging to a single group, having only two informative features. It requires fewer samples to generalize to, and thus has a lower AEq value. The dataset on the right has examples from four different groups (each indicated by a different color) having up to four informative features (indicated by 4 columns). This dataset is harder to learn and has a higher AEq value. C. Sampling bias results in the AEq of the over-represented group being significantly smaller than the AEq of the under-represented, and usually is represented by different underlying sampling frequencies. The X-axis represents sampling frequencies of a class belonging to a group ranging from 10% to 90%. D. Complexity bias results in the AEq of one group being significantly larger than the AEq

of the other group despite similar sampling frequencies. E. Label bias can be demonstrated with similar AEq values and frequencies, but different underlying generalization performances.

Healthcare Scenario

#1 Identification of Social Biases in Healthcare Cost Utilization.

In a seminal paper, Obermeyer *et al* (*4*) showed that racial bias in cost-predictive algorithms can propagate systemic forms of inequity in healthcare resource allocation. When using cost as the risk-derived metric, Black patients at a given risk score were considerably sicker than their white counterparts. This bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care led to less being spent on care for Black patients. This bias arises despite using calibration-based fairness metrics and is invariant to including an indicator variable for “race.” Thus, traditional metrics are insufficient to capture the bias. On switching from a cost-based metric to one that utilizes comorbidities, the algorithm was able to mitigate some of the bias.

We conducted an analysis utilizing AEq to replicate the results and conclusions of the dataset provided by the authors of Obermeyer *et al* (4). They showed that racial bias in cost-predictive algorithms can propagate systemic forms of inequity in health care resource allocation. Using cost as the risk-derived metric, Black patients at a given risk score were considerably sicker than their White counterparts.

We calculated the difference in AEq values between each race for each metric (**Fig 4**). The model demonstrated that when using active chronic conditions as the outcome metric, the difference in AEq between Black and White patients was not statistically significantly different from zero in the high-risk group ($P = 0.22$). In contrast, when using cost-based metrics such as avoidable costs and total costs, the difference in AEq between Black and white patients was significantly greater than 0 ($P_{total\ cost} = 6.49 \times 10^{-10}$; $P_{avoidable\ costs} = 1.67 \times 10^{-10}$), consistent with generalizability differences across different races, which fits the definition of algorithmic label bias.

We conducted additional analyses on the low-risk group and found that the differences in AEq across races of the low-risk groups were all statistically significant. We also observed that the differences in AEq were smaller when using active chronic conditions than avoidable costs or total costs, (ANOVA, $P = 1.59 \times 10^{-8}$).

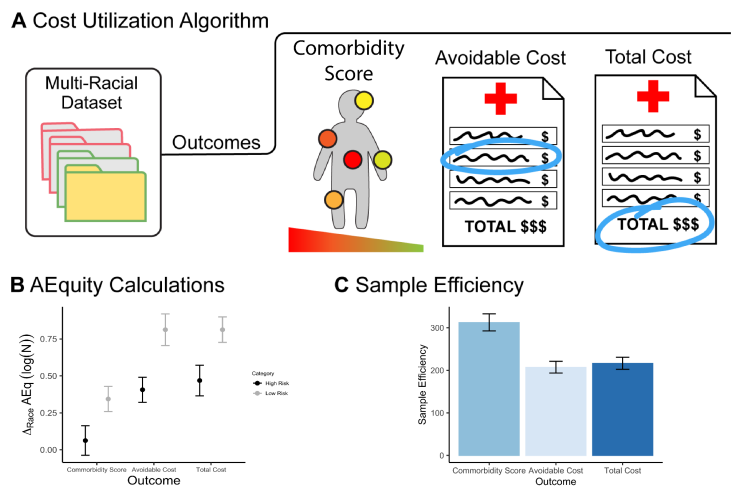

Figure 4. Detecting systemic social biases in healthcare allocation with AEq using electronic health record data. A. A dataset with claims data used to calculate comorbidity score, and cost-based metrics. B. Difference in AEq values between Black and White patients for comorbidity- and cost-based outcomes, stratified by risk level. C. Sample efficiency of AEq stratified by outcome.

#2 Identification of Predictive Biases with Deep Convolutional Networks

We show that AEq identifies the racial biases demonstrated by Seyyed-Kalantari *et al* (*3*) in the MIMIC-CXR (*11–13*) dataset. We saw generally that a higher frequency of occurrence for a diagnosis corresponded to a lower AEq value across Black and White patients (**Fig 5c**). This indicates that labels with more samples, like lung opacity (AEq ≈ 7.5), generally trend towards lower AEqs, whereas labels with fewer samples such as consolidation (AEq ≈ 9.8), enlarged cardio-mediastinum (AEq ≈ 9.9) or pneumothorax (AEq ≈ 9.95) trend towards having higher AEqs. We posit that the latter group of diagnoses may have the highest potential for bias because of difficulty in generalization. Second, we noticed that six out of the nine diagnoses, accounting for 71% of the positive samples, demonstrate higher AEq values in Black patients than in White patients. Higher AEq values indicate that, in general, models built on chest X-rays from Black patients for Black patients generalize more poorly than models built on chest X-rays from White patients for the purpose of diagnosing White patients. In other words, more samples from Black patients are required to train a model that achieves equitable performance, demonstrating the utility of AEq to provide data-driven feedback to mitigate biases.

Next, we used AEq to guide data collection and highlight the improvements in generalization performance for Black patients. In the examples provided, the number of samples required to calculate the AEq is marked by the dashed red line. Above the AEq, we simulated two data collection strategies - a naïve approach guided by population prevalence and second, a targeted approach suggested by AEq. First, we examined pneumothorax, which occurs less frequently in Black individuals and for whom AEq is substantially higher than that for White individuals. The joint AEq is less than the AEq for each race, consistent with a sampling bias. A lower joint AEq for a race-balanced dataset suggests that sampling equally from Black and White patients would have a better generalization performance. Subsequently, we see that

adding dataset diversity by equitably sampling from each race was associated with an improvement in classifier performance for Black individuals when compared to the naïve approach (**Fig 5d**, left pane). Second, we looked at edema, where the joint AEq was higher than the AEq for either race, consistent with complexity bias. Complexity bias suggests two distinct distributions for each race, and therefore to improve generalization performance for Black patients, more samples would need to be added from the Black patient population. Prioritizing data collection from Black patients better captured the group-specific presentations, and this strategy was associated with a model that generalized better to Black patients when compared to a naïve approach (**Figure 5d**, right pane).

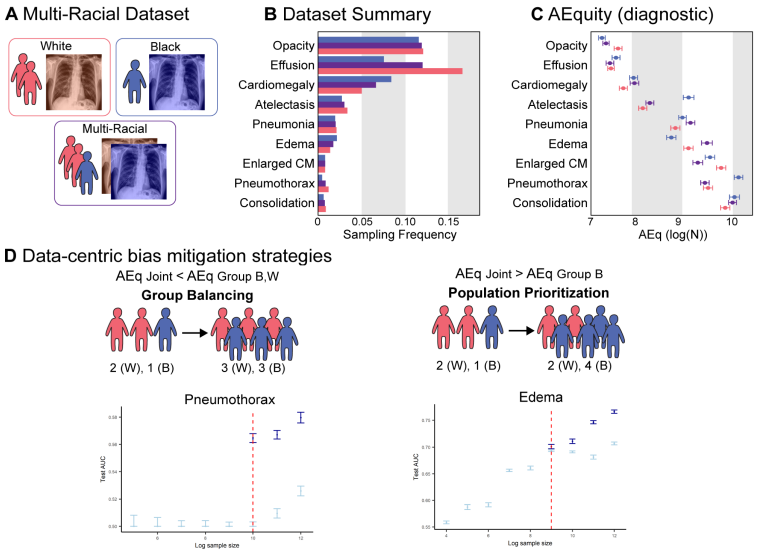

Figure 5. Application of AEq to detect and mitigate predictive biases in Chest X-rays. A. Red – White individuals; blue – Black individuals. B. Sampling frequency by diagnosis and race. C. AEq values (log(N)) by diagnosis and race. D. Data-centric mitigation strategies. Left pane – intervention to increase sampling diversity leads to a 0.02 increase in test AUC for pneumothorax for Black individuals (AUC_{Pre} = 0.60; 95% CI: (0.59, 0.60), AUC_{Post} = 0.62; 95% CI: (0.61 – 0.63), $P < 0.05$, $n_{samples}$ = 8192). The improvement was greater at smaller sample sizes. Right pane – intervention to stratify by race and add more examples from Black individuals leads to 0.07 increase in test AUC for edema for Black individuals (AUC_{Pre} = 0.72; 95% CI: (0.71, 0.73), AUC_{Post} = 0.79; 95% CI: (0.79 – 0.80), $P < 0.05$, $n_{samples}$ = 8192).

Operational Requirements

The technical and operational requirements include python 3.8.2. The required Python packages include PyTorch, SciKitLearn, Pandas and TorchVision. All the data mitigation experiments can be run on a single 4 CPU machine requiring 7 minutes for 5 bootstraps on the tabularized dataset with 149 columns. For imaging data, using the CUDA toolkit with a single NVidia GPU required only 25 minutes for a dataset of ___ chest radiographs. The tool only requires data, which can be collected and aggregated at individual institutions. The datasetmust have the input dimension, labels, and demographic information. We employ a data-centric

approach that eliminates the need to replicate proprietary models, which is a major limitation for model-centric approaches.

At the core of AEquity is an autoencoder. This provides two distinct advantages. First, auto-encoders are generalizable applicable to various data types, and we show that our method works on both tabularized data, as well as more complex image data. Second, the python package relies on pytorch and numpy, which means we can utilize seeding to ensure replicability across different hardware solutions. The code is provided at Nadkarni-Lab/AEquitas: Deep Learning Based Metric to Mitigate Dataset Bias (github.com), and we acknowledge that the tool can be distributed under the BSD 3 license. We have not patented or placed any explicit limitations on the distribution and use of this tool..

Sustainability Plan

AEquity has been developed by a cross-disciplinary team with expertise in software engineering, clinical informatics, machine learning model development and deployment. The tool is designed to be used early in the development of new machine learning models, to help mitigate the risk of biased models arising from systematic bias captured in data. When the potential for bias against a group is detected, it can help direct collection of additional data, potentially from the affected group, to mitigate such bias. Thus, it enables a process wherein the detection and mitigation of bias can take place prior to the expenditure of significant amounts of time and resources. It complements, but does not replace, ongoing monitoring of algorithmic performance, including fairness, which can be affected by dataset drift over time. Ideally, an AI/ML governance committee at each institution would develop policies and best practices for the entire lifecycle of an algorithm, from development to post-deployment monitoring.

AEquity is made available as an open-source Python library which can be easily downloaded and used locally by data scientists, minimizing the barrier to adoption. It is written in a modular manner in Python, a technology that is universally available on most systems, and with which most data scientists are familiar. No ongoing monetary costs are expected and no subscription is required.

Generalizability Plan

AEquity is model-agnostic and data-driven, which means that it can be applied across a range of different healthcare settings. We uniquely demonstrate AEquity works for both predictive biases created by deep convolutional neural networks on Chest X-rays, and social biases that are perpetuated by multivariate logistic regression.

The code and instructions for configuring the input file are provided in the github repository. We show that this method works in radiology, and more broadly electronic health record data. Because various forms of dataset bias have already been implicated in fields such as nephrology with the racial biases implicated in the calculation of estimated glomerular filtration rate (eGFR), pulmonology with pulmonary function tests providing race-insensitive measures of lung volume, and dermatology, which inadequate generalization and understanding of melanoma in darker skinned populations, this method, which uniquely targets dataset bias is anticipated to be applicable to these others fields.

While our results are promising, there are some limitations of AEq and directions for future inquiry. First, we do not examine the simultaneous application of multiple dataset interventions suggested by the application of AEq. It is possible that some of the resulting benefits may be attenuated in that scenario. Second, AEq in its current form focuses primarily on classification tasks with an unregularized latent space. A growing number of algorithms, however, are: (a) using various forms of regularization and attention to produce more

informative latent spaces(22, 23); and (b) being used for generation and representation learning rather than simply classification or regression(24). Extending AEq to these latent spaces and generative models may help investigate and mitigate bias in those settings.

While AEq can identify and help mitigate cases where different groups appear to have different distributions for a given label, it cannot explain the root cause of such a difference. These may arise from an unmeasured difference in prevalence of comorbidities caused by differential access to healthcare. We measure performance of algorithms against data obtained from clinical practice which may itself be biased. The implementation of attention modules that highlight feature importance in AEquity may provide some insight into underlying mechanisms, but this requires additional empirical research and mathematical support.

To encourage trust in AI/ML systems, we sought to further investigate the different fairness metrics including True Negative Rate, True Positive Rate, False Positive Rate, as well as further derived fairness metrics such as Precision, False Discovery Rate, and Predicted Prevalence. We see that there is a clinically and statistically significant decrease in false positive rate ($\Delta FPR = 0.10$) and a corresponding increase in true negative rate ($\Delta TNR = 0.10$), with a clinically insignificant difference in false negative rate ($\Delta FNR = -0.01$) and true positive rate ($\Delta TPR = -0.01$). This leads to an improvement in precision, FDR and predicted prevalence.

Implementation Requirements

The implementation strategy and standard operating procedures requires a diverse team including clinicians, computer scientists, healthcare administrators and industry collaborators. At the Icahn School of Medicine at Mount Sinai, as at many other institutions, the data is centralized in a single institutional repository, which in this case is called the Mount Sinai Data Warehouse. The pipeline for creating an algorithmic model requires a clinician to request data from the Mount Sinai Data Warehouse. A software engineer would subsequently pull the data from the internal data storage system and provide the required data to the clinician. The clinician and their team could train the model and then subsequently go through the administration to deploy their model silently in a rapid prospective AI/ML trial.

AEquity can be implemented both prospectively and retrospectively. In the ideal case, AEq can be prospectively calculated with respect to an outcome (like mortality) and a demographic (like race), prior to the development of a model. Statistically significant differences in AEq values by demographic category are predictive of bias.. Based on the AEq values, targeted dataset interventions can be pursued to minimize bias. These interventions may include adjusting the database query to be more inclusive, collecting additional data, or choosing a more appropriate outcome. AEq can also be applied in a retrospective manner - if an existing model is underperforming on a given subset of individuals, the hospital and administrators can request that the source of the algorithm calculate the AEq value for their proprietary dataset. The AEq can be subsequently used by the industry collaborators and healthcare administrators to either improve the model or limit the use of the model in clinical care.

AEq is widely applicable and extendable. We have implemented it for tabular EHR data as well as more complex image based models such as deep convolutional neural networks. The configuration file and code are publicly available. We have also included the ability to use multi-thread and multi-process parallel computation to allow its application to massive datasets. As AEq uses a relatively simple encoder-decoder architecture, the computational resource requirements are not expected to exceed those of downstream model development.

The key human resources as mentioned above include the clinician, who has to account for the calculated AEq values, the software engineer who implements the algorithm for their

local institution, and the administrators and industry collaborators who ultimately decide on if the model has become sufficiently equitable for clinical practice.

AEq is a data-driven solution, and therefore success of implementation can be measured via identifying changes in patterns of data-collection associated with the implementation of AEq. We can both qualitatively and quantitatively evaluate the clinician responses to the AEq by evaluating how it may have changed their decision making process on model training. We can quantitatively evaluate the trends in data collection driven by AEq to determine if AEq has made a significant difference in the kinds of data that are collected.

In terms of model performance, we can use a rapid ML/AI trial, a new set of trials that are being implemented in Mount Sinai. Rapid ML/AI trials can silently and prospectively validate AI algorithms. We can test for differences in performance on various fairness metrics including AEq, FDR, FNR, FPR, and prospectively determine if AEq has made a significant contribution to changing how data is collected and how the model is performing in certain populations.

Lessons Learned

First, AEq provides an objective metric for dataset bias. Previously, it was difficult to quantify bias. We could say that a specific dataset was bias, but it was difficult to attribute the bias to the data itself. Rather, we would say that the FNR was lower for one population than another. However, this previous statement fails to separate the model from the dataset. As a result, it was difficult to account for the data itself. AEq provides a unique quantitative and qualitative metric of the type of bias, as well as “how much”. Second, because AEq suggests data collection strategies that change the underlying structure of the data, the various types of models that require the same kinds of data - for example, segmentation, classification or unsupervised learning models that all require chest X-rays – will be impacted by AEq. Moreover, because models build of previously acquired data for methods like transfer learning, even out of domain tasks that were implicitly affected by biases present in the data may have further reduced bias. Third, because clinicians can take active steps to mitigate these biases, this would help improve trust in healthcare models. If we can routinely say that we saw this difference in AEq, and then mitigated the underlying biases in the actual dataset by collecting more samples - went out to the community and hospitals where data for under-represented indivudals is routinely collected, we would be able to build more trustworthy models by building more inclusive datasets.

Bias in healthcare algorithms is increasingly becoming a focus for regulators. For example, the Good Machine Learning Practice guidance from the FDA and EMA emphasizes the importance of ensuring that datasets are representative of the intended patient population (*17*). However, this alone is not sufficient because significant bias can arise even when standard machine learning methods are applied to diverse datasets. Integration of the AEquity metric into the algorithm development pipeline may help clarify the cause of this bias and mitigate it. The US Department of Health and Human Services has proposed a rule under Section 1557 of the Affordable Care Act to ensure that “a covered entity must not discriminate against any individual on the basis of race, color, national origin, sex, age, or disability through the use of clinical algorithms in decision-making” (*19*). The small number of samples required for AEq analyses, and the fact that it can be applied early in the algorithm-development pipeline make it an attractive complement to post-hoc prospective validation, which is currently the mainstay of bias analysis.

References

1. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).

2. L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, M. Ghassemi, CheXclusion: Fairness gaps in deep chest X-ray classifiers (2020), (available at http://arxiv.org/abs/2003.00827).

3. L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, M. Ghassemi, Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).

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

5. A. S. Rich, T. M. Gureckis, Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*. **1**, 174–180 (2019).

6. Y. Park, J. Hu, M. Singh, I. Sylla, I. Dankwa-Mullan, E. Koski, A. K. Das, Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA Netw Open*. **4**, e213909 (2021).

7. F. Galera, P. Garcia-del-Barrio, P. Mendi, Consumer surplus bias and the welfare effects of price discrimination. *J Regul Econ*. **55**, 33–45 (2019).

8. S. Jain, A. Smit, A. Y. Ng, P. Rajpurkar, Effect of Radiology Report Labeler Quality on Deep Learning Models for Chest X-Ray Interpretation. *arXiv [eess.IV]* (2021), (available at http://arxiv.org/abs/2104.00793).

9. J. R. Feiner, J. W. Severinghaus, P. E. Bickler, Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. *Anesth. Analg.* **105**, S18–S23 (2007).

10. F. F. Gulamali, A. S. Sawant, P. Kovatch, B. Glicksberg, A. Charney, G. N. Nadkarni, E. Oermann, Autoencoders for sample size estimation for fully connected neural network classifiers. *npj Digital Medicine*. **5** (2022), , doi:10.1038/s41746-022-00728-0.

11. A. E. W. Johnson, T. Pollard, R. Mark, S. Berkowitz, S. Horng, The MIMIC-CXR Database (2019), (available at https://physionet.org/content/mimic-cxr/).

12. A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. **6**, 317 (2019).

13. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and

PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*. **101** (2000), doi:10.1161/01.CIR.101.23.e215.

14. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, "ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases" in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, L. Lu, X. Wang, G. Carneiro, L. Yang, Eds. (Springer International Publishing, Cham, 2019; http://link.springer.com/10.1007/978-3-030-13969-8_18), pp. 369–392.

15. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. Conf. AAAI Artif. Intell.* **33**, 590–597 (2019).

16. R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, J. Zou, Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatol.* **157**, 1362–1369 (2021).

17. U.S. Food & Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency, Good Machine Learning Practice for Medical Device Development (2021), (available at https://www.fda.gov/media/153486/download).

18. L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, M. Ghassemi, 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** (2022), pp. 1161–1162.

19. C. Shachar, S. Gerke, Prevention of Bias and Discrimination in Clinical Practice Algorithms. *JAMA*. **329**, 283 (2023).

20. S. Kapur, Reducing racial bias in AI models for clinical use requires a top-down intervention. *Nature Machine Intelligence*. **3**, 460–460 (2021).

21. I. X. Kendi, *How to be an antiracist* (One World, New York, 2019).

22. E. K. Oikonomou, E. S. Spatz, M. A. Suchard, R. Khera, Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. *The Lancet Digital Health*. **4**, e796–e805 (2022).

23. J. K. De Freitas, K. W. Johnson, E. Golden, G. N. Nadkarni, J. T. Dudley, E. P. Bottinger, B. S. Glicksberg, R. Miotto, Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns*. **2**, 100337 (2021).

24. T. H. Kung, M. Cheatham, ChatGPT, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, V. Tseng, “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models” (preprint, Medical Education, 2022), (available at http://medrxiv.org/lookup/doi/10.1101/2022.12.19.22283643).

25. Hao, Moon, Didari, Woo, Bangert, Highly Efficient Representation and Active Learning Framework and Its Application to Imbalanced Medical Image Classification. *datacentricai.org* (available at https://datacentricai.org/neurips21/papers/107_CameraReady_GPAL_DCAI.pdf).

26. L. Deng, The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Process. Mag.* **29**, 141–142 (2012).

27. A. Krizhevsky, Learning multiple layers of features from tiny images, (available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf).

28. M. Riad, D. P. Dunham, J. R. Chua, N. Shakoor, S. Hassan, S. Everakes, J. A. Block, I. Castrejon, Health disparities among Hispanics with rheumatoid arthritis: Delay in presentation to rheumatologists contributes to later diagnosis and treatment. *J. Clin. Rheumatol.* **26**, 279–284 (2020).