How fair is your graph? Exploring fairness concerns in neuroimaging studies

Fernanda Ribeiro[*] FERNANDA.RIBEIRO@UQ.EDU.AU
School of Psychology
Queensland Brain Institute
The University of Queensland, Brisbane, QLD, Australia

Valentina Shumovskaia[*] VALENTINA.SHUMOVSKAIA@EPFL.CH
School of Engineering
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Thomas Davies T.O.M.DAVIES@SOTON.AC.UK
Electronics and Computer Science
University of Southampton
Southampton, UK

Ira Ktena IRAKTENA@DEEPMIND.COM
DeepMind
London, UK
[*] *Contributed equally*

Abstract

Recent work on neuroimaging has demonstrated significant benefits of using population graphs to capture non-imaging information in the prediction of neurodegenerative and neurodevelopmental disorders. These non-imaging attributes may not only contain demographic information about the individuals, e.g. age or sex, but also the acquisition site, as imaging protocols and hardware might significantly differ across sites in large-scale studies. The effect of the latter is particularly prevalent in functional connectomics studies, where it remains unclear how to sufficiently homogenise fMRI signals across the different sites. In addition, recent studies have highlighted the need to investigate potential biases in the classifiers devised using large-scale datasets, which might be imbalanced in terms of one or more sensitive attributes. This can be exacerbated when employing these attributes in a population graph to explicitly introduce inductive biases to the machine learning model and lead to disparate predictive performance across sub-populations. This study scrutinises such a system and aims to uncover potential biases of a semi-supervised classifier that relies on a population graph. We further explore the effect of the graph structure and stratification strategies, as well as methods to mitigate such biases and produce fairer predictions across the population.

1. Introduction

Issues related to fairness and equity in healthcare decision-making have been the focus of intense scholarly debate (Obermeyer et al., 2019; Rajkomar et al., 2018; Seyyed-Kalantari et al., 2020; Wiens et al., 2019). Even though computer-aided diagnosis (CAD) systems have integrated significant advances to assist clinicians in various tasks, including segmentation, classification, phenotype prediction, and treatment efficacy, these systems have rarely been scrutinised enough for their potential for discrimination against certain population subgroups by their deployment or adoption in clinical practice. Disparate treatment concerns can arise solely due to the composition of the training data (Larrazabal et al., 2020; Puyol-Antón et al., 2021), meaning that certain population subgroups might be underrepresented or completely missing during training of a machine learning CAD system. In healthcare applications, condition scarcity for specific subgroups can impede the curation of a balanced dataset across all sensitive attributes of interest. Similarly, the disease prevalence may vary across population subgroups (Werling and Geschwind, 2013), e.g. autism spectrum disorders (ASD) are more prevalent in males compared to females. At the same time, the clinical presentation of a disease might be completely different across subgroups. In ASD, in particular, differences have been established between neurodiverse males and females in terms of the interactions between key functional brain networks (Alaerts et al., 2016). Furthermore, a clinical decision-making system can be trained on data capturing certain demographics (e.g. young males) and then be deployed on a population with a different demographic distribution. Lastly and most importantly, these systems can perpetuate or exacerbate biases already present in the ground truth decisions used during their development.

Extensive work has been carried out to uncover fairness issues of computer vision systems that operate in the Euclidean domain and use inductive learning – that is, models trained on labeled training samples and evaluated on an unseen test dataset. However, these are still under-explored for approaches that operate in irregular domains in a transductive setting, in which models have access to the entire database while optimized on a subset of labeled samples. These have been shown to lead to significant performance improvements in neuroimaging tasks, like ASD and Alzheimer's disease prediction, by employing graph-based label propagation (Zhao et al., 2014) or equivalent techniques from graph representation learning (Parisot et al., 2018). One such example is the application of graph neural networks for semi-supervised learning on population graphs that leverage demographic or other auxiliary information (see Figure 1). Such approaches have demonstrated stark improvements compared to alternatives that do not rely on these sensitive attributes, because they capture the interactions and similarities between subjects or their individual scans, unlike more traditional classifiers (Abraham et al., 2017). However, these studies often only report overall performance metrics, such as prediction accuracy and area under the receiver-operating characteristic (AUC-ROC) curve. Therefore, there is a limited understanding of whether these methods and training strategies inadvertently improve predictive performance in one subgroup of the population at the expense of another.

Generalizable Insights about Machine Learning in the Context of Healthcare

We address this important gap in the fairness literature and explore how the population graph structure and stratification strategies during the training of graph neural networks affect the fairness of the devised semi-supervised classifier. Beyond presenting the shortcomings of existing design decisions, we further explore mitigation strategies that improve fairness in these classifiers, as defined in Hardt et al. (2016). The challenges of the ABIDE database (Di Martino et al., 2014) that we focus on in this work are two-fold: **(a)** it con-

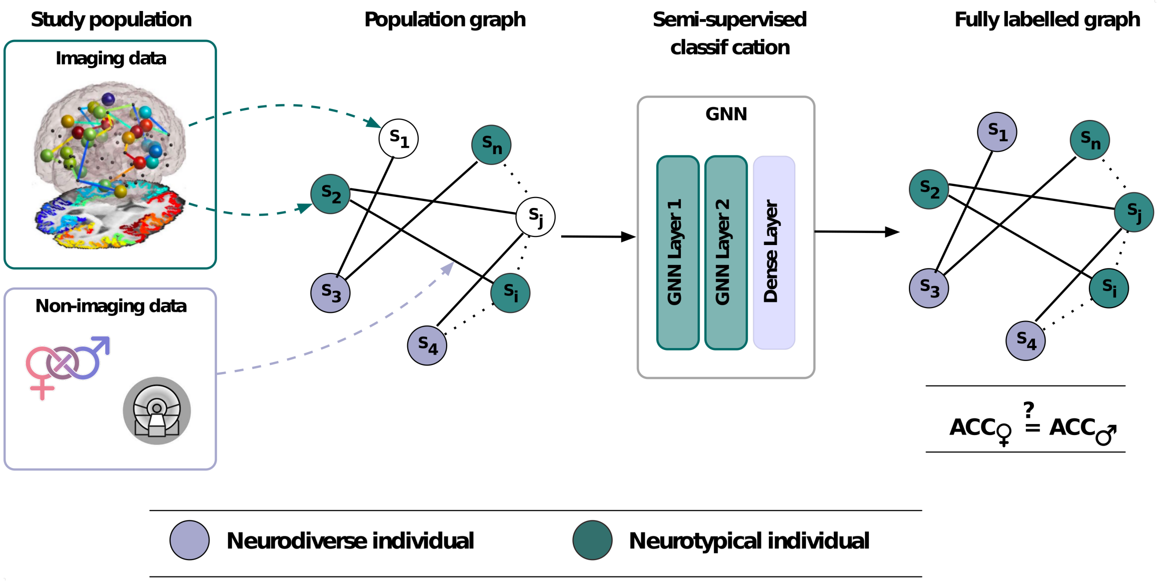

Figure 1: In a system that explicitly uses sensitive attributes to introduce inductive biases how do we ensure that the final predictions are fair across subgroups?

tains a heterogeneous set of neuroimaging data from neurodiverse and neurotypical study participants from 20 different imaging sites, while **(b)** within the same acquisition site the demographic and diagnostic distribution of individuals is highly imbalanced. We develop a rigorous and thorough evaluation framework that can be employed in multi-site studies to deal with challenges arising from imbalanced and highly heterogeneous consortia. This framework ensures that potential biases and disparate mistreatment can be uncovered, both for demographic subgroups as well as geographical locations, and should be adopted in the clinical evaluation of such approaches. We further report results with different mitigation strategies, i.e. fine-tuning and Just Train Twice, and discuss their applicability in the transductive setting. Due to the multifaceted challenges pertinent to this clinical database, we also show results with these mitigations on a simplified simulated dataset that is spared from the multi-site heterogeneity.

In brief, we found that the stratification strategy and the composition of the training set resulting from that had no significant effect on model fairness, when that is measured in terms of true positive rate differences. This contradicts most findings from prior studies that highlight the importance of the training set composition for the downstream fairness evaluation. We believe this difference is due to our experimental setting, i.e., the transductive setting. While most (if not all) previous studies have focused on the inductive setting, our model has access to and learns from the features of all samples, rather than only those of the training set alone.

2. Background

2.1. Graph neural networks for neuroimaging

Graph neural networks (GNNs) have been adopted in various neuroimaging studies for node- and graph-level predictive tasks. These approaches employ purely spatial (Ribeiro et al., 2021) or spectrally-inspired (Parisot et al., 2018) GNNs to devise convolutional filters that can be applied in irregular domains (e.g. brain connectivity or population graphs). Here, we focus on transductive learning approaches wherein the features of all subjects (training, validation, and testing) are present during training, while only the labels of the training set are available. This type of setting is useful for node-level predictive tasks like the one presented in Parisot et al. (2018), the earliest work applying GNNs on population graphs. The main advantage of this approach is that auxiliary information, e.g. the acquisition site, can be captured in the graph structure itself to introduce desired inductive biases to the model. This so-called phenotypic graph was shown to yield the highest overall performance (in terms of accuracy and AUC-ROC) in two different tasks, ASD and Alzheimer's disease prediction, compared to random, k-nn, and complete graphs, highlighting that incorporating information relevant to the disease in the population graph structure can be highly beneficial for the overall performance. Alternative graph-learning methods have since been proposed instead of hand-picking the non-imaging attributes (Cosmo et al., 2020). Other works (Vivar et al., 2020; Hett et al., 2021) explored multi-graph settings and data imputation for missing features, but we consider those less relevant to our setting.

2.2. Fairness metrics in medical applications

Different notions of fairness (or lack thereof) have been discussed in the expanding literature on algorithmic fairness. These focus on different aspects of potential discrimination in a decision-making system. ***Disparate treatment*** indicates that a system yields different outputs for different subgroups of people with the same (or very similar) features except for the sensitive feature (Barocas and Selbst, 2016). This is often referred to as *direct discrimination* and arises when the decision essentially relies on the sensitive attribute. ***Disparate impact*** characterises the scenario where a system provides outputs that benefit / hurt people sharing a sensitive attribute more frequently than others. Mitigating disparate impact is equivalent to striving for *statistical* (Corbett-Davies et al., 2017) or *demographic parity* (Dwork et al., 2012). ***Disparate mistreatment***, in turn, describes the failure of a system to achieve the same classification accuracy (or conversely, error rate) for subgroups of people sharing different values of a sensitive attribute. *Equality of opportunity* (Hardt et al., 2016) and predictive equality strive to address these limitations in a decision-making system. The key differences between these three notions lie in whether the decision-making system intentionally or inadvertently discriminates against a group characterised by a particular sensitive attribute. Disparate impact and disparate mistreatment both account for indirect unfairness, but their application scenarios differ. Disparate impact is unaware of the ground truth information about the decision (e.g. diagnosis), while that is not the case for disparate mistreatment. In the clinical setting, we do have access to the ground truth (at least for the training data) and for the aforementioned reasons, we consider equality of opportunity the most relevant to the application explored in this work.

Various approaches have been proposed in the literature to evaluate classifier disparities across subgroups in the context of binary classification. In this setting we aim to find a mapping function $f(\mathbf{x})$ between individual feature vectors $\mathbf{x} \in \mathbb{R}^d$ and class labels $y \in \{0, 1\}$ based on a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$. In Hardt et al. (2016), authors introduced a simple, interpretable, and easily verifiable notion of nondiscrimination to specified protected attributes that guarantees equal true positive rate (TPR) across the values of these protected attributes. In particular, $Pr(\hat{y} = 1|a = 0, y = 1) = Pr(\hat{y} = 1|a = 1, y = 1)$, where a is the protected attribute, y is the desired outcome label and $\hat{y}$ the predicted class. For $y = 0$, the constraint equalizes the false positive rate (FPR) between the two groups. The same metric was previously used to measure discrimination between the two sexes in different contexts by De-Arteaga et al. (2019) and Seyyed-Kalantari et al. (2020). In Seyyed-Kalantari et al. (2020), TPR disparities were investigated in relation to different sensitive attributes for classifying respiratory conditions from chest X-rays. Sensitive attributes were: sex, age, race, and insurance policy. They were able to identify disparities that could pose serious barriers to the effective deployment of these models and proposed additional changes in either dataset design and/or modeling techniques to ensure more equitable models. In Puyol-Antón et al. (2021), a cardiac segmentation task was considered instead, and fairness issues were investigated regarding gender and race. They found that the accuracy of the baseline segmentation model for each racial sub-group was correlated with its representation in the training set. In other contexts, parity in terms of overall misclassification rate (OMR), false omission rate (FOR) and false discovery rates (FDR) have been considered to evaluate the fairness of a system (Zafar et al., 2019).

Other definitions of fairness focus on calibration (Kleinberg et al., 2016) and demographic parity (Corbett-Davies and Goel, 2018). Mainly, demographic parity is the equality of outcomes for different groups, i.e., parity in the proportion of some decision D (Corbett-Davies and Goel, 2018; Olfat and Mintz, 2020). However, demographic parity is a stricter measure when there is a strong correlation between the sensitive attribute and the prediction target (Olfat and Mintz, 2020), which is the case for ASD (higher prevalence within the male population). Given the growing recognition that not all conditions can be simultaneously satisfied (Chouldechova and Roth, 2018), we focus on the TPR gap (or equality of odds) as, arguably, the most relevant to the application scenario.

2.3. Bias mitigation

Different approaches have been proposed to mitigate potential fairness issues and model bias. As these often arise from data imbalance in the training set and the underrepresentation of certain groups, pre-processing approaches play an important role here. In particular, under-sampling the prevalent class or over-sampling the underrepresented class (Larrazabal et al., 2020) can prove to be effective strategies, similarly to inverse propensity weighting (Robins et al., 2000). Along the same lines, generative models can be leveraged to fill parts of the distribution where there is missing or only limited data is available. In-processing approaches include implicit and explicit regularisation (Olfat and Mintz, 2020), as well as introducing adversarial components to the model (Madras et al., 2018; Dai and Wang, 2021) or fairness constraints (Zafar et al., 2019). Researchers have also discussed making sensitive attributes available as a means to improve fairness (Dwork et al., 2012), as well as

Table 1: Descriptive statistics from the 8 largest acquisition sites, including the number of neurotypical and neurodiverse, male and female participants. These constitute 65% of all subjects in the ABIDE study.

	Male participants		*Female participants*		
Acquisition site	Neurodiverse	Neurotypical	Neurodiverse	Neurotypical	**Total**
NYU	64	72	10	26	172
UM	26	35	8	17	86
USM	43	24	0	0	67
UCLA	31	24	6	3	64
PITT	21	22	3	4	50
MAX_MUN	16	26	3	1	46
TRINITY	19	25	0	0	44
YALE	14	11	8	8	41

different ways to leverage these attributes (Dwork et al., 2018). Post-processing mitigation strategies, on the other hand, include classification with rejection, classifier calibration (e.g., adjusting the threshold of classification for each group), and equalized odds as proposed in Hardt et al. (2016).

3. Materials and Methods

3.1. Dataset & Descriptive Statistics

Our study focuses on the *ABIDE database* described in Di Martino et al. (2014) – a consortium of several international acquisition sites comprising functional neuroimaging and phenotypic data from 1112 participants. Out of those, 871 participants met the imaging quality and phenotypic information criteria (Abraham et al., 2017), totaling 403 neurodiverse and 468 neurotypical individuals. The number of individuals participating in each acquisition site varies significantly, with the most prominent site contributing data from 172 individuals and the least prominent one from 11 individuals. Furthermore, there is significant variation in terms of the demographic and diagnosis distribution across acquisition sites (see Table 1), i.e. some sites only provide data from males and have a higher prevalence of neurodiverse participants (e.g. USM). In contrast, for other sites, diagnosis imbalance is starker for females than for males (e.g. NYU). These statistics are important to understand the challenges of this dataset, while the multi-site setting allows to test the generalisability of the approaches across sites. Table 1 reports the overall statistics describing the data of the eight most prevalent acquisition sites, regarding sex and diagnosis. Equivalent statistics for the remaining sites are provided in the Appendix.

It is important to consider that scanner and imaging protocol variations introduce additional challenges, especially when considering resting-state fMRI (rs-fMRI) sequences and their derivatives, in the form of connectivity matrices. For each individual participating in this study, their rs-fMRI data were preprocessed with the Configurable Pipeline for the Analysis of Connectomes (C-PAC) (Craddock et al., 2013). Subsequently, the mean time-series were extracted for a set of brain regions delineated by the Harvard-Oxford anatomical atlas, which comprises $R = 110$ cortical and subcortical regions of interest (Desikan et al., 2006), and normalised to zero mean and unit variance. The Fisher's transformed correlation matrix was then used to specify $\mathbf{x}_i \in \mathbb{R}^{R\times(R-1)}$, i.e. the vectorised connectivity matrix for each individual.

In summary, this database presents a particularly challenging setting in which we can explore the propensity of GNN models to be biased against underrepresented populations.

3.2. Population graph construction

As defined in Parisot et al. (2018), the phenotypic population graph is constructed by weighting the connectome similarity matrix with a phenotypic graph that captures the agreement of pairs of participants in terms of phenotypic features, i.e.,

$$W(i,j) = sim(\mathbf{x}_i, \mathbf{x}_j) \sum_{h=1}^{H} \delta(a_h(i), a_h(j)), \tag{1}$$

where $\mathbf{x}_i$, $\mathbf{x}_j$ are the vectorised functional connectomes and $A = \{a_h\}$ the set of phenotypic attributes we consider (i.e. sex and acquisition site). $sim(\mathbf{x}_i, \mathbf{x}_j)$ corresponds to a similarity metric between connectomic feature vectors. δ is the Kronecker delta function for a pair of nodes (i,j) and a non-imaging feature h:

$$\delta(a_h(i), a_h(j)) = \begin{cases} 1 & \text{if } a_h(i) = a_h(j) \\ 0 & \text{if } a_h(i) \neq a_h(j) \end{cases} \tag{2}$$

We consider four different graph structures to understand the impact of the population graph on the fairness of the target predictions: **(1)** a weighted graph based on the subjects' *sex* alone, **(2)** the *acquisition site* alone, **(3)** *both* sex and acquisition site, and **(4)** a *complete* graph that does not leverage phenotypic information. It is worth noting that manipulating the graph structure in this setting can be perceived as a pre-processing approach to mitigate bias, as it does not introduce additional constraints / regularisation during the training process.

3.3. Stratification strategy

In prior work, k-fold stratified cross-validation was used to evaluate the performance of the proposed method (Abraham et al., 2017; Parisot et al., 2018). However, stratification based on diagnosis can lead to a significantly imbalanced training set with respect to the sensitive attribute of interest, given that for certain sites no female participants were recruited. We therefore investigate the impact of stratification strategies on fairness metrics, considering sex as the sensitive attribute (a). As previous studies have shown (De-Arteaga et al., 2019; Larrazabal et al., 2020; Puyol-Antón et al., 2021), the composition of the training data can significantly impact the bias of the devised classifier. Hence, the training data bias with respect to the sensitive attribute can be further accentuated by a stratification based solely on diagnosis, as often seen in cross-validation settings, due to the demographic shift between acquisition sites. As summarised in Table 1 and Table 2, the number of individuals with specific diagnoses and sensitive attribute values varies significantly across acquisition sites. The impact of this demographic and diagnosis shift on the downstream task was previously demonstrated when reporting classifier performance on samples from unseen sites (Abraham et al., 2017).

To test for the robustness of our GNN model to distribution shifts, we consider four different stratification strategies: **(1)** based on the *target variable* – diagnosis, **(2)** based on *diagnosis* and the *sensitive attribute* (i.e., sex), **(3)** based on *diagnosis* and the *acquisition site*, and **(4)** based on the *sensitive attribute* and the *acquisition site.*

3.4. Model

Beyond the different stratification strategies and graph structures, we compare the GNN model employed in Parisot et al. (2018) and originally proposed in Defferrard et al. (2016) to a ridge classifier. The GNN model learns the parameters θ_l of the polynomial filters $g_\theta(\mathcal{L}) = \sum_{l=0}^{L} \theta_l \mathcal{L}^l$, with $\mathcal{L}$ being the normalized Laplacian matrix of the affinity graph defined as $\mathcal{L} = I_N - D^{-1/2} W D^{-1/2}$, and D is the degree matrix. This corresponds to the reparametrization introduced by Defferrard et al. (2016) yielding filters that are strictly L-localized (i.e. capture the L-hop neighbourhood around the central node) and significantly reduces the computational complexity of the convolution operator in the Fourier domain.

Without the use of this polynomial reparametrization, one would need to decompose the Laplacian matrix corresponding to the adjacency matrix W and learn an orthonormal basis of eigenvectors $U = [u_0, ...u_{N-1}]$ corresponding to real, eigenvalues $\Lambda = diag([\lambda_0, ...\lambda_{N-1}])$ with $\mathcal{L} = U \Lambda U^T$. This operation is $\mathcal{O}(N^3)$ with respect to the number of nodes in the graph. A spectral convolution of a signal defined on the nodes $\mathbf{x}$ (in our case, the connectivity profile of an individual) with a filter $g_\theta = diag(\theta)$ defined in the Fourier domain can then be defined as a multiplication in the Fourier domain, i.e. $g_\theta * \mathbf{x} = g_\theta(\mathcal{L}) \cdot \mathbf{x}$, which allows us to convolve the node feature vectors with the polynomial filters and “diffuse” information from neighbouring nodes. In our experiments we use $L = 3$ layers for the GNN, which is the optimal depth identified empirically by Parisot et al. (2018). The remaining list of values for the model and training hyperparameters are presented in the Supplementary Material.

We consider the ridge classifier as a baseline that does not capture the sensitive attribute and site information explicitly, i.e. the model only has access to the vectorised connectivity matrices, $\{\mathbf{x}_i\}_{i=1}^N$ and the affinity matrix W is not used. The GNN model learns the parameters θ_l of the polynomial filters defined as a function of the Laplacian matrix of the population graph, i.e. $g_\theta(\Lambda)$.

3.5. Mitigation techniques

Finally, we experiment with transfer learning, an in-processing mitigation strategy. In our experiments, we first train the GNN model for 150 epochs and then fine-tune the model for each sensitive group for 50, 100, and 150 epochs, generating two sensitive group models for each setting, similarly to Puyol-Antón et al. (2021). We further employ the Just Train Twice (Liu et al., 2021) method, which upweights the misclassified samples after one round of training and has been shown to improve robustness of models to distribution shifts. We apply JTT for 30 and 50 epochs while we explore the effect of the weighting for each of these settings.

3.6. Code availability

All accompanying Python source code will be made available on GitHub (`https://github.com/tomogwen/population-gcn`). Moreover, an executable code notebook hosted on Google Colab will be made available, allowing reproducibility of our experiments.

4. Results

4.1. Impact of stratification strategy

The stratification strategy inherently impacts the distribution of labels, sensitive attributes and acquisition sites in a k-fold cross-validation setup, which would render the comparison across different stratification strategies challenging and, potentially, unfair. To address this issue, we randomly select approximately 10% of the individuals as a held-out test set. We sample two male and two female participants completely at random – one diagnosed as neurotypical and one as neurodiverse of each sex – from all acquisition sites (wherever available). These individuals' labels are then excluded from all training and validation folds. This setup is very similar to bootstrapping, with the difference that we do not sample the test set with replacement for a single cross-validation seed/random state.

For each configuration, we train models with 10 different cross-validation random states to capture possible noise in our experimental observations. Overall, we obtained 100 models for each combination of model (GNNs vs. Ridge classifier – the latter being our baseline model), graph structure (as described in 3.2), and stratification strategy (described in 3.3), which were evaluated using the same held-out test set. Figure 2 illustrates the distributions of true positive rate differences between males and females across these 100 models for each combination.

We perform a repeated measures two-way ANOVA to assess the effect of the *stratification strategy* and the selected *model* (including all GNNs trained on different graph structures) on TPR differences. We found a main effect of model [$F(4, 396) = 291.005, p < 0.001$], but not stratification strategy [$F(3, 297) = 0.441, p = 0.724$]. Post-hoc test (using the Bonferroni correction to adjust p) indicated that all TPR difference distributions significantly differed from one another ($p < 0.001$), with the GNN trained on a population graph weighted by sex information having the smallest TPR difference (mean TPR difference ranges from 0.069 to 0.087 across stratification strategies). In summary, the *sex* graph leads to the lowest true positive rate difference. Most importantly, it significantly improves on fairness with respect to the baseline as well as the *complete* graph.

4.2. Impact of graph structure

Since stratification strategies did not seem to directly affect TPR differences, *all experiments reported in this section were carried out with stratification based on the diagnosis* to align with prior work (Parisot et al., 2018). In order to assess the robustness of our previous finding, we trained 10 different models (using 10 different parameter initialisation seeds) for 10 different held-out test sets, selected with a different random seed in a similar manner

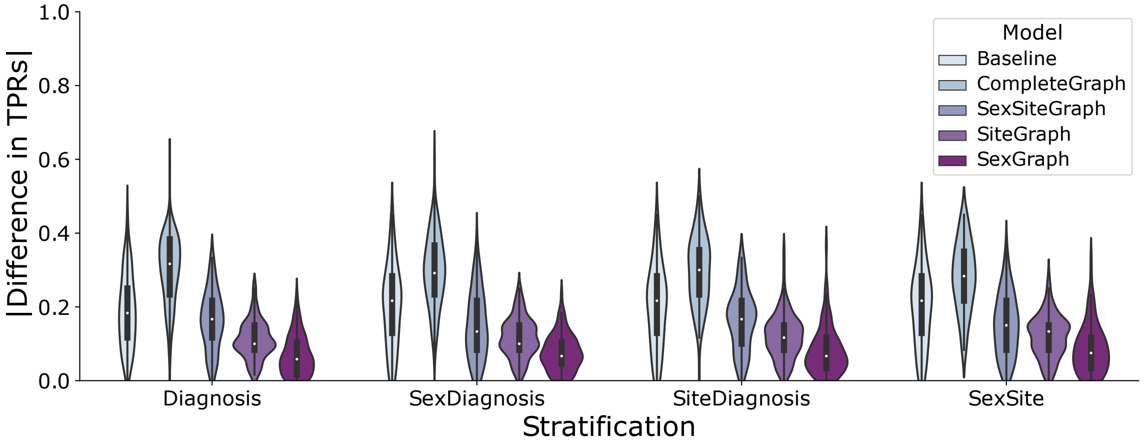

Figure 2: Absolute difference in true positive rates between males and females in the test set across graph structures and stratification strategies for a fixed held-out set.

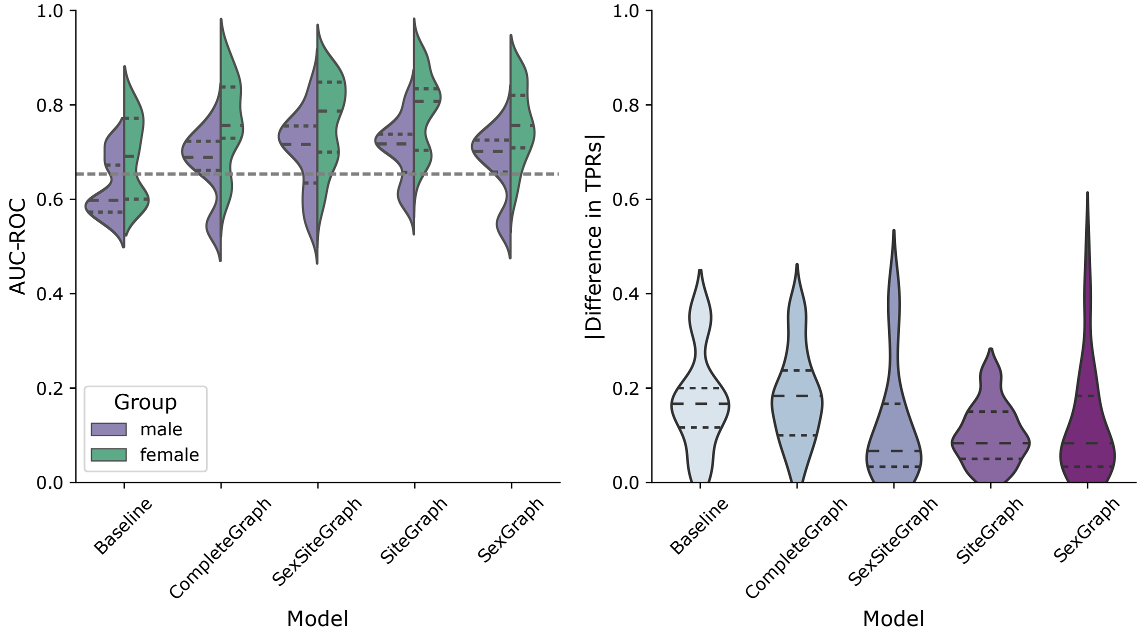

Figure 3: AUC-ROC for males (purple) and females (green) and absolute difference in true positive rate. The dashed grey line on the AUC-ROC plot indicates the average performance of the baseline model (Ridge classifier) across the population (including both males and females).

as above (i.e. ensuring an as balanced as possible test set with respect to site and sex distribution). Thus, we trained and tested 10 different models on each held-out test set, totaling 100 models. Figure 3 shows the TPR differences (left) and AUC-ROC (right). Even though we found that all population graphs resulted in higher average AUC-ROC than the baseline model (Ridge classifier), the absolute difference in TPRs substantially improved (decreased) when phenotypic information was used. **Therefore, the higher performance of GNNs employed on population graphs did not come at the cost of the higher difference in TPRs.** It is worth noting that leveraging phenotypic information in the population graph construction, including the sensitive attribute, did not disproportionately impact model performance for the under-represented group (see also Figure 8). This result aligns with prior findings in the inductive setting that highlight the benefits of fairness through awareness of the sensitive attribute (Dwork et al., 2012).

4.3. Impact of transfer learning

Finally, we investigate the potential of transfer learning for bias mitigation. Although the absolute TPR difference, our proxy of model unfairness, was not as acute as in other applications, there is room for improvement to reach an absolute TPR difference of 0 (fair model with respect to this fairness metric). Therefore, our goal in this section is to mitigate disparate performance across males and females and push true positive rate different towards 0.

4.3.1. Fine-tuning on synthetic data

In order to disentangle the effect of the multi-site heterogeneity from that of sensitive attribute imbalance, we generate a synthetic dataset that is spared from the complexity of the multi-site setting. The generated data is depicted in Fig. 4 (left), while the proportion of males and females characterised as neurotypical and neurodiverse matches that of the original dataset. After training a baseline GNN model with a complete graph and a sex graph, we subsequently fine-tuned the respective models on the male and female populations separately, yielding two specialised models for each value of the sensitive attribute in the dataset. We observe that, similar to prior findings by Puyol-Antón et al. (2021), the models fine-tuned on the male samples lead to higher AUC-ROC compared to the original model, while the same holds for the models fine-tuned on the female training samples when evaluated on the female samples in the test set. For completeness, we further evaluated the specialised models on the samples of the opposite sex that they have not been fine-tuned on.

4.3.2. Fine-tuning on the ABIDE database

We further assessed whether fine-tuning pretrained models for each sensitive group separately can generate group-specific models that perform better for a protected group than the original model trained on both groups in the ABIDE database. In this section, we show the results of such experiments for a population graph that does not rely on sensitive attributes. Figure 5 shows the TPR differences (top left), AUC-ROC (bottom left), sensitivity (top right), and specificity (bottom right). Fine-tuning on the male population did not change the results considerably across the board compared to the pretrained model (complete graph), regardless of the number of epochs. Conversely, fine-tuning the model on the female sample reduced the mean difference in TPRs (overall performance is comparable

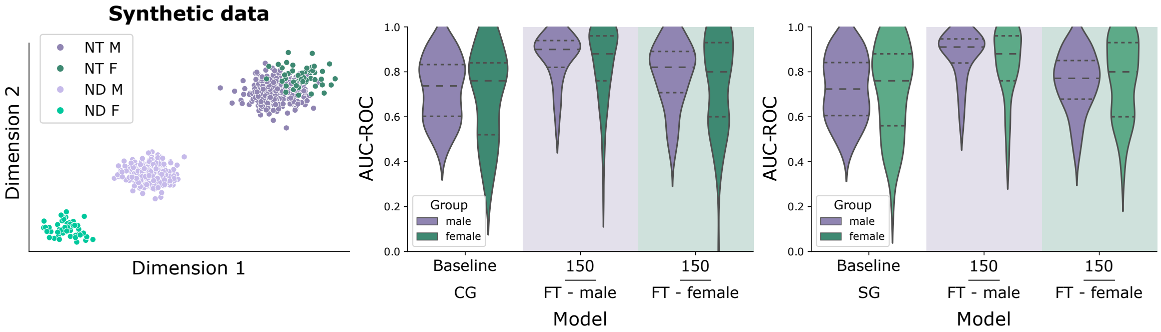

Figure 4: Synthetic data (left) and AUC-ROC for models using the complete graph (center) and the sex graph (right) for males (purple) and females (green) groups. The purple area indicates a model fine-tuned (FT) on the male population (FT-male) of the training set for 150 epochs, whereas the green area shows models fine-tuned on the female population (FT-female). The baseline is the original graph model. NT M - neurotypical males; NT F - neurotypical females; ND M - neurodiverse males; ND F - neurodiverse females; CG - **c**omplete weighted **g**raph; SG - **s**ex weighted population **g**raph.

to the population graph leveraging sex and site information). However, this improvement resulted in reduced AUC and sensitivity for both groups.

4.3.3. Just Train Twice on the ABIDE database

Finally, we explored the potential of Just Train Twice (JTT) on the ABIDE database, which upweights samples that have been misclassified by the model after one round of training. This can be considered as an alternative form of fine-tuning without explicitly focusing on a specific subgroup or sensitive attribute. Our hypothesis, however, is that the model is more likely to misclassify under-represented subgroups in the training set. Results with this technique are summarised in Figure 6. We observe that, unlike fine-tuning on a population characterised by a specific attribute, JTT marginally improves sensitivity for the sensitive group without hurting specificity after 30 epochs. Upon post-hoc analysis of the model's predictions at the end of the first round of training, we observe that the original model tends to misclassify neurotypical females at a higher rate than neurotypical males. Although we did not observe noticeable improvements in the model's specificity, JTT holds the potential to improve the model's sensitivity.

5. Discussion

In this work, we thoroughly explored different mitigation strategies to improve the fairness of ASD predictions in a multi-site cohort using publicly available data from the ABIDE study. This is the first study focusing on the impact of relying on sensitive attributes to construct the population graph in a semi-supervised setting. Our results align with other works that have shown that leveraging the sensitive attribute is important to improving

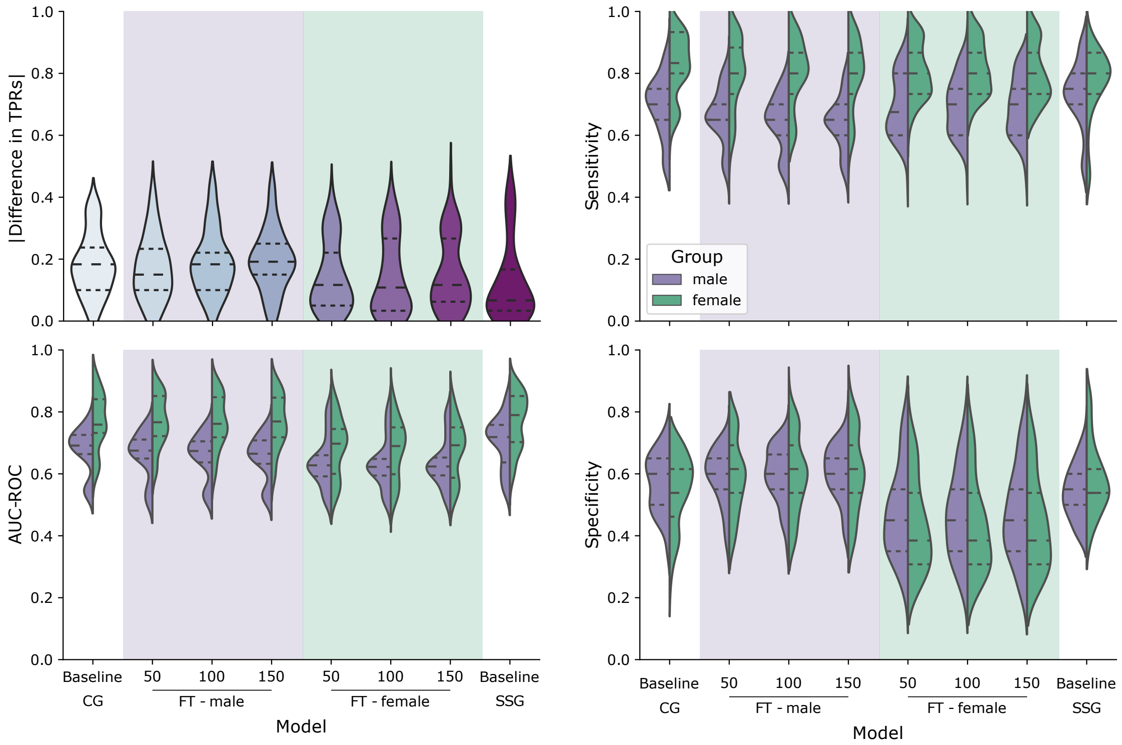

Figure 5: Absolute difference in true positive rates, sensitivity, AUC-ROC, and specificity for males (purple) and females (green) groups. The purple shaded area shows models fine-tuned on the male samples (FT-male) of the training set for 50, 100, and 150 epochs, whereas the green shaded area shows models fine-tuned on the female samples (FT-female). All fine-tuned models used the complete graph. Baselines are the pre-trained models. CG - **c**omplete **g**raph; SSG - **s**ex and acquisition **s**ite weighted population graph.

model fairness (Dwork et al., 2012, 2018) instead of discarding this information. We observe that the stratification strategy does not impact the differences in true positive rate between male and female participants (considering sex as the sensitive attribute). This shows that the impact of the underlying graph structure is more important than the composition of the training set, at least with such a small dataset. At the same time, the different patterns of functional connectivity (hyper-connectivity in females vs. hypo-connectivity in males) in individuals diagnosed with ASD (Alaerts et al., 2016) hint that it can be beneficial to learn different latent representations for the two subgroups, which is what we achieve by relying on sex for the graph construction that yields two disconnected graphs one for each sex.

Overall, our results suggest that there is no one-size-fits-all metric for evaluating potential biases in a CAD system. Even though the difference in TPRs is reduced after fine-tuning models on the female population, the fine-tuned models end up being biased towards positive predictions (hence, have lower specificity). This pattern is not observed on the synthetic data and can be attributed to the heterogeneity of the ABIDE database

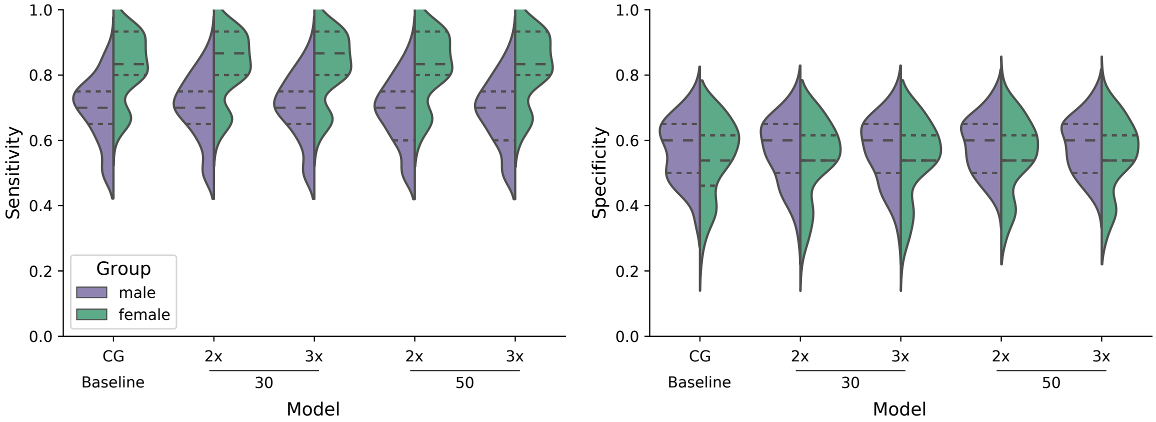

Figure 6: Sensitivity and specificity for males (purple) and females (green) groups. Models were fine-tuned for 30 and 50 epochs by upsampling (2x or 3x) the misclassified training examples. All fine-tuned models used the complete graph. Baselines are the pre-trained models (i.e. not fine-tuned). CG - **c**omplete **g**raph.

in conjunction with our evaluation setup, which makes it harder to generalise to new unseen sites. As we showed through the dataset statistics, there are multiple acquisition sites without female participants or, when those are available they might solely be part of the test set and, hence, never be encountered during training. Additionally, fine-tuning the pretrained models on each demographic group separately did not yield group-specific models that outperformed the pretrained model (in contrast to previous findings in cardiac MR image classification (Puyol-Antón et al., 2021)). This discrepancy may also stem from the fact that our models are trained in a transductive setting, in which models still have access to all data samples while they are only optimized on a subset of labeled samples. It is also worth considering that the multi-site nature of the ABIDE database makes our training and test sets highly heterogeneous. At the same time, it poses a more realistic challenge compared to other studies that focus on clinical data from a single geographical location or acquisition site. Our results suggest that transfer learning in the multi-site transductive setting might not be as effective, while techniques such as JTT that fine-tune on misclassified samples from either subgroup hold more promise.

Limitations An important limitation of the ABIDE database that we perform our experiments on is the binary nature of the sensitive attribute, i.e. sex. Even though recent studies have explored the clinical phenotype of autism in gender minority adults (Kung, 2020), fMRI data for such participants have not been made available. Another limitation is considering the prediction of diagnosis as a binary classification problem, given that ASD is inherently a spectrum disorder. Even though the same approach has been adopted in prior studies (Abraham et al., 2017; Parisot et al., 2018), reducing this to a classification problem does not consider the heterogeneity across individuals with DSM-IV-TR (fourth and text revised edition of the Diagnostic and Statistical Manual of Mental Disorders), autistic disorder, Asperger syndrome, pervasive developmental disorder-not otherwise specified

(PDD-NOS) and individuals identified as ASD but not further differentiated into specific DSM-IV-TR subtypes (Di Martino et al., 2014).

Future work should focus on replicating these observations on a different dataset and clinical classification task to verify the generalisability of our findings. Furthermore, depending on the application, as mentioned in Section 2.2 other fairness metrics can be explored and reported as these can shed light on different aspects of bias in the devised classifier. Finally, other mitigation strategies could be investigated in a similar transductive setting, such as training the classifier as multi-objective optimization problem (Martinez et al., 2020) or using augmentation techniques (Beinecke and Heider, 2021).

References

Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017.

K Alaerts, SP Swinnen, and N Wenderoth. Sex differences in autism: a resting-state fmri investigation of functional brain connectivity in males and females. *Social cognitive and affective neuroscience*, 11(6):1002–1016, 2016.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Jacqueline Beinecke and Dominik Heider. Gaussian noise up-sampling is better suited than smote and adasyn for clinical decision making. *BioData Mining*, 14:1–11, 2021.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM KDD*, pages 797–806, 2017.

Luca Cosmo, Anees Kazi, Seyed-Ahmad Ahmadi, Nassir Navab, and Michael Bronstein. Latent-graph learning for disease prediction. In *MICCAI*, pages 643–653. Springer, 2020.

Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Front Neuroinform*, 42:10–3389, 2013.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information, 2021.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAT**, pages 120–128, 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3837–3845, 2016.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

A Di Martino, CG Yan, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6): 659–667, 2014.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *FAT**, pages 119–133. PMLR, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Kilian Hett, Vinh-Thong Ta, Ipek Oguz, José V Manjón, Pierrick Coupé, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-scale graph-based grading for alzheimer's disease prediction. *Medical Image Analysis*, 67:101850, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Karson TF Kung. Autistic traits, systemising, empathising, and theory of mind in transgender and non-binary adults. *Molecular autism*, 11(1):1–8, 2020.

AJ Larrazabal, N Nieto, V Peterson, DH Milone, and E Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *PNAS*, 117(23):12592–12594, 2020.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, pages 6781–6792. PMLR, 2021.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, pages 3384–3393. PMLR, 2018.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *ICML*, pages 6755–6764. PMLR, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

Matt Olfat and Yonatan Mintz. Flexible Regularization Approaches for Fairness in Deep Learning. pages 3389–3394, 2020. ISBN 9781728174471. doi: 10.1109/CDC42340.2020.9303736.

S Parisot, SI Ktena, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48: 117–130, 2018.

E Puyol-Antón, B Ruijsink, et al. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *MICCAI*, pages 413–423. Springer, 2021.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Fernanda L Ribeiro, Steffen Bollmann, and Alexander M Puckett. Predicting the retinotopic organization of human visual cortex from anatomy using geometric deep learning. *NeuroImage*, 244:118624, 2021. ISSN 1053-8119.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

L Seyyed-Kalantari, G Liu, M McDermott, IY Chen, and M Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021*, pages 232–243. World Scientific, 2020.

Gerome Vivar, Anees Kazi, Hendrik Burwinkel, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi. Simultaneous imputation and disease classification in incomplete medical datasets using Multigraph Geometric Matrix Completion (MGMC). *arXiv preprint arXiv:2005.06935*, 2020.

Donna M Werling and Daniel H Geschwind. Sex differences in autism spectrum disorders. *Current opinion in neurology*, 26(2):146, 2013.

J Wiens, S Saria, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

Mingbo Zhao, Rosa HM Chan, Tommy WS Chow, and Peng Tang. Compact graph based semi-supervised learning for medical diagnosis in alzheimer's disease. *IEEE signal processing letters*, 21(10):1192–1196, 2014.

Appendix A.

Dataset statistics

In this section we provide the summary statistics for the remaining sites that recruited 35% of the participants in the ABIDE database.

Table 2: Descriptive statistics from the remaining 12 acquisition sites. These constitute 35% of all subjects in the ABIDE study. These include the number of neurotypical and neurodiverse, male and female participants.

	Male participants		*Female participants*		
Acquisition site	Neurodiverse	Neurotypical	Neurodiverse	Neurotypical	**Total**
KKI	9	15	3	6	33
UM_2	12	20	1	1	34
LEUVEN_1	14	14	0	0	28
LEUVEN_2	9	12	3	4	28
OLIN	11	12	3	2	28
SDSU	8	13	0	6	27
SBL	12	14	0	0	26
STANFORD	9	9	3	4	25
OHSU	12	13	0	0	25
UCLA_2	11	8	0	2	21
CALTECH	4	6	1	4	15
CMU	4	3	1	4	11

Hyperparameters

Table 3 provides the set of hyperparameter values for our experiments on the ABIDE database. For the synthetic data experiments we reduced the learning rate to 0.002.

Table 3: Hyperparameters for GNN model.

Parameter name	parameter value
learning rate	0.005
weight decay	$5e^{-4}$
number of epochs	150
dropout	0.3
number of hidden units	16
number hidden layers	1
polynomial degree	3
number of input features	2000

Other fairness metrics

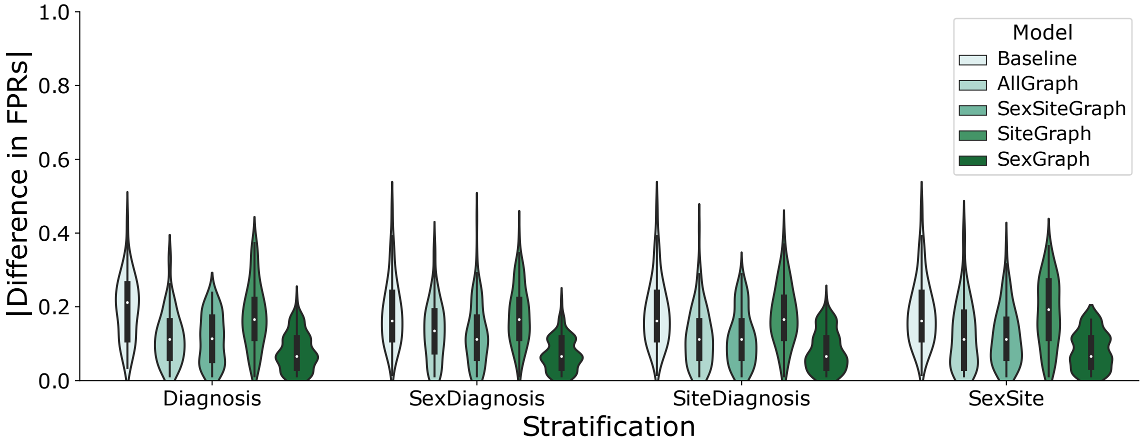

Figure 7: False positive rate difference between males and females in the test set across graph structures and stratification strategies as in Figure 2.

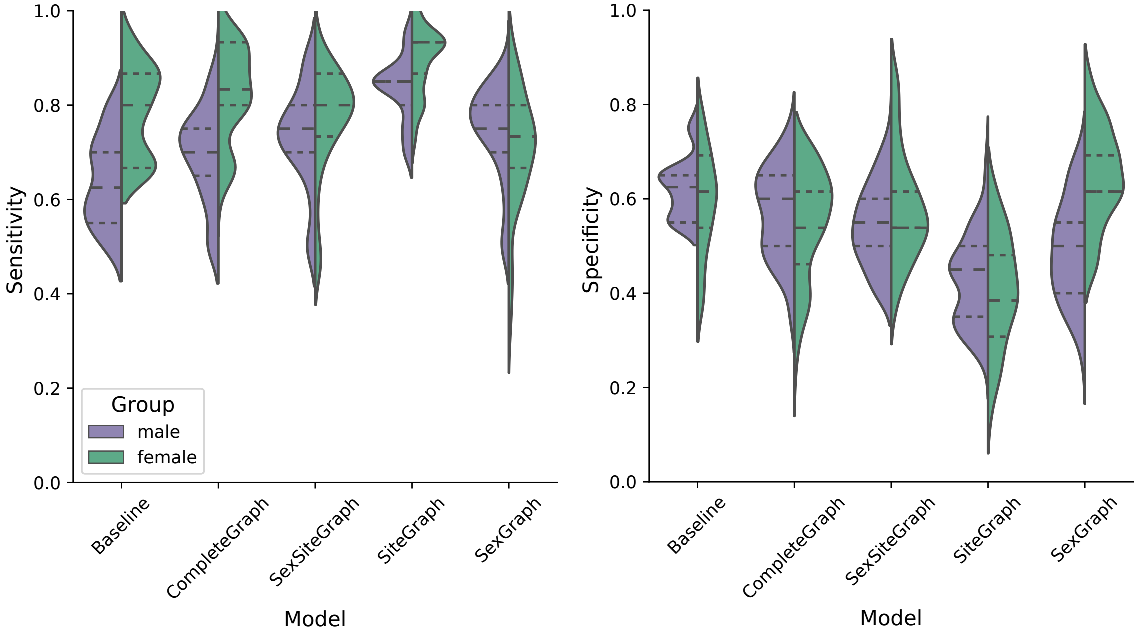

Figure 8: Sensitivity and specificity for males (purple) and females (green) across graph structures and models as in Figure 3.