Biology of Sex Differences

RESEARCH

Open Access

Modeling brain sex in the limbic system as phenotype for female-prevalent mental disorders

Gloria Matte Bon[1,2*], Dominik Kraft[1], Erika Comasco[2], Birgit Derntl[1,3] and Tobias Kaufmann[1,3,4*]

Abstract

Background Sex differences exist in the prevalence and clinical manifestation of several mental disorders, suggesting that sex-specific brain phenotypes may play key roles. Previous research used machine learning models to classify sex from imaging data of the whole brain and studied the association of class probabilities with mental health, potentially overlooking regional specific characteristics.

Methods We here investigated if a regionally constrained model of brain volumetric imaging data may provide estimates that are more sensitive to mental health than whole brain-based estimates. Given its known role in emotional processing and mood disorders, we focused on the limbic system. Using two different cohorts of healthy subjects, the Human Connectome Project and the Queensland Twin IMaging, we investigated sex differences and heritability of brain volumes of limbic structures compared to non-limbic structures, and subsequently applied regionally constrained machine learning models trained solely on limbic or non-limbic features. To investigate the biological underpinnings of such models, we assessed the heritability of the obtained sex class probability estimates, and we investigated the association with major depression diagnosis in an independent clinical sample. All analyses were performed both with and without controlling for estimated total intracranial volume (eTIV).

Results Limbic structures show greater sex differences and are more heritable compared to non-limbic structures in both analyses, with and without eTIV control. Consequently, machine learning models performed well at classifying sex based solely on limbic structures and achieved performance as high as those on non-limbic or whole brain data, despite the much smaller number of features in the limbic system. The resulting class probabilities were heritable, suggesting potentially meaningful underlying biological information. Applied to an independent population with major depressive disorder, we found that depression is associated with male–female class probabilities, with largest effects obtained using the limbic model. This association was significant for models not controlling for eTIV whereas in those controlling for eTIV the associations did not pass significance correction.

Conclusions Overall, our results highlight the potential utility of regionally constrained models of brain sex to better understand the link between sex differences in the brain and mental disorders.

*Correspondence:
Gloria Matte Bon
gloria.matte-bon@uni-tuebingen.de
Tobias Kaufmann
tobias.kaufmann@med.uni-tuebingen.de
Full list of author information is available at the end of the article

Highlights

- We assessed sex differences and heritability of limbic and non-limbic volumes.
- Limbic volumes showed stronger sex differences and higher heritability overall.
- We trained brain sex classification models on limbic or non-limbic volumes.
- Performance was high and the sex class probabilities were heritable for all models.
- In females, major depression diagnosis was associated with higher limbic estimates compared to healthy controls.

Keywords Brain sex classification, Machine learning, Female mental health, Neuroimaging, Limbic system

Plain language summary

Psychiatric disorders have different prevalence between sexes, with women being twice as likely to develop depression and anxiety across the lifespan. Previous studies have investigated sex differences in brain structure that might contribute to this prevalence but have mostly focused on a single-structure level, potentially overlooking the interplay between brain regions. Sex differences in structures responsible for emotional regulation (limbic system), affected in many psychiatric disorders, have been previously reported. Here, we apply a machine learning model to obtain an estimate of brain sex for each participant based on the volumes of multiple brain regions. Particularly, we compared the estimates obtained with a model based solely on limbic structures with those obtained with a non-limbic model (entire brain except limbic structures) and a whole brain model. To investigate the genetic determinants of the models, we assessed the heritability of the estimates between identical twins and fraternal twins. The estimates of all our models were heritable, suggesting a genetic component contributing to brain sex. Finally, to investigate the association with mental health, we compared brain sex estimates in healthy subjects and in a depressed population. We found an association between depression and brain sex in females for the limbic model, but not for the non-limbic model. No effect was found in males. Overall, our results highlight the potential utility of machine learning models of brain sex based on relevant structures to better understand the link between sex differences in the brain and mental disorders.

Background

Common mental disorders occur at different prevalence rates between sexes [1]. In particular, women are twice as likely to develop anxiety and depression across the lifespan compared to men [1–4]. This difference arises after puberty [4, 5], suggesting the involvement of sex-specific factors in the development of such disorders [6]. To identify these factors, neuroimaging studies have investigated sex differences in brain structure and function [7–11], mostly using univariate analyses. However, discordant findings have been reported [7, 10, 11]. In a recent meta-analysis, Ritchie et al. [10] found generally larger volumes in males, while other studies reported larger volumes in females for different regions [8, 11]. Possible explanations might be differences in the normalization and segmentation processes [7, 12], as well as the effects of total brain volume [12]. Sex differences in total brain volume has been shown to drive many structural and volumetric differences in the brain [12, 13], leading to discordant results depending on the correction method applied. In addition, variations in structural and functional MRI according to the menstrual cycle and hormonal contraceptive use have been reported in the literature for many structures, such as hippocampus, amygdala, prefrontal cortex, cingulate cortex, and insula [14, 15]. Of note, many of the structures with notable sex differences and hormonal effects are part of the limbic system [16–20]. The limbic system is strongly involved in emotional processing, learning and memory, functions typically altered in mental and neurological disorders [21]. Due to its involvement in such functions, the limbic system has consequently been proposed as a key player in mood disorders such as major depression [22–25].

Recently, multivariate approaches have been developed to study sex differences in the brain. Machine learning models that classify for sex based on brain structural or functional magnetic resonance images (MRI) yield class probabilities that can be used as an imaging-derived multivariate phenotype to study sex differences on a continuum from female- to male-like brains [26–29]. Conceptually similar approaches have already been used extensively to predict brain age [30–36], where machine learning models deliver a continuous phenotype reflecting apparent aging effects. While most brain age studies to date built models based on data from the whole brain, regionally constrained models may identify

region-specific associations with mental health, such as frontal brain age alterations in schizophrenia or subcortical alterations in Alzheimer's disease [34]. In a recent study, Sanford and colleagues [36] investigated sex differences in local brain age gaps (i.e. difference between regionally constrained neuroimaging-predicted age and chronological age) in young adults. Compared to males, females showed significantly lower local brain age gap in the frontal region and insula, while they had significantly higher local brain age gap in the posterior regions. However, on a global scale the authors did not report any differences in brain age gaps, suggesting finer-grained (i.e., regional specific) models having a higher sensitivity to sex differences. Translating this finding into the field of sex classification, leveraging regional constrained models may provide estimates more sensitive to sex-specific phenotypes. Whereas Weis and colleagues [29] have classified sex based on the whole brain connectome and different functional brain networks separately, the potential of regionally constrained models to study brain sex based on structural MRI has yet to be investigated.

Here, we investigate whether a regionally constrained model based on brain volumes of the limbic system can correctly classify sex, and whether the obtained regional class probabilities are sensitive to mental health. We compared limbic brain sex to non-limbic brain sex, aiming to investigate relevant biological differences in brain sex determination as well as the possible clinical association with major depression. Specifically, the present study sought: (i) to compare sex differences at a univariate (i.e., single structure) level between limbic and non-limbic structures, (ii) to validate regionally constrained machine learning models trained either on limbic or non-limbic feature sets as compared to a whole brain model, (iii) to test for an association between obtained class probabilities and major depressive disorder (MDD) diagnosis. We hypothesize that (i) the regionally constrained models, much like whole brain models, are able to correctly predict sex from structural imaging data, (ii) the estimates (i.e., class probabilities) contain biologically meaningful variation (tested via heritability analysis), and that (iii) the limbic estimates have a stronger association with depression than estimates from other models.

Methods

Participant selection

As illustrated in Fig. 1, structural MRI data from the Human Connectome Project (HCP) [37] was used for univariate analysis of brain features, and for training of machine learning models in a healthy sample. For the HCP, the subject selection criteria (see Supplementary Figure S1, Additional File 1) aimed to (I) maintain an equal female-male ratio (based on biological sex as provided by the data), while (II) limiting possible confounding effects such as hormonal fluctuation, and (III) maximizing the sample size for machine learning. To achieve this, after excluding 14 individuals following quality control of the imaging data, female subjects were first selected according to the hormonal information available with the goal to limit the potential effects of irregular cycle and hormonal alterations such as presence of hypo- or hyperthyroidism and Thyroid Stimulating Hormone (TSH) levels out of the normal range (0.4–4.0 mU/L, as define by the HCP-YA data dictionary), yielding n = 391 females. These females were then matched to an equal number of males according to age and Euler number. Finally, in order to maximize the sample size, the remaining males (n = 105) were matched for age and Euler number with an equal number of randomly selected females from the subjects excluded in the first step. This procedure led to a total sample size of N = 992 subjects (50% females), with an age range 22–38 years old (females: mean age = 28.98, sd = 3.63; males: mean age = 27.90, sd = 3.60).

The Queensland Twin Imaging (QTIM) study was used as an independent healthy validation dataset [38] to replicate univariate analyses and to test the HCP-trained machine learning models. After excluding 20 outliers based on imaging quality control, only subjects in the age range 18–30 years were selected, to ensure a similar age range with the HCP. The final QTIM sample comprised N = 1017 subjects (61.6% females, age range 18–30 years old, females: mean age = 22.49, sd = 2.84; males: mean age = 22.29, sd = 2.86).

To test for clinical associations with the derived brain sex class probabilities, data from the Strategic Research Program for the Promotion of Brain Science (SRPBS) [39] was used. Due to data availability and the differences in prevalence across sexes [1], we focus on MDD. After excluding 10 outliers following quality control, we selected healthy controls (HC) and individuals with MDD, based on the diagnosis variable available in the dataset. To reduce scan site confounds, we only included HC data for sites in which MDD data was available. This yielded a sample of N = 844 subjects (HC: N = 595, 54.5% female, age range 18–80 years old, females: mean age = 41.34, sd = 14.99, males: mean age = 38.24, sd = 16.28; MDD: N = 249, 47.8% female, age range 18–75 years old, females: mean age = 43.39, sd = 12.87, males: mean age = 41.77, sd = 11.04).

Image segmentation, quality control and features selection

Raw T1-weighted MRI scans were preprocessed in FreeSurfer v7 and automated cortical and subcortical reconstruction were performed. To obtain a more precise segmentation of the limbic system, we

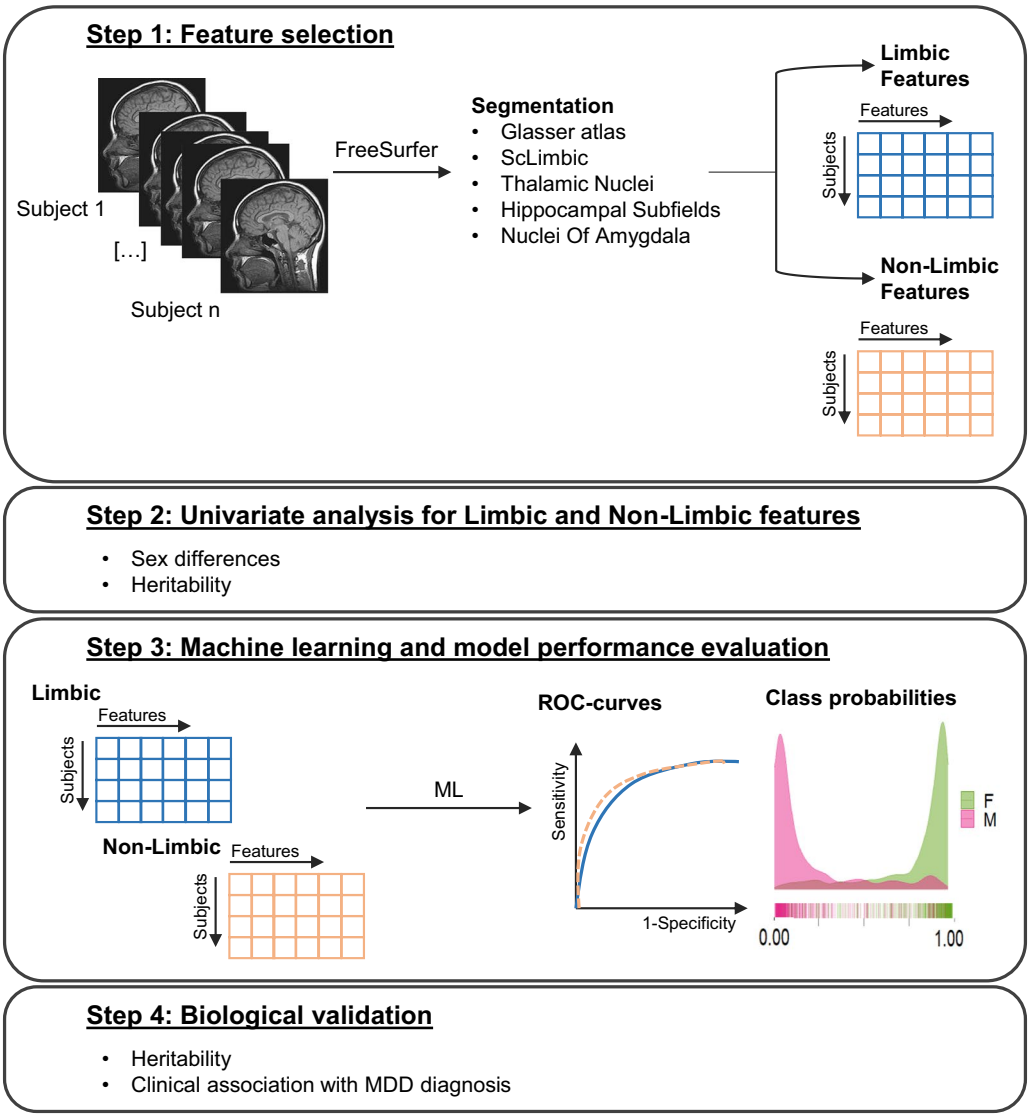

Fig. 1 Schematic overview of the study design. Step 1: After combining different segmentations in FreeSurfer, each structure was assigned either to the limbic or to the non-limbic feature set. Step 2: Univariate analysis of sex differences and feature heritability were applied to the limbic and non-limbic feature set. Step 3: A machine learning model for sex classification was trained on each feature set (regionally constrained models). Model performance was assessed via AUC-ROC and the class probabilities for each participant were stored. Step 4: The heritability and the clinical association with major depressive disorder (MDD) diagnosis was assessed for the class probabilities obtained for each model

combined volumes obtained with different segmentation approaches. The cerebral cortex was segmented using a multimodal parcellation described in Glasser et al. [40], which returns 180 features for each hemisphere. For subcortical regions, we combined the classic set of features from FreeSurfer with additional segmentations of subcortical limbic structures [17], hippocampus [41], amygdala [42], and thalamus subfields [43]. For all segmented cortical and subcortical regions, we calculated the corresponding volumes (mm^3). We then assigned the structures derived from each segmentation process to a limbic or non-limbic feature set based on the common definition of limbic system as found in the literature [17–20], carefully avoiding overlap between different segmentation approaches. The final feature set comprises 493 regions (358 features from Glasser multimodal segmentation [40], 17 subcortical features from FreeSurfer classical segmentation, 12 features from the subcortical limbic segmentation [17], 38 features from hippocampus segmentation [41], 18 features from the amygdala segmentation [42] and 50 features from thalami subfields segmentation [43]), of which we assigned 160 regions to

the limbic (see Supplementary Table 1, Additional File 1) and 333 regions to the non-limbic feature set.

To correct for differences in head size that could affect the machine learning models' ability to classify sex, all analysis were repeated accounting for the estimated total intracranial volume (eTIV). For each raw brain imaging feature, we regressed out the eTIV and used the eTIV corrected data as features in the respective eTIV-adjusted machine learning models. As further control, we compared this eTIV residualisation approach to another approach, the power-corrected proportion method (PCP) (D. [44, 45], which assumes an exponential relationship between the raw volume and the eTIV. We correlated the features obtained with the two methods (see Supplementary figure S2, Additional File 1), and since they converged on very similar results, we used the eTIV residualisation approach in all eTIV-accounting analyses.

We used Euler numbers, a proxy of image data quality [46] for quality control of the imaging data. We averaged Euler numbers from the left and right hemisphere and excluded subjects with an average Euler number lower than three standards deviations from the sample mean.

Sex differences and heritability analyses of single brain volumes

We first investigated differences between limbic and non-limbic structures using univariate analyses for each structure of interest. For each region, we tested for sex differences using linear models that accounted for age and Euler number. Subsequently, we assessed the differences in effect size distributions between limbic and non-limbic structures using a t-test. To account for differences in head size, we repeated the same univariate analysis introducing the eTIV as additional covariate together with age and Euler number. Next, we computed the broad sense heritability (see Heritability analysis) of volumes for each region and subsequently tested for heritability differences between limbic and non-limbic structures using a t-test. To look at the association between the two measures, we computed a correlation between the sex differences and the heritability for limbic and non-limbic structures. We tested for significant differences between correlation coefficients using a Fisher's test for the comparisons of independent correlations.

Sex classification models and clinical associations

For each set of features (limbic, non-limbic and whole brain), two sex classification models were fitted, one using the raw volumes, and one using the residuals from the same features after regressing out the estimated total intracranial volume as follow:

$$lm(Volume \sim Estimated\ Total\ Intracranial\ Volume)$$

To avoid confounding factors, our training sample was balanced for sex and subjects were matched for age and image quality. We trained our models on HCP data, using gradient tree boosting as implemented in the *xgboost* package in R (version 4.2.2). Sex was coded in the training sample as binary variable with 0 assigned to males and 1 assigned to females. The resulting class probability ranges between 0 (male-like) and 1 (female-like). The learning rate was set at $\eta = 0.01$ and the initial number of rounds to 1000, and we performed fivefold cross-validation within the training sample. For each iteration the prediction error was assessed and used to determine the optimal iteration number, used to train the final models on the full set of data. We then applied these models to the test samples to classify brain sex. The resulting class probabilities were extracted and used for further analysis. To evaluate each model performance, both the accuracy and the area under the receiving operating characteristic curves (AUC) were calculated. As further control for the training procedure, we trained our models in a sex-balanced subsample of QTIM, matching the participants according to age and Euler number (N = 780, 50% females, females: mean age = 22.47, sd = 2.89, males: mean age = 22.28, sd = 2.86) and evaluated the performances.

To test the statistical significance of the models, we performed permutation testing. For each raw and eTIV-controlled set of features, the classification labels (sex of the participants) were randomly permuted 5000 times, while maintaining the feature sets unchanged, resulting in the randomized association between the feature matrix and the labels. For each permutation, fivefold cross-validation was applied. The accuracies were stored and used to build a null distribution that was then compared against the accuracies of the true models. Both a cut-off of 50% (chance level) for the accuracy and AUC and significant statistical results in the permutation testing were considered to evaluate the model as successful in classifying sex.

For external validation, we applied the HCP-trained model to independent data from the QTIM and SBRPS samples and computed the class probabilities in these unseen datasets. In the SBRPS data, we tested for the association with major depression using a linear model considering the diagnosis as independent variable, class probability obtained with each model as the dependent variable and accounting for sex, age, Euler number and site, as follow

$$lm(Class\ probabilities \sim Diagnosis + sex + age + Euler\ number)$$

To overcome potential issues with a continuum of class probabilities originating from two distinct distributions (males, females), we repeated the same analysis within females and males separately, accounting for age and Euler number as covariates as follow:

$$lm(Class\ probabilities \sim Diagnosis + age + Euler\ number).$$

As further control, we repeated the same analyses (general and within each sex separately) on a subset of subjects with an equal number of HC and MDD subjects matched by age and sex (HC: N = 249; MDD: N = 249; females = 48%). We consider diagnosis as an independent variable and sex, age and Euler number as covariates for the analysis in the general sample, while accounting only for age and Euler number as covariates in the within sex analyses, following the models stated above. Effect sizes were assessed using Cohen's d [47].

Heritability analysis

Monozygotic and dizygotic twin couples were selected for the HCP and QTIM data. Only couples of twins with both twins in the data were selected in both HCP and QTIM dataset. Due to the sample, in the HCP all included dizygotic couples were of the same sex. In the QTIM sample, both same sex and discordant sex couples of twins were selected. Siblings were excluded from the analysis in both datasets. The total sample sizes for heritability analysis were N = 378 (MZ = 236, DZ = 142) for HCP and N = 674 (MZ = 302, DZ = 372) for QTIM data. The heritability analyses were run at a single-structure level and on the predicted class probabilities from the machine learning models. An AE model was used for both, returning the variance explained by the additive genetic component (A) and non-shared environment component (E). Sex, age and Euler number were accounted as covariates. For the single feature analyses, we repeated the same analysis on the raw volumes and on the residuals after regressing out the covariates and the total intracranial volume to control for head size. The heritability analyses were carried out with the *twinlm* function of the *mets* package in R (version 1.3.1).

Results

Limbic structures showed greater sex differences and heritability than non-limbic structures

The degree of sex differences of a given FreeSurfer-derived feature to its heritability in two independent samples (HCP and QTIM) is depicted in Fig. 2. Limbic structures showed significantly greater sex differences as compared to non-limbic structures in both the HCP ($t = 4.89$, $p < 0.001$) and QTIM ($t = 3.87$, $p < 0.001$) data. These results were replicated when accounting for the eTIV in both samples (HCP: $t = 5.72$, $p < 0.001$; QTIM: $t = 5.59$, $p < 0.001$), suggesting overall greater sex differences independent of head size in the limbic system. The limbic volumes were also more heritable than the non-limbic volumes, both with and without controlling for eTIV, in HCP (Raw features: $t = 4.85$, $p < 0.001$; eTIV-controlled: $t = 4.84$, $p < 0.001$) and QTIM sample (raw features: $t = 3.36$, $p < 0.001$, eTIV-controlled: $t = 3.75$, $p < 0.001$).

For both, limbic and non-limbic structures, we observed a significant positive association between sex difference and heritability, indicating that the most heritable features had also greater sex-differences (range across the eight regression lines depicted in Fig. 2: $r = 0.238$, $p = 0.003$ to $r = 0.633$ $p < 2.2e{-}16$) (see Supplementary Table 2, Additional File 2, for the complete list of values. Positive t-values for sex differences reflect greater volumes in males.). Slopes were not statistically different between limbic and non-limbic features, except for eTIV-corrected features in the HCP sample, where the slope difference between limbic and non-limbic reached statistical significance ($z = -2.125$, $p = 0.034$). However, caution is warranted interpreting this difference as this finding could not be replicated in QTIM.

Regionally constrained models successfully classify sex

We used the different feature sets (limbic, non-limbic, whole brain) from the HCP sample to train a machine learning model that classified sex from volumetric imaging data. We first assessed model performance within HCP using fivefold cross-validation, indicating that all three models were able to successfully classify sex (please see Fig. 3 for illustration). The limbic model achieved an accuracy of 87% and AUC of 0.935, while the non-limbic achieved an accuracy of 84.4% and AUC of 0.92. Both models performed approximately as well as the whole brain model (accuracy = 86.6%, AUC = 0.941). When controlling for eTIV, the models were still capable to correctly classify sex, although with a lower accuracy and AUC (limbic (eTIV): accuracy = 70.9%, AUC = 0.778; non-limbic (eTIV): accuracy = 74.6%, AUC = 0.819; whole brain (eTIV): accuracy = 77.2%, AUC = 0.861). No significant difference in the performance between the three models was found.

Permutation testing with 5000 permutations per model indicated that no permutation-based accuracy was better than the accuracy obtained with the true models, confirming the validity of our models (see Supplementary Figure S3, Additional File 1). Notably, when looking into

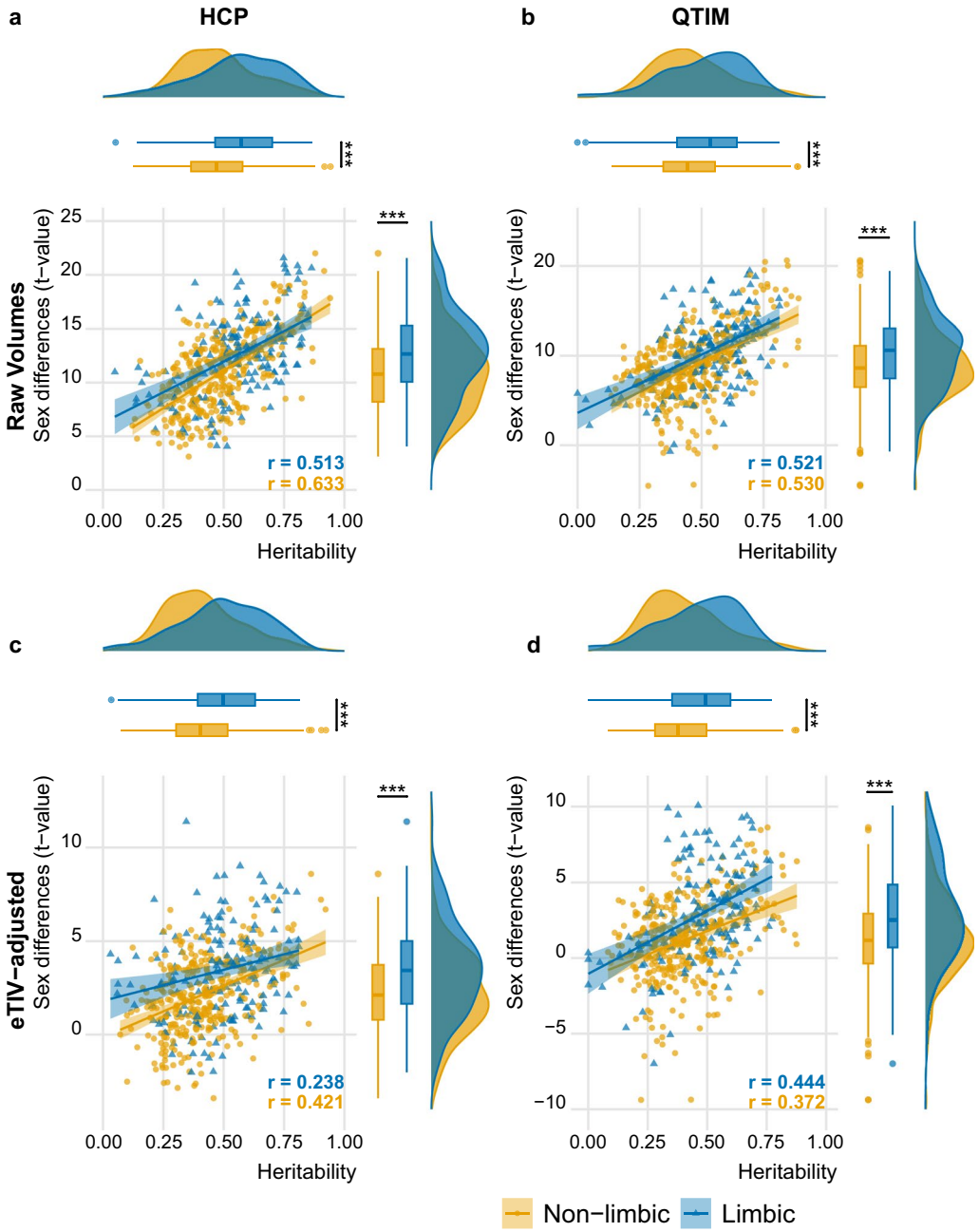

Fig. 2 Limbic volumes show stronger sex differences and higher heritability than non-limbic volumes. The analysis was initially conducted in HCP (panel **a**, **c**) and thereafter replicated in the independent QTIM sample (panel **b**, **d**). The upper row (**a**, **b**) presents results from raw volumes whereas the lower row (**c**, **d**) shows results after controlling for estimated total intracranial volume. Each data point reflects the effect of one brain region. Regression lines illustrate the direction of effect across all limbic or non-limbic regions, indicating that the most heritable features had also grater sex differences. In both samples, the identified effects from raw volumes were smaller yet still significant when accounting for the total intracranial volume. ***$p < 0.001$

the feature importance for the raw and eTIV-corrected whole brain models, the main contributors to both models belong to the limbic system (see Supplementary Figure S4, Additional File 1).

Next, we applied the HCP-trained models to QTIM data for external validation, confirming solid performance in independent data (with an accuracy of 82.3%, 77.6% and 79.7% and an AUC of 0.905, 0.885 and 0.920

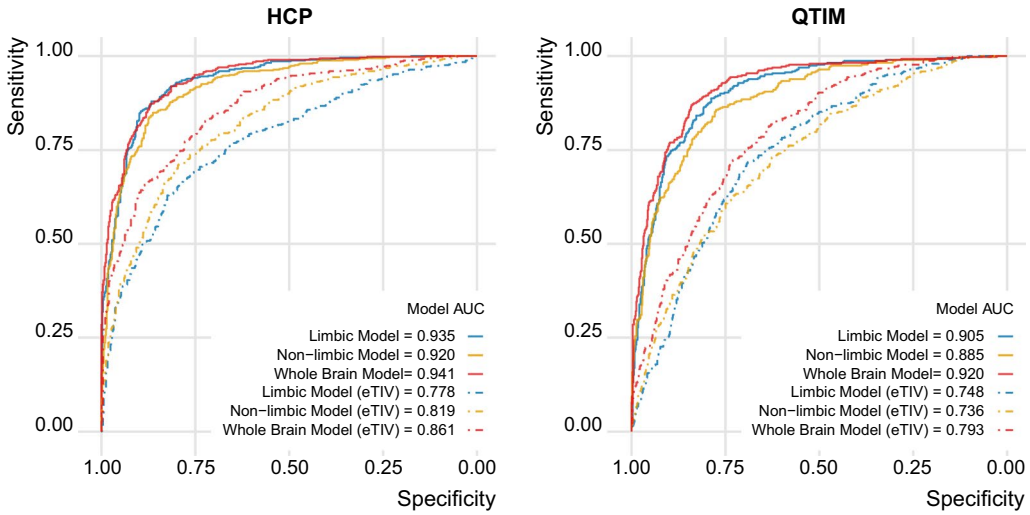

Fig. 3 The limbic model achieved performances similar to models based on non-limbic and whole brain data. The ROC curves and the AUC for each model in the HCP sample (cross-validation within the training set) and in the QTIM sample (validation in independent test data) show high performances for all models. Regressing eTIV from the features (dashed lines) reduced performance compared to raw volumes (solid lines) yet all models still performed well above chance

for limbic, non-limbic and whole brain model, respectively). For eTIV accounted models we also observed performances similar to those obtained within HCP (accuracy: 68.4%, 66.8%, 71.6%, AUC: 0.748, 0.736, 0.793 AUC for limbic, non-limbic and whole brain model, respectively).

When training the data in QTIM, the performances were in line with those obtained from the training in HCP and the independent testing in QTIM, with accuracies of 85.9%, 82.1% and 87.6% and AUC of 0.933, 0.902 and 0.943 for the limbic, non-limbic and whole brain model respectively. When correcting for eTIV, the models achieved lower but still relevant performances (accuracies: 73.7%, 74.0%, 77.2%, AUC: 0.806, 0.830, 0.865 for limbic (eTIV), non-limbic (eTIV), and whole brain (eTIV)) (see Supplementary Figure S5, Additional File 1). These results support the generalizability of our HCP trained model since there was only a minor improvement in performance metrics when the model was trained and validated in QTIM compared to the performance we achieved by applying our independently trained HCP model to QTIM.

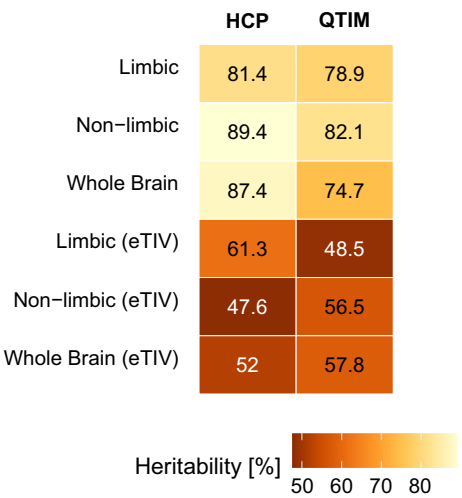

Fig. 4 The class probabilities from all models were heritable in both samples, indicating meaningful underlying biological information. Each cell shows the broad sense heritability in percent, with highest values for those derived from models that do not account for eTIV. Similar values are obtained when repeating the analyses within each sex (see Supplementary Figure S6, Additional File 1)

Brain sex class probabilities are heritable

The results of the broad sense heritability estimated from twin data of the class probabilities obtained from each of the machine learning models is shown in Fig. 4. Sex class probabilities were heritable for all models, both in the HCP (limbic: 81.4%, non-limbic: 89.4%, whole brain: 87.4%) and the QTIM data (limbic: 78.9%, non-limbic: 82.1%, whole brain: 74.7%). When accounting for the total intracranial volume, the broad sense heritability decreased in both samples yet was still substantial (HCP: limbic (eTIV): 61.3%, non-limbic (eTIV): 47.6%, whole brain (eTIV): 51.98%; QTIM: limbic (eTIV): 48.5%, non-limbic (eTIV): 56.5%, whole brain (eTIV): 57.8%).

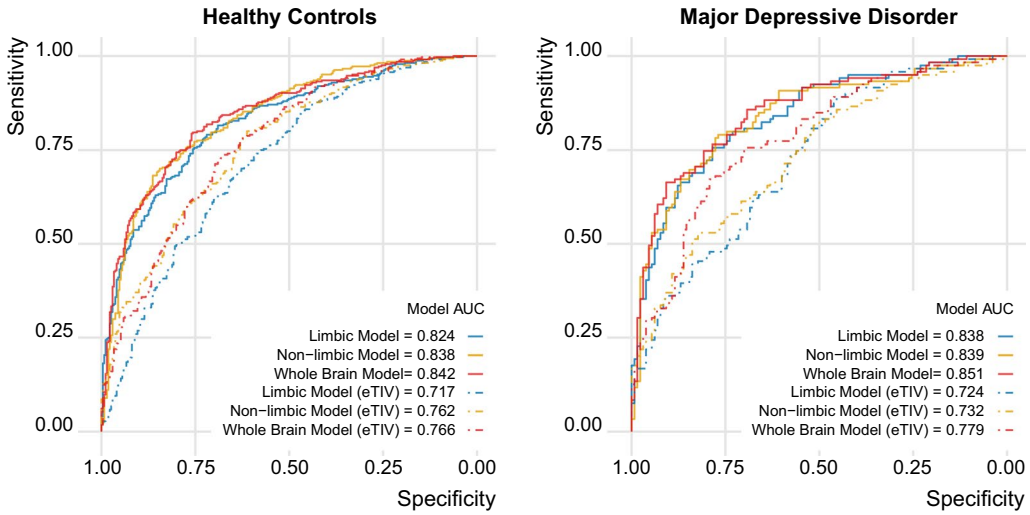

Fig. 5 All models achieve high performance in the SRPBS sample. The ROC curves and the AUC show the performance of each model for healthy controls and MDD patients. In healthy controls the accuracies were 75%, 73.3%, and 74.1% for the limbic, non-limbic and whole brain model, respectively. When accounting for eTIV the accuracy and the AUC decreased in all models, yielding 65.2%, 69.2% and 70.3% accuracy for limbic (eTIV), non-limbic (eTIV) and whole brain (eTIV), respectively. For the MDD group the accuracies were slightly lower for all models (raw models: 71.9%, 69.9% and 70.7%; eTIV-corrected: 63.5%, 64.7%, 68.3% for limbic, non-limbic and whole brain, respectively)

Higher class probabilities are associated with major depression

To investigate the association with MDD in a clinical sample, we applied the models to the SRPBS sample, considering healthy controls and depressed patients. After verifying the accuracy and the AUC of each model in healthy control and MDD patients (see Fig. 5), we extracted the class probabilities and associated them with the diagnosis. Our results indicated that the class probabilities in the clinical sample were overall higher (i.e. in the direction of a female brain phenotype) as compared to the healthy controls. These results were significant in all models, with strongest effect for the limbic model (limbic: t-value = 2.81, p = 0.005, Cohen's d = 0.21; non-limbic: t = 2.57, p = 0.011, Cohen's d = 0.2; whole: t = 2.40 p = 0.016, Cohen's d = 0.18). When accounting for eTIV, these results were no longer significant (please see Fig. 6 for details). Interestingly, when analyzing the association between class probabilities and depression within each sex, no effect was found in males, while only the limbic (t = 3.11, p = 0.002, Cohen's d = 0.34) and whole brain (t = 2.27, p = 0.023, Cohen's d = 0.25) models showed significant differences between HC and MDD in females. When accounting for eTIV, none of the associations within sex were significant. However, similar patterns of stronger effects in females compared to males was found for the three models (females: limbic (eTIV): t = 1.94, p = 0.053, Cohen's d = 0.21; non-limbic (e-TIV): t = 0.15, p = 0.881, Cohen's d = 0.02; whole brain (eTIV): t = 1.57, p = 0.118, Cohen's d = 0.17; males: limbic (eTIV): t = 0.13, p = 0.895, Cohen's d = 0.01; non-limbic (eTIV): t = 1.20, p = 0.231, Cohen's d = 0.13; whole brain (eTIV): t = 0.27, p = 0.787, Cohen's d = 0.03). When repeating the analyses in the age-matched HC-MDD subsample, only the limbic model in females showed significant greater class probabilities in MDD (t = 2.14, p = 0.033, Cohen's d = 0.28), while no significant association was found for the general sample or males or when correcting for eTIV (see Supplementary Figure S7, Additional File 1).

Discussion

The present study investigated whether regionally constrained machine learning models can correctly classify sex from T1-based volumetric MRI data, and if regional class probabilities are more sensitive to a female-prevalent mental disorder compared to whole brain models, the current standard in the field. Known for its involvement in various emotional and cognitive abilities as well as its central role for mental conditions, we here focused on the limbic system to investigate brain sex as a putative phenotype to study female-prevalent mental disorders. Our results based on univariate analysis indicate that limbic structures show substantially greater sex differences compared to non-limbic structures. Limbic structures were also more heritable, which may indirectly suggest that their sex differences are mediated by genetic components. These univariate findings supported the utility of the limbic system as a target for regional constrained

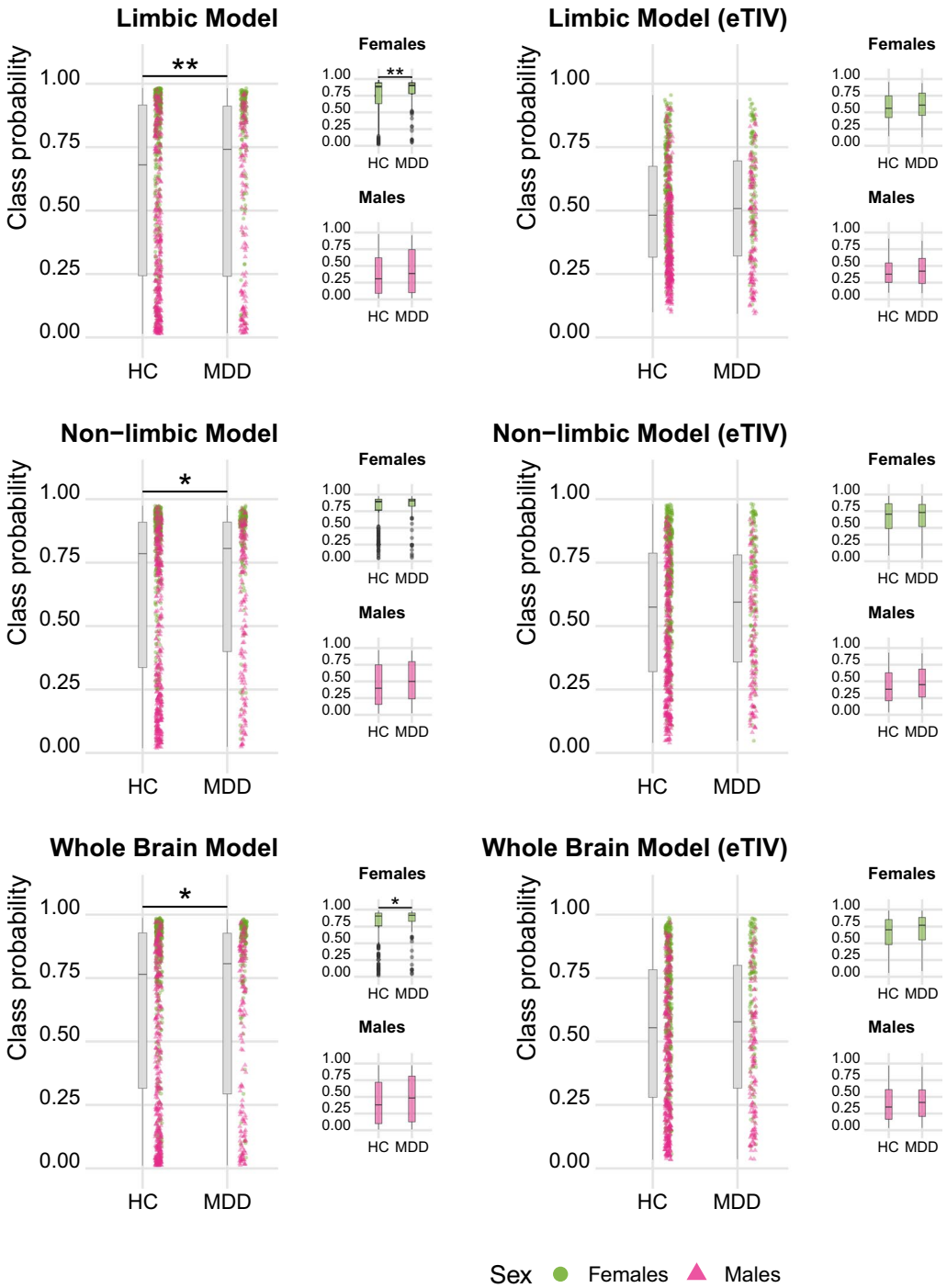

Fig. 6 MDD diagnosis is associated with a more female-like brain. Associations between class probabilities and diagnosis for major depression disorders. Higher class probabilities are associated with MDD diagnosis for all raw models. When accounting for eTIV these effects are no longer significant. *$p < 0.05$; **$p < 0.01$

brain sex prediction. Our multivariate machine learning models using limbic, non-limbic and whole brain features, respectively, were able to classify sex, yielding heritable class probabilities that were associated with major depression diagnosis, suggesting meaningful underlying biological variance.

Limbic structures showed greater sex differences and heritability than non-limbic structures

In line with previous studies showing associations of limbic structures with sex and hormonal status [7, 9, 14, 15], our univariate analyses showed that limbic structures were characterized by stronger sex differences compared to non-limbic structures. Likewise, we found stronger heritability for limbic compared to non-limbic structures, which resembles a previous report of high heritability in several limbic structures, including hippocampus, amygdala, and nucleus accumbens [48], and contributes to the global efforts to disentangle genetic contribution to brain structure and function [48–53]. The degree to which genetic contributions to the anatomy of the limbic system are influenced by sex is an understudied topic with mixed results thus far. While some studies find no sex differences in brain volumes [48], others report lower heritability in females, suggesting as possible explanation a greater influence of environmental factors such as the hormonal status in females [51, 54]. Supporting the latter hypothesis, many limbic structures express receptors for sex hormones [15, 55], contributing to plasticity mechanisms and structural changes. Here, we however found the same pattern of stronger heritability in limbic structure compared to non-limbic structures in males and females, indicating that the heritability patterns were not driven by one sex. Notably though, our results indicate a strong positive association between sex differences and heritability of brain structure, indicating that the structures showing greater sex differences are also the most genetically determined structures, even after controlling for head size.

Regionally constrained models successfully classify sex

The direct comparison and interpretation of sex effects on specific brain regions using univariate frameworks may be limited by the diversity in terms of protocols and parameters used by different studies [7, 15]. Thus, multivariate analysis in a machine learning framework can provide the advantage of condensing the information from a set of brain features into a single score (the class probability), which could be used as variable for further analysis avoiding difficult comparisons. Here, we demonstrated that sex can be classified from regionally constrained feature sets without performance loss. In particular, our limbic and non-limbic models were both able to classify brain sex with accuracies and AUC similar to those obtained with the whole brain model. It is worth noting that the limbic model achieved descriptively the same level of accuracy and AUC with much less features than the other models (limbic: 160 features, non-limbic: 333, whole brain: 493). It must be noted that the cross-validation procedure implemented in HCP did not account for the presence of family members in the splitting procedure, thus potentially biasing the classification performance via twin similarity. However, the external validation in two independent samples achieved high accuracies for all models, suggesting that the presence of twins in the model training did not induce substantial bias.

The importance of external validation for machine learning approaches to improve generalizability and reproducibility has been largely discussed in the literature [56–59]. In fact, while internal validation by splitting the data in training and test sets is an important tool to ensure reproducibility in a similar sample, it can also be affected by different forms of data leakage [56], inflating the performance of the model. In this context, external validation is key to ensure generalizability in independent data and in samples with different characteristics and acquired under different conditions [59]. Here, we externally validated the model in two different sample, proving the ability of our models to generalize under different conditions.

Brain sex class probabilities are heritable

Since deviations in class assignment can reflect both methodological error or biological variance we assessed the degree to which the class probabilities returned by each of the models capture biologically meaningful variance, by first assessing their heritability. Although previous work has investigated heritability of different structural and functional brain measures using univariate analyses [49, 51, 53, 54], heritability studies of multivariate estimates of brain sex are scarce. A recent study obtained heritable sex scores from a classification of whole brain data [60], in line with our results. Here, we extend these findings to regional constrained estimates of brain sex. We found that the class probabilities were heritable for all models, including those controlling for total intracranial volume, supporting their interpretability by pointing at underlying genetic factors.

Higher class probabilities are associated with major depression

Building upon the heritability results indicating biological meaningful variance, we furthermore investigated the clinical association with major depression under the hypothesis that class probabilities of limbic features may serve as a putative phenotype for the investigation of sex-prevalent mental health conditions. Based on data-availability, we focused on MDD, known to be more prevalent in females [1]. Previous univariate analyses have shown alterations in neuroimaging phenotypes in MDD [22–25, 61–63], with particular focus on the limbic system. Studies investigating structural changes in brain MRI have

displayed reduction in brain volume of several limbic regions in depressed subjects, underling the possible involvement of the limbic system in onset and maintenance of depression [22–25, 63]. However, many of these studies do not account for possible sex differences in the effects, reporting an overall smaller volume of cortical and subcortical structures in MDD patients compared to HC [22, 24, 25]. Multivariate classification models deliver probabilities at the single subject level, facilitating the investigation of sex specific effects in disorder associations. Comparing clinical data both across sex and within sex allowed us to quantify the degree to which MDD associations with brain imaging data are sex specific. Previous studies attempted to associate brain sex estimates with other common mental disorders and symptoms with sex deviant prevalence [64, 65]. However, these studies only focused on whole brain estimates. We extend this by showing that in all three models (limbic, non-limbic, whole brain) a more female-like brain is associated with MDD diagnosis. Interestingly, the association was strongest for the limbic model, complementing previous univariate findings on the relevance of the limbic system and supporting that regionally constrained estimates might represent sensitive markers to study brain–mental health associations. Moreover, when analyzing the two sexes separately, the effect survived only in females and only in the limbic and whole brain model, highlighting the importance of investigating sex differences in clinical associations with brain sex. Nevertheless, it must be noticed that when controlling for eTIV none of the models were significantly associated with depression diagnosis. Thus, from the data at hand we cannot rule out that the observed clinical associations were driven by sex differences in total brain size.

Methodological considerations and future directions

Potential limitations may stem from the fact that hormonal status, by acting on brain plasticity, might affect brain structure. Thus, the class probabilities obtained with our model might change according to the phase of the menstrual cycle, the intake of hormonal contraceptives, and the age-related hormonal status of the participants. As noted, many limbic structures are particularly sensitive to these changes [14, 15, 50]. We attempted to mitigate the hormonal effects by matching the subjects in the training data based on the available menstrual cycle information. However, the lack of precise data on the hormonal status of the participants hinders the effort to control for the variability due to hormonal effects. Women's health factors such as menstrual cycle, hormonal contraceptives use, pregnancy, and menopause are still largely overlooked in neuroimaging research [66]. Thus, further neuroimaging data and research considering hormonal levels across the menstrual cycle and different life stages is needed to provide more insights on brain sex classification models and to move the field forward. Moreover, while we placed strong emphasis on replicating our results in independent data, lending credibility in both the univariate analyses (heritability, sex differences) as well as in the multivariate analysis (model performance, heritability of class probability), data availability limited us in the replication of clinical associations. We thus deem it important to validate our clinical associations in another dedicated study, which would also test our results generalizability to different technical and clinical characteristics, such as differences in scan protocols or confounds due to the known impact of antidepressant treatment on brain plasticity [22, 63]. Finally, although the effects in clinical association analyses with MDD were in the same direction in all models (see Fig. 6), those controlling for eTIV did not reach statistical significance. Sex differences in total brain volume have been previously associated with structural differences in regional volume [12, 13, 45, 67]. Here, we corrected our analysis by regressing out the eTIV from each feature. For additional validation, we applied the power-corrected proportion approach and correlated the resulting features with features from the eTIV regression. The substantial correlation between both approaches ($r > 0.98$) corroborates that our eTIV correction is robust (see Supplementary Figure S2, Additional File 1). However, other methods beyond those tested could still yield different results [45, 67]. Indeed, the discussion on how to account for such differences is still ongoing. Recent studies showed how applying different correction methods for total intracranial volume based on features transformation might lead to different results in both univariate and multivariate analyses for sex differences [45, 67]. Other approaches acting at a subject selection level (e.g. matching participants of different sexes based on the total intracranial volume to ensure a balanced sample) might be more successful in limiting the effect of brain size in multivariate analyses [68]. Studies exploiting future developments in the area of eTIV correction might therefore add further insight into the impact of eTIV differences.

Conclusion

In conclusion, we here show in two independent data sets that sex can be classified from T1-based MRI volumes using regionally constrained models, by integrating prior knowledge into the selection of machine learning model features. Our limbic model achieved as high accuracy as the whole brain model using only one third of the features of the latter and the respective class probabilities displayed the strongest associations with major depression diagnosis. Heritability analysis further supports

that these probabilities capture biologically meaningful information.

Perspectives and significance

Previous studies have deployed machine learning models to classify sex from brain imaging data and derived male–female class probabilities at the individual level. However, the degree to which these probabilities vary across subsets of brain regions remains largely unexplored. Here, we study sex differences in limbic vs. non-limbic brain features and found strongest association of limbic sex probabilities with clinical data. These findings highlight the potential utility of regionally constrained models to investigate the link between brain sex and mental disorders and call for future investigations into other mental disorders with sex differences in prevalence and symptom profiles.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13293-024-00615-1.

Additional File 1: Supplementary Figures S1-S7, Supplementary Table 1.

Additional File 2: Supplementary Table 2.

Acknowledgements

This work was performed as part of the International Research Training Group: Women's Mental Health Across the Reproductive Years (IRTG 2804). The study was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). The authors acknowledge support from the Open Access Publication Fund of the University of Tübingen

Author contributions

Gloria Matte Bon: Conceptualization, Formal analysis, Data Curation, Data Interpretation, Visualization, Writing—Original Draft, Writing -Review and Editing. Dominik Kraft: Data Interpretation, Writing—Original Draft, Writing—Review and Editing. Erika Comasco: Supervision, Funding Acquisition, Writing—Review and Editing. Birgit Derntl: Supervision, Funding Acquisition, Writing—Review and Editing. Tobias Kaufmann: Conceptualization, Data Curation, Data Interpretation, Supervision, Funding Acquisition, Writing—Original Draft, Writing—Review and Editing.

Funding

Open Access funding enabled and organized by Projekt DEAL. TK received funding from the Interfaculty Graduate Program AI4Med-BW and the Fortüne Program (2660-0-0) from the Faculty of Medicine at University of Tübingen, the German Research Foundation (IRTG 2804), the Research Council of Norway (#323961) and the European Research Council (ERC CoG, #101086793, HealthyMom). TK is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764. BD received funding from the German Research Foundation (IRTG 2804, DE2319/9-1). EC received funding from SciLifeLab.

Availability of data and materials

HCP data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. In addition, the authors used data from the Queensland Twin IMaging (QTIM, https://openneuro.org/datasets/ds004169/versions/1.0.4) and the Strategic Research Program for Brain for the Promotion of Brain Science (SRPBS, https://bicr-resource.atr.jp/srpbs1600/). Data collection and sharing for the SRPBS data was provided by the DecNef Department at the Advanced Telecommunication Research Institute International, Kyoto, Japan. Analysis code is available at https://github.com/gloriamatte/Limbic_Model.

Declarations

Ethics approval and consent to participate

The HCP study obtained the ethical approvement of the IRB of Washington University. The QTIM study was approved by the Human Research Ethics Committee at the QIMR Berghofer Medical Research Institute (Ref#P701). All participants including a parent or guardian for those aged under 18 years provided written informed consent. The SRPBS was conducted in accordance with the Declaration of Helsinki and the study has been approved by the institutional review boards for each contributing institution. Written informed consent was obtained for all participants [39].

Consent for publication

Written informed consent was obtained by each study providing the data.

Competing interests

The authors declare no conflict of interest.

Author details

[1]Department of Psychiatry and Psychotherapy, Tübingen Center for Mental Health, University of Tübingen, Calwerstraße 14, 72076 Tübingen, Germany. [2]Department of Women's and Children's Health, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [3]German Center for Mental Health (DZPG), Partner Site Tübingen, Tübingen, Germany. [4]Centre for Precision Psychiatry, Division of Mental Health and Addiction, Institute of Clinical Medicine, University of Oslo, Oslo, Norway.

Received: 10 August 2023 Accepted: 29 April 2024
Published online: 15 May 2024

References

1. World Health Organization. World mental health report: Transforming mental health for all. Geneva: World Health Organization; 2022.
2. Kaczkurkin AN, Raznahan A, Satterthwaite TD. Sex differences in the developing brain: insights from multimodal neuroimaging. Neuropsychopharmacology. 2019;44(1):71–85. https://doi.org/10.1038/s41386-018-0111-z.
3. Pinares-Garcia P, Stratikopoulos M, Zagato A, Loke H, Lee J. Sex: a significant risk factor for neurodevelopmental and neurodegenerative disorders. Brain Sci. 2018;8(8):154. https://doi.org/10.3390/brainsci8080154.
4. Rubinow DR, Schmidt PJ. Sex differences and the neurobiology of affective disorders. Neuropsychopharmacology. 2019;44(1):111–28. https://doi.org/10.1038/s41386-018-0148-z.
5. Slavich GM, Sacher J. Stress, sex hormones, inflammation, and major depressive disorder: extending social signal transduction theory of depression to account for sex differences in mood disorders. Psychopharmacology. 2019;236(10):3063–79. https://doi.org/10.1007/s00213-019-05326-9.
6. DeCasien AR, Guma E, Liu S, Raznahan A. Sex differences in the human brain: a roadmap for more careful analysis and interpretation of a biological reality. Biol Sex Differ. 2022;13(1):43. https://doi.org/10.1186/s13293-022-00448-w.
7. Hillerer KM, Slattery DA, Pletzer B. Neurobiological mechanisms underlying sex-related differences in stress-related disorders: effects of neuroactive steroids on the hippocampus. Front Neuroendocrinol. 2019;55:100796. https://doi.org/10.1016/j.yfrne.2019.100796.
8. Liu S, Seidlitz J, Blumenthal JD, Clasen LS, Raznahan A. Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans. Proc Natl Acad Sci. 2020;117(31):18788–98. https://doi.org/10.1073/pnas.1919091117.

9. Pletzer B. Sex hormones and gender role relate to gray matter volumes in sexually dimorphic brain areas. Front Neurosci. 2019;13:592. https://doi.org/10.3389/fnins.2019.00592.
10. Ritchie SJ, Cox SR, Shen X, Lombardo MV, Reus LM, Alloza C, Harris MA, Alderson HL, Hunter S, Neilson E, Liewald DCM, Auyeung B, Whalley HC, Lawrie SM, Gale CR, Bastin ME, McIntosh AM, Deary IJ. Sex differences in the adult human brain: evidence from 5216 UK Biobank participants. Cereb Cortex. 2018;28(8):2959–75. https://doi.org/10.1093/cercor/bhy109.
11. Ruigrok ANV, Salimi-Khorshidi G, Lai M-C, Baron-Cohen S, Lombardo MV, Tait RJ, Suckling J. A meta-analysis of sex differences in human brain structure. Neurosci Biobehav Rev. 2014;39:34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004.
12. Eliot L, Ahmed A, Khan H, Patel J. Dump the "dimorphism": comprehensive synthesis of human brain studies reveals few male-female differences beyond size. Neurosci Biobehav Rev. 2021;125:667–97. https://doi.org/10.1016/j.neubiorev.2021.02.026.
13. Dhamala E, Ooi LQR, Chen J, Kong R, Anderson KM, Chin R, Yeo BTT, Holmes AJ. Proportional intracranial volume correction differentially biases behavioral predictions across neuroanatomical features, sexes, and development. Neuroimage. 2022;260: 119485. https://doi.org/10.1016/j.neuroimage.2022.119485.
14. Dubol M, Epperson CN, Sacher J, Pletzer B, Derntl B, Lanzenberger R, Sundström-Poromaa I, Comasco E. Neuroimaging the menstrual cycle: a multimodal systematic review. Front Neuroendocrinol. 2021;60: 100878. https://doi.org/10.1016/j.yfrne.2020.100878.
15. Rehbein E, Hornung J, Sundström Poromaa I, Derntl B. Shaping of the female human brain by sex hormones: a review. Neuroendocrinology. 2021;111(3):183–206. https://doi.org/10.1159/000507083.
16. Catenaccio E, Mu W, Lipton ML. Estrogen- and progesterone-mediated structural neuroplasticity in women: evidence from neuroimaging. Brain Struct Funct. 2016;221(8):3845–67. https://doi.org/10.1007/s00429-016-1197-x.
17. Greve DN, Billot B, Cordero D, Hoopes A, Hoffmann M, Dalca AV, Fischl B, Iglesias JE, Augustinack JC. A deep learning toolbox for automatic segmentation of subcortical limbic structures from MRI images. Neuroimage. 2021;244: 118610. https://doi.org/10.1016/j.neuroimage.2021.118610.
18. Grodd W, Kumar VJ, Schüz A, Lindig T, Scheffler K. The anterior and medial thalamic nuclei and the human limbic system: tracing the structural connectivity using diffusion-weighted imaging. Sci Rep. 2020;10(1):10957. https://doi.org/10.1038/s41598-020-67770-4.
19. Roxo MR, Franceschini PR, Zubaran C, Kleber FD, Sander JW. The limbic system conception and its historical evolution. Sci World J. 2011;11:2427–40. https://doi.org/10.1100/2011/157150.
20. Yamagata B, Murayama K, Black JM, Hancock R, Mimura M, Yang TT, Reiss AL, Hoeft F. Female-specific intergenerational transmission patterns of the human corticolimbic circuitry. J Neurosci. 2016;36(4):1254–60. https://doi.org/10.1523/JNEUROSCI.4974-14.2016.
21. Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF. The brain basis of emotion: a meta-analytic review. Behav Brain Sci. 2012;35(3):121–43. https://doi.org/10.1017/S0140525X11000446.
22. Koolschijn PCMP, van Haren NEM, Lensvelt-Mulders GJLM, Hulshoff Pol HE, Kahn RS. Brain volume abnormalities in major depressive disorder: a meta-analysis of magnetic resonance imaging studies. Hum Brain Mapp. 2009;30(11):3719–35. https://doi.org/10.1002/hbm.20801.
23. Sacher J, Neumann J, Fünfstück T, Soliman A, Villringer A, Schroeter ML. Mapping the depressed brain: a meta-analysis of structural and functional alterations in major depressive disorder. J Affect Disord. 2012;140(2):142–8. https://doi.org/10.1016/j.jad.2011.08.001.
24. Videbech P. Hippocampal volume and depression: a meta-analysis of MRI studies. Am J Psychiatry. 2004;161(11):1957–66. https://doi.org/10.1176/appi.ajp.161.11.1957.
25. Zheng R, Zhang Y, Yang Z, Han S, Cheng J. Reduced brain gray matter volume in patients with first-episode major depressive disorder: a quantitative meta-analysis. Front Psych. 2021;12: 671348. https://doi.org/10.3389/fpsyt.2021.671348.
26. Kim K, Joo YY, Ahn G, Wang H, Moon S, Kim H, Ahn W, Cha J. The sexual brain, genes, and cognition: a machine-predicted brain sex score explains individual differences in cognitive intelligence and genetic influence in young children. Hum Brain Mapp. 2022;43(12):3857–72. https://doi.org/10.1002/hbm.25888.
27. Tunç B, Solmaz B, Parker D, Satterthwaite TD, Elliott MA, Calkins ME, Ruparel K, Gur RE, Gur RC, Verma R. Establishing a link between sex-related differences in the structural connectome and behaviour. Phil Trans R Soc B Biol Sci. 2016;371(1688):20150111. https://doi.org/10.1098/rstb.2015.0111.
28. Vosberg DE, Syme C, Parker N, Richer L, Pausova Z, Paus T. Sex continuum in the brain and body during adolescence and psychological traits. Nat Hum Behav. 2020;5(2):265–72. https://doi.org/10.1038/s41562-020-00968-8.
29. Weis S, Patil KR, Hoffstaedter F, Nostro A, Yeo BTT, Eickhoff SB. Sex classification by resting state brain connectivity. Cereb Cortex. 2020;30(2):824–35. https://doi.org/10.1093/cercor/bhz129.
30. Cole JH, Franke K. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends Neurosci. 2017;40(12):681–90. https://doi.org/10.1016/j.tins.2017.10.001.
31. de Lange A-MG, Barth C, Kaufmann T, Anatürk M, Suri S, Ebmeier KP, Westlye LT. The maternal brain: region-specific patterns of brain aging are traceable decades after childbirth. Hum Brain Mapp. 2020;41(16):4718–29. https://doi.org/10.1002/hbm.25152.
32. de Lange A-MG, Kaufmann T, van der Meer D, Maglanoc LA, Alnæs D, Moberget T, Douaud G, Andreassen OA, Westlye LT. Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proc Natl Acad Sci. 2019;116(44):22341–6. https://doi.org/10.1073/pnas.1910666116.
33. Franke K, Ziegler G, Klöppel S, Gaser C. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage. 2010;50(3):883–92. https://doi.org/10.1016/j.neuroimage.2010.01.005.
34. Kaufmann T, van der Meer D, Doan NT, Schwarz E, Lund MJ, Agartz I, Alnæs D, Barch DM, Baur-Streubel R, Bertolino A, Bettella F, Beyer MK, Bøen E, Borgwardt S, Brandt CL, Buitelaar J, Celius EG, Cervenka S, Conzelmann A, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat Neurosci. 2019;22(10):1617–23. https://doi.org/10.1038/s41593-019-0471-7.
35. Popescu SG, Glocker B, Sharp DJ, Cole JH. Local brain-age: a U-Net model. Front Aging Neurosci. 2021;13: 761954. https://doi.org/10.3389/fnagi.2021.761954.
36. Sanford N, Ge R, Antoniades M, Modabbernia A, Haas SS, Whalley HC, Galea L, Popescu SG, Cole JH, Frangou S. Sex differences in predictors and regional patterns of brain age gap estimates. Hum Brain Mapp. 2022;43(15):4689–98. https://doi.org/10.1002/hbm.25983.
37. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: an overview. Neuroimage. 2013;80:62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
38. Strike LT, Blokland GAM, Hansell NK, MartinNG, Toga AW, Thompson PM, De Zubicaray, GI, McMahon KL, Wright MJ. Queensland Twin IMaging (QTIM). Openneuro. 2022. https://doi.org/10.18112/OPENNEURO.DS004169.V1.0.6
39. Tanaka SC, Yamashita A, Yahata N, Itahashi T, Lisi G, Yamada T, Ichikawa N, Takamura M, Yoshihara Y, Kunimatsu A, Okada N, Hashimoto R, Okada G, Sakai Y, Morimoto J, Narumoto J, Shimada Y, Mano H, Yoshida W, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. Sci Data. 2021;8(1):227. https://doi.org/10.1038/s41597-021-01004-8.
40. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC. A multi-modal parcellation of human cerebral cortex. Nature. 2016;536(7615):171–8. https://doi.org/10.1038/nature18933.
41. Iglesias JE, Augustinack JC, Nguyen K, Player CM, Player A, Wright M, Roy N, Frosch MP, McKee AC, Wald LL, Fischl B, Van Leemput K. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. Neuroimage. 2015;115:117–37. https://doi.org/10.1016/j.neuroimage.2015.04.042.
42. Saygin ZM, Kliemann D, Iglesias JE, van der Kouwe AJW, Boyd E, Reuter M, Stevens A, Van Leemput K, McKee A, Frosch MP, Fischl B, Augustinack JC. High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. Neuroimage. 2017;155:370–82. https://doi.org/10.1016/j.neuroimage.2017.04.046.
43. Iglesias JE, Insausti R, Lerma-Usabiaga G, Bocchetta M, Van Leemput K, Greve DN, van der Kouwe A, Fischl B, Caballero-Gaudes C, Paz-Alonso PM.

A probabilistic atlas of the human thalamic nuclei combining ex vivo MRI and histology. Neuroimage. 2018;183:314–26. https://doi.org/10.1016/j.neuroimage.2018.08.012.
44. Liu D, Johnson HJ, Long JD, Magnotta VA, Paulsen JS. The power-proportion method for intracranial volume correction in volumetric imaging analysis. Front Neurosci. 2014. https://doi.org/10.3389/fnins.2014.00356.
45. Sanchis-Segura C, Ibañez-Gual MV, Adrián-Ventura J, Aguirre N, Gómez-Cruz ÁJ, Avila C, Forn C. Sex differences in gray matter volume: How many and how large are they really? Biol Sex Differ. 2019;10(1):32. https://doi.org/10.1186/s13293-019-0245-7.
46. Rosen AFG, Roalf DR, Ruparel K, Blake J, Seelaus K, Villa LP, Ciric R, Cook PA, Davatzikos C, Elliott MA, Garcia de La Garza A, Gennatas ED, Quarmley M, Schmitt JE, Shinohara RT, Tisdall MD, Craddock RC, Gur RE, Gur RC, Satterthwaite TD. Quantitative assessment of structural image quality. Neuroimage. 2018;169:407–18. https://doi.org/10.1016/j.neuroimage.2017.12.059.
47. Cohen J. Statistical power analysis for the behavioral sciences (2nd ed). L. Erlbaum Associates, 1988.
48. den Braber A, Bohlken MM, Brouwer RM, van't Ent D, Kanai R, Kahn RS, de Geus EJC, Hulshoff Pol HE, Boomsma DI. Heritability of subcortical brain measures: a perspective for future genome-wide association studies. Neuroimage. 2013;83:98–102. https://doi.org/10.1016/j.neuroimage.2013.06.027.
49. Adhikari BM, Jahanshad N, Shukla D, Glahn DC, Blangero J, Fox PT, Reynolds RC, Cox RW, Fieremans E, Veraart J, Novikov DS, Nichols TE, Hong LE, Thompson PM, Kochunov P. Comparison of heritability estimates on resting state fMRI connectivity phenotypes using the ENIGMA analysis pipeline. Hum Brain Mapp. 2018;39(12):4893–902. https://doi.org/10.1002/hbm.24331.
50. Barth C, Steele CJ, Mueller K, Rekkas VP, Arélin K, Pampel A, Burmann I, Kratzsch J, Villringer A, Sacher J. In-vivo dynamics of the human hippocampus across the menstrual cycle. Sci Rep. 2016;6(1):32833. https://doi.org/10.1038/srep32833.
51. Batouli SAH, Sachdev PS, Wen W, Wright MJ, Ames D, Trollor JN. Heritability of brain volumes in older adults: the older Australian Twins Study. Neurobiol Aging. 2014;35(4):937.e5-937.e18. https://doi.org/10.1016/j.neurobiolaging.2013.10.079.
52. Chiang M-C, McMahon KL, de Zubicaray GI, Martin NG, Hickie I, Toga AW, Wright MJ, Thompson PM. Genetics of white matter development: A DTI study of 705 twins and their siblings aged 12 to 29. Neuroimage. 2011;54(3):2308–17. https://doi.org/10.1016/j.neuroimage.2010.10.015.
53. Kochunov P, Jahanshad N, Marcus D, Winkler A, Sprooten E, Nichols TE, Wright SN, Hong LE, Patel B, Behrens T, Jbabdi S, Andersson J, Lenglet C, Yacoub E, Moeller S, Auerbach E, Ugurbil K, Sotiropoulos SN, Brouwer RM, et al. Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data. Neuroimage. 2015;111:300–11. https://doi.org/10.1016/j.neuroimage.2015.02.050.
54. Lukies MW, Watanabe Y, Tanaka H, Takahashi H, Ogata S, Omura K, Yorifuji S, Tomiyama N, the Osaka University Twin Research Group. Heritability of brain volume on MRI in middle to advanced age: a twin study of Japanese adults. PLoS ONE. 2017;12(4): e0175800. https://doi.org/10.1371/journal.pone.0175800.
55. Sundström-Poromaa I, Comasco E, Sumner R, Luders E. Progesterone—friend or foe? Front Neuroendocrinol. 2020;59: 100856. https://doi.org/10.1016/j.yfrne.2020.100856.
56. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun. 2024;15(1):1829. https://doi.org/10.1038/s41467-024-46150-w.
57. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol. 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.
58. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5(1):48. https://doi.org/10.1038/s41746-022-00592-y.
59. Varoquaux G, Colliot O. Evaluating machine learning models and their diagnostic value. In: Colliot O, editor. Machine learning for brain disorders, vol. 197. New York: Springer, US; 2023. p. 601–30.
60. van Eijk L, Zhu D, Couvy-Duchesne B, Strike LT, Lee AJ, Hansell NK, Thompson PM, de Zubicaray GI, McMahon KL, Wright MJ, Zietsch BP. Are sex differences in human brain structure associated with sex differences in behavior? Psychol Sci. 2021;32(8):1183–97. https://doi.org/10.1177/0956797621996664.
61. Grieve SM, Korgaonkar MS, Koslow SH, Gordon E, Williams LM. Widespread reductions in gray matter volume in depression. NeuroImage Clin. 2013;3:332–9. https://doi.org/10.1016/j.nicl.2013.08.016.
62. Harris MA, Cox SR, de Nooij L, Barbu MC, Adams MJ, Shen X, Deary IJ, Lawrie SM, McIntosh AM, Whalley HC. Structural neuroimaging measures and lifetime depression across levels of phenotyping in UK biobank. Transl Psychiatry. 2022;12(1):157. https://doi.org/10.1038/s41398-022-01926-w.
63. Schmaal L, Veltman DJ, van Erp TGM, Sämann PG, Frodl T, Jahanshad N, Loehrer E, Tiemeier H, Hofman A, Niessen WJ, Vernooij MW, Ikram MA, Wittfeld K, Grabe HJ, Block A, Hegenscheid K, Völzke H, Hoehn D, Czisch M, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. Mol Psychiatry. 2016;21(6):806–12. https://doi.org/10.1038/mp.2015.69.
64. Phillips OR, Onopa AK, Hsu V, Ollila HM, Hillary RP, Hallmayer J, Gotlib IH, Taylor J, Mackey L, Singh MK. Beyond a binary classification of sex: an examination of brain sex differentiation, psychopathology, and genotype. J Am Acad Child Adolesc Psychiatry. 2019;58(8):787–98. https://doi.org/10.1016/j.jaac.2018.09.425.
65. van Eijk L, Zietsch BP. Testing the extreme male brain hypothesis: is autism spectrum disorder associated with a more male-typical brain? Autism Res. 2021;14(8):1597–608. https://doi.org/10.1002/aur.2537.
66. Taylor CM, Pritschet L, Jacobs EG. The scientific body of knowledge—whose body does it serve? A spotlight on oral contraceptives and women's health factors in neuroimaging. Front Neuroendocrinol. 2021;60:100874. https://doi.org/10.1016/j.yfrne.2020.100874.
67. Sanchis-Segura C, Ibañez-Gual MV, Aguirre N, Cruz-Gómez ÁJ, Forn C. Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. Sci Rep. 2020;10(1):12953. https://doi.org/10.1038/s41598-020-69361-9.
68. Wiersch L, Hamdan S, Hoffstaedter F, Votinov M, Habel U, Clemens B, Derntl B, Eickhoff SB, Patil KR, Weis S. Accurate sex prediction of cisgender and transgender individuals without brain size bias. Sci Rep. 2023;13(1):13868. https://doi.org/10.1038/s41598-023-37508-z.

Publisher's Note