Title: Analyzing brain data by sex: Are we asking the right question?

Abbreviated title: Analyzing brain data by sex

Authors: Nitay Alon (MA)[1], Isaac Meilijson (PhD)[1], Daphna Joel (PhD)[2,3*]

[1] School of Mathematical Sciences, Tel Aviv University. Tel-Aviv, 6997801, Israel.

[2] School of Psychological Sciences, Tel-Aviv University. Tel-Aviv, 6997801, Israel.

[3] Sagol School of Neuroscience, Tel-Aviv University. Tel-Aviv, 6997801, Israel.

*Corresponding author: djoel@tauex.tau.ac.il

Conflict of interest statement: The authors declare no competing financial interests.

Acknowledgments: This work has been conducted using the UK Biobank Resource under Application 42111, and supported by the Israel Science Foundation (grant No. 217/16 to DJ and IM).

Abstract

The decades old hypothesis that sex effects on the brain result in 'female' and 'male' phenotypes governs conventional analyses by sex. In these (e.g., Student's t-test), the null hypothesis is that males and females belong to a single population (or phenotype), and the alternative hypothesis is that they belong to two different populations/phenotypes. Yet, evidence that sex effects may be opposite under different conditions raises a third hypothesis – that both females and males may manifest each of the two phenotypes of a brain measure. Here we applied a mixture analysis, which can test this latter hypothesis, and Student's t-test to 289 MRI-derived measures of grey and white matter from 23,935 human brains. Whereas Student's t-test yielded significant sex/gender differences in 225 measures, the mixture analysis revealed that 282 brain measures were better described by the hypothesis that women and men sample from the same two phenotypes, and that, for the most part, they do so with quite similar probabilities. A further analysis of 41 brain measures for which there were a 'female-favored' and a 'male-favored' phenotype, revealed that brains do not consistently manifested the male-favored (or the female-favored) phenotype. Last, considering the relations between all brain measures, the brain architectures of women and men were remarkably similar. These results do not support the existence of 'female' and 'male' brain phenotypes but are consistent with other lines of evidence suggesting that sex category explains a very small part of the variability in human brain structure.

Introduction

In vivo and *in vitro* studies of laboratory animals reveal that sex-related factors affect many aspects of the developing and mature brain (for review see Joel et al., 2020; McCarthy, 2020; McEwen and Milner 2017). These findings are typically taken as evidence for the existence of a female and a male phenotype for specific brain measures and for the brain as a whole (e.g., the masculinization hypothesis, McCarthy, 2020). The assumption that such phenotypes exist, is also manifested in the almost unanimous interpretation of the call to consider sex as a biological variable as a call to assess sex differences in specific brain measures (e.g., Diester et al., 2019; Prager, 2017; Ritz et al., 2014; Shansky & Woolley, 2016). In such an analysis, the null hypothesis is that females and males belong to a single population (or phenotype), and if this hypothesis is rejected, the conclusion is that males and females belong to two different populations/phenotypes.

Studies in laboratory animals reveal, however, that sex effects on a specific brain measure may be opposite under different environmental conditions, so that the phenotype typical of females under one set of conditions may be typical of males under another set of conditions, and vice versa (for review see Joel, 2011; Joel et al., 2020). Such observations suggest another type of relationship between sex and the brain, namely, that brain measures exhibit two phenotypes, which may be manifested by both males and females, albeit in different proportions or under different conditions. This latter hypothesis cannot be tested using a conventional analysis by sex, such as Student's t-test, but may be tested using a mixture analysis, which tests which hypothesis better describes the data – the hypothesis that females and males sample[1], respectively, from a female and a male phenotypes (the 'pure-types'

[1] References to 'sample' here and elsewhere in the text are used with the following meaning: Each phenotype contains a range of possible values which occur with different frequencies, and the specific value of a brain measure for an individual brain is sampled from this distribution. This value may be sampled from one of two phenotypes, and the major question of the present study concerns the relations

hypothesis, Fig. 1A) or the hypothesis that females and males sample from the same two phenotypes ('mixed-types' hypothesis, Fig. 1B-E).

Here we compared the results of Student's t-test and mixture analysis applied to 289 MRI-derived measures of grey and white matter from the brains of ~24,000 women and men. The mixture analysis was conducted under the assumption that two phenotypes (distributions) underlie the observed distribution of each brain measure. To test whether women and men sample from two different phenotypes ('pure-types') or from the same two phenotypes ('mixed-types'), the expectation-maximization algorithm (Dempster et al., 1997) was used to compute maximum likelihood estimators of the parameters (mean and variance) of each of a measure's two underlying distributions as well as the proportions of men and women who sample from each of these distributions. For mixed-types measures, we further tested whether women and men sample from the two phenotypes with same or different probabilities (Fig. 1C-D). The analyses were performed under the working assumption that the underlying distributions are Gaussian. Appendix I explains that our approach yields essentially the same answers under a broader class of distributions, and provides data regarding the suitability of the Gaussian assumption for the data analyzed here.

Finally, for the brain measures for which most women sample from one phenotype ('female-favored' phenotype, e.g., dashed line in Fig. 1D) while most men sample from the other ('male-favored' phenotype, e.g., solid line in Fig. 1D), we tested whether sampling is consistent across brain features (e.g., consistently sampling from the male-favored phenotype). We computed for each individual and each of these measures the posterior[2]

between a person's sex category (female, male) and the phenotype from which the value of each of their brain measures is sampled. 'Sample' as used in this paper does not imply an intentional process on the part of brains or humans.

[2] 'Posterior probability' refers to the probability that this measure was sampled from the 'male-favored' phenotype, given the observed value of this brain measure in this brain and the underlying distributions parameters.

probabilities to select from the male-favored distribution, and assessed the correlation between these posterior probabilities for all possible pairs of brain measures of the same type (e.g., volume). If brains are consistent in the phenotype from which they sample, then high positive correlations are expected between all pairs of measures, whereas if brains are 'mosaics' (Joel, 2011, 2021; Joel et al., 2015) – unique combinations of features, some in the

male-favored phenotype and some in the female-favored phenotype - then most correlations are expected to be low.

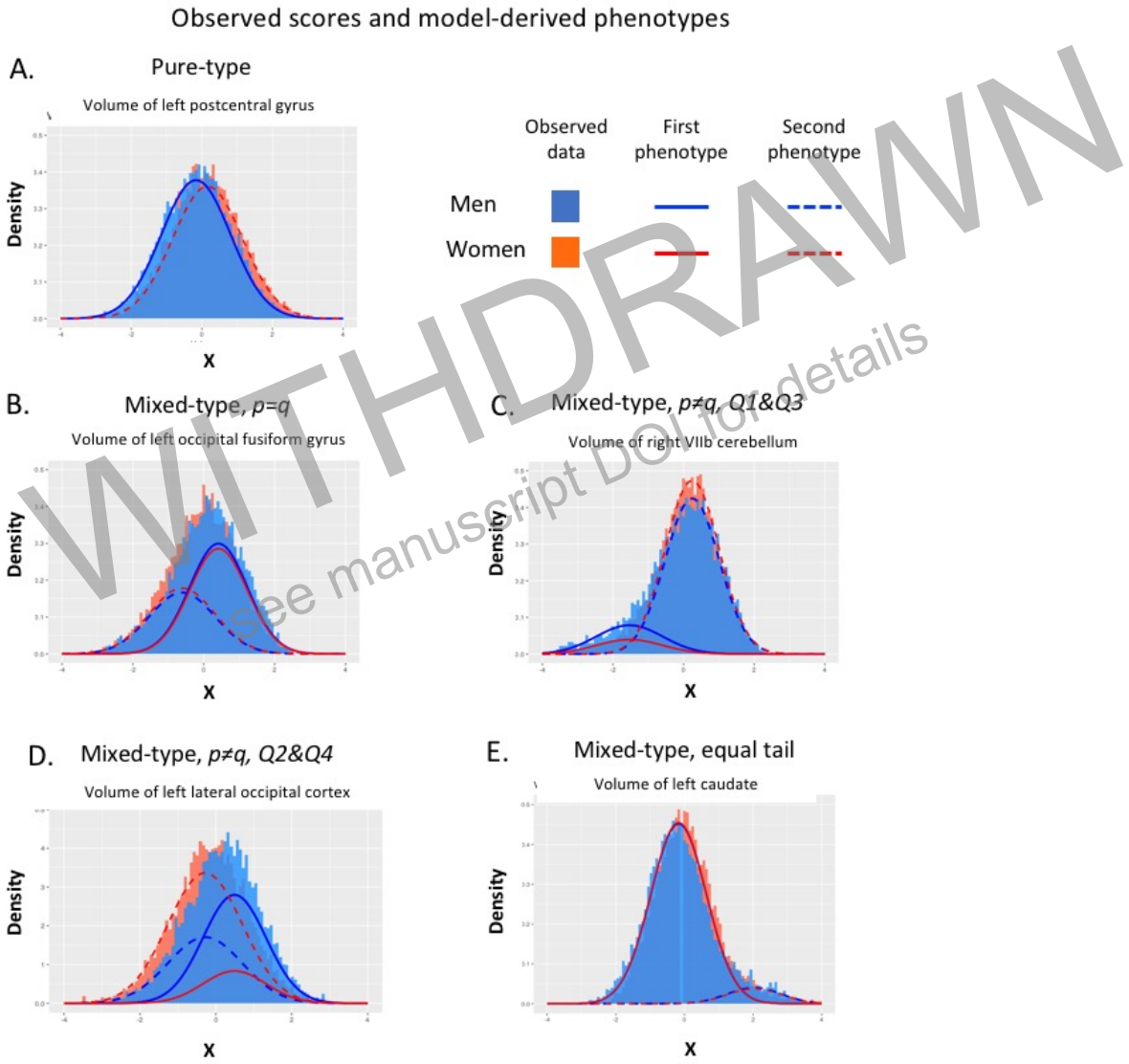

Figure 1. Observed scores and model-derived phenotypes. (A-E) Frequency distributions of the observed scores of women (orange) and men (light blue) for specific brain measures, and of the scores of women (red) and men (blue) on the model-fitted underlying phenotypes - one presented with a dashed line and the other with a solid line. (A) An example of a pure-types measure. Of the seven measures for which the pure-types hypothesis was not rejected, the volume of the left postcentral gyrus showed the largest sex/gender difference in the observed data (Cohen's d = 0.273). (B) An example of a mixed-types measure

in which men and women sample with the same probability from the two model-fitted phenotypes ($p = q$). (C) An example of a mixed-types measure in which both men and women 'favor' the same model-fitted phenotype, but with significantly different probabilities ($p \neq q$, Q1&Q3). (D) An example of a mixed-types measure in which men 'favor' one model-fitted phenotype and women 'favor' the other ($p \neq q$, Q2&Q4). Of the 41 such measures, the volume of the left lateral occipital cortex showed the largest sex/gender difference in sampling probabilities (Cohen's h = 0.764) and in the observed data (Cohen's d = 0.338). (E) An example of a mixed-types measure with a "tail" – only a small proportion of humans sample from one of the model-fitted phenotypes. In this example, men and women sample with the same probability from the two model-fitted phenotypes ($p = q$).

MATERIALS AND METHODS

Data collection and preparation for analysis

The present study was conducted as part of UK Biobank application 42111, and made use of imaging-derived native space measures generated by an image-processing pipeline developed and ran on behalf of UK Biobank (Alfaro-Almagro et al., 2018; Miller et al., 2016). Data were derived from the brains of 12,466 women (mean age = 65.16 years, SD = 7.28) and 11,469 men (mean age = 66.52 years, SD = 7.56). The following measures were included in the analysis: the volume of 139 regions of grey matter; the mean diffusivity (MD) and fractional anisotropy (FA) of 48 tracts defined using the Tract-Based Spatial Statistics analysis; and the weighted-mean MD and FA of a set of 27 major tracts, derived using probabilistic tractography-based analysis (for details of the acquisition protocols, image processing pipeline, and derived measures, see https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf). Because these measures are correlated with total brain volume (Sanchis-Segura et al., 2018; Takao et al., 2011; Vos et al., 2011) and there is a large sex/gender difference in the latter (Cohen's d = 1.36 in the present

sample), all analyses were conducted with brain volume taken into account using the power method (Liu et al., 2014; Sanchis-Segura et al., 2018). Total intracranial volume (TIV) was calculated as the sum of the following two variables from the UK Biobank dataset: volume of grey and white matter and volume of ventricular cerebrospinal fluid. For each brain measure, a linear regression of the logarithm of its value versus log-TIV was fitted, and the residuals were used in all subsequent analyses.

Experimental Design and Statistical Analysis

Student's t-test and effect size (Cohen's d)

For each brain measure we computed Student's t-test and Cohen's d.

Mixture analysis

Expectation-Maximization Algorithm (EM)

The EM algorithm (Dempster et al., 1997) is applied to data assumed to be generated from a mixture of parametric distributions with unknown parameters and unknown mixture probabilities. The working assumption is that the data are sampled from two Gaussian distributions, one with parameters, μ_1, σ_1, and the other with parameters, μ_2, σ_2. The proportions of men and women who sample from the high-mean distribution are p and q, respectively. The EM method was used to compute maximum likelihood estimators (MLE) for both the distribution parameters (μ_1, σ_1, μ_2, σ_2) and the proportions (p and q), with the following equations for the expectation (E) and maximization (M) steps. Let I be the indicator variable that a man's feature is sampled from the high-mean distribution, and let J be the indicator variable that a woman's feature is sampled from the high-mean distribution.

Let μ_1, σ_1 be the parameters (mean and standard deviation) of the high-mean Gaussian distribution and let μ_2, σ_2 be the parameters of the other Gaussian component. Let p and q be the proportion of men and women, respectively, sampling from the high-mean distribution. Let x, y denote generically the feature values of men and women and let m, n be the corresponding sizes of the two groups. At the i^{th}iteration of the algorithm, the E-step is computed using the updates from the $i-1$ iteration as:

$$\hat{I}(x) = E(I|x;\mu_1,\mu_2,\sigma_1,\sigma_2,p,q) = \frac{p * N(x;\mu_1,\sigma_1)}{p * N(x;\mu_1,\sigma_1) + (1-p) * N(x;\mu_2,\sigma_2)}$$

$$\hat{J}(y) = E(J|y;\mu_1,\mu_2,\sigma_1,\sigma_2,p,q) = \frac{q * N(y;\mu_1,\sigma_1)}{q * N(y;\mu_1,\sigma_1) + (1-q) * N(y;\mu_2,\sigma_2)}$$

The M-step parameters are updated in terms of $\hat{I} = \hat{I}(x), \hat{J} = \hat{J}(y)$ as:

$$\mu_{1,new} = \frac{\Sigma\hat{I} * x + \Sigma\hat{J} * y}{\left(\sum\hat{I} + \sum\hat{J}\right)}$$

$$\mu_{2,new} = \frac{\sum x + \sum y - \sum\hat{I} * x + \Sigma\hat{J} * y}{m + n - \left(\sum\hat{I} + \sum\hat{J}\right)}$$

$$\sigma^2_{1,new} = \frac{\sum\hat{I} * \left(x - \mu_{1,new}\right)^2 + \sum\hat{J}\left(y - \mu_{2,new}\right)^2}{\left(\sum\hat{I} + \sum\hat{J}\right)}$$

$$\sigma^2_{2,new} = \frac{\sum\left(1-\hat{I}\right) * \left(x - \mu_{2,new}\right)^2 + \sum\left(1-\hat{J}\right)\left(y - \mu_{2,new}\right)^2}{m + n - \left(\sum\hat{I} + \sum\hat{J}\right)}$$

Until convergence.

A well-known pathology of the EM algorithm stems from its ability to inflate log-likelihood by assigning a single observation to one Gaussian and decrease this Gausssian's variance to zero (Bishop, 2006). To prevent this, the ratio of the variance of the two Gaussians was constrained to be between 0.2 and 5. However, these limits were not met - the minimal and maximal variance ratios over the 289 features analyzed were 0.3 and 4.7, respectively.

The EM-computed MLE's were supported by a bootstrap simulation (Efron, 1979) for nine randomly selected brain measures (three for which $p = q$, three for which $p \neq q$, but both p and q are larger than 0.5; and three for which $p \neq q$, with $p > 0.5$ and $q < 0.5$). For these nine

measures, the EM algorithm was applied over 200 bootstrap samples. The MLE's standard deviation of the parameters was small (0.0526-0.0530), supporting the main findings of this paper (see Figure 6A in Appendix I for a scattergram of the bootstrap estimated p and q for three measures, one of each type).

A power analysis (described in Appendix I) confirmed that the size of the sample in the present study is adequate for the aims of the present study (Figure 6B).

Hypothesis testing

Pure-types or mixed-types?

For each brain measure, the null hypothesis is that the pair (p, q) is either (0,1) or (1,0) - that is, that women sample from a 'female' distribution and men sample from a 'male' distribution. The alternative hypothesis is that there exist two latent distributions with parameters μ_1,σ_1 and μ_2,σ_2, from which men and women sample with proportions above 0 and below 1. Equations 1 and 2 describe the mixed-types model: Letting N stand for normal density, and p and q stand for the proportion of men and women, respectively, sampling from the high-mean distribution, the density of men's features (X) and women's features (Y) is

$$(1)\ f_X(x,\Theta) = pN(x;\mu_1,\sigma_1) + (1-p)N(x;\mu_2,\sigma_2)$$

$$(2)\ f_Y(x,\Theta) = qN(x;\mu_1,\sigma_1) + (1-q)N(x;\mu_2,\sigma_2)$$

For each brain measure, the MLE of these parameters was estimated by the EM method. A log-likelihood ratio test was conducted to test the null hypothesis of pure-types versus the alternative hypothesis of mixed-types.

Is there a sex/gender difference in the probability of sampling from the two phenotypes?

Measures for which the pure-types hypothesis was rejected are those for which the observed data belong to a non-Gaussian distribution best described by a mixture of two Gaussians that are sampled by both women and men. To test whether men and women differ in their probabilities of sampling from the two Gaussians, a similar analysis was conducted, but this time the null hypothesis was that $p = q$, and the alternative hypothesis was that $p \neq q$. For brain features for which p was significantly different from q, the size of the difference was estimated using Cohen's h (Cohen, 1988).

Assessing sampling consistency across brain measures.

Correlation matrices between the EM responsibilities (i.e., the posterior probabilities of each individual to sample from a reference distribution given his/her sex category, Hastie et al., 2001) of two sets of measures were evaluated: I. for the 41 mixed-types measures for which $p \neq q$ and the female-favored distribution was different from the male-favored distribution (e.g., Fig. 1D), we assessed correlations between the posterior probabilities to select from the male-favored distribution; II. for all 282 mixed-types measures, we assessed correlations between the posterior probabilities to select from the high-mean distribution. In both sets, Pearson correlation coefficients were computed only between measures of the same type (i.e., separately for measures of volume, mean FA, weighted mean FA, mean MD, and weighted mean MD).

In all analyses, the p-value was computed using Wilks' theorem (van der Vaart, 1998; Wilks, 1938) and adjusted for FWER using the Benjamini-Hochberg correction (Benjamini and

Hochberg, 1995). Adjusted p-values smaller than 0.05 were considered statistically significant.

Predicting sex category on the basis of brain structure

Following Chekroud and colleagues (2016), a logistic model was fitted to the residuals of the 289 brain measures to predict the probability of an individual belonging to the male or female sex category. The model fitted was Elastic-net (Zou and Hastie, 2005) regression using R glmnet package (Hastie et al., 2001). The model was fitted using the 10-fold schema on 75% of the observations and then validated on the remaining 25%.

Code

The data in this work were analyzed using the R programming language (R Core Team, 2013). The equal probability EM was computed using the mixtools package (Benaglia et al., 2009). Cohen's d was estimated using the effsize package (Torchiano, 2020). The code for the paper is available at https://github.com/nitayalon/biobank_data_analysis.

RESULTS

Student's t-test analysis

Out of the 289 brain measures analyzed, there was a significant sex/gender difference in 225 measures. The size of the sex/gender differences ranged between -0.232 To +0.338 (Fig. 2A), with the average absolute size of the significant differences being 0.11.

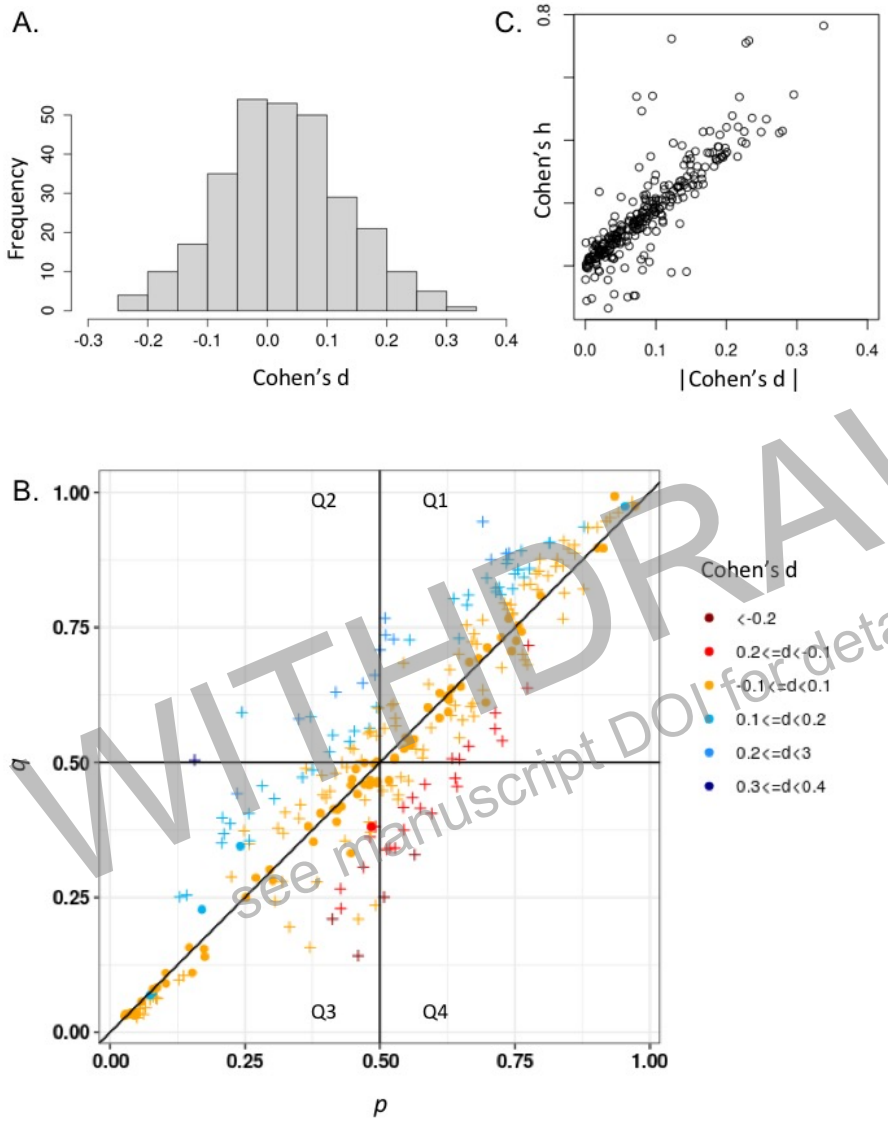

Figure 2. A. Sex/gender differences in the observed data. A histogram of the size (Cohen's d) of the sex/gender difference in the observed means for each of the 289 brain measures. **B. Probability of sampling from the high-mean distribution.** The probability that a man (*p,* X axis) and a woman (*q,* Y axis) will sample from the high-mean distribution of the 282 mixed-types measures. The color-code represents the effect size (Cohen's d) of the sex/gender difference in the observed data; Measures for which there was a significant sex/gender difference in the probability of sampling from the model-fitted underlying phenotypes are marked with a plus symbol. **C. Sex/gender differences in sampling probabilities and in the observed**

data. The size of the sex/gender difference in the observed means (|Cohen's d|, x axis) and in the sampling probabilities (Cohen's h, y axis) for each of the 282 brain measures.

Mixture analysis

Out of the 289 brain measures analyzed, in 282 the pure-types hypothesis was rejected – that is, women and men sampled from the two underlying distributions in positive proportions.

Brain measures better described by a pure-types model

The seven measures for which the pure-types hypothesis was not rejected are listed in Table 1. In four of these measures, the sex/gender difference in the observed data was trivial (|Cohen's d| < 0.04) and in three of these, also not statistically significant. It therefore seems safe to conclude that each of these four measures is best described as reflecting a single phenotype. In the remaining three pure-types measures – the volumes of: the anterior division of the left postcentral gyrus (Fig. 1A), the right cingulate gyrus, and the right planum polare – there was a (small) sex/gender difference in the observed data (0.17 < |Cohen's d| < 0.28), suggesting the existence of a female and a male phenotype.

Brain measures better described by a mixed-types model

Figure 2B displays for each of the remaining 282 brain measures the probability that a man (*p,* X axis) and a woman (*q,* Y axis) will sample from the model-fitted distribution with the higher mean. Figure 2C displays for each of these 282 measures, the size of the sex/gender difference in sampling probabilities (Cohen's h, Y axis) and in the observed means (|Cohen's d|, X axis), which were highly correlated ($r_{Pearson}$ = 0.76). In 84 of the measures, men and women sampled

from the two distributions with the same probabilities ($p = q$, e.g., Fig. 1B,E). In the remaining 198 measures (marked with a plus symbol in Fig. 2B), women and men sampled with significantly different probabilities ($p \neq q$, e.g., Fig. 1C,D). In 104 measures, the sex/gender difference in sampling probabilities was small (Cohen's h < 0.2), in 86 moderate (0.2 < Cohen's h < 0.5), and in eight large (0.5 < Cohen's h < 0.765; Fig. 1D presents the distributions of women and men for the volume of the left lateral occipital cortex, which showed the largest sex/gender difference in sampling probabilities (Cohen's h = 0.764) and in means in the observed data (Cohen's d = 0.338)). To further appreciate the magnitude of the sex/gender differences in sampling probability, we compared the likelihood that a woman and a man would sample from the same phenotype of a brain measure to the likelihood that two women or two men would sample from the same phenotype. The ratio between the first likelihood and the other two was, on average, 0.979 (range, 0.699-1.170). For comparison, for the 84 p=q measures, the corresponding ratio was, on average, 0.998 (range, 0.902-1.078).

Is sampling consistent across brain measures?

There were 41 mixed-types features for which $p \neq q$ and one phenotype was sampled mainly by women (female-favored phenotype) *and* the other mainly by men (male-favored phenotype). These brain measures are located at Quadrants 2 and 4 of the graph in Fig. 2B, and listed in Table 2. Figure 1D presents the distributions of women and men for one such measure, which showed the largest difference in sampling probabilities. Figures 3A-C present the correlations between the posterior probabilities that a measure was sampled from the male-favored phenotype, in women (lower triangle) and men (upper triangle), separately for each type of measure. Measures of volume (Fig. 3A) were largely uncorrelated, indicating that with respect to regional volume, brains are not consistent in sampling from the male-favored

phenotype. In contrast, the correlation coefficients between FA (Fig. 3B) and MD measures (Fig. 3C) occupied a wider range, with both positive and negative moderate correlations. Negative correlations reflect a situation in which sampling the male-favored phenotype in one region correlates with sampling the female-favored phenotype of another region. To better understand these unexpected negative correlations, we assessed the correlations in the same set of measures, but this time between the posterior probabilities to select from the high-mean distribution (Fig. 3D-F). This analysis resulted in positive correlations only (as was also the case in the correlations between the posterior probabilities to select from the high-mean distribution for all other measures of mean FA and mean MD; data not shown), indicating that value (high versus low) is more important than sex/gender category in explaining variability in FA and MD even for measures for which the majority of men sampled from one phenotype and the majority of women sampled from the other. This may also be true for the few moderate-to-large positive correlations between measures of volume (Fig. 3A), which are positive also in Fig. 3D – these correlations may reflect not the consistent effects of sex but rather the consistent effects of other factors. Indeed, the strongest correlations were found between homologous regions in the two hemispheres (the right and left hippocampus; the right and left cuneal cortex; and the right and left superior frontal gyrus).

The most remarkable observation over the images presented in Figure 3 is that the correlation matrices in men and women are almost identical (the numerical values of the correlations are given in Figure 4). This remarkable similarity is also evident when considering the correlations between all possible same-type pairs of the 282 mixed-types measures (Fig. 5A. Note that the

correlations here are between the posterior probabilities of each individual to select from the high-mean distribution, rather than from the male-favored distribution).

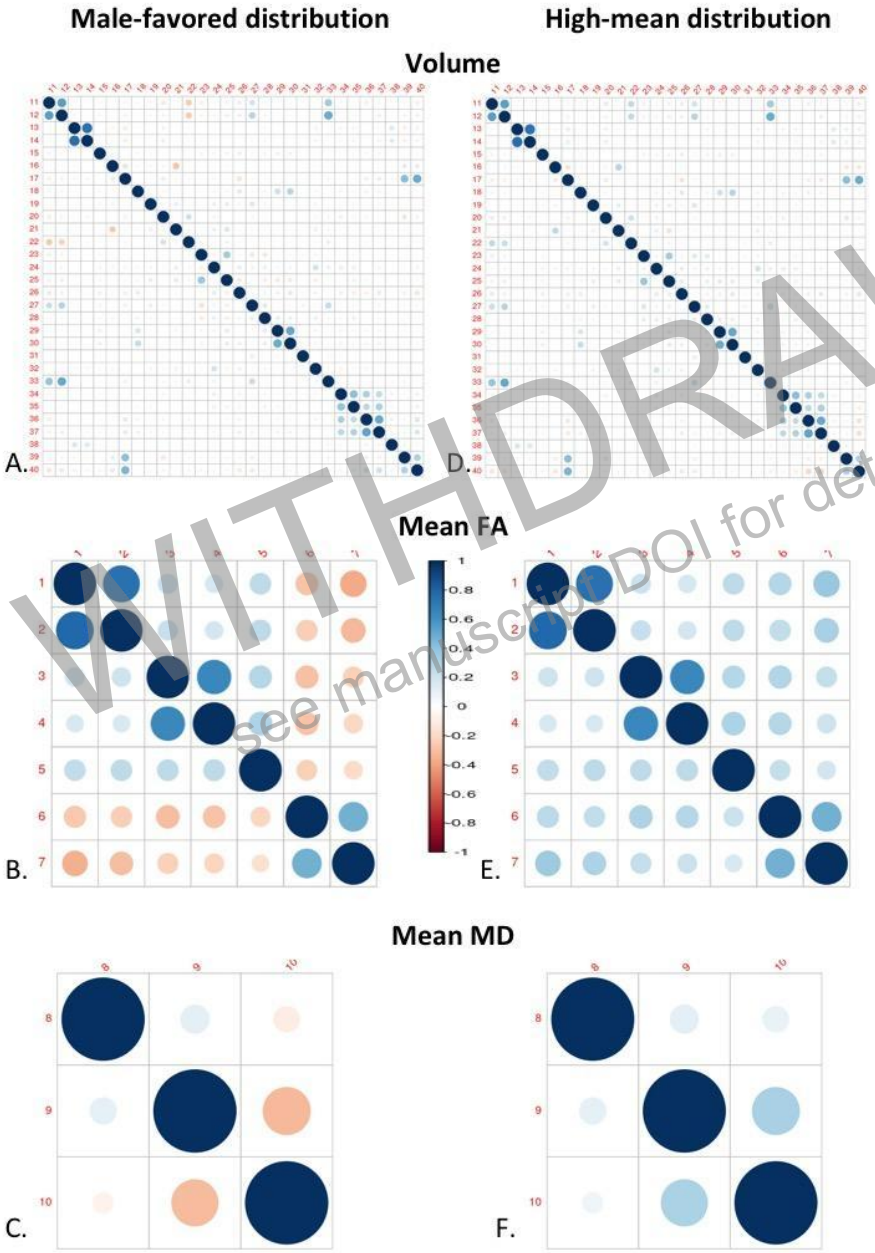

Figure 3. Assessing internal consistency. (A-C) The correlation coefficients in women (lower triangle) and men (upper triangle) between the posterior probabilities of each individual to select from the male-favored distribution in all possible pairs of the 41 measures that have a female-favored and a male-favored distribution, separately for three types of brain measures - (A) volume, (B) mean FA, and (C) mean MD. Correlation strength is represented using a red (-1) – white – blue (+1) color scale, and the absolute size is

also represented by the size of the dot. The numbers correspond to the number of each measure in Table 2. (D-F) Same as A-C but for the posterior probabilities of each individual to select from the high-mean distribution.

A. **Volume**

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
11	1	0.55	0.03	0.01	0.03	0.02	-0.07	0.05	0.04	-0.13	-0.01	-0.25	-0.09	0.05	-0.08	0.1	0.21	0.04	0.04	0.05	-0.02	0.03	0.35	-0.04	-0.01	-0.1	-0.06	0.03	-0.05	-0.09
12	0.54	1	0.04	0.02	0.04	0	-0.08	0.03	0.05	-0.08	0.01	-0.26	-0.06	0.06	-0.1	0.09	0.3	0.06	0.03	0.03	0	0.02	0.49	0	-0.04	-0.03	-0.1	0.03	-0.06	-0.09
13	0.05	0.04	1	0.73	0.01	-0.06	-0.11	-0.02	0.01	0.09	0.03	-0.03	-0.01	-0.03	0.03	0.02	-0.03	-0.07	-0.01	-0.02	0	-0.08	0.02	0.07	0.07	0.02	0.03	0.15	-0.1	-0.05
14	0.04	0.03	0.76	1	0	-0.05	-0.11	-0.01	0	0.1	0.02	-0.03	-0.03	-0.02	-0.01	0.01	-0.04	-0.04	-0.01	-0.02	0	-0.07	0.01	0.04	0.07	0.03	0.08	0.14	-0.09	-0.07
15	0.07	0.06	0.03	0.01	1	-0.01	-0.01	0.05	0.05	-0.01	0.01	-0.01	-0.07	0.06	-0.07	0.1	0.05	0.08	0.09	0.08	0.02	0.03	0	-0.03	-0.04	-0.05	-0.06	0	-0.01	-0.04
16	-0.02	-0.05	-0.05	-0.03	-0.02	1	0.16	0.05	-0.04	-0.07	-0.28	-0.02	0	-0.04	-0.04	-0.03	0.01	-0.03	0.02	0.03	0	-0.03	-0.01	-0.03	0.02	-0.01	0.01	-0.03	0.13	0.12
17	-0.09	-0.09	-0.06	-0.06	-0.02	0.11	1	-0.01	0.01	-0.11	-0.05	0.03	0.02	0.07	0.03	-0.16	-0.04	0.03	0	0	-0.02	0	-0.07	-0.02	0.03	0.1	0.11	-0.06	0.42	0.49
18	0.06	0.05	0	0	0.05	0.03	-0.01	1	0	-0.06	-0.02	0.01	-0.02	0.06	-0.07	0.14	0.05	0.07	0.21	0.29	0.01	0.06	0.01	-0.06	-0.06	-0.09	-0.07	0.02	-0.01	-0.04
19	0.01	0.02	0.03	0.01	0.04	-0.07	-0.01	-0.01	1	-0.05	0.11	-0.01	-0.01	0.1	0.12	0.06	0.16	0.06	0.02	0	-0.01	0.08	0.01	-0.03	-0.06	-0.03	-0.06	-0.01	-0.02	0.01
20	-0.11	-0.08	0.03	0.06	0	-0.04	-0.09	-0.04	-0.02	1	-0.03	0.21	0.02	-0.08	0.01	-0.03	-0.06	-0.02	0	0	-0.03	-0.05	-0.07	0.08	-0.05	0.05	0.05	0.01	-0.14	-0.08
21	-0.01	0.02	0.01	-0.01	0.02	-0.26	-0.04	-0.05	0.09	-0.01	1	0.05	0.02	0.1	0.17	0.03	0.04	0.02	-0.03	-0.03	-0.01	0.09	0.06	0	-0.01	-0.02	-0.02	0.04	-0.04	-0.03
22	-0.27	-0.23	-0.04	-0.04	-0.03	0	0.06	0.01	0	0.2	0.05	1	0.02	0.03	0.08	-0.04	-0.02	0.02	0	-0.01	-0.04	0.04	-0.11	0.01	-0.02	0.04	-0.03	0.02	0.02	0.04
23	-0.1	-0.08	-0.03	-0.03	-0.05	0.01	0.04	-0.03	0.02	0.04	0.04	0.04	1	-0.07	0.34	0.06	-0.14	-0.14	-0.04	-0.05	-0.05	-0.07	-0.01	0.03	0.05	-0.02	0	-0.03	0.03	0.05
24	0.05	0.07	0	-0.01	0.06	-0.07	0.06	0.05	0.06	-0.04	0.11	0.06	-0.08	1	0.02	0.12	0.07	0.09	0.06	0.06	0.01	0.2	0.09	-0.11	-0.1	0	-0.08	0.03	0.04	0.07
25	-0.09	-0.1	0.02	-0.01	-0.07	0.01	0.04	-0.07	0.11	0.03	0.13	0.06	0.38	0	1	-0.13	-0.04	-0.11	-0.13	-0.12	-0.03	-0.11	-0.03	0.02	0.08	-0.05	0.01	0	0.04	0.04
26	0.12	0.09	0.01	0.02	0.1	-0.04	-0.11	0.11	0.04	0.04	0.04	-0.02	-0.04	0.12	-0.11	1	0.02	0.07	0.09	0.13	0.05	0.08	0.02	-0.1	-0.09	-0.09	-0.13	0.03	-0.09	-0.15
27	0.24	0.3	-0.01	-0.03	0.07	-0.02	-0.09	-0.02	0.11	-0.05	0.04	-0.03	-0.17	0.06	-0.1	0.01	1	0.15	0	0.03	0.07	0.03	0.25	-0.03	-0.07	-0.01	-0.06	0.03	-0.07	-0.07
28	0.03	0.04	-0.03	-0.01	0.05	-0.06	-0.01	0.07	0.05	-0.03	0.03	-0.01	-0.12	0.1	-0.11	0.08	0.12	1	0.12	0.1	0.01	0.1	0.04	-0.05	-0.09	0	-0.04	-0.01	0	-0.01
29	0.04	0.03	0.02	0.02	0.04	0.01	-0.02	0.19	0.03	0.02	-0.06	0.01	-0.06	0.06	-0.13	0.03	0.02	0.11	1	0.51	0.02	0.07	-0.03	-0.07	-0.06	-0.06	-0.05	0.02	-0.02	-0.04
30	0.05	0.04	0.01	0	0.09	-0.01	-0.03	0.25	0.01	0.03	-0.05	0.01	-0.07	0.04	-0.13	0.09	0.03	0.12	0.5	1	0.03	0.06	-0.03	-0.07	-0.05	-0.07	-0.08	-0.02	-0.02	-0.06
31	0.01	0.01	0.01	0	0.04	-0.01	0	0.02	0.01	-0.04	0.02	-0.01	-0.03	0.02	-0.04	0	0.11	0.01	0.04	0.03	1	0.01	-0.06	-0.04	-0.03	-0.02	-0.01	-0.03	0	-0.03
32	0.03	0.03	-0.05	-0.07	0.03	-0.07	0	0.01	0.07	-0.06	0.09	0.04	-0.03	0.2	-0.07	0.08	0.03	0.05	0.03	0.03	0	1	0.05	-0.07	-0.04	-0.07	-0.02	0.03	-0.02	0
33	0.39	0.51	0.03	0.04	0.03	-0.03	-0.08	0.01	-0.02	-0.06	0.06	-0.12	-0.03	0.09	-0.04	0.03	0.24	0.02	-0.03	-0.03	-0.01	0.05	1	0.02	0.02	0.03	0.01	0.05	-0.07	-0.04
34	-0.07	-0.03	0.02	0	-0.03	-0.01	-0.01	-0.06	-0.04	0.08	0.01	0	0.02	-0.11	0.01	-0.08	-0.04	-0.04	-0.05	-0.05	-0.04	-0.07	-0.01	1	0.42	0.33	0.26	-0.08	-0.03	0.07
35	0	-0.04	0.01	0.03	-0.03	0.02	0.04	-0.06	-0.05	0.02	0.01	-0.04	0.03	-0.09	0.08	-0.07	-0.05	-0.08	-0.07	-0.06	-0.01	-0.03	0	0.4	1	0.28	0.33	-0.05	0.01	0.09
36	-0.11	-0.03	0.01	0.01	-0.04	-0.03	0.14	-0.07	-0.02	0.05	0.03	0.06	-0.03	0.02	-0.05	-0.08	-0.02	-0.01	-0.06	-0.04	-0.01	-0.07	0	0.32	0.25	1	0.5	-0.08	0.06	0.17
37	-0.07	-0.1	0.02	0.05	-0.05	0.03	0.12	-0.08	-0.06	0.04	0	-0.03	-0.01	-0.06	0.01	-0.1	-0.07	-0.03	-0.07	-0.08	-0.03	-0.01	-0.02	0.27	0.3	0.53	1	-0.06	0.06	0.17
38	0.02	0.03	0.17	0.16	0.04	-0.02	-0.04	0.02	-0.01	0.02	0.02	0.03	-0.03	0.03	0	0.01	0.04	0.01	0	-0.01	0.01	0.02	0.06	-0.05	-0.04	-0.08	-0.07	1	-0.07	-0.05
39	-0.06	-0.07	-0.06	-0.06	-0.02	0.11	0.41	-0.02	-0.02	-0.14	-0.04	0.04	0.05	0.04	0.05	-0.06	-0.09	-0.02	-0.02	-0.02	0.01	0	-0.08	-0.01	0.03	0.09	0.08	-0.04	1	0.32
40	-0.15	-0.13	-0.07	-0.06	-0.07	0.13	0.45	-0.02	-0.02	-0.06	-0.03	0.06	0.06	0.04	0.07	-0.13	-0.12	-0.04	-0.05	-0.06	-0.02	0	-0.09	0.06	0.09	0.21	0.21	-0.05	0.31	1

B. **Mean FA**

	1	2	3	4	5	6	7
1	1	0.75	0.21	0.18	0.25	-0.28	-0.38
2	0.78	1	0.22	0.18	0.25	-0.25	-0.33
3	0.21	0.21	1	0.65	0.29	-0.29	-0.24
4	0.17	0.16	0.64	1	0.31	-0.29	-0.21
5	0.25	0.25	0.25	0.26	1	-0.23	-0.19
6	-0.26	-0.25	-0.3	-0.28	-0.22	1	0.48
7	-0.35	-0.3	-0.24	-0.21	-0.17	0.48	1

C. **Mean MD**

	10	8	9
10	1	-0.06	-0.32
8	-0.1	1	0.12
9	-0.33	0.1	1

Figure 4. Assessing internal consistency. (A-C) The correlation coefficients in women (lower triangle) and men (upper triangle) between the posterior probabilities of each individual to select from the male-favored distribution in all possible pairs of the 41 measures that have a female-favorable and a male-favorable distribution, separately for three types of brain measures - (A) volume, (B) mean FA, and (C) mean MD.

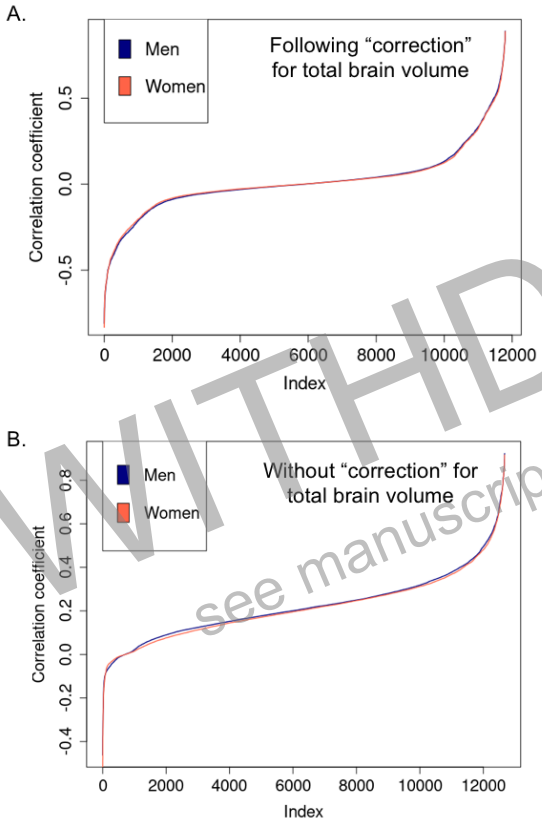

Figure 5. Correlation coefficients between the posterior probabilities to select from the high-mean distribution. (A) The correlation coefficients in women (red) and men (blue) between the posterior probabilities of each individual to select from the high-mean distribution in all possible same-type pairs of the 282 mixed-types brain measures. The correlations in men are sorted from lowest to highest, and the

correlations in women are presented in the men's order. (B) Same as (A), but for the 287 mixed-types "uncorrected" brain measures.

Predicting sex category on the basis of brain structure

Logistic regression over the 289 "corrected" brain measures accurately predicted the sex category of brains' owners in 75% of cases.

DISCUSSION

A conventional analysis (Student's t-test) of 289 MRI-derived measures of the brains of ~24,000 humans, yielded (after FDR correction for multiple comparisons) significant sex/gender differences in 225 measures. These results would have led to the conclusion that for these 225 brain measures, there is a female and a male phenotype, whereas for the remaining 64 measures, there is a single phenotype. A mixture analysis, which allows a third alternative – that women and men sample from the same two phenotypes, revealed that this is the best description for the vast majority of brain measures (282 out of 289). Of the remaining seven measures, three were best described as having a female and a male phenotype, and four as having a single phenotype.

A further mixture analysis of the 282 measures revealed that in 84 brain measures, women and men sampled from each phenotype with similar proportions, in 104 measures, the sex/gender difference in sampling probabilities was small (Cohen's $h < 0.2$), in 86 moderate ($0.2 <$ Cohen's $h < 0.5$), and in eight large ($0.5 <$ Cohen's $h < 0.765$). The overall small sex/gender differences in sampling probabilities would contradict also a "soft" version of the 'male and female phenotype hypothesis', if this existed, according to which the large

majority of men sample from one phenotype whereas the large majority of women sample from the other phenotype.

We would like to note that the results of the mixture analysis do not contradict evidence that sex-related genes and hormones affect specific brain measures or that there are sex/gender differences in the brain. Rather, they challenge the dominant assumption that these observations reflect the existence of a male and a female phenotype. Our results demonstrate that if brain measures are described as reflecting two underlying Gaussian-shaped phenotypes (rather than, for example, one non-Gaussian-shaped phenotype, or three Gaussian-shaped phenotypes), then for the vast majority of brain measures, women and men sample from both phenotypes, and for the most part do so with quite similar probabilities.

The present results also contradict another common assumption regarding sex effects on the brain – that these are consistent across brain measures within a single brain. In the present context this should be evident in consistently sampling the male-favored (or the female-favored) phenotype across the 41 brain measures for which such phenotypes exist (in the remaining 241 brain measures, the phenotype sampled by most women was also sampled by most men). For the relevant measures of volume, the correlations between the posterior probabilities of each individual to select from the male-favored distribution were mainly around zero, indicating that sampling from the male-favored phenotype for one region provided no information on whether the male-favored or the female-favored phenotype of another region was sampled. For the relevant MD and FA measures, some of these correlations were negative, suggesting that sampling from the male-favored phenotype for one region was associated with sampling from the female-favored phenotype for another region. Both patterns of correlations are consistent with our previous observations that human brains are most often comprised of a mosaic of female-typical and male-typical brain features

(Joel et al., 2015, 2020) and suggest the existence of factor(s) that are more important than sex category in explaining variability in brain measures.

The possibility that sex category is not a major predictor of variability in human brain structure is further supported by the almost identical correlations in women and men between the posterior probabilities to select from the high-mean distribution for all possible pairs of same-type brain measures (Fig. 5A). Very similar correlations in men and women were also observed when the same analysis was conducted without "correcting" for total brain volume (Fig. 5B, Appendix I). This remarkable similarity suggests that in spite of the large difference in total brain volume (Cohen's d = 1.36 in the present sample), the same principles are governing brain architecture in women and men.

The conclusion that sex category is not a major predictor of variability in human brain structure does not contradict evidence that supervised machine learning algorithms may use sex-related variability in brain structure to predict the sex category of a brain's owner (Chekroud et al., 2016; Del Giudice et al., 2016; Joel et al., 2016; Sanchis-Segura et al., 2020). Indeed, using one such approach (logistic regression, as in Chekroud et al., 2016) we accurately predicted the sex category of brains' owners in 75% of cases (the lower classification rate compared to previous studies (Chekroud et al., 2016; Del Giudice et al., 2016; Joel et al., 2016) is expected given that we used data "corrected" for total brain volume, Sanchis-Segura et al., 2020).

What our present and previous studies (Joel et al., 2015, 2018, 2020) challenge is the common assumption that sex-related effects consistently add up in individual brains so that the brains of women are meaningfully different from the brains of men (e.g., Chekroud et al., 2016; Del Giudice et al., 2016; Ingalhalikar et al., 2014). Different analytical approaches demonstrate that this is not the case. Thus, unsupervised machine learning algorithms applied

to the entire brain revealed that the brain architectures typical of women are also typical of men and vice versa; large sex/gender differences were found only in the prevalence of some rare brain architectures (Joel et al., 2018). Recent studies, which assessed the contribution of several factors to variability in brain function (measured using functional MRI), reported that sex/gender category explained only a small fraction of this variability (Mitricheva et al., 2019; Kersey et al., 2019). Finally, an assessment of the relations between the number of sex/gender differences in functional MRI studies and sample size did not reveal the positive correlation expected if brains of women and men were meaningfully different (David et al., 2018).

The present study has several limitations. The sample is quite ethnically homogeneous (92.2% Caucasians) and restricted age-wise (all participants are over 42 years old), limiting the generalizability of our conclusions across ethnicity and age. On the other hand, this relative homogeneity would have increased the chances of finding consistent sex effects, if these were present, as other studies have shown that sex/gender differences in brain structure may differ across age (e.g., Jancke et al., 2015; Lenroot and Giedd, 2010) and across countries differing in their ethnicity composition (e.g., Joel et al., 2015; Zilles et al., 2001). Another limitation of the present study is that we analyzed only MRI-derived brain measures, which show smaller sex/gender differences compared to some post mortem-derived measures (e.g., number of neurons in specific hypothalamic nuclei). It is unlikely, however, that a large enough dataset of the latter type of measures would be available to enable the analyses conducted here.

Conclusions

The decades old hypothesis that sex effects on the brain result in a female and a male phenotype for specific brain measures and for the brain as a whole governs the ways we analyze brain data by sex, and consequently, the type of answers we may receive. In particular, this assumption leads to framing questions regarding the relations between sex and the brain in terms of similarity and difference – are the brains of females and males the same or different? The present analysis revealed that for most (282 out of 289) MRI-derived measures of human brain structure the answer is – neither. These measures are better described by a third type of relations between sex and the brain, namely, that both women and men may manifest each of the two phenotypes of a brain measure. This description is consistent with other lines of evidence suggesting that sex-related variables are a part of a large set of factors that interact to create a highly heterogenous population of human brains, and that sex/gender category explains a very small part of this variability (for a recent review see Joel, 2021). There is therefore a need to develop new methods for studying the human brain and its relations with sex-related variables that go beyond the common practice of comparing a group of females to a group of males.

REFERENCES

Alfaro-Almagro F et al. (2018) Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166:400-424.

Benaglia T, Chauveau D, Hunter DR, Young D (2009) mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software 32.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. . Journal of the Royal Statistical Society: Series B (Methodological) 57:289-300.

Bishop C (2006) Pattern Recognition and Machine Learning: Springer.

Chang ML (2010) Shortchanged: why women have less wealth and what can be done about it. Oxford ; New York: Oxford University Press.

Chang ML (2010) Shortchanged: why women have less wealth and what can be done about it. Oxford ; New York: Oxford University Press.

Chekroud AM, Ward EJ, Rosenberg MD, Holmes AJ (2016) Patterns in the human brain mosaic discriminate males from females. P Natl Acad Sci USA 113:E1968-E1968.

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. , 2nd Edition. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Crenshaw K (1989) Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum:Article-8.

Crenshaw K (1989) Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum:Article-8.

David SP, Naudet F, Laude J, Radua J, Fusar-Poli P, Chu I, Stefanick ML, Ioannidis JPA (2018) Potential Reporting Bias in Neuroimaging Studies of Sex Differences. Sci Rep 8:6082.

de Vries GJ, Södersten P (2009) Sex differences in the brain: The relation between structure and function. Hormones and Behavior 55:589-596.

de Vries GJ, Södersten P (2009) Sex differences in the brain: The relation between structure and function. Hormones and Behavior 55:589-596.

Del Giudice M, Booth T, Irwing P (2012) The Distance Between Mars and Venus: Measuring Global Sex Differences in Personality. PLoS ONE 7:e29265.

Del Giudice M, Booth T, Irwing P (2012) The Distance Between Mars and Venus: Measuring Global Sex Differences in Personality. PLoS ONE 7:e29265.

Del Giudice M, Lippa RA, Puts DA, Bailey DH, Bailey JM, Schmitt DP (2016) Joel et al.'s method systematically fails to detect large, consistent sex differences. P Natl Acad Sci USA 113:E1965-E1965.

Dempster AP, Laird NM, Rubin DB (1997) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39:1–38.

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. . The Annals of Statistics 7:1-26.

Eliot L (2020) Sex/gender differences in the brain and their relationship to behavior. In: Cambridge International Handbook on Psychology of Women. (Cheung F, Halpern D, eds). Cambridge, UK: Cambridge University Press.

Fitzgerald LF, Cortina LM (2018) Sexual Harassment in Work Organizations: A View from the Twenty-First Century. In: Handbook on the Psychology of Women (White JW, Travis C, eds), p 40. Washington, D.C.: American Psychological Association.

Fitzgerald LF, Cortina LM (2018) Sexual Harassment in Work Organizations: A View from the Twenty-First Century. In: Handbook on the Psychology of Women (White JW, Travis C, eds), p 40. Washington, D.C.: American Psychological Association.

Fjell AM, Westlye LT, Greve DN, Fischl B, Benner T, van der Kouwe AJ, Salat D, Bjornerud A, Due-Tonnessen P, Walhovd KB (2008) The relationship between diffusion tensor imaging and volumetry as measures of white matter properties. Neuroimage 42:1654-1668.

Griffiths PE, The Hegeler I (2002) What Is Innateness? Monist 85:70-85.

Griffiths PE, The Hegeler I (2002) What Is Innateness? Monist 85:70-85.

Harper J, O'Donnell E, Sorouri Khorashad B, McDermott H, Witcomb GL (2021) How does hormone transition in transgender women change body composition, muscle strength and haemoglobin? Systematic review with a focus on the implications for sport participation. Br J Sports Med:bjsports-2020-103106.

Harper J, O'Donnell E, Sorouri Khorashad B, McDermott H, Witcomb GL (2021) How does hormone transition in transgender women change body composition, muscle strength and haemoglobin? Systematic review with a focus on the implications for sport participation. Br J Sports Med:bjsports-2020-103106.

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2nd Edition: Springer.

Hess C (2020) Providing Unpaid Household and Care Work in the United States: Uncovering Inequality. In: Institute for Women's Policy Research.

Hess C (2020) Providing Unpaid Household and Care Work in the United States: Uncovering Inequality. In: Institute for Women's Policy Research.

Hilton EN, Lundberg TR (2021) Transgender Women in the Female Category of Sport: Perspectives on Testosterone Suppression and Performance Advantage. Sports Med 51:199-214.

Hilton EN, Lundberg TR (2021) Transgender Women in the Female Category of Sport: Perspectives on Testosterone Suppression and Performance Advantage. Sports Med 51:199-214.

Ingalhalikar M, Smith A, Parker D, Satterthwaite TD, Elliott MA, Ruparel K, Hakonarson H, Gur RE, Gur RC, Verma R (2014) Sex differences in the structural connectome of the human brain. Proc Natl Acad Sci U S A 111:823-828.

Jancke L, Merillat S, Liem F, Hanggi J (2015) Brain Size, Sex, and the Aging Brain. Hum Brain Mapp 36:150-169.

Joel D (2011) Male or Female? Brains are Intersex. Front Integr Neurosci 5:57.

Joel D (2012) Genetic-gonadal-genitals sex (3G-sex) and the misconception of brain and gender, or, why 3G-males and 3G-females have intersex brain and intersex gender. Biol Sex Differ 3:27.

Joel D (2020) Beyond sex differences and a male-female continuum: Mosaic brains in a multidimensional space. In: Handbook of Clinical Neurology (Lanzenberger R, Kranz GS, Savic I, eds). Amsterdam: Elsvier.

Joel D (2021) Beyond the binary: Rethinking sex and the brain. Neurosci Biobehav Rev 122:165-175.

Joel D, Berman Z, Tavor I, Wexler N, Gaber O, Stein Y, Shefi N, Pool J, Urchs S, Margulies DS, Liem F, Hanggi J, Jancke L, Assaf Y (2015) Sex beyond the genitalia: The human brain mosaic. P Natl Acad Sci USA 112:15468-15473.

Joel D, Fausto-Sterling A (2016) Beyond sex differences: new approaches for thinking about variation in brain structure and function. Philos T R Soc B 371.

Joel D, Garcia-Falgueras A, Swaab D (2020) The Complex Relationships between Sex and the Brain. Neuroscientist:1073858419867298.

Joel D, McCarthy MM (2017) Incorporating Sex As a Biological Variable in Neuropsychiatric Research: Where Are We Now and Where Should We Be? Neuropsychopharmacology 42:379-385.

Joel D, Persico A, Hanggi J, Pool J, Berman Z (2016) Reply to Del Giudice Et Al., Chekroud Et Al., and Rosenblatt: Do Brains of Females and Males Belong to Two Distinct Populations? P Natl Acad Sci USA 113:E1969-E1970.

Joel D, Persico A, Salhov M, Berman Z, Oligschlager S, Meilijson I, Averbuch A (2018) Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa. Front Hum Neurosci 12:399.

Jones L (2019) Women's Progression in the Workplace. In. King's College London: Global Institute for Women's Leadership.

Jones L (2019) Women's Progression in the Workplace. In. King's College London: Global Institute for Women's Leadership.

Jordan-Young RM (2011) Brain storm: the flaws in the science of sex differences. Cambridge, Mass.: Harvard Univ. Press.

Jordan-Young RM (2011) Brain storm: the flaws in the science of sex differences. Cambridge, Mass.: Harvard Univ. Press.

Juraska JM, Fitch JM, Henderson C, Rivers N (1985) Sex differences in the dendritic branching of dentate granule cells following differential experience. Brain Res 333:73-80.

Kersey AJ, Csumitta KD, Cantlon JF (2019) Gender similarities in the brain during mathematics development. NPJ Sci Learn 4:19.

Lenroot RK, Giedd JN (2010) Sex differences in the adolescent brain. Brain Cogn 72:46-55.

Liu D, Johnson HJ, Long JD, Magnotta VA, Paulsen JS (2014) The power-proportion method for intracranial volume correction in volumetric imaging analysis. Front Neurosci 8:356.

McCarthy MM (2020) Origins of Sex Differentiation of Brain and Behavior. In: Developmental Neuroendocrinology (Wray S, Blackshaw S, eds), pp 393-412: Springer Nature Switzerland.

McCarthy MM, Arnold AP (2011) Reframing sexual differentiation of the brain. Nat Neurosci 14:677-683.

Meyerowitz JJ (2002) How Sex Changed: A History of Transsexuality in the United States. Cambridge, Massachusetts: Harvard University Press.

Meyerowitz JJ (2002) How Sex Changed: A History of Transsexuality in the United States. Cambridge, Massachusetts: Harvard University Press.

Miller KL et al. (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523-1536.

Mitricheva E, Kimura R, Logothetis NK, Noori HR (2019) Neural substrates of sexual arousal are not sex dependent. Proc Natl Acad Sci U S A 116:15671-15676.

Prager EM (2016) Addressing sex as a biological variable. Journal of Neuroscience Research 95:11

Reich CG, Taylor ME, McCarthy MM (2009) Differential effects of chronic unpredictable stress on hippocampal CB1 receptors in male and female rats. Behav Brain Res 203:264-269.

Rhode DL (2017) Women and leadership. New York: Oxford University Press.

Rhode DL (2017) Women and leadership. New York: Oxford University Press.

Rippon G, Jordan-Young R, Kaiser A, Fine C (2014) Recommendations for sex/gender neuroimaging research: key principles and implications for research design, analysis, and interpretation. Front Hum Neurosci 8.

Rippon G, Jordan-Young R, Kaiser A, Fine C (2014) Recommendations for sex/gender neuroimaging research: key principles and implications for research design, analysis, and interpretation. Front Hum Neurosci 8.

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 (2018) Sex Differences in the Adult Human Brain: Evidence from 5216 UK Biobank Participants. Cereb Cortex 28:2959-2975.

Rudman LA, Glick P (2010) The social psychology of gender: how power and intimacy shape gender relations, Paperback ed Edition. New York: Guilford.

Rudman LA, Glick P (2010) The social psychology of gender: how power and intimacy shape gender relations, Paperback ed Edition. New York: Guilford.

Sanchis-Segura C, Ibanez-Gual MV, Adrian-Ventura J, Aguirre N, Gomez-Cruz AJ, Avila C, Forn C (2019) Sex differences in gray matter volume: how many and how large are they really? Biol Sex Differ 10:32.

Sanchis-Segura C, Ibanez-Gual MV, Aguirre N, Gomez-Cruz AJ, Forn C (2020) Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. Sci Rep 10:12953.

Stock K (2018) Changing the concept of "woman" will cause unintended harms. In: The Economist, p 8.

Stock K (2018) Changing the concept of "woman" will cause unintended harms. In: The Economist, p 8.

Takao H, Hayashi N, Inano S, Ohtomo K (2011) Effect of head size on diffusion tensor imaging. Neuroimage 57:958-967.

Tjaden P, Thoennes N (1998) Prevalence, incidence and consequences of violence against women: Findings from the National Violence Against Women Survey. In: U.S. Department of Justice, National Institute of Justice.

Tjaden P, Thoennes N (1998) Prevalence, incidence and consequences of violence against women: Findings from the National Violence Against Women Survey. In: U.S. Department of Justice, National Institute of Justice.

van der Vaart AW (1998) Asymptotic Statistics. Cambridge: Cambridge University Press.

Vos SB, Jones DK, Viergever MA, Leemans A (2011) Partial volume effect as a hidden covariate in DTI analyses. Neuroimage 55:1566-1576.

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. . Ann Math Statist 9:60- 62.

Zilles K, Kawashima R, Dabringhaus A, Fukuda H, Schormann T (2001) Hemispheric shape of European and Japanese brains: 3-D MRI analysis of intersubject variability, ethnical, and gender differences. Neuroimage 13:262-271.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67:301-320.

Table 1. List of measures better described by a pure-types model

Brain measure	Cohen's d	Adjusted p value
Weighted-mean MD in tract corticospinal tract (right)	0.0002	0.989
Insular Cortex (left)	0.013	0.442
Mean MD in cingulum hippocampus on FA skeleton (left)	0.030	0.069
Insular Cortex (right)	-0.037	0.017
Cingulate Gyrus, anterior division (right)	0.180	0.000
Planum Polare (right)	0.212	0.000
Postcentral Gyrus (left)	0.273	0.000

Cohen's d and p value of the sex/gender difference in the means in the observed data

Table 2. List of mixed-types measures with a female-favored and a male-favored phenotype (Quadrants 2 & 4 in Fig. 2B)

Region number	Region name	*p*	*q*	Cohen's d	Cohen's h
Volume of grey matter					
11	Cuneal Cortex (left)	0.407	0.520	0.119	0.227
12	Cuneal Cortex (right)	0.478	0.553	0.078	0.15
13	Hippocampus (left)	0.533	0.431	-0.068	0.204
14	Hippocampus (right)	0.522	0.457	-0.051	0.13
15	Inferior Frontal Gyrus, pars opercularis (right)	0.445	0.527	0.097	0.164
16	Inferior Temporal Gyrus, temporooccipital part (left)	0.539	0.472	-0.067	0.134
17	IX Cerebellum (right)	0.500	0.708	0.279	0.429
18	Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) (right)	0.466	0.504	0.049	0.076
19	Lateral Occipital Cortex, superior division (left)	0.157	0.504	0.338	0.764
20	Lingual Gyrus (left)	0.519	0.341	-0.173	0.362
21	Middle Temporal Gyrus, temporooccipital part (left)	0.413	0.551	0.165	0.277
22	Occipital Pole (right)	0.643	0.455	-0.188	0.38

23	Parietal Operculum Cortex (right)	0.534	0.440	-0.084	0.188
24	Planum Polare (left)	0.418	0.630	0.225	0.428
25	Planum Temporale (left)	0.528	0.341	-0.189	0.38
26	Precentral Gyrus (left)	0.497	0.600	0.096	0.207
27	Precuneous Cortex (right)	0.372	0.584	0.177	0.428
28	Subcallosal Cortex (left)	0.494	0.550	0.049	0.112
29	Superior Frontal Gyrus (left)	0.444	0.538	0.119	0.188
30	Superior Frontal Gyrus (right)	0.483	0.557	0.087	0.148
31	Superior Parietal Lobule (left)	0.481	0.563	0.107	0.164
32	Superior Temporal Gyrus, posterior division (right)	0.451	0.524	0.074	0.146
33	Supracalcarine Cortex (right)	0.430	0.500	0.086	0.14
34	Temporal Fusiform Cortex, anterior division (left)	0.575	0.415	-0.140	0.321
35	Temporal Fusiform Cortex, anterior division (right)	0.508	0.251	-0.218	0.537
36	Temporal Pole (left)	0.564	0.329	-0.208	0.477
37	Temporal Pole (right)	0.513	0.339	-0.148	0.354
38	Thalamus (right)	0.596	0.405	-0.137	0.384
39	Vermis VIIb Cerebellum	0.491	0.662	0.216	0.348
40	Vermis X Cerebellum	0.350	0.580	0.257	0.465

41	Weighted-tract cingulate gyrus part of cingulum (left)	0.553	0.466	-0.095	0.174
Mean FA					
1	Cingulum cingulate gyrus on FA skeleton (left)	0.583	0.459	-0.130	0.249
2	Cingulum cingulate gyrus on FA skeleton (right)	0.560	0.435	-0.134	0.251
3	Corticospinal tract on FA skeleton (left)	0.543	0.417	-0.131	0.253
4	Corticospinal tract on FA skeleton (right)	0.544	0.375	-0.191	0.341
5	Medial lemniscus on FA skeleton (right)	0.639	0.471	-0.186	0.34
6	Posterior limb of internal capsule on FA skeleton (left)	0.494	0.602	0.139	0.217
7	Retrolenticular part of internal capsule on FA skeleton (left)	0.468	0.647	0.201	0.362
Mean MD					
8	Pontine crossing tract on FA skeleton	0.453	0.558	0.131	0.21
9	Posterior limb of internal capsule on FA skeleton (right)	0.244	0.593	0.122	0.725
10	Uncinate fasciculus on FA skeleton (left)	0.505	0.448	-0.072	0.114

P and *q* are the probabilities of sampling from the high-mean distribution; Cohen's d of the sex/gender difference in the observed data; Cohen's h of the difference between *p* and *q*.

Appendix I

The suitability of the assumption that the underlying distributions are Gaussian

The MLE method yields essentially the same answers under a much broader class of exponential-type distributions

Consider an exponential-type family of distributions, with density or probability function of the form $f(x;\theta)=h(x)\exp(\theta x–b(\theta))$. Examples of exponential-type families are Gaussian with fixed variance, Gamma with fixed shape parameter, Negative-binomial, Poisson, Binomial. It is well known that the function $b(\theta)$ is convex, its first derivative with respect to θ is E[X] and its second derivative is Var[X].

The mixture model sets the densities of the feature values for men and women as $g_0(x;\theta) = (1-p)f(x;\theta_0) + pf(x;\theta_1)$ and $g_1(x;\theta) = (1-q)f(x;\theta_0) + qf(x;\theta_1)$. These functions can be expressed as the product $g(x) = h(x)[(1-r)*\exp(\theta_0 x-b(\theta_0)) + r*\exp(\theta_1 x-b(\theta_1))]$ of a function h free of θ and a sigmoidal function. Thus, the function h plays no role in the maximization of the likelihood function, in essence the same sigmoidal function for all exponential-type families. The Gaussian representative is convenient, for which the parameters are clear-cut to interpret.

As an example, consider the Gamma distribution with density $\lambda^{\gamma}x^{\gamma-1}\exp(-\lambda x)/\Gamma(\gamma)$ for fixed shape parameter γ. If a Gamma-distributed variable X with scale parameter λ_0 is normalized as $X=\gamma/\lambda_0 – (\sqrt{\gamma}/\lambda_0)Z$, the density of Z (with shape parameter λ) becomes proportional to

$\exp\{[\sqrt{\gamma}\,(\lambda/\lambda_0 -1)]\, z\} - \gamma[(\lambda/\lambda_0 -1) – \log(\lambda/\lambda_0)]\}$, that is close to the Gaussian

$\exp\{[\sqrt{\gamma}\,(\lambda/\lambda_0 -1)]\, z\} - [\sqrt{\gamma}\,(\lambda/\lambda_0 -1)]^2/2\} = \exp\{\theta z-\theta^2/2\}$ as long as $(\lambda/\lambda_0 -1)$ and $\sqrt{\gamma}\,(\lambda/\lambda_0 -1)$ are small enough, and not only for the large values of γ that would make the Gamma distribution close to Gaussian.

The fit between the observed data and Gaussian distributions

As is evident in Figure 2B, for some brain measures (located in the upper and lower extremes of the main diagonal in Fig. 2B), the two distributions differ considerably in the proportion of participants sampling from each (i.e., the large majority of participants sampled from one distribution). This occurs when the observed data belong to an asymmetric distribution, which, in the context of the method applied, is best described by a mixture of two Gaussians - one accounts for most of the observations and the other for the fatter tail (e.g., Fig. 1C, 1E). This is supported by the observation that the Kolmogorov-Smirnov distances between the empirical and the fitted data for these brain measures (dotted line in Fig. 6C) are larger than the Kolmogorov-Smirnov distances expected for symmetric distributions (solid line in Fig. 6C). (Women and men were similarly likely to sample from the "tail" distribution in 37% of these 'asymmetric' measures (e.g., Fig. 1E), men were more likely than women to sample from the "tail" distribution in 37% of these measures (e.g., Fig. 1C), and women were more likely to sample from the "tail" distribution than men, in 26% of these measures.)

In contrast to the deviation from symmetry for these brain measures, the Kolmogorov-Smirnov distances for the brain measures located in the main part of the main diagonal in Fig. 2B (dashed line in Fig. 6C) were not larger than expected for symmetric distributions. This observation is particularly important in the context of the relations between the underlying phenotypes and sex category. This is because this group of brain measures includes the 41 measures for which one phenotype was sampled mainly by men and the other mainly by women. The observation that these measures may be appropriately descried by Gaussian distributions strengthens the conclusions derived from our analysis of these features.

Power computation

For a given significance level (such as 0.05), it is possible to compute the power (complementary Type-II error probability) of the likelihood ratio test. Wilks' theorem (van der Vaart, 1998; Wilks, 1938), claims that twice the log-likelihood ratio is distributed under the null hypothesis according to the χ^2 distribution with properly defined degrees of freedom. The null hypothesis data for the simulation was generated from a standard Gaussian. The solid line in Figure 6B is the almost identical superposition of the χ^2 distribution and the empirical distribution under each of three applied sample sizes (each based on 10,000 simulation runs). The alternative hypothesis data were generated using the estimated parameters of a feature with a small sex/gender difference (Cohen's d in the observed data = 0.1, μ_1 = -0.328, σ_1 = 0.93, μ_2= 0.732, σ_2= 0.722, p = 0.742, q = 0.645), under three choices of sample size.

The vertical dotted line in Figure 6B marks the rejection threshold for $\alpha = 0.05$ (the vertical distance between 1 and the solid line). The vertical distances from 1 to the three dashed lines (0.25, 0.5, and 0.7) are the power of the test for sample sizes 1,000, 5,000, and 8,000, respectively. Thus, the sample sizes of the UK Biobank data (12,466 women and 11,469 men) are powerful enough to distinguish between pure-types and mixed-types.

Analysis of the "uncorrected data"

We conducted the "pure-types or mixed-types" analysis on the log of each of the 289 MRI-derived brain measures. Because a few scores on some of the brain measures were very low, preventing the conduction of the EM analysis, we left out from this analysis all scores below or above 5 standard deviations from the mean (the median number of scores left out per brain measure was 1, the average was 4.875, and the 75 percentile was 5.75). For 287 brain

measures, the pure-types hypothesis was rejected in favor of the mixed-types hypothesis. Next, we calculated for all scores (including those left out in the EM stage) of these 287 mixed-types measures, the posterior probability that the score was sampled from the high-mean distribution. Finally, Pearson correlation coefficients were computed for all possible pairs of same-type measures, separately for women and for men.

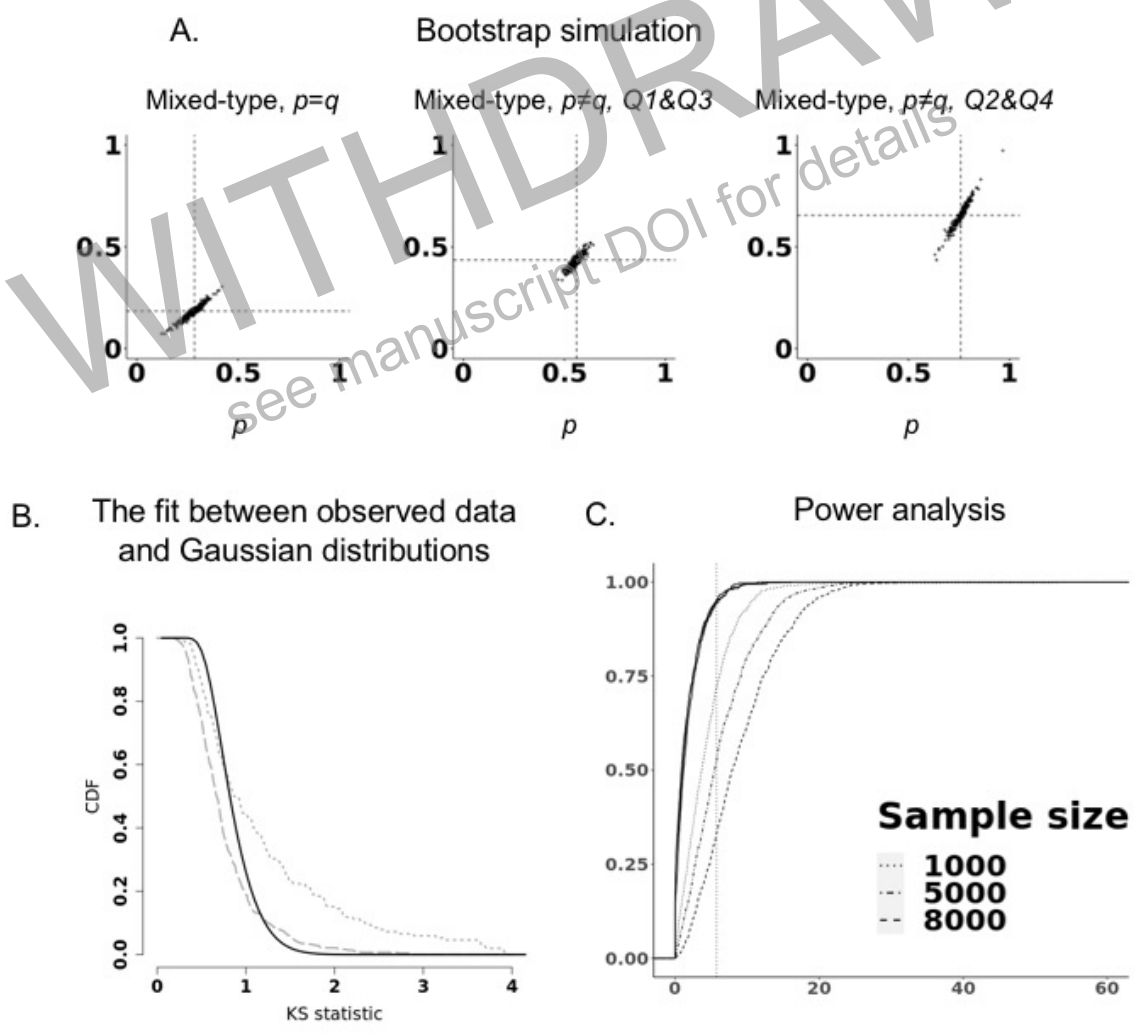

Figure 6. A. Results of the bootstrap simulation. A scattergram of the bootstrap estimated p and q for three mixed-types measures: Left: $p = q$, Middle: $p \neq q$, located at Q1 in Fig. 3, right: $p \neq q$, located at Q4 in Fig. 3. The dashed lines mark the value of the p and q of that brain measure in the reported analysis.

B. The fit between the observed data and Gaussian distributions. Kolmogorov-Smirnov distance for all the brain measures for which *p* and *q* are both larger than 0.75 or smaller than 0.25 (dotted line), for all other brain measures (dashed line), and the distances expected for symmetric distributions (solid line). **C. Power analysis of various sample sizes.** Thick lines are the CDF of the log-likelihood ratio under the null hypothesis (Chi square with 2 degrees of freedom), the dashed lines are the simulated CDF of the log-likelihood ratio under the alternative hypothesis. The vertical line marks the rejection threshold.