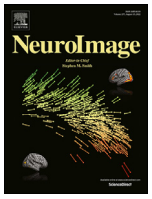
NeuroImage 257 (2022) 119343

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Beyond "sex prediction": Estimating and interpreting multivariate sex differences and similarities in the brain

Carla Sanchis-Segura*, Naiara Aguirre, Álvaro Javier Cruz-Gómez, Sonia Félix, Cristina Forn

Departament de Psicologia Bàsica, Clínica i Psicobiologia, Universitat Jaume I, Avda. Sos Baynat, SN., Castelló 12071, Spain

ARTICLE INFO

Keywords:
Sex differences
Sex similarities
MRI
Machine learning
Effect size
Gray matter
TIV-adjustment
Robust statistics

ABSTRACT

Previous studies have shown that machine-learning (ML) algorithms can "predict" sex based on brain anatomical/functional features. The high classification accuracy achieved by ML algorithms is often interpreted as revealing large differences between the brains of males and females and as confirming the existence of "male/female brains". However, classification and estimation are different concepts, and using classification metrics as surrogate estimates of between-group differences may result in major statistical and interpretative distortions. The present study avoids these distortions and provides a novel and detailed assessment of multivariate sex differences in gray matter volume (GMVOL) that does not rely on classification metrics. Moreover, appropriate regression methods were used to identify the brain areas that contribute the most to these multivariate differences, and clustering techniques and analyses of similarities (ANOSIM) were employed to empirically assess whether they assemble into two sex-typical profiles. Results revealed that multivariate sex differences in GMVOL: (1) are "large" if not adjusted for total intracranial volume (TIV) variation, but "small" when controlling for this variable; (2) differ in size between individuals and also depends on the ML algorithm used for their calculation (3) do not stem from two sex-typical profiles, and so describing them in terms of "male/female brains" is misleading.

1. Introduction

The study of sex differences in the brain is rather unique because it arouses great interest and heated debates within and outside the scientific realm (Maney, 2015; O'Connor and Joffe, 2014). In this regard, some researchers argue that differences between females and males are so substantial and widespread that brains can be considered "sexually dimorphic" or "male/ female brains" (e.g., Cahill 2014, 2006, Ingalhalikar et al. 2014, Lombardo et al. 2012). Conversely, other researchers think males and females are more similar than different in most, if not all, brain features, and that the existing sex differences lack the necessary internal consistency to constitute two types of brains, one typical of males and other typical of females (e.g., Eliot et al. 2021, Joel 2021, 2011, Rippon et al. 2014). However, most, if not all, scientists agree that sex differences in the brain exist, and that the use of neuroimaging techniques to investigate them will help to understand their possible differences in behaviorally relevant domains (e.g., when trying to know why the prevalence, course, and prognosis of many neurodevelopmental, psychiatric, and neurological disorders differ between females and males; Clayton, 2018; McCarthy, 2015; Pinares-Garcia et al., 2018).

Machine-learning (ML) offers new and informatively rich methods for multivariate exploration of the increasingly large and complex datasets from human neuroimaging studies (Bzdok, 2017). In the sex differences field, most ML applications have focused on classification tasks. These studies have effectively shown that ML algorithms can exploit anatomical and/ or functional features to ascertain whether a brain belongs to a male or to a female with an ≈80–90% accuracy (Anderson et al., 2018; Chekroud et al., 2016; Feis et al., 2013; Joel et al., 2018a; Luo et al., 2019; Rosenblatt, 2016; Sanchis-Segura et al., 2020; Sepehrband et al., 2018; Van Putten et al., 2018; Wang et al., 2012; Xin et al., 2019; Zhang et al., 2021, 2018). Because this kind of "prediction" only informs us about what is already known (to which sex category each particular sampled brain belongs), the in-

Abbreviations: %CC, percent of correctly classified cases; AAL, automated anatomical labeling; ANOSIM, analysis of similarities; CDF, cumulative distribution function; CI, confidence interval; CSF, cerebrospinal fluid; GM, gray matter; GMVOL, gray matter volume; KDE, kernel density estimate; ICC, intraclass correlation coefficient; IQR, interquartile range; LDA, linear discriminant analysis; LR, logistic regression; MARS, multiple adaptive regression splines; ML, machine learning; PCAM, probability of being classified as male; PCP, power-corrected proportions; POMP, percentage of the maximum possible; PS, probability of superiority; RF, random forest; RM, ranking mean; RRP, rank of ranks' products; SVM, support vector machine; TIV, total intracranial volume; VOI, volume of interest; WCC, within-class consistency; WM, white matter.

* Corresponding author.
E-mail address: csanchis@uji.es (C. Sanchis-Segura).

https://doi.org/10.1016/j.neuroimage.2022.119343.
Received 15 February 2022; Received in revised form 26 May 2022; Accepted 29 May 2022
Available online 30 May 2022.

terest and relevance of these findings resides more in the brain's ability to be classified than in the classifications obtained. Thus, classification is rarely the true goal of these studies, and the metrics obtained (i.e., the percentage of properly classified cases,%CC) are employed to indirectly estimate the degree of statistical distinctiveness and/ or separateness of the brains of females and males at the multivariate level. In this regard, a common interpretation of sex classification studies is that, because many and distinct ML algorithms are able to very accurately identify sex from brain features, all these algorithms *must be* identifying a reproducible constellation of brain differences that assemble into two clearly distinguishable and sex-specific brain types ("male/ female brains"; e.g., Chekroud et al. 2016, Luo et al. 2019, Rosenblatt 2016, Sepehrband et al. 2018, Wang et al. 2012).

However, as an approach to assess differences and similarities in the brains of females and males, classification and its associated metrics present several statistical and conceptual shortcomings. Most of these weaknesses arise from the fact that binary classification requires dichotomizing a continuous scoring output provided by ML algorithms. In this regard, the dichotomization of a continuous variable is rarely justified from either a conceptual or statistical perspective (Altman and Royston, 2006; Cohen, 1983; Harrell, 2015; MacCallum et al., 2002), given that it is well known that it results in a "*loss of information about individual differences as well as havoc with regard to estimation and interpretation of relationships among variables*" (MacCallum et al., 2002, pages 19–20). In a similar vein, "*classification through forced up-front dichotomization in an attempt to simplify the problem results in arbitrariness and major information loss*" (Harrell, 2015,page 4).

More specifically, as an approach to assess differences, classification through dichotomization is problematic because it requires predefining and applying a threshold from which all the cases above it are assigned to a category and all the cases below it are assigned to the alternative category. Thus, even when using a non-arbitrary threshold, classification through dichotomization results in arbitrariness because individuals who are close to the cutoff but on opposite sides it are treated as if they were totally different (when they are quite similar), whereas all the individuals located on each side of the threshold are treated as if they were identical (when they are not) (Altman and Royston, 2006; Harrell, 2015; MacCallum et al., 2002). This removes all the information about individual differences within each category, but also about the between-categories separation (Cohen, 1983; Harrell, 2015; MacCallum et al., 2002). In fact, because classification replaces quantitative information with nominal labels, any numerical information or relationship is lost, and subsequent inferences are very much limited to those based on the relative frequency of each label/ category. All these shortcomings are clearly reflected in the most commonly used classification metric, the%CC, which is considered a "*very insensitive and statistically inefficient measure*" (Harrell, 2015, page 258) that provides a count for the cases above/ below the threshold, regardless of how far these cases are from the threshold or their actual scores on any previously evaluated variable. As such, the%CC does not allow (and, when used as a single summary metric, it precludes) to describe the original outcome's distribution, summarizing the individuals' scores or estimating the actual size of between-group differences.

All these limitations are overcome if dichotomization is avoided and the scores obtained from ML algorithms (e.g., class probabilities) are used as a continuous dependent variable on which individual and group differences can be assessed. This alternative use of the ML algorithms' output is long-known and it can be traced back to the "gender diagnosticity" approach developed by Lippa & Connelly, who were probably the first to propose *"to compute diagnostic probabilities of class membership that can then serve as individual difference measures*" (Lippa and Connelly, 1990, page 1054). Since then, this approach has been profusely exploited in psychological research, but we are only aware of one study that has employed it in the neuroimaging field (Zhang et al., 2021; for conceptually-related approaches, see Phillips et al. 2019, van Eijk et al. 2021). The scarcity of studies using this approach is surprising because, when the aim is to assess sex (or other between-group) differences, it has at least two main advantages over the classification approach. First, it treats females and males as two empirical distributions that spread at different probability levels within particular ranges of the outcome's continuum (instead of as two nominal categories), hence allowing to explore and quantify within- and between-sex variation. Second, the divergences between these male and female distributions can be explored with a wide variety of statistical methods that can provide quantitative and meaningful effect sizes.

Therefore, in the present study, we adopted this approach to assess the possible multivariate sex differences in gray matter volume (GMVOL) without resorting to classification or its metrics. Because the sizes of univariate volumetric sex differences (Sanchis-Segura et al., 2019; van Eijk et al., 2021; Williams et al., 2021) and sex-classification accuracy (More et al., 2020; Sanchis-Segura et al., 2020) are strongly influenced by total intracranial volume (TIV), this assessment was conducted with raw estimates of GMVOL and after adjusting these estimates with the well-validated (Sanchis-Segura et al., 2020, 2019) power-corrected proportions method (Liu et al., 2014; PCP). More specifically, the raw and PCP-adjusted GMVOL estimates of the 116 brain areas defined by the AAL atlas (Tzourio-Mazoyer et al., 2002) were introduced as features of five different classification algorithms, which were trained and tested in two independent, sex-balanced, samples ($n = 288$ and $n = 150$ per group, respectively) in order to obtain the individuals' class probabilities (in this case, operationalized as the *probability of being classified as male;* PCAM). PCAM scores allowed to map females and males into a [0,1] continuum, and their distributional differences were thoroughly explored with robust statistical and graphical methods (Callaert, 1999; Handcock, 1998; Rousselet et al., 2017; Wilcox and Rousselet, 2018) to quantify the size of their multivariate differences in GMVOL. In a second step, the brain areas that contributed the most to the PCAM scores yielded by each algorithm in each dataset were identified by means of boosted-beta regressions (Schmid et al., 2013). These and other complementary analyses made it possible to assess whether or not different algorithms provide similar outcomes and identify the same brain architectures as typical of females and males and, therefore, evaluate whether there are "male/ female brains" at the neuroanatomical level.

2. Materials and methods

2.1. Participants

The present study was conducted using data from the 1200 Subject Release of the Human Connectome Project (HCP), which includes structural Magnetic Resonance Imaging (MRI) data from 1113 healthy young adult participants (Van Essen et al., 2013). The HCP dataset contains an unequal number of females ($n = 606$) and males ($n = 507$) who differ in age ($Mean_{females}$=29.56, $Mean_{males}$=27.90, t_{1111}=7.63, $p<4.94^{-14}$). Therefore, we used a self-built algorithm to randomly select a sex-balanced sample of participants (438 females, 438 males) that did not differ in age. This sample was subsequently split into the so-called training and testing subsamples (see below).

2.2. Imaging and data preprocessing

2.2.1. MRI acquisition and images preprocessing

The MRI acquisition details for the HCP-sample can be found in the reference manual of the S1200 release of the HCP (https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf).

Images were preprocessed with the CAT12 toolbox (http://www.neuro.uni-jena.de/cat/, version r1184) of the SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/, version 6906) software. CAT12 preprocessing was conducted following the standard default procedure suggested in the manual. Briefly, it includes the following steps: (1)

segmentation of the images into gray matter, white matter, and cerebrospinal fluid; (2) registration to a standard template provided by the International Consortium of Brain Mapping (ICBM); (3) DARTEL normalization of the gray matter segments to the MNI template; (4) modulation of the normalized data via the "affine + non-linear" algorithm; and (5) data quality check (in which no outliers or incorrectly aligned cases were detected). Images were not smoothed because we were only interested in the modulated images.

After applying this procedure, which does not include any correction for overall head size, voxels were mapped into 116 regions according to the Automated Anatomical Labeling atlas (Tzourio-Mazoyer et al., 2002) by calculating the total gray matter volume for each region of interest (VOI) and participant via a MATLAB script (https://www0.cs.ucl.ac.uk/staff/g.ridgway/vbm/get_totals.m). TIV was estimated using native-space tissue maps obtained in the segmentation step. More specifically, TIV was calculated as the sum of GM, WM, and CSF total values multiplied by voxel size and divided by 1000 to obtain a milliliter (ml) measurement. The estimates of GMVOL in these 116 regions were employed as the predictors for the machine-learning algorithms described below.

2.2.2. Training/ testing subsets and data standardization

Following current recommendations (Bzdok and Ioannidis, 2019; Hastie et al., 2009), classification algorithms were fitted and tested in two separate and sex-balanced groups of participants, thus allowing an honest evaluation of the models' performance and avoiding classification distortions due to between-class imbalance (Ali et al., 2013; García et al., 2007). However, because the HCP sample includes data from twins, non-twin siblings, and unrelated individuals, constructing the training and testing subsamples through simple random assignment would result in splitting very similar pairs of individuals (e.g., monozygotic twins) across these two subsamples, artificially reducing bias, and overestimating the models' performance. Therefore, in the present study, the training and testing subsamples were constructed using a three-step procedure: (1) Families were first randomly grouped into two non-overlapping sets; (2) Male-female pairs were randomly extracted from each set to create a sex-balanced "training pre-sample" and a sex-balanced "testing pre-sample" with appropriate relative sizes (2/3 and 1/3, respectively); and (3) To ensure that the final training and testing subsamples were sex-balanced and had a similar age mean and distribution, some individuals in these provisional subsamples were replaced through random sampling from the remaining pool of subjects. Thus, the final training subsample included 288 females and 288 males, whereas the testing subsample included 150 females and 150 males of similar ages (see further details in Supplementary Table 1A and B). These two subsamples were also largely free of any relatedness bias because only 23 members (out 300 subjects, 7.66%) of the testing subsample belonged to families that were also included in the training subsample."

Before being used as predictors, all volumetric variables were transformed into *z*-scores to avoid distortions due to their different ranges (Ali and Smith-Miles, 2006; Hastie et al., 2009). Standardization was initially performed in the *training subset*, and the exact same scaling parameters were subsequently used to standardize the *testing subset*.

2.2.3. TIV adjustment: the raw and the PCP datasets

Previous studies have shown that the estimates of univariate and multivariate sex differences are largely dependent on TIV variation and that not all the currently used methods are equally effective and valid for removing TIV-variation (Sanchis-Segura et al., 2020, 2019). Therefore, in the present study, all analyses were conducted twice in the same subjects, without introducing any TIV adjustment ("raw" dataset) and after removing TIV variation with the well-validated *power-corrected proportions* (PCP) method (Liu et al., 2014). The PCP method improves the traditional proportions approach by introducing an exponential correcting parameter in the denominator. More specifically, the adjusted volume of interest (VOI) is calculated as $VOI_{adj} = VOI/TIV^{b}$, where the b parameter corresponds to the slope value of the LOG(VOI) ~ LOG(TIV) regression line (Liu et al., 2014). As when standardizing the data, in this study, the LOG(VOI) ~ LOG(TIV) regression lines were calculated using the data of the individuals in the training subset, and the same b parameters obtained in this subset were used to adjust TIV-related variation in the testing subset.

2.3. Machine-learning algorithms

We report and compare the outcomes of five classification algorithms that differ in their assumptions (Supplementary Table 1C) and that provide an adequate representation of the principal "families" of machine-learning classifiers.

Testing several ML algorithms is important because algorithms' internal operations are very much dependent on these assumptions (Hastie et al., 2009; Kiang, 2003) and may potentially lead to different outcomes. Thus, comparing and/or combining the outcomes of different ML algorithms should lead to more robust and generalizable findings and conclusions (Breiman, 2001; Hancox-Li, 2020). The outcomes considered in this study included common classification metrics (such as the percentage of correctly classified cases,%CC), but also novel and alternative ones that were obtained by using the posterior classification probabilities obtained from ML algorithms (in this case, operationalized as the *probability of being classified as male,* PCAM) as a continuous variable (see details in Section 2.4).

All the ML classifiers were implemented and cross-validated (5 folds; 10 repeats) using the interface provided by the *caret* package for R. In alphabetical order, the predictive algorithms used in the present study were:

- *Linear Discriminant Analysis (LDA):*Implemented by the default options of the *lda* function from the *MASS* package (Venables and Ripley, 2002).
- *Logistic Regression (LR):* Implemented by the *glm* function (family= "binomial") of the *stats* package natively included in R Core Team (2020).
- *Multiple Adaptive Regression Splines (MARS):* Implemented by the *earth* function of the *earth* package for R (Milborrow, 2019). The hyper-parameters of the model were determined by a cross-validated grid search assessing 30 possible combinations (degree: 1–3, nprune=2–116, length.out=10).
- *Random Forest (RF):* Implemented by the *rf* function of the *randomForest* package (Liaw and Wiener, 2002), built up by aggregating 500 classification trees, each of them using 10 randomly selected predictors.
- Support Vector Machine *with a radial kernel (SVM):* Implemented using the *svmRadial* function of the *kernlab* package for R (Karatzoglou et al., 2004). The *tune* function (tenfold cross-validation) was used to automatically select the optimal values for the regularization and kernel-width hyper-parameters.

2.4. Statistical analyses

All statistical analyses were conducted in the testing subsamples of the raw and the PCP datasets using different packages for R (R Core Team, 2020). Statistical analyses focused on description and effect sizes estimation rather than merely testing statistical significance (Wasserstein and Lazar, 2016). All effect size estimates were accompanied by 95% confidence intervals (CI), and, when appropriate, these effects were also reported in terms of their percentage of the maximum possible (POMP) score. POMP scores were calculated using the POMP formula=[(*observed value-minimum possible value*)/(*maximum possible value -minimum possible value*)]*100 (see further details in Cohen et al. 1999), Grissom and Kim 2012). Moreover, when statistical significance was tested, *p* values were corrected for multiple comparisons with the FDR (Benjamini and Hochberg, 2018) or -when comparing deciles (see below)- with the Hochberg method (Hochberg, 1988).

2.4.1. Algorithms' performance: predictive accuracy

Algorithms' performance was initially measured as the percentage of correctly classified cases (%CC) and its 95% CI. These%CC scores were initially compared with the chance-expected value of 0.5 with one-sided binomial tests and with each other by means of the McNemar's test. Classification bias (whether females or males had higher chances of being misclassified) was also assessed using the McNemar's test. The details of these analyses are presented in the Supplementary Material.

2.4.2. Assessing multivariate sex differences in GMVOL

As originally proposed by Lippa and Connelly (1990), we used the classification probabilities obtained from ML algorithms (in this case, operationalized as the *probability of being classified as male,* PCAM) as a continuous dependent variable on which individual and between-group differences can be quantified. At this respect, it is important to note that, although some algorithms are probabilistic (e.g. logistic regression) and others are not (e.g., SVMs), classification probabilities are obtainable from probabilistic and non-probabilistic algorithms (Chen et al., 2021). It should be also noted that in the present study the PCAM is used as a mere quantitative score that allows ranking the cases from the most probable to the least probable member of one class but that is devoid of any predictive aim. Therefore, this use of PCAM scores is free of calibration-related concerns (Chen et al., 2021; Niculescu-Mizil and Caruana, 2005).

The males and females' PCAM distributions obtained from each algorithm in each dataset were first described through bootstrap estimates of appropriate statistics (skewness, kurtosis, deciles, inter-quantile range and variance; repetitions=10,000). Differences between the overall PCAM distributions yielded by different algorithms were tested through a series of independent Kolmogorov-Smirnoff tests. In a second step, PCAM scores were used to quantify the multivariate sex differences in GMVOL at different levels.

Sex differences in PCAM scores: single-measure estimates. Possible sex differences in PCAM dispersion measures (variances and inter-quantile ranges, IQR) were assessed through the original version and a customized version of the *comvar2* function included in the freely accessible Rallfun-v38 file (https://dornsife.usc.edu/labs/rwilcox/software/).

The overall degree of similarity between the PCAM density distributions for males and females was quantified using the η overlap index. The η index measures the area intersected by two probability density functions, and it is conceptually related to other measures of overlap, such as the Kullback-Leibler divergence and the Bhattacharyya's distance. However, unlike these overlap metrics, $\hat{\eta}$ can be estimated in the absence of symmetry, unimodality, or any other distributional assumption (Pastore and Calcagnì, 2019). In the present study, kernel density estimation (KDE) and $\hat{\eta}$ were obtained through the *boot.overlap* (10,000 repetitions) function of the *overlapping* package for R (Pastore, 2018). A second and complementary estimate of these sex differences at the distribution level was obtained by calculating the probability of superiority (PS). The PS is defined as the probability that a randomly sampled member of group A will have a higher score than the score attained by a randomly sampled member of group B. More specifically, the probability that males' PCAM scores would be higher (PS_M), equal to, or lower than those of females (PS_F), along with the Cliff's δ statistic (Cliff, 1993) and its 95%CI, was obtained through the *cidv2* function of the *rogme* package (Rousselet et al., 2017) .

Sex differences in PCAM scores: relative distribution methods and quantile differences. Because no single score can properly summarize the differences between two distributions (Callaert, 1999; Cook et al., 2016; Del Giudice, 2019; Grissom and Kim, 2012; Handcock and Morris, 1999; Rousselet et al., 2017), male-female differences in the PCAM continuum were characterized by comparing their cumulative distribution functions (CDF; Callaert, 1999; Grissom and Kim, 2012). CDFs make it possible to directly estimate the proportion of cases in each group with PCAM values equal to or lower than any possible cutoff, but also the proportion of subjects in one group have PCAM values equal or lower than a given proportion of cases in another group (Callaert, 1999; Grissom and Kim, 2012). Within each CDF, sex-based comparisons were conducted at each decile with the *shifthd_pbci* function (bootstrap: 10,000 repetitions) of the *rogme* package (Rousselet et al., 2017). The *shifthd_pbci* -and other functions of this package described below- use the Harrell-Davis quantile estimator in conjunction with a percentile bootstrap approach to calculate the deciles and the between-groups differences at those deciles. Unlike traditional parametric methods, this approach ensures that the estimates fall within the bounds of the PCAM distribution [0,1], thus preventing inappropriate inferences. Moreover, during the calculation of these deciles' differences, the corresponding 95% CIs are calculated, and the significance level is adjusted for multiple comparisons using the Hochberg method. Thus, when one of these CIs does not include the zero value, the difference might be declared statistically significant at $\alpha<0.05$ without being concerned about Type I error (Rousselet et al., 2017).

With the decile estimates obtained, the so-called shift functions (Wilcox and Rousselet, 2018) were also calculated. The shift-function plots the between-groups decile differences against the deciles of one group, thus providing a complete picture of how, and by how much, the score distribution of one group should be re-arranged to match that the scores' distribution of another group (for a detailed description, see Rousselet et al. 2017). Finally, we also compared whether the estimated size of the female-male differences at D5 (median) differed between algorithms and within the deciles of the PCAM distributions obtained with each algorithm. These comparisons were conducted with the original *bwquantile* function (see acknowledgements section) and with a customized version of this function, respectively.

Following current recommendations (Rousselet et al., 2017; Wilcox and Rousselet, 2018), we also estimated the size of the typical difference between *any* given male and *any* given female at each PCAM distribution of the raw and PCP datasets. These bootstrapped estimations were conducted using the *allpdiff_hdpbci* function (bootstrap: 10,000 repetitions) of the *rogme* package (Rousselet et al., 2017), which computes through the Harrell-Davis estimator the deciles (and their 95%CI) of the empirical distribution of all (in this case, 22,500) pair-wise differences between the members of two independent groups. We also calculated the CDFs for these pair-wise differences in the raw and PCP datasets, and then between-datasets decile-based comparisons were conducted with the *shiftdhd_pbci* function (bootstrap: 10,000 repetitions) of the *rogme* package (Rousselet et al., 2017). Finally, we employed the *Dqcomhd* function (bootstrap: 50,000 repetitions) of the *WRS2* package (Mair and Wilcox, 2018) to ascertain whether the deciles of these male-female pair-wise differences significantly varied between algorithms in each dataset.

2.4.3. Interpreting multivariate sex differences in GMVOL

Interpretability has become a major issue in ML applications (Carvalho et al., 2019; Hancox-Li, 2020; Ribeiro et al., 2016). In the particular case of the study of multivariate sex differences, knowledge about the brains of females and males is only gained when the complex and numerical output of ML algorithms is decomposed and the brain features that contribute the most to the groups' distinguishability are identified. To provide this information, we extracted global, post-hoc, model-agnostic explanations of the five ML algorithms tested in this study by modeling their outputs through the use of interpretable surrogate models (for a discussion about the different types of interpretability and their associated methods, see Carvalho et al. 2019, Lipton 2018, Ribeiro et al. 2016). More specifically, we employed boosted beta regression procedures to identify the brain features that best predicted the PCAM scores observed with each algorithm in each dataset.

Boosted beta regression analyses and between-algorithms' agreement. In the statistical literature, beta regression has been established as a powerful

and readily interpretable procedures to model bounded [0,1] distributions (Ferrari and Cribari-Neto, 2004). However, because the outcome of classical beta regression procedures might be challenging when using a large number of predictors, boosted beta regression models have been developed (Schmid et al., 2013). Boosted beta regression is based on the gamboostLSS algorithm, which performs a reliable variable selection during the iterative fitting process (for a comprehensive description of boosted beta regression, see Schmid et al. 2013). This feature selection process results in simpler and more interpretable models but it may potentially reduce the generalizability or robustness of the obtained explanations (for an ample discussion of the prediction/ explanation tensions and tradeoffs, see Breiman 2001, DelGiudice 2021, Hancox-Li 2020).

In the present study, boosted beta regressions were implemented through the *betaboost* package for R (Mayr et al., 2018), using the PCAM scores observed with each algorithm in each dataset as the response variable and the volumetric scores of the testing subsample in the raw/PCP datasets as predictors. The number of iterations that most reduced the risk was established through cross-validation and the contribution of each predictor was estimated by using the obtained mu-coefficient values and by constructing a relative importance measure: relative importance= 100*(accumulated risk reduction attributable to a predictor/ total risk reduction in the model).

In a second step, the degree of agreement between the boosted beta regression models obtained was assessed. These comparisons were conducted between datasets and between algorithms within each dataset. For each of these two sets of comparisons, we first assessed whether boosted beta regression models included the same brain features as relevant predictors. More specifically, R-wise agreement (coincidence between all models) was estimated by means of the Hubert's Kappa index (Hubert, 1977) and the multi-rater delta index (which, unlike other more commonly used agreement metrics, is not affected by the ratings' marginal distributions; see, Andrés and Marzo 2004, Andrés and Hernández 2019). These two agreement indexes were calculated using software specifically developed for this purpose and freely available at https://www.ugr.es/~bioest/software/cmd.php?seccion=agreement.

We also assessed the degree of agreement on the relative importance attributed to each predictor by using Lin's concordance correlation coefficient (Lin, 1989), Kendall's W agreement coefficient (Kendall and Smith, 1939), the mean of bivariate Spearman's rho rank correlations (Gamer et al., 2019), and the intraclass-correlation coefficient (two-way ANOVA, random effects, single and average ratings; Hallgren, 2012; Koo and Li, 2016). In the case of comparisons of algorithms within each dataset, agreement was assessed at the interval and at the ordinal level by inputting the obtained coefficients' values and their ordinal rank positions, respectively. In the case of datasets comparisons, agreement was assessed by using each predictor's ranking mean (RM) or its position in a multiplicative "rank of ranks" (RRP, Tofallis, 2014) across algorithms. These two measures were also used in a correlational analysis assessing whether the relative importance of these predictors was associated with TIV variation. Thus, as previously done (Sanchis-Segura et al., 2020, 2019), linear regression analyses were conducted to obtain an estimate (r^2) of the TIV-explained variance in each brain region, and the r^2 scores corresponding to the brain regions identified as relevant predictors in each dataset were correlated with their corresponding RM and RRP values. Finally, because calculating the average importance of each variable across several models is expected to allow more accurate and reliable inferences (Breiman, 2001; DelGiudice, 2021; Hancox-Li, 2020), the predictor's ranking mean was also used to identify the 10 predictors exhibiting the highest average importance across all algorithms in the raw and in the PCP dataset.

Between-algorithms PCAM variation. To assess the degree of similarity of the PCAM scores obtained with different algorithms, three complementary approaches were used. First, for each individual, its minimum PCAM score was subtracted from its maximum PCAM score within each dataset, and the CDFs depicting this maximum PCAM variation in each dataset were built up and described by several summary statistics (minimum, average, deciles, and maximum). Second, the same statistics were used to describe the degree of PCAM variation for each algorithms' pair within the raw and the PCP datasets. Finally, zero-order Spearman's rho between-algorithms' correlation matrices in the raw and PCP datasets were calculated. Because we corroborated a significant contribution of TIV to PCAM scores in the raw dataset, this correlational analysis was also conducted using partial (-TIV) Spearman correlations.

Hierarchical clustering and ANOSIM analyses. As described above, the high accuracy observed in previous sex classification studies has often been interpreted as showing the ability of ML algorithms to identify two clearly distinguishable brain types, one typical of males and the other typical of females. However, proving that these brain types actually exist requires confirming that the brains of females and males substantially differ in a specific and reproducible pattern of brain features. To assess whether or not these distinctive brain profiles could be found in our data, agglomerative hierarchical clustering methods (average linkage) were used.

Specifically, hierarchical clustering analyses were performed with the *hclust* function of the *stats* package (R Core Team, 2020). Initially, the included features were the volumetric z-scores of those brain areas identified as relevant predictors of the PCAM scores yielded by each ML algorithm in each dataset (see 2.4.3.1). Dissimilarity was measured in terms of Euclidean and Spearman distances; Euclidean distances served to quantify the individuals' disparity in terms of the magnitude of their accumulated differences in GMVOL, whereas Spearman distances measured the discordance in the shape of their brain profiles. Each resulting dendrogram was cut at appropriate heights to obtain 2 to 10 clusters and in each of these alternative partitions the size (number of subjects) and composition (proportion of females) of the resulting clusters were assessed. The robustness of the obtained results was corroborated by repeating the same analyses with: (1) the top five predictors of the PCAM scores yielded by each ML algorithm in each dataset; (2) the 10 predictors exhibiting the highest average importance across all algorithms in each dataset; and (3) the volumetric z-scores of the 116 brain areas from the AAL atlas.

Complementarily, a series of analysis of similarity (ANOSIM) were conducted. ANOSIM is an ANOVA-like, non-parametric test that operates on distance matrices and assesses the null hypothesis that distances between members of two or more predefined groups (in this case, males/females) is the same as between the members of these groups (Clarke, 1993). Because in the present study this assessment involved a large number of instances, statistical significance was almost guaranteed and, consequently, it was not truly informative (Lindley, 1957). Therefore, we focused on estimating the value of R statistic (and its 95%CI), which compares the mean of ranked dissimilarities between groups to the mean of ranked dissimilarities within groups, and whose meaningful range lies between 0 (when the similarity within groups is the same as between-groups) and 1 (when all samples within groups are less dissimilar to each other than to any pair of samples from different groups; see Clarke 1993). More specifically, Euclidean and Spearman distance matrices were calculated in 10,000 bootstrap samples using the *get_dist* function of the *factoextra* package (Kassambara and Mundt, 2020). In each of these distance matrices, an ANOSIM test was conducted with the default options of the *anosim* function of the *vegan* package (Oksanen et al., 2020) and the corresponding R value was obtained. In a second step, the 95% confidence intervals of these estimates (normal approximation and percentile method) were obtained through the *boot.ci* function of the *boot* package (Canty and Ripley, 2020). Again, these calculations were performed using as features (1) the volumetric z-scores of those brain areas identified as relevant predictors of the PCAM scores yielded by each ML algorithm in each dataset; (2) the top five predictors of the PCAM scores yielded by each ML algorithm in each dataset; (3) the volumetric z-scores of the 116 brain areas from the AAL

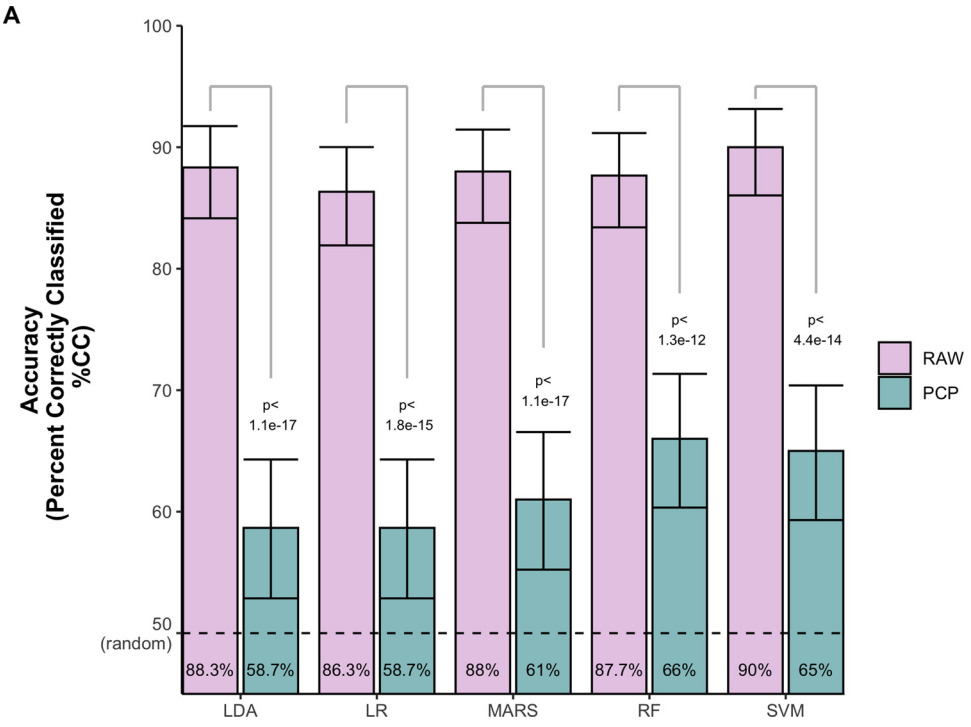


Fig. 1. Sex classification accuracy. Bars depict the percent of correctly classified cases (%CC) achieved by each algorithm (LDA=linear discriminant analysis, LR= logistic regression, MARS=multiple adaptive regression splines, RF=Random Forest, SVM=Support Vector Machine) in the raw dataset and in the PCP dataset. See Supplementary Table 2A-D for a complete statistical output.

atlas; and 4) the 10 predictors exhibiting the highest average importance across all algorithms in each dataset.

3. Results

3.1. Classification accuracy

Fig. 1 displays the classification accuracy (%CC) achieved by each algorithm in the testing subsamples ($n = 150$ per group) of the raw and PCP datasets. As previously observed (More et al., 2020; Sanchis-Segura et al., 2020), the proportion of subjects correctly classified was much higher in the raw dataset (%CC average= 88.06) than in the PCP dataset (%CC average=61.86; Supplementary Table 2A and B). Fig. 1 also shows that the%CC varied slightly between algorithms. In the raw dataset, none of these differences achieved statistical significance (Supplementary Table 2C). In the PCP dataset, LDA and LR exhibited lower accuracy than RF and SVM, but the statistical significance of these differences was lost after correcting for multiple comparisons (p_{adj}= 0.06 in all cases; Supplementary Table 2D). Therefore, it can be concluded that prediction accuracy clearly differs between datasets; however, within each dataset, all the algorithms seem to yield very similar%CC values.

3.2. Assessing multivariate sex differences in GMVOL

Fig. 2 depicts the kernel density estimates (KDE) of the PCAM distributions yielded by each algorithm in each dataset. Based on these graphical representations, it is apparent that PCAM distributions differed between males and females, but also between datasets and between algorithms within each dataset. More specifically, in the raw dataset, both males and females exhibited non-normal and very skewed distributions, with most of the females accumulating near of the lower bound of the PCAM continuum, and most of the males accumulating near of the upper bound. These distributions were also very long-tailed, with a few scattered individuals spreading beyond their respective sex clusters and virtually occupying the entire PCAM range. These distributional characteristics significantly varied depending on the algorithm (e.g., LDA vs. RF; Supplementary Table 3A-B, Supplementary Fig.1). On the other hand, in the PCP dataset, all the PCAM distributions were non-skewed, but they also showed major differences between them. Thus, the PCAM distributions obtained with the LDA, LR and MARS algorithms were rather flat, with females and males spreading at similar rates across almost the entire PCAM range. In contrast, the RF and SVM associated distributions were more clearly peaked and extended within a shorter span around the center of the PCAM range (Supplementary Table 3A-C, Supplementary Fig.1).

Based on the PCAM distributions, overall estimates of the multivariate sex differences in GMVOL were obtained. These measurements indicated that, despite the previously mentioned between-algorithm variations, sex differences were “large” in the raw dataset. Thus, the overlap between the males/females’ distributions was “small” (≈12–18%; Supplementary Table 4A), and the chance that a randomly chosen male would have a PCAM score higher than that of a randomly chosen female (PS_M) exceeded 90% in all cases (Supplementary Table 4B). Conversely, in the PCP dataset differences were much smaller, with high levels of overlap (range: 55.7–71.3%) and lower PS_M scores (ranges: 0.64–0.7; Supplementary Table 4D and E). Additional estimates of the multivariate sex differences in GMVOL were obtained by comparing location and dispersion measures. No differences in variances or inter-quantile ranges were observed (Supplementary Table 4C and D), suggesting that -when considered on the PCAM continuum- the variability of males and females does not significantly differ. Conversely, the PCAM medians of males and females were significantly different, and although the size of these differences varied between algorithms, they can all be considered “large” and “small” in the raw and PCP datasets, respectively (see details in Section 3.2.1 and in Supplementary Table 5A and B).

However, all these measures provide a single effect size estimate that may not be very informative or could even be misleading with regard to the possible complex differences between two distributions (Grissom and Kim, 2012; Handcock and Morris, 1999; Rousselet et al., 2017). To fully represent and compare distributions robust statistical and informatively-rich graphical methods such as relative distribution methods (Callaert, 1999) and the shift-function (Rousselet et al., 2017; Wilcox and Rousselet, 2018) are required (Callaert, 1999;

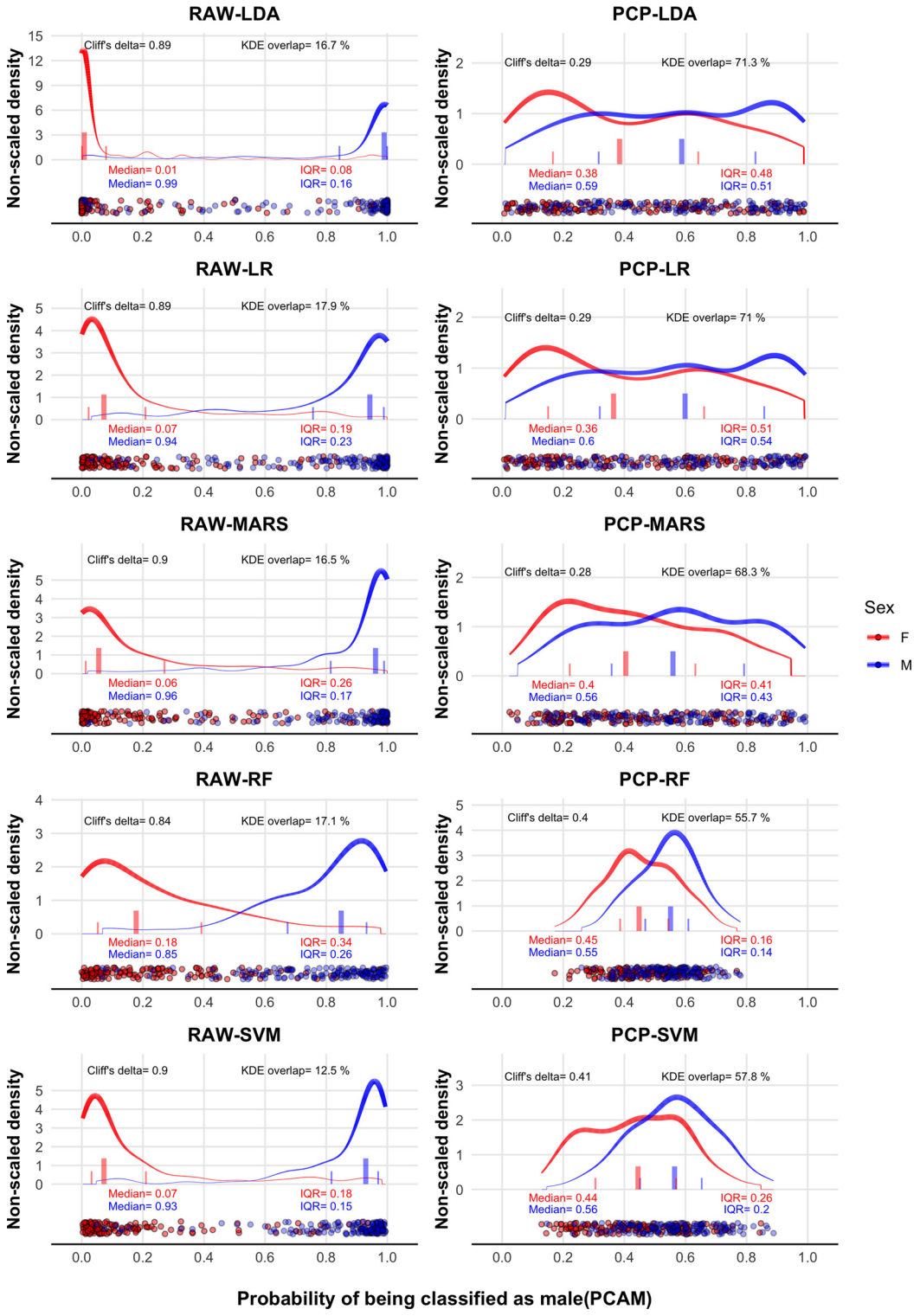


Fig. 2. The PCAM continuum. Plots depict the strip charts (bottom) and the non-scaled density functions (top) of the PCAM scores of females (red) and males (blue) yielded by each algorithm in each dataset. The thickness of the lines is directly proportional to the scaled density of each distribution. Plots also include the medians and inter-quantile ranges of each group (vertical bars) and estimates of their similarities/ differences at the distribution level (overlap/ Cliff's delta, respectively). For a complete statistical output, see Supplementary Table 4A–D. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

Del Giudice, 2019; Handcock and Morris, 1999; Rousselet et al., 2017). Therefore, we used these methods to extend our analyses and provide two complementary perspectives (Rousselet et al., 2017) of the multivariate sex differences in GMVOL.

3.2.1. How do males and females compare to each other?

Fig. 3 displays the cumulative density functions (CDFs) and the deciles' distribution for the PCAM scores of males and females yielded by each algorithm in each dataset, making it possible to compare males and females in three different ways: 1) by directly contrasting the proportion of cases in each group with PCAM scores equal to or lower than any possible cutoff; 2) by estimating how many subjects in one group have PCAM values equal to or lower than a given proportion of cases in the other group; and 3) by comparing the PCAM values at the deciles of the females/ males' distributions.

All these comparisons confirmed that, when PCAM scores are obtained from multivariate composites of raw GMVOL, males and females are quite different. For instance, in the raw dataset, 10% of males with the lowest scores (D1) had PCAM values that were higher than or equal to those observed in ≈80% of the females. However, Fig. 3 also shows that the size of these sex differences varied across individuals. Thus, taking the LDA outcomes as an example, sex differences in PCAM scores were already "large" at D1 (30.6% of the maximum possible; POMP), but they were much larger at the medians (D5, POMP=98.1) and tended to decrease thereafter (D9, POMP=40.72). These inter-decile variations were statistically significant (Supplementary Table 5A and C), and resulted in clearly non-monotonic shift-functions (Supplementary Fig. 2), suggesting that differences at center locations might lead to inappropriate inferences about the differences observed at more distal locations of the same distribution. On the other hand, the estimated size of sex differences also varied between algorithms (LDA>LR=MARS=SVM>RF). Thus, for example, the RF estimated difference at D5 was 66.5 POMP, which is 31.6% lower than the estimate provided by the LDA algorithm. This and other similar between-algorithm variations were statistically significant (Supplementary Table 5E).

Conversely, when the effects of TIV-variation are ruled out, males and females are much more similar to each other. Thus, in the PCP dataset, 10% of males with the lowest scores (D1) had PCAM values that were higher than or equal to those of just 20–30% of the females. Moreover, and also in contrast to what had been observed in the raw dataset, sex differences were approximately constant across deciles (≈10–20 POMP), and resulted in almost flat shift-functions (Supplementary Table 5B and D; Supplementary Fig. 2). In addition, distinct algorithms provided different estimates of the size of these sex differences. The relative magnitudes of these variations were similar to those observed in the raw dataset, but in this case, they did not reach statistical significance (Supplementary Table 5F).

3.2.2. What is the typical difference between any given female and any given male?

When the interest is not as much to describe and compare males and females, but rather to estimate the size of the typical difference between any given male and any given female, the distribution of all their pair-wise differences can be calculated and directly analyzed (Rousselet et al., 2017; Wilcox and Rousselet, 2018). Thus, Fig. 4A depicts the KDE of all the pair-wise differences between males and females in each algorithm and dataset, whereas panel B depicts their corresponding CDFs. Thus, in this case, CDFs make it possible to: (1) estimate the empirical probability of finding a pairwise male-female difference whose size is equal to or lower than a given reference value; (2) estimate the size of the pairwise differences for any given proportion of cases; and (3) compare the estimates of these pairwise differences provided by different algorithms in each dataset or by the same algorithm in the raw and PCP datasets.

As Fig. 4 shows, in the raw dataset, pairwise male-female differences extended over a very wide range, but they were also very asymmetrically distributed and favored males in more than 90% of the cases (i.e., D1 values>0 in all algorithms; Supplementary Table 6A). The size of these differences depended on the algorithm used to calculate the PCAM scores (medians' range= 0.58–0.93; Supplementary Table 6B), although they were generally "large" as compared to the possible maximum (average POMP difference= 67.2). Consequently, the multivariate estimates of raw GMVOL in a randomly picked male-female pair are expected to clearly differ, leading to PCAM scores substantially (POMP difference >30%) larger in males than in females in around 80–90% of the cases.

By contrast, when the influence of TIV-variation is statistically controlled, pair-wise male-female differences show an algorithm-dependent range but they are always quasi-symmetrically distributed around their median values (Supplementary Table 6C). These median values differed between algorithms (range= 0.09–0.16; Supplementary Table 6D), although all of them indicated that pairwise male-female differences were "small" as compared to the possible maximum (average POMP difference= 12.9) and significantly smaller than the differences observed in the raw dataset (Supplementary Table 6E). Accordingly, the multivariate estimates of TIV-adjusted GMVOL of randomly picked male-female pairs are expected not to differ much, and the females' PCAM scores should be higher than or equal to males' scores in 30–40% of the cases.

3.3. Interpreting multivariate sex differences in GMVOL

If simplified to the maximum, the results described in the previous section indicate that the multivariate sex differences in raw measures of GMVOL are "large", but also that these differences become "small" when the effects of TIV-variation are statistically controlled. This conclusion is similar to the one that could be obtained after examining%CC scores. However, this parallelism should not lead to the interpretation that%CC and PCAM-based measures are equivalent. In fact, within each dataset,%CC scores were uncorrelated or even inversely correlated with the estimated size of the multivariate sex differences in GMVOL (see Supplementary Figs. 3 and 5, respectively). Moreover: (1) Whereas%CC primarily relates to algorithm's performance, PCAM-based measures describe and compare individuals and groups, making it possible to quantify between- and within-sex variation; (2) Whereas%CC is a "*very insensitive and statistically inefficient measure*" (Harrell, 2015; page 258) that did only vary between-datasets, PCAM-based measures are sensitive enough to reveal that multivariate sex differences in GMVOL also differ between algorithms and between subjects; (3) Whereas%CC scores are obtained from pre-imposed criteria, PCAM-based measures are fully empirical and conceptually unrestricted.

On this last point, previous sex classification studies (e.g., Anderson et al. 2018, Chekroud et al. 2016, Rosenblatt 2016, Sepehrband et al. 2018, Xin et al. 2019, Zhang et al. 2021) have reinvigorated sex binary views according to which human brains can be categorized into two "types", one typical of males and the other typical of females. More specifically, the high%CC scores observed in these studies are often understood as the objective result of the ML algorithms' ability to "*predict sex*" (Sepehrband et al., 2018; page 217) and as "*demonstrating that multivariate patterns of gray matter are reliably dimorphic between sexes*", unraveling their "*essential discriminability*" (Anderson et al., 2018; pages 1502–1503) and definitively proving that "*brains are indeed typically male or typically female*" (Rosenblatt, 2016). However, finding two- and only two- brain types is not as much an empirical result as it is a pre-imposed requirement (Harrell, 2015) of these sex classification studies. Thus, because it is obtained through the application of a cutoff on a continuous score and no gray zone is included, classification is less of a prediction than a decision that results from an "*artificial forced choice*" (Harrell, 2015; page 5) between two predefined and mutually-exclusive alternatives. Therefore, it can be argued that classification is not uncovering two brain types; instead, it is actually forcing the number of possible types to be just two, and it does this by: (1) ignoring that there are a number of cases (those situated around the cutoff) for which predictors

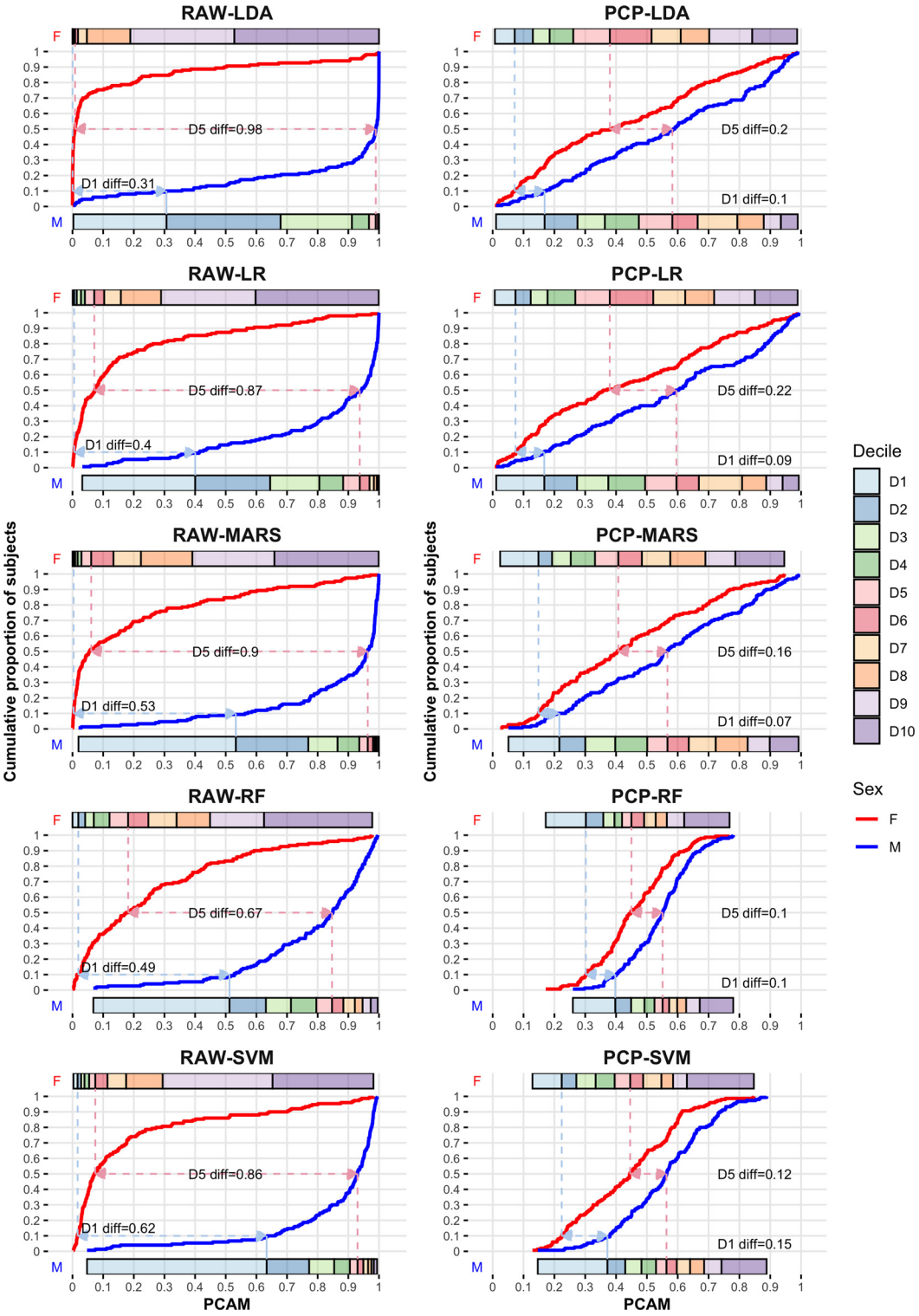


Fig. 3. Comparing females and males on the PCAM continuum. Plots depict the cumulative density functions of the PCAM scores for females (red) and males (blue). Horizontal color bars depict the tenths of the females' (top) and males' (bottom) PCAM distributions. As the provided examples illustrate, these plots allow to compare the PCAM values at the deciles of the females/ males' distributions but also to compare the proportion of cases in each group with PCAM scores equal to or lower than any possible cutoff, and how many subjects in one group have PCAM values equal to or lower than a given proportion of cases in the other group. See further details and analyses in Supplementary Table 5A–F and in the accompanying Supplementary Fig. 2. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

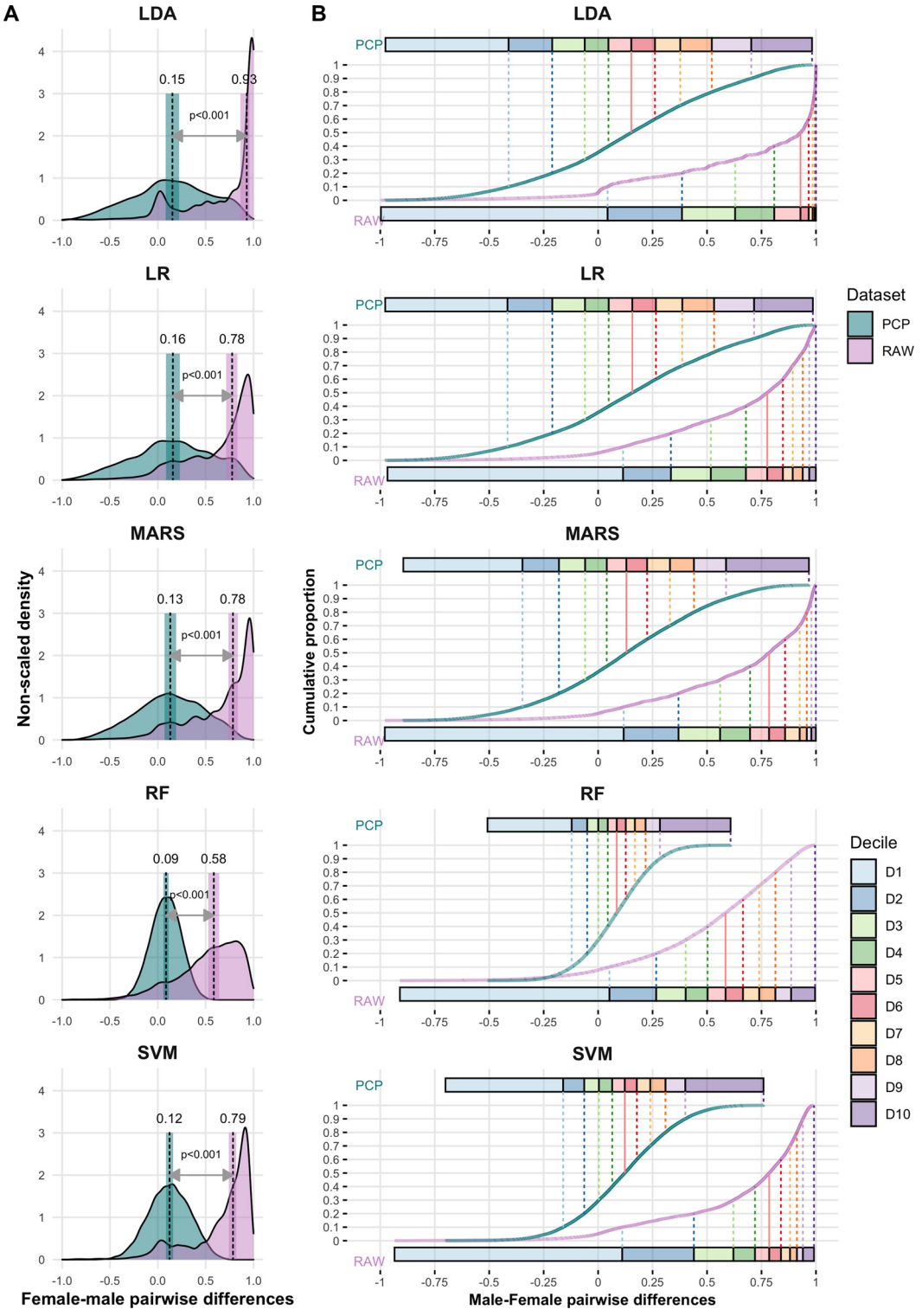


Fig. 4. What is the typical difference between any female and any male?(A) Estimated density functions of all pairwise differences between males and females in each algorithm and dataset. (B) CDFs of these pair-wise differences allow to (1) estimate the empirical probability of finding a pairwise male-female difference whose size is equal to or lower than any predesignated value; (2) estimate the size of the pairwise differences for any given proportion of cases; (3) compare the estimates of these pairwise differences provided by different algorithms in each dataset or by a single algorithm in the raw and PCP datasets. See further details in Supplementary Table 6A–E and in the accompanying Supplementary Fig. 4.

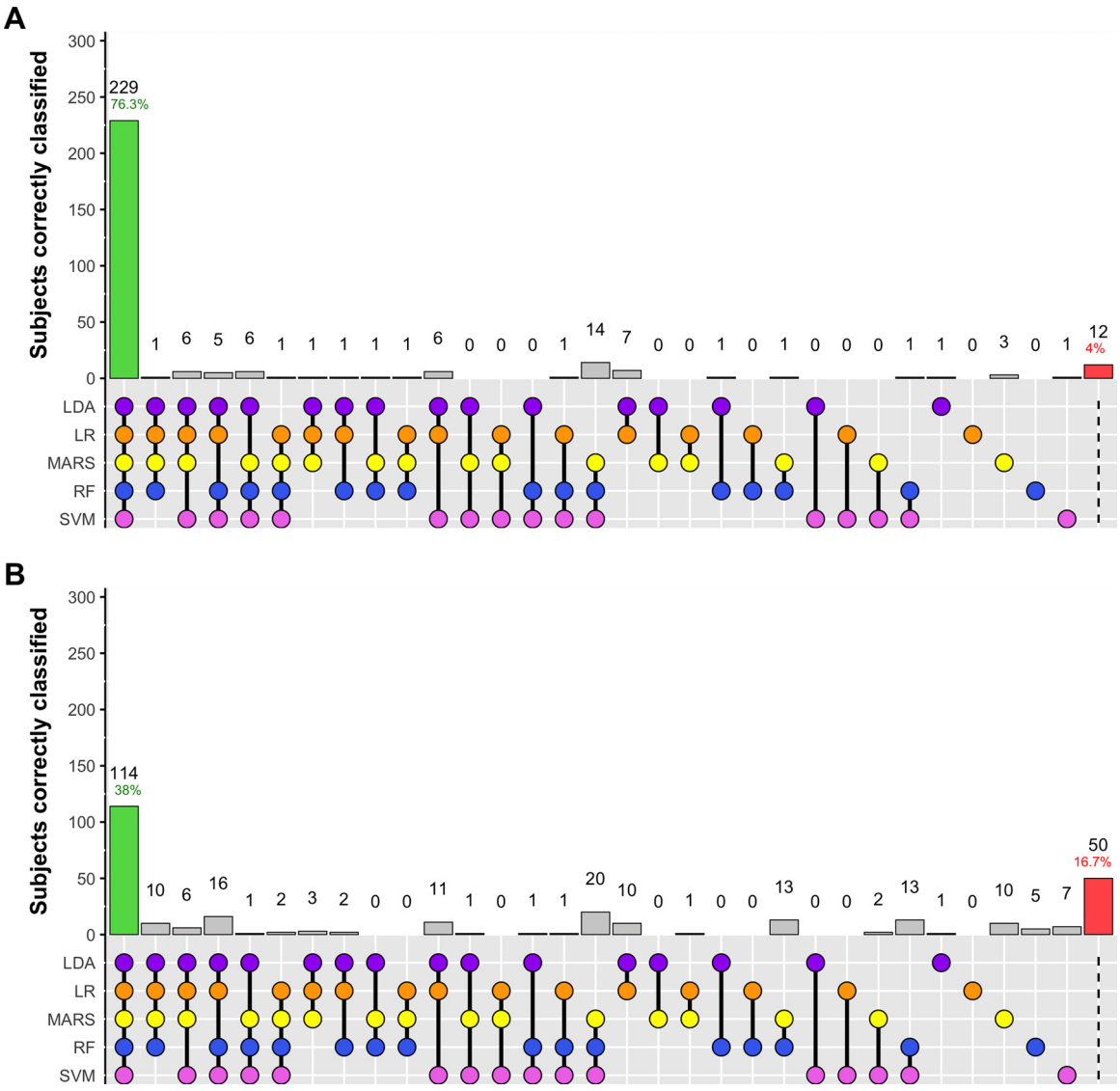


Fig. 5. Intersection analysis of the cases correctly classified by different algorithms. Panels include upSet plots illustrating the intersections (bottom) and the intersection sizes (top) of the cases correctly classified by the LDA, LR, MARS, RF, and SVM algorithms in the raw (A) and PCP (B) datasets, respectively. Within each panel, columns correspond to distinct possible intersections: the line-joined circles identify which algorithms are part of the intersection, whereas the bars on the top show the number of correctly classified cases found at that intersection (the intersection's cardinality). Thus, for example, the first column starting from the left illustrates the number of cases that were correctly classified by all five algorithms, the next one illustrates the number of cases that were correctly classified by the LDA, LR, MARS, and RF (but not by the SVM) algorithms, and so on. Cardinality is also numerically reported in terms of absolute frequency for all intersections, but in the columns corresponding to the intersection between all sets (green bar) and the complement (those cases that were not correctly classified by any algorithm, red bar), cardinality is also reported as a percentage. Because the same subjects were tested by all the algorithms (sets), horizontal bars commonly included to depict set sizes in upSet plots are omitted. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

do not provide enough information to make an informed decision; and (2) disregarding the differences between the individuals assigned to the same category (which are treated as identical exemplars of a class or type, even though there is no a priori guarantee that all of them share a single and distinctive pattern of brain features).

Related to this, the fact that different algorithms are able to correctly identify sex with 80–90% accuracy does not ensure that these algorithms are providing the same outcomes or identifying the same underlying reality (Eliot et al., 2021; Hancox-Li, 2020; Joel, 2021). In fact, we observed that the%CC ranged between 86.3–90% and 58.7–65% (averages: 88.1 and 61.8%) in the raw and PCP datasets, but the actual percentages of subjects correctly classified by all five algorithms were 76.3 and 38.0%, respectively (Fig. 5). Moreover, 4% and 16.7% of the subjects were not correctly classified by any algorithm, whereas 19.7% and 45.3% of the subjects were only correctly classified by one or more algorithms in the raw and PCP datasets, respectively. These results indicate that the%CC scores provided by different algorithms are numerically similar but obtained on partially different subsets of individuals. The same results also indicate that each algorithm is correctly classifying several "kinds" of individuals (e.g., those properly recognized by all algorithms, those solely recognized by that particular algorithm, etc.) and, therefore, assigning the same label ("female" or "male") to more than one "kind" of subject (and, by implication, to several patterns of brain features or "brain types"). Of note, this within- and between-algorithm classification plurality is observed in both the raw and PCP datasets, although it seems to be quantitatively more important after statistically controlling the effects of TIV variation.

Finally, it should be noted that interpreting any%CC score in terms of "brain types" requires a great conceptual leap because%CC scores do not provide any information about the brains of females and

males or about which brain features could be considered the hallmarks of these alleged "male/ female" brain types (Eliot et al., 2021; Joel, 2021). This limitation is also shared by PCAM-based measures, which quantify *how different* the brains of males and females are, although they do not directly provide information about *where* these differences take place or about how they group together. Therefore, to answer these two questions two additional sets of analyses were conducted.

3.3.1. Which brain areas contribute the most to multivariate sex differences in GMVOL?

Boosted beta regression procedures were used to identify and quantify the relative importance of the brain features contributing the most to the PCAM scores yielded by each algorithm in each dataset. The degree of agreement between datasets and between algorithms in each dataset, regarding the number, identity, and relative importance of the predictors identified as relevant, was also assessed.

Overview and between-datasets agreement. As could be expected (More et al., 2020) and as Figs. 6 and 7 illustrate, the number, identity, and relative importance of the PCAM predictors identified in the raw and PCP datasets were clearly divergent. Specifically, up to 64 predictors were included between the raw and the PCP datasets, but only 22 of them were present in both. Accordingly, poor levels of mutual nominal agreement were observed (Δ=0.330 [0.157, 0.502]). In addition, the relative importance of these 64 predictors varied greatly depending on the dataset, resulting in agreement levels that were virtually zero (Supplementary Table 7F). When this comparison was performed only with the 22 predictors included in both datasets, higher but still "poor" levels of agreement were observed (e.g., ICC <0.5; see other agreement metrics in Supplementary Table 7G). Of note, in the raw dataset, the predictors that consistently showed higher importance were brain areas in which TIV explains a large amount of variance (see r^2 values in Fig. 6 and Supplementary Table 7H). This relationship was corroborated by statistically significant correlations (rho= 0.592 and rho= 0.571, $p<0.001$) between these r^2 values and two estimates of the predictors' relative importance across algorithms (ranks' averages and a multiplicative "rank of ranks"; Fig. 6 and Supplementary Table 7H). Conversely, in the PCP dataset, TIV-variation did not account for any variance in GMVOL (Fig. 7 and Supplementary Table 7H), and the predictors' relative importance was unrelated to these non-statistically significant r^2 values (rho= −0.119 and rho= −0.123, $p>0.05$; Fig. 7 and Supplementary Table 7I). Taken together, these results reveal that the most relevant brain areas in predicting the PCAM scores in the raw dataset were not the same, and, when they were, they did not have the same relative importance as in the PCP dataset, thus confirming that raw and TIV-adjusted measures of GMVOL provide information about two distinct constructs.

Between-algorithm agreement in the raw and in the PCP dataset. Major differences in the predictors' number, identity, and relative importance were also observed when comparing different algorithms within each dataset. Thus, in the raw dataset (Fig. 6), a total of 32 brain areas were identified as relevant predictors of PCAM scores. Different models included a different number of predictors (range: 9–19), and only three of them (AAL#21, left olfactory cortex; AAL#56 right Fusiform gyrus; AAL#98 right cerebellum 4,5) were included in all of them. Thus, although absolute nominal agreement was moderate (K_{HR}=0.505 [0.385, 0.625], multi-rater Δ=0.727 [0.639, 0.814]), the inspection of the within-class consistencies (WCC) revealed that agreement was primarily observed for predictors excluded from (WCC=0.809) the different regression models -and not from the predictors included in (WCC=0.182) them (Supplementary Table 7J). In a similar vein, the predictors' relative importance varied considerably between algorithms. In fact, the values of the predictors' coefficients in different models showed agreement/ reliability levels that were virtually zero (Supplementary Table 7L). Agreement increased but remained low when the predictors' relative importance was considered at the ordinal level. More specifically, the reliability of these ordinal estimates was larger than 0, but "poor" when based on the results of a single algorithm (ICC single-rating=0.289 [0.141, 0.478]), and only the average of these ordinal estimates yielded levels of agreement that can be considered "moderate"(Portney and Watkins, 2009) (ICC average-rating=0.671[0.450, 0.821], $p<0.001$; Supplementary Table 7L).

In the PCP dataset (Fig. 7), a total of 54 predictors were identified as relevant. As in the raw dataset, different algorithms included a different number of predictors (range: 9–41), and only four of them (AAL#9, left frontal mid orbital gyrus; AAL#72, right caudate nucleus; AAL#79, left Heschl gyrus; AAL#93, cerebellum crus 1) were included in all the regression models. Consequently, estimates of absolute agreement were low (K_{HR}=0.323 [0.215, 0.431], multi-rater Δ=0.508 [0.399, 0.615]) and primarily concerned those predictors excluded from (WCC=0.589) the different regression models -and not for the predictors included in (WCC=0.141) them (Supplementary Table 7K). As also observed in the raw dataset, the agreement between the values of the predictors' coefficients in different models did not statistically differ from zero (Supplementary Table 7M). Agreement between predictors increased, but remained low, when their relative importance was considered at the ordinal level. Thus, again paralleling the results observed in the raw dataset, the ordinal estimates obtained with any single algorithm showed "poor" reliability (ICC single-rating=0.312 [0.187, 0.457]; Supplementary Table 7M), and only the average of these ordinal estimates yielded agreement levels that can be considered "moderate"(Portney and Watkins, 2009) (ICC average-rating=0.694 [0.520, 0.812], $p<0.001$; Supplementary Table 7M).

Implications of the obtained results. We observed that the most relevant brain areas in predicting the PCAM scores yielded by distinct algorithms within each dataset were not the same, and, when they were, their relative importance was also quite different. These between-algorithm discrepancies may explain why different algorithms correctly classify different subsets of individuals (Fig. 5), and why different algorithms yield different-shaped PCAM distributions (Fig. 2) that result in multivariate sex differences that vary in size (Figs. 3 and 4). More specifically, because they differ in their statistical assumptions and operations, distinct algorithms rely on distinct brain features (Figs. 6 and 7) and assign different PCAM scores to the same subjects (Fig. 8A, Supplementary Table 8A-B; for dyadic between-algorithm comparisons, see Supplementary Table 8C-F). As these PCAM-variations are highly idiosyncratic (Fig. 8B), each individual occupies a different relative position in each PCAM distribution (Fig. 8C and D), and these distributions end up spreading dissimilarly within the PCAM range (Fig. 2). In the raw dataset, between-algorithm discrepancies are partially concealed by male-female differences in TIV that push their respective scores towards opposite sides of the PCAM range and boost sex differences and, hence, between-algorithm classification similarities (Fig. 5A) and correlations (Supplementary Fig. 6). However, when TIV effects are statistically controlled (as in the PCP dataset or by partialling out the TIV effects), between-algorithm discrepancies in classification (Fig. 5B) and PCAM scores (Fig. 8D; Supplementary Fig. 6) become evident.

Thus, despite working with identical data from the same individuals, the different algorithms tested in the present study do not provide directly exchangeable outcomes or identify a single, coherent, and reproducible subset of brain features as the source of the males-females multivariate differences in GMVOL (neither in the raw dataset or in the PCP dataset). These observations are consistent with the lack of agreement observed between the few studies that tried to identify the neuroanatomical features that could best distinguish the brains of females and males (Anderson et al., 2018; Sepehrband et al., 2018; Xin et al., 2019; Zhang et al., 2021; for a comparative review, see Eliot et al. 2021). Together, these sources of empirical evidence directly challenge a common interpretation of classification studies according to which, if distinct ML algorithms are able to very accurately "predict" sex from brain

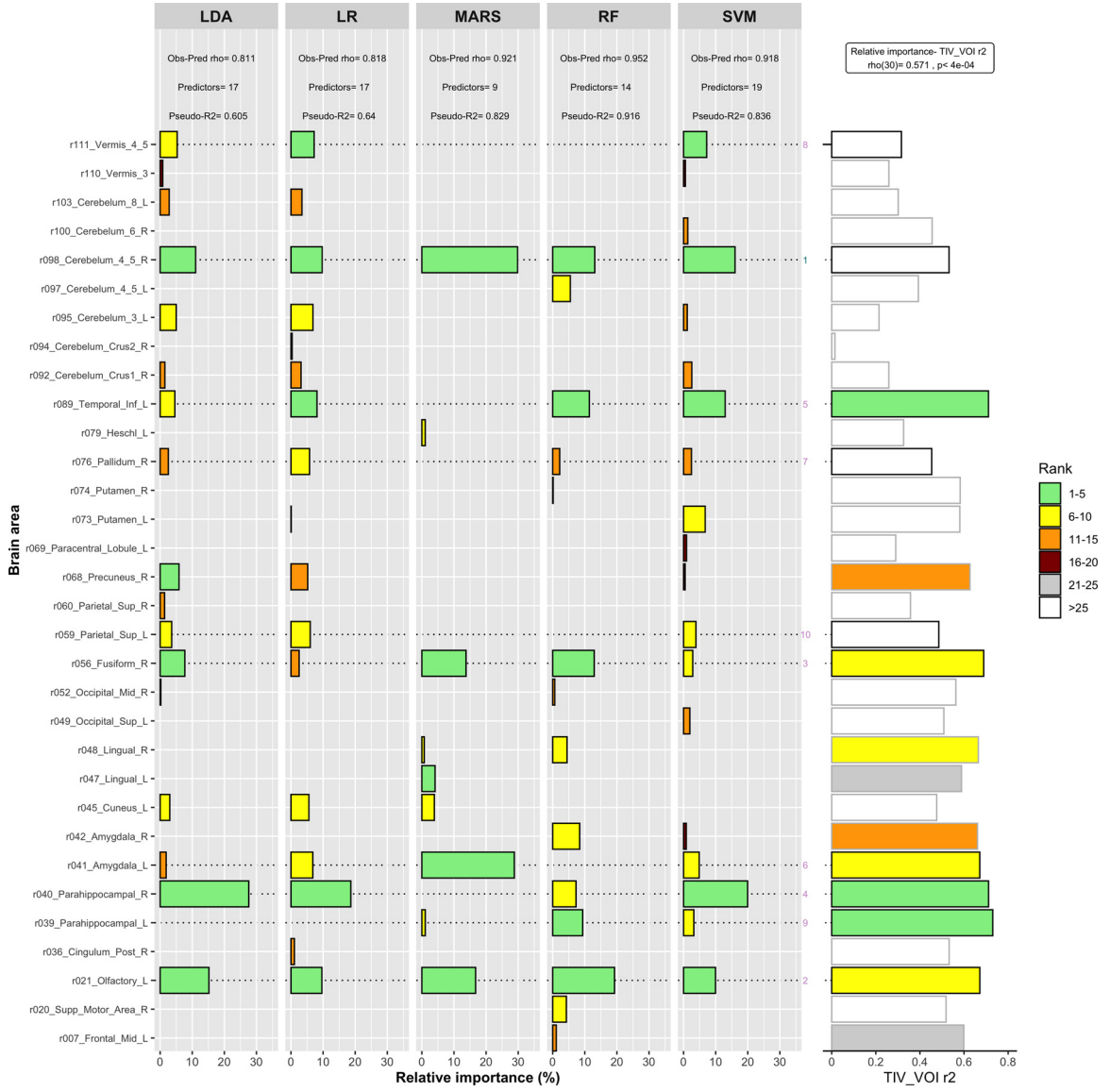


Fig. 6. Relative importance of PCAM predictors in the raw dataset. Plots depict the brain areas identified as relevant predictors of the PCAM scores yielded by each algorithm and their relative importance. The figure also depicts the top 10 predictors across all five algorithms according to their average rank values (the only predictor also found in the top 10 of the PCP dataset is highlighted in green). The right side of the plot depicts the proportion of variance (r2 value) explained by TIV in each of these brain regions. The correlation between the two sets of data and additional parameters of the boosted beta regressions employed to identify these predictors are also included. Additional details are provided in Supplementary Table 7A,C, E-G, H,J, and L.

features, all these algorithms *must be* identifying a reproducible constellation of brain differences that reliably assemble into two clearly distinguishable brain types ("male/ female brains"). In fact, this common assumption is unwarranted and, as our results clearly show, it is at odds with the currently available empirical evidence.

More specifically, just because different algorithms provide high and similar%CC scores, it cannot be directly assumed that these algorithms are identifying a single "sex-dimorphic" pattern of neuroanatomical features that reliably differs between males and females (and, therefore, as proving that "male/ female brains" actually exist). In fact, different sets of brain features are identified as the most relevant predictors of the PCAM yielded by different algorithms (within the present study but also across independent studies). This observation could be reflecting that there is no single a universal pattern of features that allow distinguishing the brain of females and males. However, these divergences could also be due to the inherent instability of the explanations provided by complex classification/ regression models (that is, they could be due to the "Rashomon effect"; see Breiman 2001, page 206; Hancox-Li 2020). Therefore, although the results of our regression analyses argue against interpreting the results of sex classification studies in terms of "male/female brains", these results cannot be interpreted as directly or definitively proving that these two brain types do not exist either (for a detailed discussion, see DelGiudice 2021).

In this regard, the "Rashomon effect" poses a major complication when trying to gain insight into how the predictors and the outcome are actually related (Fisher et al., 2019; Hancox-Li, 2020; Marx et al., 2019), and, therefore, when trying to ascertain whether or not there is a universal pattern of brain features that reliably distinguish the brains of females and males. However, the consequences of the "Rashomon effect" can be mitigated and more reliable inferences about the "true" predictors-outcome relationship can be obtained by combining (e.g., averaging) the explanations provided by different algorithms (for a more ample discussion, see Breiman 2001, DelGiudice 2021, Fisher et al. 2019, Hancox-Li 2020). Attending to this fact, as well as to the obtained ICC scores (which were substantially higher when obtained from average ratings, see Section 3.3.1.2), the ten predictors that

Fig. 7. Relative importance of PCAM predictors in the PCP dataset. Plots illustrate the relative importance of the brain areas identified as relevant predictors of the PCAM scores yielded by each algorithm. The figure also depicts the top 10 predictors across all five algorithms according to their average rank values (the only predictor also found in the top 10 of the raw dataset is highlighted in purple). The proportion of variance (r2 value) explained by TIV in each of these brain regions is depicted by the bars of the right side of the plot. The correlation between the two sets of data as well as other parameters of the boosted beta regressions employed to identify these predictors are also included. Further details are provided in Supplementary Table 7B,D, E-G, I, K, and M.

exhibited the highest average importance across all five algorithms in each dataset were identified (see Figs. 6 and 7) and included in our subsequent analyses.

3.3.2. Do multivariate sex differences in GMVOL stem from differences between a "male brain" and a "female brain"?

In the preceding section, we concluded that, due to the so-called "Rashomon effect" (Breiman, 2001; Hancox-Li, 2020), the observation that different ML algorithms rely on different sets of brain features to "predict sex" (in different samples or even within a single sample) may not suffice to definitely prove that the proposed "male/ female" brain types do not exist. However, synergistic evidence to reject these sex-specific brain types is provided by the observation that a single ML algorithm relies on different brain features when classifying different females or males of a single sample, hence assigning the same class label (Fig. 5) or even virtually identical PCAM scores to individuals exhibiting very different brain profiles (see examples in Figs. 9A and 10A; see converging evidence in Joel et al. (2018b).

More specifically, when accumulated differences in raw GMVOL (Euclidean distances) at the brain areas identified as relevant predictors of the LDA-PCAM scores in the raw dataset are considered, the differences between members of different sex categories are larger than those observed between members of the same sex category (ANOSIM $R = 0.455$ [0.363, 0.540], Supplementary Table 9A), and males and females tend to group into two separate clusters (Fig. 9B). These two clusters are homogeneous and robust, and they are only minimally perturbed when additional partitions are imposed (Fig. 9D). However, these clusters do not correspond to two "brain types" or sex-specific brain profiles. In fact, when the same individuals are partitioned based on the dissimilarity of their brain profiles (Spearman's distances), subjects cluster in a sex-unrelated manner (Fig. 9C and E) because the brain profiles similarities observed between members of different sex categories are equivalent to those observed between the members of each single sex category (ANOSIM $R = 0.049$ [0.006, 0.080], Supplementary Table 9B). In other words, males and females are clearly different and form two well-defined clusters regarding the total amount of GMVOL at the brain areas

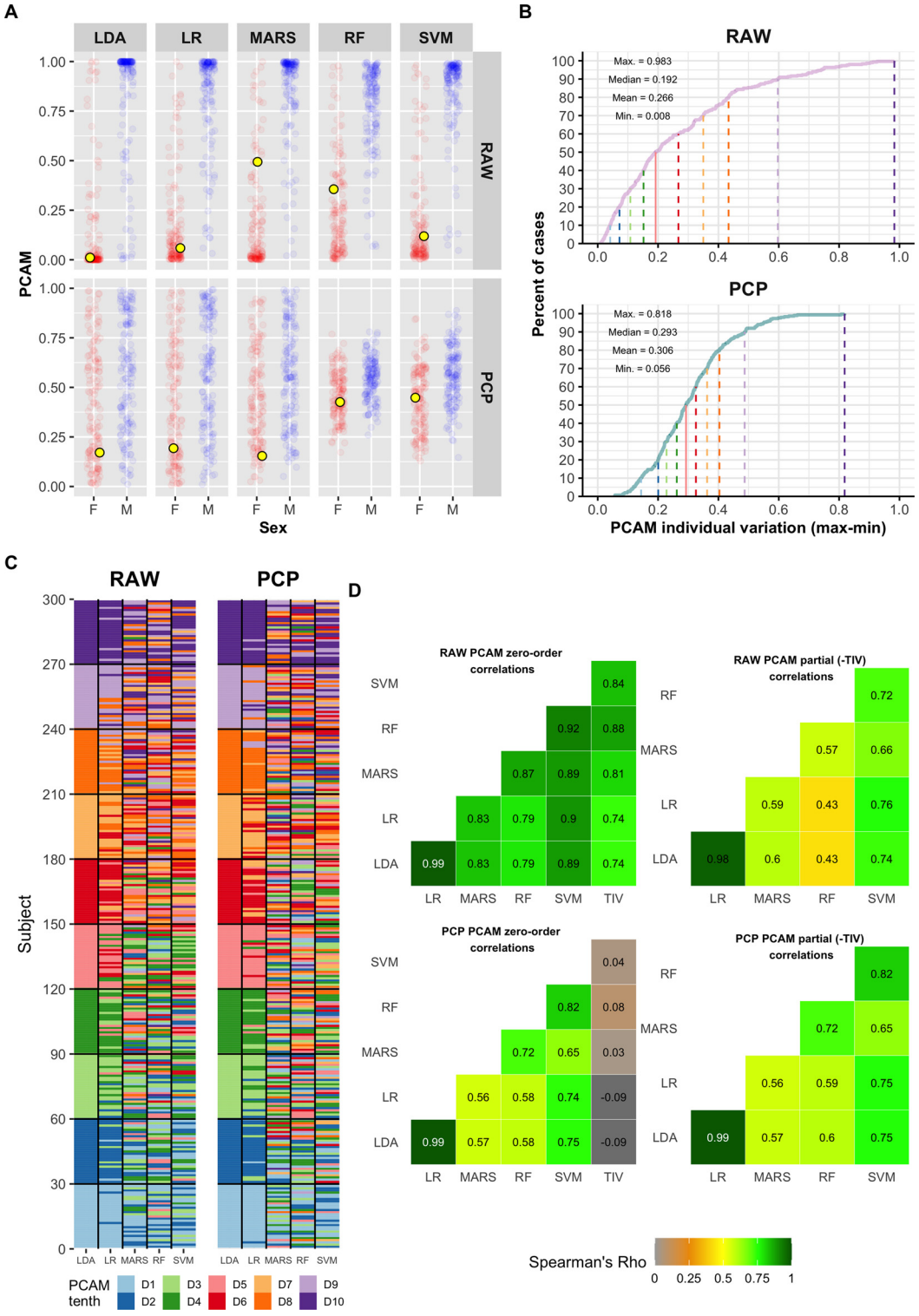


Fig. 8. Individual PCAM variation. A) Scatterplot showing (in the background) the female/ male (red/ blue, respectively) individual PCAM values yielded by each algorithm and dataset. To illustrate how different algorithms provide different PCAM scores, the values of a single subject are highlighted (yellow filled circles). B) Cumulative density functions of the maximum PCAM variation in the raw and the PCP datasets. C) Tiled heatmap illustrating how the same subject occupies different relative positions in each PCAM distribution of the raw/ PCP dataset. Each subject is depicted as a horizontal line colored according to the tenth on which it is located in each distribution. To ease visualization, in each dataset, subjects (vertical axis) were ordered according to their LDA-PCAM scores. D) Zero-order and partial (-TIV) Spearman correlations between the PCAM scores provided by each algorithm in the raw and PCP datasets. The full output of these analyses is provided in Supplementary Table 8A-F and in the accompanying Supplementary Fig. 6. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

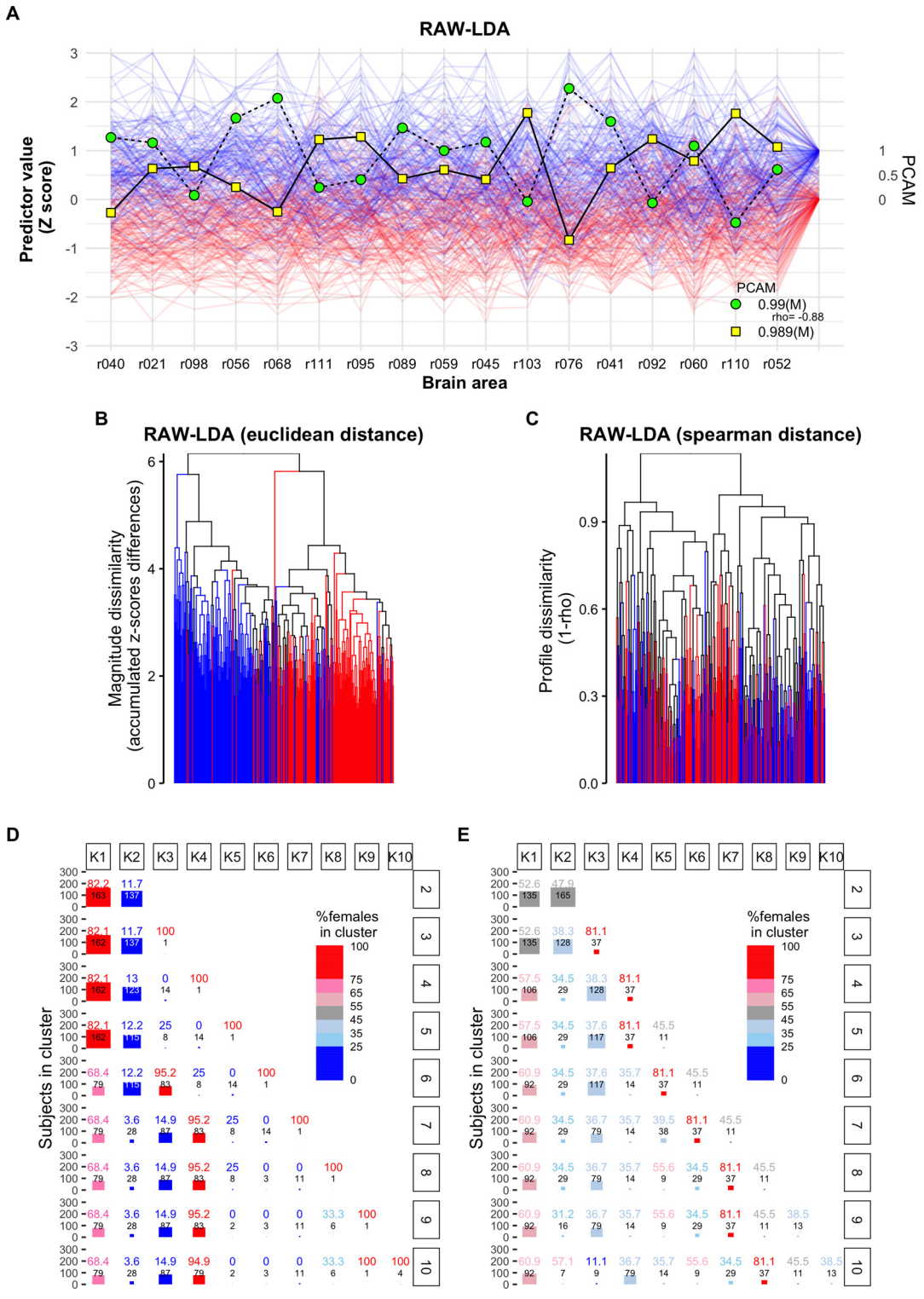


Fig. 9. Brain profiles in the raw dataset. (A) In the background, a "spaghetti" plot displays the individual values (females in red, males in blue) in all the brain areas identified as relevant predictors of PCAM scores (see Fig. 6). Two cases (green/ yellow dots) are highlighted to illustrate how the same classification category (and even virtually identical PCAM scores) can be achieved by different and brain profiles. (B-C) Hierarchical agglomerative clustering based on Euclidean and Spearman distances, respectively. Branches are colored according to the sex composition of the emerging aggregations (red=only females, blue=only males, black=males and females). (D-E) Clusters' size and composition. Dendrograms displayed in panels B and C were cut at appropriate heights to obtain 2–10 clusters. The composition of these clusters (K2-K10, horizontal axis) was analyzed in terms of the proportion of females (rectangles' color; large numbers) and the cluster's size (rectangle area; small black numbers). Similar results were obtained when using the other algorithms and with a larger (116)/ smaller (5) number of predictors (Supplementary Figs. 7–11). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

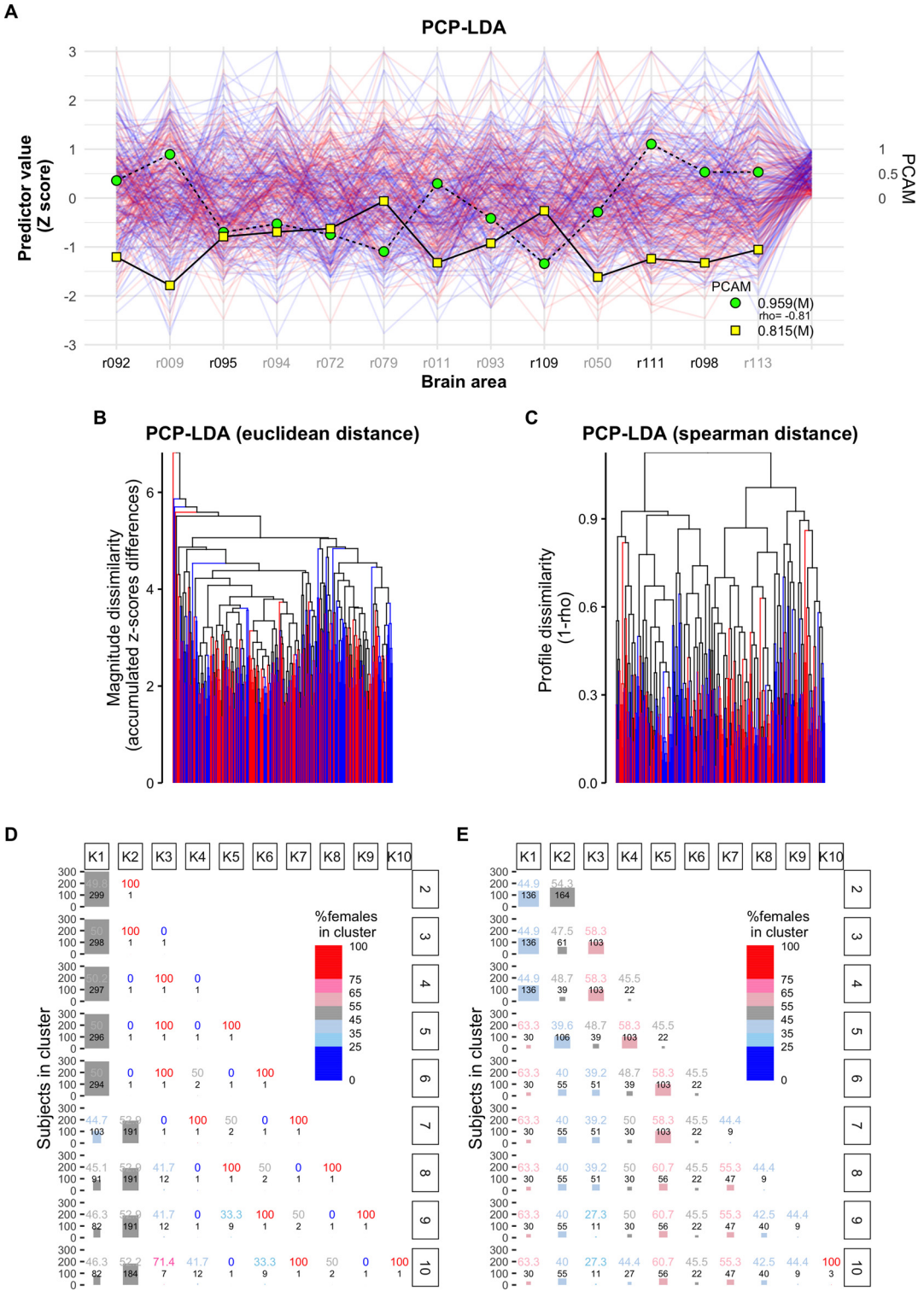


Fig. 10. Brain profiles in the PCP dataset. (A) In the background, a "spaghetti" plot displays the individual values (females in red, males in blue) in all the brain areas identified as relevant predictors of PCAM scores (see Fig. 7; gray labels highlight predictors with negative regression weights, see Supplementary Table 7B). To illustrate how the same classification category and similar PCAM scores can be achieved by different brain profiles, two cases (green/ yellow dots) are highlighted. (B-C) Hierarchical agglomerative clustering based on Euclidean and Spearman distances, respectively. Branches are colored according to the sex composition of the emerging aggregations (red=only females, blue=only males, black=males and females). (D-E) Clusters' size and composition. Dendrograms displayed in panels B and C were cut to appropriate heights to obtain 2–10 clusters. The composition of these clusters (K2-K10, horizontal axis) was analyzed in terms of the proportion of females (rectangles' color; large numbers) and the cluster's size (rectangle area; small black numbers). Similar results were obtained when using the other algorithms and with a larger (116)/ smaller (5) number of predictors (Supplementary Figs. 12–16). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

identified as relevant predictors of LDA-PCAM scores (Euclidean distances), but neither males nor females show a homogeneous profile for the relative distribution of GMVOL in these brain areas. Consequently, as measured by their Spearman's distances, the brain profiles of a randomly picked female and a randomly picked male resemble each other as much as they resemble the brain profiles of other females or males, respectively.

As illustrated in Supplementary Figs. 7–11 and Supplementary Table 9A-H, this pattern of results is reproduced for each ML algorithm tested in the present study. Moreover, the same results are also observed when using: (1) only the five areas more directly related to the PCAM scores provided by each algorithm as predictors (hence ruling out that the reduced similarity between the brain profiles of members of the same sex category resulted from a "too large" number of relevant predictors in some of the models); (2) the 10 predictors exhibiting the highest average importance across all algorithms (which should allow more accurate and reliable inferences; see the last paragraph of Section 3.3.1.3 and also Breiman (2001, DelGiudice (2021), Hancox-Li (2020); and (3) the 116 brain areas of the AAL atlas (thus assessing brain profiles' similarities in a way that is totally independent from the predictors' importance estimated in the regression analyses). Similar results were also observed when TIV-variation was statistically controlled (PCP dataset; Fig. 10; Supplementary Figs. 12–16 and Supplementary Table 9A–H), although in this case, both the accumulated differences and the brain profiles' similarities are basically unrelated to sex categories (e.g., PCAM-LDA in the PCP dataset: ANOSIM $R_{Euclidean}$= 0.017 [−0.018, 0.004]; ANOSIM $R_{Spearman}$= 0.025 [−0.014, 0.050]; Supplementary Table 9A-B).

Therefore, it can be concluded that: (1) As shown throughout the present study, the size of multivariate sex differences in raw and TIV-adjusted GMVOL are "large" and "small", respectively; and (2) Regardless of their size, these differences do not arise from divergences between a "typical male" and a "typical female" brain profile, but from divergences between multiple and idiosyncratic brain profiles that seem to be loosely related to sex categories.

4. Conclusions

When the output of ML algorithms is not discretized, multivariate information about the brains of females and males can be condensed in a single continuum (Lippa and Connelly, 1990; Zhang et al., 2021). The present study shows that, by assessing how females and males differentially occupy this empirically-defined unidimensional space, the size of their multivariate differences can be estimated on a standardized [0,1] scale or in ordinal/ distributional terms. Used in this way, the PCAM continuum -and other similar ones (Phillips et al., 2019; van Eijk et al., 2021)- offers an alternative strategy to investigate sex effects in the brain that can be easily applied to data from different neuroimaging modalities (or extended to other research domains; e.g., the assessment of multivariate differences between healthy and clinical populations).

In the present study, PCAM measures were employed to thoroughly describe males' and females' distributions and quantify the size of their multivariate differences in GMVOL. As mentioned in the introduction section, PCAM measures make it possible to treat females and males as two empirical distributions that can be explored with a wide variety of statistical techniques, among which we highlight the shift-function and the CDF-based analyses illustrated in the present study. The informational richness of these methods to describe (Fig. 2) and compare males' and females' distributions (Fig. 3) and their differences (Fig. 4) contrasts with the informational emptiness of%CC scores (Fig. 1), which do not really describe or compare females and males or quantify the size of their differences (for a broader discussion, see the opening of Section 3.3). Moreover, the methods adopted here do not use means or any other summary statistics, but rather they explore the entire female and male distributions to quantify how and by how much they differ in location, variability, and shape. This approach avoids reducing the study of sex differences and similarities to simple comparisons between "the average male" and "the average female" (which far too often are used to make unwarranted generic statements and conclusions about all females and all males), and instead it offers "[…] *the opportunity to get a deeper, more accurate and more nuanced understanding of data*" (Rousselet et al., 2017, page 1738).

On the other hand, all our statistical analyses indicated that TIV variation is a major determinant of the size of the multivariate sex differences in GMVOL. Thus, although the algorithm used to calculate the PCAM scores influenced the exact value of these estimates, sex differences were consistently "large" in the raw dataset, but "small" once TIV-variation was statistically controlled (e.g., the across-algorithm averages for the typical male-female pairwise difference were 67.2 and 12.9 POMP, respectively; see Fig. 4). More specifically, we observed that male-female differences in TIV push their respective scores towards opposite sides of the PCAM continuum (Fig. 2), thus boosting the estimated size of their multivariate differences in GMVOL (Figs. 3 and 4) and, thereby, the%CC scores (Fig. 1) as well as between-algorithm classification similarities and PCAM inter-correlations (Figs. 5 and 8). These observations align with previous findings indicating that TIV-adjustment greatly reduces the size of univariate sex differences in GMVOL (e.g., Fjell et al. 2009, Pintzka et al. 2015, Ritchie et al. 2018, Sanchis-Segura et al. 2019), as well as the neuroanatomical distinctiveness of the brains of females and males at the multivariate level (More et al., 2020; Sanchis-Segura et al., 2020). In this regard, it should be noted that there seems to be an increasing consensus about the consideration of TIV variation as a confound and, therefore, the need to remove its influence before measuring sex differences (e.g., Barnes et al. 2010, Fjell et al. 2009, Jäncke et al. 1997, Leonard et al. 2008, Pintzka et al. 2015, Ritchie et al. 2018). However, this position is not unanimous, and, in fact, most previous sex classification studies (e.g., Anderson et al. 2018, Luo et al. 2019, Rosenblatt 2016, Sepehrband et al. 2018) have used raw volumetric measurements as predictors, a methodological decision that seems to have influenced their results (i.e., very high%CC scores) and conclusions. Therefore, here we propose that, whenever possible, neuroanatomical sex differences should be estimated in both raw and TIV-adjusted measures. This dual assessment provides a more complete perspective than what can be obtained when only using raw or TIV-adjusted GMVOL estimates and it makes possible to parse out the relative contribution of "direct" or "local" sex effects from those attributable to gross morphological differences between females and males.

It is noteworthy that, although raw and TIV-adjusted data provide a very different portrait of the multivariate sex differences in GMVOL, they both show that these differences do not arise from a specific pattern of differences in a few key brain areas. Consequently, they both lead to the conclusion that there are not two brain "types". Thus, depending on whether raw or TIV-adjusted GMVOL estimates are considered, groups of males and females may differ greatly (or not) in their respective amounts of GMVOL in specific brain areas, and, when the accumulated differences in all these areas (Euclidean distances) are calculated by adding them up, the results of these sums may (or may not) clearly differentiate males and females into two separate clusters. However, when the results of these overall sums do not differentiate males and females (TIV-adjusted data; Fig. 10, panels B and D), but also when they clearly do (raw data; Fig. 9, panels B and D), the involved summands seem to be weighted very differently for the different members of each of these two sex categories. Thus, neither males nor females show a homogeneous profile for the relative distribution of GMVOL in these brain areas (Panel A of Figs. 9 and 10), and the resemblance (Spearman distances) between the brain profiles of randomly picked male or female pairs is not higher than the similarity observed between randomly picked female-male pairs (Figs. 9 and 10, panels C and E). Therefore, and because the same pattern of results was observed under 24 experimental conditions involving different algorithms, datasets, and/or predictors, we conclude that the multivariate male-female differences in GMVOL do not represent divergences between a "typical male" and a

"typical female" neuroanatomical profile reproduced in all or in most individuals in each sex category. Accordingly, we also conclude that, at least at the neuroanatomical level, the brains of females and males are not sexually dimorphic (literally "two shapes"), and there are no "male/female brains" (for a similar conclusion, see Eliot et al. 2021, Joel 2021, Joel et al. 2018a, Weis et al. 2020, Zhang et al. 2021). Similarly, and although multivariate male/female differences in the brain can be summarized with a single and continuous score, the brains of females and males are not aligned along a "female-male" continuum, at least not one that univocally maps to discrete neuroanatomical features (for a similar conclusion, see Joel 2021, 2020).

Given these observations, we think the PCAM-based estimates of the multivariate sex differences in GMVOL are probably better interpreted as a summary of heterogenous patterns of differences between several subsets of males and females that diverge in distinct brain features. By implication, we conclude that the PCAM continuum -and other similar measures- provides a reduced metric space that is useful for comparing females and males and estimating their brain differences, but a single continuum is clearly insufficient to properly describe and adequately conceptualize the complex and highly idiosyncratic sex-associated effects in the brains of females and males. Therefore, just like other multivariate effect size indexes (Del Giudice, 2019), PCAM-based measures might be more useful when summarizing a reduced, coherent, and theoretically-justified set of variables (whose within-profile differences can be interpreted) than when calculated from a large number of loosely related/ arbitrarily chosen brain features.

Finally, our results also show that -because they differ in their a priori assumptions and internal operations, different ML algorithms may produce different outcomes. Therefore, comparing or combining the results of several algorithms should lead to more reliable and valid conclusions than those extracted from just one. Moreover, the algorithm/s chosen becomes a critical methodological decision that should be reported in detail and carefully considered when summarizing the results of different studies. A similar caution should be also applied to the results and conclusions of the present study because they cannot (and should not) be unlinked from the methods used to obtain them. Thus, although we tried to assess our hypotheses under a wide range of conditions and employing different methods with the aim of providing convergent evidence, any of our methodological decisions can (and probably should) be regarded as a potential limitation to the generalizability of the findings and conclusions presented.

Ethics approval and consent to participate

This study was carried out in accordance with the recommendations of the ethical standards of the American Psychological Association. In accordance with the Declaration of Helsinki, all subjects gave written informed consent prior to participating.

Data availability statement

This study was primarily conducted using data from the open source 1200 Subject Release (S1200) of the Human Connectome Project (HCP). The access to this sample should be directly requested to the Washington University - University of Minnesota Consortium of the Human Connectome Project (WU-Minn HCP).

Declaration of Competing Interest

The authors declare no competing interests.

Credit authorship contribution statement

Carla Sanchis-Segura: Visualization, Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Naiara Aguirre:** Project administration, Writing – review & editing. **Álvaro Javier Cruz-Gómez:** Project administration, Writing – review & editing. **Sonia Félix:** Project administration, Writing – review & editing. **Cristina Forn:** Visualization, Conceptualization, Writing – original draft, Writing – review & editing.

Acknowledgments

Data were provided 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.

We are very grateful to Dr. Rand R. Wilcox for his expert advice in quantile comparisons and for developing the *bwquantile* (*between by within based in quantiles*) function specifically for the present study. We also thank Dr. A. Martin for his guidance on agreement measures and for developing the software that allowed the calculation of the delta coefficient of agreement.

This research was supported by a grant (PID2019–106793RB-I00/AEI / 10.13039/501100011033) provided by Ministerio de Ciencia e Innovación to CF and CS-S and a grant (UJI B2020–02) awarded to CF and CS-S. N.A was supported by an FPU grant from the Ministerio de Educacion [FPU16/01525] and S.F. by an FPI grant from UJI (PREDOC/2020/22). These funding sources did not play any role in designing the study or in the collection, analysis, and interpretation of the data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119343.

References

Ali, A., Shamsuddin, S.M., Ralescu, A.L., 2013. Classification with class imbalance problem: a review. Int. J. Adv. Soft Comput. Appl..

Ali, S., Smith-Miles, K.A., 2006,. Improved support vector machine generalization using normalized input space. In: Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/11941439-40.

Altman, D.G., Royston, P., 2006. The cost of dichotomising continuous variables. Br. Med. J. doi:10.1136/bmj.332.7549.1080.

Anderson, N.E., Harenski, K.A., Harenski, C.L., Koenigs, M.R., Decety, J., Calhoun, V.D., Kiehl, K.A., 2018. Machine learning of brain gray matter differentiates sex in a large forensic sample. Hum. Brain Mapp. doi:10.1002/hbm.24462.

Andrés, A.M., Marzo, P.F., 2004. Delta: a new measure of agreement between two raters. Br. J. Math. Stat. Psychol. doi:10.1348/000711004849268.

Andrés, M., Hernández, Á., 2021 Multi-rater delta: extending the delta nominal measure of agreement between two raters to many raters. J Stat Comput Simul, doi:10.1080/00949655.2021.2013485

Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S.M.D., Lehmann, M., Hobbs, N., Clarkson, M.J., MacManus, D.G., Ourselin, S., Fox, N.C., 2010. Head size, age and gender adjustment in MRI studies: a necessary nuisance? Neuroimage doi:10.1016/j.neuroimage.2010.06.025.

Benjamini, Y., Hochberg, Y., 2018. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16, 199–231.

Bzdok, D., 2017. Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. doi:10.3389/fnins.2017.00543.

Bzdok, D., Ioannidis, J.P.A., 2019. Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci. doi:10.1016/j.tins.2019.02.001.

Cahill, L., 2014. Equal ≠ the same: sex differences in the human brain. Cerebrum Dana Forum Brain Sci. 2014, 5.

Cahill, L., 2006. Why sex matters for neuroscience. Nat. Rev. Neurosci. doi:10.1038/nrn1909.

Callaert, H., 1999. Nonparametric hypotheses for the two-sample problem. J. Stat. Educ. 7.

Canty, A., Ripley, B., 2020. boot: bootstrap R (S-Plus) functions.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: a survey on methods and metrics. Electron doi:10.3390/electronics8080832.

Chekroud, A.M., Ward, E.J., Rosenberg, M.D., Holmes, A.J., 2016. Patterns in the human brain mosaic discriminate males from females. Proc. Natl. Acad. Sci. U. S. A. doi:10.1073/pnas.1523888113.

Chen, I.Y., Joshi, S., Ghassemi, M., Ranganath, R., 2021. Probabilistic machine learning for healthcare. Annu. Rev. Biomed. Data Sci. 4, 393–415. doi:10.1146/annurev-biodatasci-092820-033938.

Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143. doi:10.1111/j.1442-9993.1993.tb00438.x.

Clayton, J.A., 2018. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol. Behav. 187, 2–5. doi:10.1016/J.PHYSBEH.2017.08.012.

Cliff, N., 1993. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol. Bull. doi:10.1037/0033-2909.114.3.494.

Cohen, J., 1983. The cost of dichotomization. Appl. Psychol. Meas. doi:10.1177/014662168300700301.

Cohen, P., Cohen, J., Aiken, L.S., West, S.G., 1999. The problem of units and the circumstance for POMP. Multivar. Behav. Res. doi:10.1207/S15327906MBR3403_2.

Cook, Di., Lee, E.K., Majumder, M., 2016. Data visualization and statistical graphics in big data analysis. Annu. Rev. Stat. Appl. doi:10.1146/annurev-statistics-041715-033420.

Del Giudice, M., 2019. Measuring sex differences and similarities, in: vanderLaan, D.P.; Wong, W.I. (Ed.), Gender and Sexuality Development: contemporary Theory and Research. New York. ISBN: 303084272X

DelGiudice, M., 2021. The prediction-explanation fallacy: a pervasive problem in scientific applications of machine learning. PsyArXiv 2021, 1–18. doi:10.31234/OSF.IO/4VQ8F.

Eliot, L., Ahmed, A., Khan, H., Patel, J., 2021. Dump the "dimorphism": comprehensive synthesis of human brain studies reveals few male-female differences beyond size. Neurosci. Biobehav. Rev. doi:10.1016/j.neubiorev.2021.02.026.

Feis, D.L., Brodersen, K.H., von Cramon, D.Y., Luders, E., Tittgemeyer, M., 2013. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. Neuroimage 70, 250–257. doi:10.1016/j.neuroimage.2012.12.068.

Ferrari, S.L.P., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. J. Appl. Stat. doi:10.1080/0266476042000214501.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 1–81.

Fjell, A.M., Westlye, L.T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Agartz, I., Salat, D.H., Greve, D.N., Fischl, B., Dale, A.M., Walhovd, K.B., 2009. Minute effects of sex on the aging brain: a multisample magnetic resonance imaging study of healthy aging and Alzheimer's disease. J. Neurosci. doi:10.1523/JNEUROSCI.0115-09.2009.

Gamer, M., Lemon, J., Fellows, I., Singh, P., 2019. irr: various coefficients of interrater reliability and agreement.R package version 0.84.1.

García, V., Sánchez, J.S., Mollineda, R.A., Sotoca, R.A.J.M., 2007. The class imbalance problem in pattern classification and learning. Data Eng..

Grissom, R.J., Kim, J.J., 2012. Effect sizes for research: univariate and multivariate applications. Effect Sizes for Research: Univariate and Multivariate Applications, 2nd ed. Routledge Second EditionMultivariate application tests. doi:10.4324/9780203803233.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. doi:10.20982/tqmp.08.1.p023.

Hancox-Li, L., 2020. Robustness in machine learning explanations: does it matter? Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pages 640-647. doi:10.1145/3351095.3372836.

Handcock, M.S., Morris, M., 1999. Relative distribution methods in the social sciences, Springer, New York, NY. ISBN:978-0-387-98778-1

Harrell, F.E., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic. Springer International Publishing Switzerland.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning The Elements of Statistical Learningdata Mining, Inference, and Prediction, 2nd ed. Springer Series in Statistics doi:10.1007/978-0-387-84858-7.

Hochberg, Y., 1988. A sharper bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802. doi:10.1093/biomet/75.4.800.

Hubert, L., 1977. Kappa revisited. Psychol. Bull. doi:10.1037/0033-2909.84.2.289.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C., Verma, R., 2014. Sex differences in the structural connectome of the human brain. Proc. Natl. Acad. Sci. U. S. A. 111, 823–828. doi:10.1073/PNAS.1316909110.

Jäncke, L., Staiger, J.F., Schlaug, G., Huang, Y., Steinmetz, H., 1997. The relationship between corpus callosum size and forebrain volume. Cereb. Cortex doi:10.1093/cercor/7.1.48.

Joel, D., 2021. Beyond the binary: rethinking sex and the brain. Neurosci. Biobehav. Rev. doi:10.1016/j.neubiorev.2020.11.018.

Joel, D., 2020. Beyond sex differences and a male–female continuum: mosaic brains in a multidimensional space. Handb. Clin. Neurol. doi:10.1016/B978-0-444-64123-6.00002-3.

Joel, D., 2011. Male or female? Brains are intersex. Front. Integr. Neurosci. doi:10.3389/fnint.2011.00057.

Joel, D., Persico, A., Averbuch, A., Meilijson, I., Oligschläger, S., Salhov, M., Berman, Z., 2018a. Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa. Front. Hum. Neurosci. doi:10.3389/fnhum.2018.00399.

Joel, D., Persico, A., Salhov, M., Berman, Z., Oligschläger, S., Meilijson, I., Averbuch, A., 2018b. Analysis of human brain structure reveals that the brain "types" typical of males are also typical of females, and vice versa. Front. Hum. Neurosci. doi:10.3389/fnhum.2018.00399.

Karatzoglou, A., Hornik, K., Smola, A., Zeileis, A., 2004. kernlab - an S4 package for kernel methods in R. J. Stat. Softw. doi:10.18637/jss.v011.i09.

Kassambara, A., Mundt, A., 2020. factoextra: extract and visualize the results of multivariate data analyses.

Kendall, M.G., Smith, B.B., 1939. The problem of m rankings. Ann. Math. Stat. doi:10.1214/aoms/1177732186.

Kiang, M.Y., 2003. A comparative assessment of classification methods. Decis. Support Syst. doi:10.1016/S0167-9236(02)00110-0.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. doi:10.1016/j.jcm.2016.02.012.

Leonard, C.M., Towler, S., Welcome, S., Halderman, L.K., Otto, R., Eckert, M.A., Chiarello, C., 2008. Size matters: cerebral volume influences sex differences in neuroanatomy. Cereb. Cortex doi:10.1093/cercor/bhn052.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News.

Lin, L.I.K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics doi:10.2307/2532051.

Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.

Lippa, R., Connelly, S., 1990. Gender diagnosticity: a new bayesian approach to gender-related individual differences. J. Pers. Soc. Psychol. 59, 1051–1065. doi:10.1037/0022-3514.59.5.1051.

Lipton, Z.C., 2018. The mythos of model interpretability. Commun. ACM doi:10.1145/3233231.

Liu, D., Johnson, H.J., Long, J.D., Magnotta, V.A., Paulsen, J.S., 2014. The power-proportion method for intracranial volume correction in volumetric imaging analysis. Front. Neurosci. doi:10.3389/fnins.2014.00356.

Lombardo, M.V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., Bullmore, E.T., Baron-Cohen, S., 2012. Fetal testosterone influences sexually dimorphic gray matter in the human brain. J. Neurosci. 32, 674–680. doi:10.1523/JNEUROSCI.4389-11.2012.

Luo, Z., Hou, C., Wang, L., Hu, D., 2019. Gender identification of human cortical 3-D morphology using hierarchical sparsity. Front. Hum. Neurosci. 13, 29. doi:10.3389/fnhum.2019.00029.

Handcock, M., 1998. Relative distribution methods. Sociol. Methodol. doi:10.1111/0081-1750.00042.

MacCallum, R.C., Zhang, S., Preacher, K.J., Rucker, D.D., 2002. On the practice of dichotomization of quantitative variables. Psychol. Methods doi:10.1037/1082-989X.7.1.19.

Mair, P., Wilcox, R., 2018. Robust statistical methods using WRS2. J. Stat. Softw. VV, forthcoming.

Maney, D.L., 2015. Just like a circus: the public consumption of sex differences. Curr. Top. Behav. Neurosci. doi:10.1007/7854_2014_339.

Marx, C.T., Calmon, F.D.P., Ustun, B., 2019. Predictive multiplicity in classification. In: Proceedings of the 37th International Conference Machine Learning ICML 2020 PartF16814, pp. 6721–6730. doi:10.48550/arxiv.1909.06677.

Mayr, A., Weinhold, L., Hofner, B., Titze, S., Gefeller, O., Schmid, M., 2018. The betaboost package—a software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data. Int. J. Epidemiol. doi:10.1093/ije/dyy093.

McCarthy, M.M., 2015. Incorporating sex as a variable in preclinical neuropsychiatric research. Schizophr. Bull. doi:10.1093/schbul/sbv077.

More, S., Eickhoff, S.B., Caspers, J., Patil, K.R., 2020. Confound removal and normalization in practice: a neuroimaging based sex prediction case study. Lecture Notes Computer Science (Including Subseries Lecture Notes Artificial Intelligence Lecture Notes Bioinformatics) 12461 LNAI, 3–18. doi:10.1007/978-3-030-67670-4_1.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning ICML, pp. 625–632. doi:10.1145/1102351.1102430.

O'Connor, C., Joffe, H., 2014. Gender on the brain: a case study of science communication in the new media environment. PLoS One doi:10.1371/journal.pone.0110830.

Oksanen, J., Blanchet, G., Friendly, M., Kondt, R., Legendre, P., McGlin, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.., Szoecs, E., Wagner, H., 2020. vegan: community ecology package.

Pastore, M., 2018. Overlapping: a R package for estimating overlapping in empirical distributions. J. Open Source Softw. doi:10.21105/joss.01023.

Pastore, M., Calcagnì, A., 2019. Measuring distribution similarities between samples: a distribution-free overlapping index. Front. Psychol. doi:10.3389/fpsyg.2019.01089.

Phillips, O.R., Onopa, A.K., Hsu, V., Ollila, H.M., Hillary, R.P., Hallmayer, J., Gotlib, I.H., Taylor, J., Mackey, L., Singh, M.K., 2019. Beyond a binary classification of sex: an examination of brain sex differentiation. Psychopathol. Genotype. J. Am. Acad. Child Adolesc. Psychiatry doi:10.1016/j.jaac.2018.09.425.

Pinares-Garcia, P., Stratikopoulos, M., Zagato, A., Loke, H., Lee, J., 2018. Sex: a significant risk factor for neurodevelopmental and neurodegenerative disorders. Brain Sci. 8. doi:10.3390/BRAINSCI8080154, 2018Page 154 8, 154.

Pintzka, C.W.S., Hansen, T.I., Evensmoen, H.R., Håberg, A.K., 2015. Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a HUNT MRI study. Front. Neurosci. doi:10.3389/fnins.2015.00238.

Portney, L.G., Watkins, M.P., 2009. Foundations of Clinical Research: Applications to Practice, 3rd ed. Pearson/Prentice Hall, Upper Saddle River, NJ.

R Core Team, 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing,. Vienna, Austria. URL http://www-R-project.org.

Ribeiro, M.T., Singh. S., Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv Prepr. arXiv1606.05386.

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. doi:10.3389/fnhum.2014.00650.

Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., Harris, M.A., Alderson, H.L., Hunter, S., Neilson, E., Liewald, D.C.M., Auyeung, B., Whalley, H.C., Lawrie, S.M., Gale, C.R., Bastin, M.E., McIntosh, A.M., Deary, I.J., 2018. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. Cereb. Cortex doi:10.1093/cercor/bhy109.

Rosenblatt, J., 2016. Multivariate revisit to "sex beyond the genitalia. Proc. Natl. Acad. Sci. U. S. A. doi:10.1073/pnas.1523961113.

Rousselet, G.A., Pernet, C.R., Wilcox, R.R., 2017. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. Eur. J. Neurosci. doi:10.1111/ejn.13610.

Sanchis-Segura, C., Ibañez-Gual, M.V., Adrián-Ventura, J., Aguirre, N., Gómez-Cruz, Á.J., Avila, C., Forn, C., 2019. Sex differences in gray matter volume: how many and how large are they really? Biol. Sex Differ. doi:10.1186/s13293-019-0245-7.

Sanchis-Segura, C., Ibañez-Gual, M.V., Aguirre, N., Gómez-Cruz, Á.J., 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. doi:10.1038/s41598-020-69361-9.

Schmid, M., Wickler, F., Maloney, K.O., Mitchell, R., Fenske, N., Mayr, A., 2013. Boosted beta regression. PLoS One doi:10.1371/journal.pone.0061623.

Sepehrband, F., Lynch, K.M., Cabeen, R.P., Gonzalez-Zacarias, C., Zhao, L., D'Arcy, M., Kesselman, C., Herting, M.M., Dinov, I.D., Toga, A.W., Clark, K.A., 2018. Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. Neuroimage 172, 217–227. doi:10.1016/j.neuroimage.2018.01.065.

Milborrow S. . Derived from mda:mars by Trevor Hastie, With, R.T.U.A.M.F. utilities, Wrapper., T.L. leaps, 2019. earth: multivariate adaptive regression splines. R package version 5.1.2.

Tofallis, C., 2014. Add or multiply? A tutorial on ranking and choosing with multiple criteria. INFORMS Trans. Educ. doi:10.1287/ited.2013.0124.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage doi:10.1006/nimg.2001.0978.

van Eijk, L., Zhu, D., Couvy-Duchesne, B., Strike, L., Lee, A., Hansell, N., Thompson, P., de Zubicaray, G.I., McMahon, K.L., Wright, M., Zietsch, B., 2021. Are sex differences in human brain structure associated with sex differences in behaviour? Psychol. Sci. doi:10.31234/osf.io/8fcve.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. Neuroimage doi:10.1016/j.neuroimage.2013.05.041.

Van Putten, M.J.A.M., Olbrich, S., Arns, M., 2018. Predicting sex from brain rhythms with deep learning. Sci. Rep. doi:10.1038/s41598-018-21495-7.

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S fourth edition by, WORLD. 10.2307/2685660

Wang, L., Shen, H., Tang, F., Zang, Y., Hu, D., 2012. Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain: an MVPA approach. Neuroimage 61, 931–940. doi:10.1016/j.neuroimage.2012.03.080.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p -values: context, process, and purpose. Am. Stat. 70, 129–133. doi:10.1080/00031305.2016.1154108.

Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T.T., Eickhoff, S.B., 2020. Sex classification by resting state brain connectivity. Cereb. Cortex 30, 824–835. doi:10.1093/cercor/bhz129.

Wilcox, R.R., Rousselet, G.A., 2018. A guide to robust statistical methods in neuroscience. Curr. Protoc. Neurosci. doi:10.1002/cpns.41.

Williams, C.M., Peyre, H., Toro, R., Ramus, F., 2021. Neuroanatomical norms in the UK biobank: the impact of allometric scaling, sex, and age. Hum. Brain Mapp. 42, 4623–4642. doi:10.1002/hbm.25572.

Xin, J., Zhang, Y., Tang, Y., Yang, Y., 2019. Brain differences between men and women: evidence from deep learning. Front. Neurosci. 13, 185. doi:10.3389/fnins.2019.00185.

Zhang, L., Huang, et al., 2021. The human brain is best described as being on a female/male continuum: evidence from a neuroimaging connectivity study. Cereb. Cortex doi:10.1093/cercor/bhaa408.

Zhang, C., Dougherty, C.C., Baum, S.A., White, T., Michael, A.M., 2018. Functional connectivity predicts gender: evidence for gender differences in resting brain connectivity. Hum. Brain Mapp. 39, 1765–1776. doi:10.1002/hbm.23950.