Beyond “Sex Prediction”: Estimating and Interpreting Multivariate Sex Differences and Similarities in the Brain

Carla Sanchis-Segura (csanchis@uji.es)
Universitat Jaume I. Castelló
Naiara Aguirre
Universitat Jaume I. Castelló
Álvaro Javier Cruz-Gómez
Universitat Jaume I. Castelló
Sonia Félix
Universitat Jaume I. Castelló
Cristina Forn
Universitat Jaume I. Castelló

Research Article

Keywords: Sex differences, sex similarities, MRI, machine learning, effect size, gray matter, TIV-adjustment, robust statistics

Posted Date: July 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-741734/v1

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 quite different concepts, and using classification metrics as surrogate estimates of between-group differences results in major statistical and interpretative distortions. The present study illustrates 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, modeling and clustering techniques and analyses of similarities (ANOSIM) were used to identify the brain areas that contribute the most to these multivariate differences, and 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.

Introduction

Machine-learning (ML) offers new and informatively rich methods for multivariate exploration of the increasingly large and complex datasets from human neuroimaging studies[1]. In the sex differences field, most ML applications have focused on classification tasks. These studies have effectively shown that several predictive 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[2–14]. Because this kind of "prediction" only informs us about what is already known (to which sex category each particular sampled brain belongs), the interest and relevance of these findings resides more in the brains' 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., 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.

However, replacing a direct and quantitative evaluation of multivariate sex differences with a nominal classification has many statistical and conceptual shortcomings. Most of these weaknesses arise from the fact that classification requires an artificial dichotomization of the continuous posterior classification probabilities provided by ML algorithms. Thus, as described below, dichotomization results in a "*loss of information about individual differences as well as havoc with regard to estimation and interpretation of relationships among variables*" ([15], pages 19–20), and it is rarely justified from either a conceptual or statistical perspective[15–17]. Consequently, "*classification through forced up-front dichotomization in an attempt to simplify the problem results in arbitrariness and major information loss*", and it is always "*inferior to probability modeling for driving the development of a predictive instrument or for estimation or for hypothesis testing*" ([17],page 4). Moreover, dichotomous classification predefines and statistically treats males and females as two mutually exclusive categories with zero within-group variance, hence precluding their empirical description, impeding any direct estimation of the size of their differences, and imposing an unwarranted binary interpretative framework.

With these limitations in mind, the present study was designed to assess multivariate sex differences in gray matter volume (GMVOL) without resorting to classification metrics. Because the sizes of univariate volumetric sex differences[18,19] and sex-classification accuracy[5,20] are strongly influenced by total intracranial volume (TIV), this assessment was conducted with raw estimates of GMVOL as well as after adjusting these estimates with the well-validated[5,18] power-corrected proportions method[21] (PCP). More specifically, the raw and PCP-adjusted GMVOL estimates of the 116 brain areas defined by the AAL atlas[22] 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 valid estimates of the *probability of being classified as male* (PCAM). The PCAM was used in subsequent analyses as a continuous dependent variable in which females and males were treated as empirically-mapped distributions (instead of as nominal categories), and their distributional divergences were thoroughly explored with robust statistical and graphical methods[23–26] to properly quantify the size of the

multivariate sex 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[27]. 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. In this regard, a common interpretation of classification studies is that, because distinct ML algorithms are able to very accurately "predict" sex, all these algorithms *must be* identifying a reproducible constellation of brain differences that assemble into two clearly distinguishable brain types ("male/ female brains"; e.g.,[2,3,6,8,10,12,13]). However, it could also well be that different algorithms rely on different brain features to provide their classification labels and outcomes, and, if so, these proposed brain types would become method-specific and, therefore, largely devoid of scientific utility[28].

Results And Discussion

Classification accuracy: Results and limitations.

Figure 1A 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[5,20], 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 Tables 2A and 2B). Figure 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.

As mentioned above, the %CC is regularly used as an estimate of the overall degree of multivariate distinctiveness or separateness of the brains of females and males. However, the calculation of this index requires dichotomizing the continuous output of ML algorithms, and it is well known that dichotomization leads to major statistical and interpretative distortions[15–17]. More specifically, dichotomization results in an overestimation of effect sizes[15], a loss of one-fifth to two-thirds of the outcome's variance[29], a reduction in statistical power equivalent to discarding between one-third and two-thirds of the sample[29], and a higher risk of finding false positive results[16]. However, the main problem of dichotomization is that, in the absence of any knowledge about the outcome's distribution, there is no *a priori* reason to suppose that there is an underlying dichotomy, and if there is, there is no reason this dichotomy should correspond to an outcome's half-split[15,16,29]. Therefore, although the use of this cutoff might be justified when the main goal is to obtain a binary classification, its use when estimating between-group differences is essentially arbitrary and it may abet equally arbitrary conclusions[15–17,29].

These dichotomization-induced distortions can be illustrated by comparing the results in Fig. 1A to those in Fig. 1B, which -as in other previous studies[14,30] - divides the originally continuous outcome into three, instead of two, equal segments. This simple modification reduces the %CC (Supplementary Tables 3A and 3B), thus illustrating how dichotomization leads to larger between-group differences through a suppression of the within-groups variation. These %CC changes were non-uniform, given that dichotomization and trichotomization yielded similar results in some cases (e.g. LDA in the raw dataset, %CC difference = ± 4.63%; Supplementary Table 3C) but very disparate results in others (e.g. RF algorithm in the PCP dataset, %CC difference = ± 52.67%; Supplementary Table 3D). These discrepancies illustrate that, because dichotomization singularly affects to individuals close to but on opposite sides of the cutoff (treating them as totally different when they are quite similar), dichotomization effects on the %CC are dependent on the underlying scores' distribution, and, consequently, they are variable and fundamentally unpredictable. This, together with the dichotomization-induced reduction in statistical power, can mask the effects of moderator variables (e.g., the algorithm used). Indeed, whereas the %CC obtained after dichotomization did not significantly differ between algorithms in any dataset, those obtained after trichotomization differed in both datasets (Supplementary Tables 3E and 3F).

Although trichotomization has some advantages over dichotomization[16,17,31], the results described above should not be interpreted as advocating for it. What these results show is that any discretization of a continuous outcome introduces a certain

degree of arbitrariness in the results obtained, and also that by pre-imposing a number of categories that may or may not exist in the data, discretization leads to preconditioned and discretionary inferences[15,16,29]. Thus, for example, by looking at the PCP-RF outcomes in Fig. 1A it would be possible to conclude that 66% of individuals have "male/female brains", whereas looking at the PCP-RF results in Fig. 1B -which were obtained in the same subjects, algorithm, and dataset-, the conclusion would be that 84.33% of individuals have "sex-undefined brains". These inferences are radically different, and yet -in the absence of additional information- both of them are equally arbitrary.

Assessing multivariate sex differences in GMVOL

Another major limitation of using the %CC as an estimate of multivariate differences is that it is not descriptive. In fact, the %CC is a "*very insensitive and statistically inefficient measure*" ([17], page 258), where all the cases above/ below the predefined threshold are rated the same, regardless of their actual values in the outcome variable (e.g. $P = 0.51$ is equalized to $P = 0.99$ and P = 0.01 to P = 0.49). Consequently, %CC scores preclude describing the outcome's distribution, summarizing the individuals' scores or estimating the actual size of between-group differences. These limitations are overcome when the continuous output of ML algorithms is not discretized, and when females and males are treated as empirically-established distributions that spread at different probability levels within particular ranges of the outcome's continuous space[14,32–34]. Therefore, we first described how males and females mapped onto the continuous output provided by ML algorithms (the PCAM continuum) and then we assessed their distributional divergences to obtain proper estimates of the size of the male-female multivariate differences in GMVOL

Figure 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 varied depending on the algorithm (e.g., LDA vs. RF; Supplementary Tables 4A-4B, Supplementary Fig. 1), which explains why their %CC estimates differentially varied in response to the dichotomizing or trichotomizing cutoffs (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 (Supplementary Tables 4A-4C, Supplementary Fig. 1). 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. Again, these distributional singularities explain why %CC estimates disparately varied in response to different pre-imposed cutoffs (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 invariably "large" in the raw dataset. Thus, the overlap between the males/females' distributions was "small" (≈12–18%; Supplementary Table 4D), 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 4E). 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 Tables 4D and 4E). 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 Tables 4F and 4G), suggesting that -when considered on the PCAM continuum- the variability of males and females does not significantly differ. Conversely, median comparisons confirmed "large"/ "small" sex differences in the raw and PCP datasets, respectively (see below).

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[23,35,36]. To fully represent and compare distributions, robust statistical and informatively-rich graphical methods such as the cumulative distribution function (CDF)[25] and the shift-

function[23,24] are required[23,25,36–39]. Consequently, we used these methods to provide two complementary perspectives[23] of the multivariate sex differences in GMVOL.

How do males and females compare to each other?

Figure 3 displays the CDFs for the PCAM scores 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[25,35]; and 3) by comparing the PCAM values at the deciles of the females/ males' distributions[23,24].

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 twice that size at the medians (D5, POMP=98.1) and tended to decrease thereafter (D9, POMP=40.72). These inter-decile variations were statistically significant (Supplementary Tables 5A and 5C), 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 Tables 5B and 5D; 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).

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[22,26]. 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 = 12.93) 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.

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*" ([17], 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 and, to some extent, arbitrary criteria, PCAM-based measures are fully empirical and conceptually unrestricted.

On this last point, previous sex classification studies[4,6,8,10,12,14,40] 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. However, finding two- and only two- brain types is not as much an empirical result as it is a pre-imposed requirement[17] of these sex classification studies. Moreover, the fact that different algorithms are able to correctly identify sex with a 80–90% accuracy does not ensure that these algorithms are providing the same outcomes or identifying the same underlying reality[28,41,42]. Thus, in the present study, the %CC ranged between 86.3–90% and 58.7–65% in the raw and in the PCP datasets, but the actual percentages of subjects correctly classified by all five algorithms were 76.3 and 38.0%, respectively. On the other hand, 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[41,4241,42]. 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[41,43]. Therefore, in the present study, we used boosted beta regression procedures to identify and quantify the relative importance of the brain features that contribute the most to the PCAM scores yielded by each algorithm in each dataset, and we used hierarchical clustering procedures and ANOSIM analyses to assess how these features assemble.

As it could be expected[20] and Figs. 5 and 6 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 (D = 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. 5 and Supplementary Table 7H). This relationship was corroborated by statistically significant correlations ($rho_{(30)} = 0.592$ and $rho_{(30)} = 0.571$, $p < 0.001$) between these r^2 values and two estimates of the predictors' relative importance across algorithms (ranks' averages and

a multiplicative “rank of ranks”; Fig. 5 and Supplementary Table 7H). Conversely, in the PCP dataset, TIV-variation did not account for any variance in GMVOL (Fig. 6 and Supplementary Table 7H), and the predictors’ relative importance was unrelated to these non-statistically significant r^2 values ($rho_{(52)}$=-0.119 and $rho_{(52)}$=-0.123, $p > 0.05$; Fig. 6 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.

Similarly, 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. 5), 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 were included in all of them, thus resulting in low levels of absolute nominal agreement (K_{HR}=0.505 [0.385, 0.625], multi-rater D = 0.727 [0.639, 0.814]). Moreover, 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 as “moderate”[44] (ICC average-rating = 0.671[0.450, 0.821], $p < 0.001$; Supplementary Table 7L).

In the PCP dataset (Fig. 6), 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 were included in all the regression models. Consequently, estimates of absolute agreement were low (K_{HR}=0.323 [0.215, 0.431], multi-rater D = 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”[44] (ICC average-rating = 0.694 [0.520, 0.812], $p < 0.001$; Supplementary Table 7M).

These between-algorithms’ discrepancies explain why different algorithms yield differently-shaped PCAM distributions (Fig. 2) that result in multivariate sex differences that vary in size (Figs. 3 and 4). Thus, because they differ in their statistical assumptions and operations, distinct algorithms rely on distinct brain features (Figs. 5 and 6) and assign different PCAM scores to the same subjects (Fig. 7A, Supplementary Tables 8A-8B; for dyadic between-algorithm’s comparisons, see Supplementary Tables 8C-8F; for a case-by-case examination, see Interactive Fig. 1). Because these PCAM-variations are highly idiosyncratic (Fig. 7B), each individual occupies a different relative position in each PCAM distribution (Fig. 7C and 7D), and these distributions end up spreading dissimilarly within the PCAM range (Fig. 2). In the raw dataset, between-algorithms’ 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-algorithms correlations (Supplementary Fig. 6). However, when TIV effects are statistically controlled (as in the PCP dataset or by partialling out the TIV effects), between-algorithms discrepancies in PCAM scores become evident (Fig. 7D; Supplementary Fig. 6).

Therefore, it is apparent that -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 clearly 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[4,10,12,14] (for a comparative review, see[41]), and with evidence suggesting that ML algorithms rely on different brain features when classifying different subpopulations of females and males[44]. Together, these sources of empirical evidence directly challenge the binary sex views of human brains based on a global interpretation of %CC scores. As mentioned above, these views assume that, because distinct ML algorithms are able to correctly "predict" sex from neuroanatomical features in 80–90% of the cases, all these algorithms' *must be* identifying two distinct brain types in the human species, one typical of males and the other typical of females ("male/ female brains") [2,3,6,8,10,12,13]. However, these universal "brain types" do not seem to really exist, given that different algorithms identify distinct brain features as the landmarks of "male/ female brains" in different samples of females and males and when applied to the same subjects.

In fact, even when just taking into account the outcomes of a single ML algorithm in a single sample, these proposed "male/ female" brain types do not seem to exist, either, because the same class label or even virtually identical PCAM scores can be achieved by individuals exhibiting very different brain profiles (see examples in Figs. 8A and 9A). More specifically, when accumulated differences in raw GMVOL (Euclidean distances) at the brain areas identified as relevant PCAM predictors 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. 8B). These two clusters are homogeneous and robust, and they are only minimally perturbed when additional partitions are imposed (Fig. 8D). However, these clusters do not correspond to two "brain types" or sex-specific brain profiles. Thus, when the same individuals are partitioned based on the dissimilarity of their brain profiles (Spearman's distances) rather than on the orderless sum of their differences, subjects cluster in a sex-unrelated manner (Fig. 8C and 8E) 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). This pattern of results was corroborated for each algorithm and also when using the 116 brain areas of the AAL atlas or just only the five areas more directly related to the PCAM scores provided by each algorithm as predictors (Supplementary Figs. 7–11). Similar results were also observed when TIV-variation was statistically controlled (PCP dataset; Fig. 9; Supplementary Figs. 12–16), although in this case, both the accumulated differences and the brain profiles' similarities are basically unrelated to sex categories (ANOSIM $R_{Euclidean}$= 0.017 [-0.018, 0.004]; ANOSIM $R_{Spearman}$= 0.025 [-0.014, 0.050]; Supplementary Tables 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; 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. Accordingly, there are no "male/ female brains" (for similar conclusions, see[9,14,40,41,45]), 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, either, at least not on one that can be univocally translated to their neuroanatomical features (for a similar conclusion and a more ample discussion, see [42])

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[14,32–34]. 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- offers a more informative, nuanced, and accurate way to investigate multivariate relationships between sex and brain features than what is offered by the classification approach.

Our results also reveal that, at least at the neuroanatomical level, there are not two brain "types", and overall brain sex differences in GMVOL do not stem from a specific pattern of differences in a few "key" brain areas. Accordingly, the PCAM-based estimates of these multivariate group differences 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. Therefore, like other

multivariate effect size indexes[37], 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. 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 (for a similar conclusion, see[41,43]).

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.

Materials And Methods

Participants

The present study was conducted using data from the 1,200 Subject Release of the Human Connectome Project (HCP), which includes structural Magnetic Resonance Imaging (MRI) data from 1113 healthy young adult participants [47]. 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) not differing in age. This sample was subsequently split into the so-called training and testing subsamples (see below).

Imaging and data preprocessing

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 (AAL,[22]) 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 1,000 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.

Training/ testing subsets and data standardization

Following current recommendations [48,49], classification algorithms were fitted and tested in two separated groups of participants randomly extracted from the previously described main sample. The *training subsample* included 288 females and

288 males, whereas the *testing subsample* included 150 females and 150 males, hence avoiding classification distortions due to between-class imbalance [50,51]. In both subsets, females and males were very similar in age (means' range = 28.4–28.6) and showed similar differences in TIV (Cohen's d = 1.8 in both cases; see further details in Supplementary Tables 1A and 1B).

Before being used as predictors, all volumetric variables were transformed into *z*-scores to avoid distortions due to their different ranges [48,52]. Standardization was initially performed in the *training subset*, and the exact same scaling parameters were subsequently used to standardize the *testing subset*.

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, but also that not all the currently used methods are equally effective and valid for removing TIV-variation [5,18]. 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[21]. 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 calculated for the entire sample of participants [21].

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 [48,53] and may potentially lead to different outcomes. Moreover, it is necessary to compare the outcomes of different ML algorithms (see *PCAM variation* subheading) in order to ensure that the results obtained describe method-independent findings and that the conclusions drawn are truly generalizable. The outcomes considered 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 yielded by ML algorithms (in this case, operationalized as the *probability of being classified as male*, PCAM) as a continuous variable (see details in the *Assessing multivariate sex differences in GMVOL* subheading).

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 [54].

-*Logistic Regression (LR)*: Implemented by the *glm* function (family= "binomial") of the *stats* package natively included in R [55].

-*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 [57], 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 [58]. The *tune* function (tenfold cross-validation) was used to automatically select the optimal values for the regularization and kernel-width hyper-parameters.

Statistical analyses

All statistical analyses were conducted in the testing subsamples of the raw and the PCP datasets using different packages for R[55]. Statistical analyses focused on description and effect sizes estimation rather than merely testing statistical significance[59]. 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 (see [35,60]). Moreover, when statistical significance was tested, *p* values were corrected for multiple comparisons with the FDR[61] or -when comparing deciles (see below)- with the Hochberg[62] method.

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.

As discussed in the main text, the calculation of the %CC scores requires a dichotomization of the continuous outcome provided by ML algorithms. To illustrate the statistical and interpretative distortions[15–17] introduced by this dichotomization, we calculated new %CC scores after dividing the PCAM this outcome into three thirds (as in some previous studies[14,30]) instead of into two halves. These scores were compared to the chance-expected value of 0.33 by means of chi-squared tests. The McNemar's test was used to compare the dichotomization- and trichotomization-associated %CC scores, as well as the trichotomization %CC scores yielded by each algorithm in each dataset.

Assessing multivariate sex differences in GMVOL

As in some previous studies[14,32–34], we used the *a posteriori* probabilities yielded by ML algorithms (in this case, operationalized as the *probability of being classified as male*, PCAM) as a continuous and informatively-rich dependent variable. The males and females' PCAM distributions yielded by 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). In a second step, PCAM scores were used to quantify the multivariate sex differences in GMVOL at different levels. More specifically:

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 h overlap index. The h 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[63]. 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[64]. 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 d statistic[65] and its 95%CI, was obtained through the *cidv2* function of the *rogme* package[23] .

Because no single score can properly summarize the differences between two distributions[23,25,35–39], male-female differences in the PCAM continuum were characterized by comparing their cumulative distribution functions (CDF; [25,35]). 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[25,35]. Within each CDF, sex-based comparisons were conducted at each decile with the *shifthd_pbci* function (bootstrap:

10,000 repetitions) of the *rogme* package[23]. 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 a < 0.05 without being concerned about Type I error[23].

With the decile estimates obtained, the so-called shift functions[24] 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[23]). 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[23,24], 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 [23], 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[23]. Finally, we employed the *Dqcomhd* function (bootstrap: 50,000 repetitions) of the *WRS2* package for R to ascertain whether the deciles of these male-female pair-wise differences significantly varied between algorithms in each dataset.

Interpreting multivariate sex differences in GMVOL

For obvious reasons, interpretability has become a major issue in ML applications [66]. 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 [66–68]). 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[69]. 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[27]. 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[27]). In the present study, boosted beta regressions were implemented through the *betaboost* package for R[70], 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[71] and the multi-rater delta index (which, unlike other more commonly used agreement metrics, is not affected by the ratings' marginal distributions[72,73]). 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 [74], Kendall's W agreement coefficient [75], the mean of bivariate Spearman's rho rank correlations [76], and the intraclass-correlation coefficient (two-way ANOVA, random effects,w single and average ratings [77,78]). 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, [79]) 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 in [5,18], 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.

Between-algorithm 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 maximum PCAM score was subtracted from its maximum PCAM score within each dataset, and the CDFs depicting this maximal 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, to assess the degree of ordinal similarity between the PCAM scores yielded by different algorithms, 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

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[55]. 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 *beta boosted regression* section). 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 the volumetric z-scores of the 116 brain areas from the AAL atlas and also with the top five predictors of the PCAM scores yielded by each ML algorithm in each dataset.

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[80]. Because in the present study this assessment involved a large number of instances (44,850), statistical significance was almost guaranteed and, consequently, it was not truly informative[81]. 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)[80]. More specifically, Euclidean and Spearman distance matrices were calculated in 10,000 bootstrap samples using the *get_dist* function of the *factoextra*[82] package. In each of these distance matrices, an ANOSIM test was conducted with the *anosim* function of the *vegan* package[83] 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[84]. Again, these calculations were performed using as features the volumetric z-scores of the 116 brain areas from the AAL atlas, and of all or just the top five predictors of the PCAM scores yielded by each ML algorithm in each dataset.

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

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

Declarations

Acknowledgements

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]. These funding sources did not play any role in designing the study or in the collection, analysis, and interpretation of the data.

Contributions

C.S.-S. and C.F. designed the study. N.A., A.J.C.-G. and S.F. processed the scan images on which C.S.-S. conducted the statistical analyses. C.S.-S. and C.F. wrote the manuscript. All authors contributed to manuscript revision and read, and approved the submitted version.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Bzdok, D. Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience* **11**, 543 (2017).
2. Wang, L., Shen, H., Tang, F., Zang, Y. & Hu, D. Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain: An MVPA approach. *Neuroimage* **61**, 931–940 (2012).
3. Luo, Z., Hou, C., Wang, L. & Hu, D. Gender Identification of Human Cortical 3-D Morphology Using Hierarchical Sparsity. *Front. Hum. Neurosci.* **13**, 29 (2019).
4. Anderson, N. E. *et al.* Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping* (2018). doi:10.1002/hbm.24462

5. Sanchis-Segura, C., Ibañez-Gual, M. V., Aguirre, N., Gómez-Cruz, Á. J. & Forn, C. Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Sci. Rep.* **10**, (2020).
6. Rosenblatt, J. . Multivariate revisit to 'sex beyond the genitalia'. *Proceedings of the National Academy of Sciences of the United States of America* (2016). doi:10.1073/pnas.1523961113
7. Feis, D. L., Brodersen, K. H., von Cramon, D. Y., Luders, E. & Tittgemeyer, M. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *Neuroimage* **70**, 250–257 (2013).
8. Chekroud, A. M., Ward, E. J., Rosenberg, M. D. & Holmes, A. J. Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences of the United States of America* (2016). doi:10.1073/pnas.1523888113
9. Joel, D. *et al.* Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa. *Front. Hum. Neurosci.* (2018). doi:10.3389/fnhum.2018.00399
10. Sepehrband, F. *et al.* Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. *Neuroimage* **172**, 217–227 (2018).
11. Van Putten, M. J. A. M., Olbrich, S. & Arns, M. Predicting sex from brain rhythms with deep learning. *Sci. Rep.* (2018). doi:10.1038/s41598-018-21495-7
12. Xin, J., Zhang, Y., Tang, Y. & Yang, Y. Brain Differences Between Men and Women: Evidence From Deep Learning. *Front. Neurosci.* **13**, 185 (2019).
13. Zhang, C., Dougherty, C. C., Baum, S. A., White, T. & Michael, A. M. Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Hum. Brain Mapp.* **39**, 1765–1776 (2018).
14. Zhang, Y., Luo, Q., Huang, C. C., Lo, C. Y. Z., Langley, C., Desrivières, S., ... & IMAGEN consortium. (2021). The Human Brain Is Best Described as Being on a Female/Male Continuum: Evidence from a Neuroimaging Connectivity Study. *Cereb. Cortex.* (2021). doi:https://doi.org/10.1093/cercor/bhaa408
15. MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. On the practice of dichotomization of quantitative variables. *Psychol. Methods* (2002). doi:10.1037/1082-989X.7.1.19
16. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *British Medical Journal* (2006). doi:10.1136/bmj.332.7549.1080
17. Harrell, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic*. (Springer International Publishing Switzerland, 2015).
18. Sanchis-Segura, C. *et al.* Sex differences in gray matter volume: How many and how large are they really? *Biol. Sex Differ.* (2019). doi:10.1186/s13293-019-0245-7
19. Williams, C. M., Peyre, H., Toro, R. & Ramus, F. Neuroanatomical norms in the <scp>UK</scp> Biobank: The impact of allometric scaling, sex, and age. *Hum. Brain Mapp.* hbm.25572 (2021). doi:10.1002/HBM.25572
20. More, S., Eickhoff, S. B., Caspers, J. & Patil, K. R. Confound Removal and Normalization in Practice: A Neuroimaging Based Sex Prediction Case Study. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **12461 LNAI**, 3–18 (2020).
21. Liu, D., Johnson, H. J., Long, J. D., Magnotta, V. A. & Paulsen, J. S. The power-proportion method for intracranial volume correction in volumetric imaging analysis. *Front. Neurosci.* (2014). doi:10.3389/fnins.2014.00356
22. Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* (2002). doi:10.1006/nimg.2001.0978
23. Rousselet, G. A., Pernet, C. R. & Wilcox, R. R. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *Eur. J. Neurosci.* (2017). doi:10.1111/ejn.13610
24. Wilcox, R. R. & Rousselet, G. A. A Guide to Robust Statistical Methods in Neuroscience. *Curr. Protoc. Neurosci.* (2018). doi:10.1002/cpns.41
25. Callaert, H. Nonparametric hypotheses for the two-sample problem. *J. Stat. Educ.* **7**, (1999).

26. M. Handcock, M. M. Relative distribution methods. *Sociol. Methodol.* (1998). doi:10.1111/0081-1750.00042
27. Schmid, M. *et al.* Boosted Beta Regression. *PLoS One* (2013). doi:10.1371/journal.pone.0061623
28. Hancox-Li, L. Robustness in Machine Learning Explanations: Does It Matter? (2020). doi:10.1145/3351095.3372836
29. Cohen, J. The Cost of Dichotomization. *Appl. Psychol. Meas.* (1983). doi:10.1177/014662168300700301
30. Joel, D. *et al.* Sex beyond the genitalia: The human brain mosaic. *Proc. Natl. Acad. Sci.* (2015). doi:10.1073/pnas.1509654112
31. Gelman, A. & Park, D. K. Splitting a predictor at the upper quarter or third and the lower quarter or third. *Am. Stat.* (2009). doi:10.1198/tast.2009.0001
32. Lippa, R. & Connelly, S. Gender Diagnosticity: A New Bayesian Approach to Gender-Related Individual Differences. *J. Pers. Soc. Psychol.* **59**, 1051–1065 (1990).
33. Phillips, O. R. *et al.* Beyond a Binary Classification of Sex: An Examination of Brain Sex Differentiation, Psychopathology, and Genotype. *J. Am. Acad. Child Adolesc. Psychiatry* (2019). doi:10.1016/j.jaac.2018.09.425
34. Eijk, L. van *et al.* Are sex differences in human brain structure associated with sex differences in behaviour? doi:10.31234/OSF.IO/8FCVE
35. Grissom, R. J. & Kim, J. J. *Effect sizes for research: Univariate and multivariate applications, second edition*. *Effect Sizes for Research: Univariate and Multivariate Applications, Second Edition* (Routledge, Multivariate application tests, 2012). doi:10.4324/9780203803233
36. Handcock, Mark S.; Morris, M. *Relative Distribution Methods in the Social Sciences*. *Relative Distribution Methods in the Social Sciences* (1999). doi:10.1007/b97852
37. Del Giudice, M. Measuring sex differences and similarities. in *Gender and sexuality development: Contemporary theory and research.* (ed. VanderLaan, D.P.; Wong, W. I.) (2019).
38. Tukey, J. W. *Exploratory Data Analysis*. (Addison-Wesley, 1977). doi:10.1007/978-1-4419-7976-6
39. Cook, Di., Lee, E. K. & Majumder, M. Data Visualization and Statistical Graphics in Big Data Analysis. *Annual Review of Statistics and Its Application* (2016). doi:10.1146/annurev-statistics-041715-033420
40. Del Giudice, M., Lippa, R. A., Puts, P. D. A. & Bailey, Drew H J. Bailey, Michael P. Schmitt, D. Mosaic Brains? A Methodological Critique of Joel et al. (2015) Online document. (2015). doi:DOI: 10.13140/RG.2.1.1038.8566.
41. Joel, D. Beyond the binary: Rethinking sex and the brain. *Neuroscience and Biobehavioral Reviews* (2021). doi:10.1016/j.neubiorev.2020.11.018
42. Eliot, L., Ahmed, A., Khan, H. & Patel, J. Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neurosci. Biobehav. Rev.* (2021). doi:10.1016/j.neubiorev.2021.02.026
43. Joel, D. Beyond sex differences and a male–female continuum: Mosaic brains in a multidimensional space. in *Handbook of Clinical Neurology* (2020). doi:10.1016/B978-0-444-64123-6.00002-3
44. Portney, L. G. . & Watkins, M. P. *Foundations of Clinical Research: Applications to Practice*. (Pearson/ Prentice Hall, 2009).
45. Joel, D. *et al.* Analysis of human brain structure reveals that the brain "types" typical of males are also typical of females, and vice versa. *Front. Hum. Neurosci.* (2018). doi:10.3389/fnhum.2018.00399
46. Weis, S. *et al.* Sex Classification by Resting State Brain Connectivity. *Cereb. Cortex* **30**, 824–835 (2020).
47. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *Neuroimage* (2013). doi:10.1016/j.neuroimage.2013.05.041
48. Hastie, Trevor, Tibshirani, Robert, Friedman, J. *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. *Springer series in statistics* (2009). doi:10.1007/978-0-387-84858-7
49. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences* (2019). doi:10.1016/j.tins.2019.02.001

50. Ali, A., Shamsuddin, S. M. & Ralescu, A. L. Classification with class imbalance problem: A review. *Int. J. Adv. Soft Comput. its Appl.* (2015).
51. García, V., Sánchez, J. S., Mollineda, R. A. & Sotoca, R. A. J. M. The class imbalance problem in pattern classification and learning. *Data Eng.* (2007).
52. Ali, S. & Smith-Miles, K. A. Improved support vector machine generalization using normalized input space. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2006). doi:10.1007/11941439-40
53. Kiang, M. Y. A comparative assessment of classification methods. *Decis. Support Syst.* (2003). doi:10.1016/S0167-9236(02)00110-0
54. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S Fourth edition by*. *World* (2002). doi:10.2307/2685660
55. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,. (2020).
56. Stephen Milborrow. Derived from mda:mars by Trevor Hastie, With, R. T. U. A. M. F. utilities & Wrapper., T. L. leaps. earth: Multivariate Adaptive Regression Splines. R package version 5.1.2. (2019).
57. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* (2002).
58. Karatzoglou, A., Hornik, K., Smola, A. & Zeileis, A. kernlab - An S4 package for kernel methods in R. *J. Stat. Softw.* (2004). doi:10.18637/jss.v011.i09
59. Wasserstein, R. L. & Lazar, N. A. The ASA's Statement on p -Values: Context, Process, and Purpose. *Am. Stat.* **70**, 129–133 (2016).
60. Cohen, P., Cohen, J., Aiken, L. S. & West, S. G. The problem of units and the circumstance for POMP. *Multivariate Behav. Res.* (1999). doi:10.1207/S15327906MBR3403_2
61. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (2018).
62. Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988).
63. Pastore, M. & Calcagnì, A. Measuring distribution similarities between samples: A distribution-free overlapping index. *Front. Psychol.* (2019). doi:10.3389/fpsyg.2019.01089
64. Pastore, M. Overlapping: a R package for Estimating Overlapping in Empirical Distributions. *J. Open Source Softw.* (2018). doi:10.21105/joss.01023
65. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* (1993). doi:10.1037/0033-2909.114.3.494
66. Ribeiro, M. T., Singh, S., & Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv Prepr. arXiv1606.05386.* (2016).
67. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* (2018). doi:10.1145/3233231
68. Carvalho, D. V., Pereira, E. M. & Cardoso, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)* (2019). doi:10.3390/electronics8080832
69. Ferrari, S. L. P. & Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* (2004). doi:10.1080/0266476042000214501
70. Mayr, A. *et al.* The betaboost package—a software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data. *Int. J. Epidemiol.* (2018). doi:10.1093/ije/dyy093
71. Hubert, L. Kappa revisited. *Psychol. Bull.* (1977). doi:10.1037/0033-2909.84.2.289
72. Andrés, M. & Hernández, Á. Multi-rater delta: extending the delta nominal measure of agreement between two raters to many raters. *arXiv* (2019).
73. Andrés, A. M. & Marzo, P. F. Delta: A new measure of agreement between two raters. *Br. J. Math. Stat. Psychol.* (2004). doi:10.1348/000711004849268
74. Lin, L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* (1989). doi:10.2307/2532051

75. Kendall, M. G. & Smith, B. B. The Problem of m Rankings. *Ann. Math. Stat.* (1939). doi:10.1214/aoms/1177732186
76. Gamer, M; Lemon, J; Fellows, I; Singh, P. irr: Various Coefficients of Interrater Reliability and Agreement.R package version 0.84.1. (2019).
77. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* (2016). doi:10.1016/j.jcm.2016.02.012
78. Hallgren, K. A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods Psychol.* (2012). doi:10.20982/tqmp.08.1.p023
79. Tofallis, C. Add or Multiply? A Tutorial on Ranking and Choosing with Multiple Criteria. *INFORMS Trans. Educ.* (2014). doi:10.1287/ited.2013.0124
80. CLARKE, K. R. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18**, 117–143 (1993).
81. Lindley, D. V. A statistical paradox. *Biometrika* **44**, 187–192 (1957).
82. Kassambara, A; Mundt, A. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2020).
83. 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.M; Szoecs, E.; Wagner, H. vegan: Community Ecology Package. (2020).
84. Canty, A.; Ripley, B. boot: Bootstrap R (S-Plus) Functions. (2020).

Figures

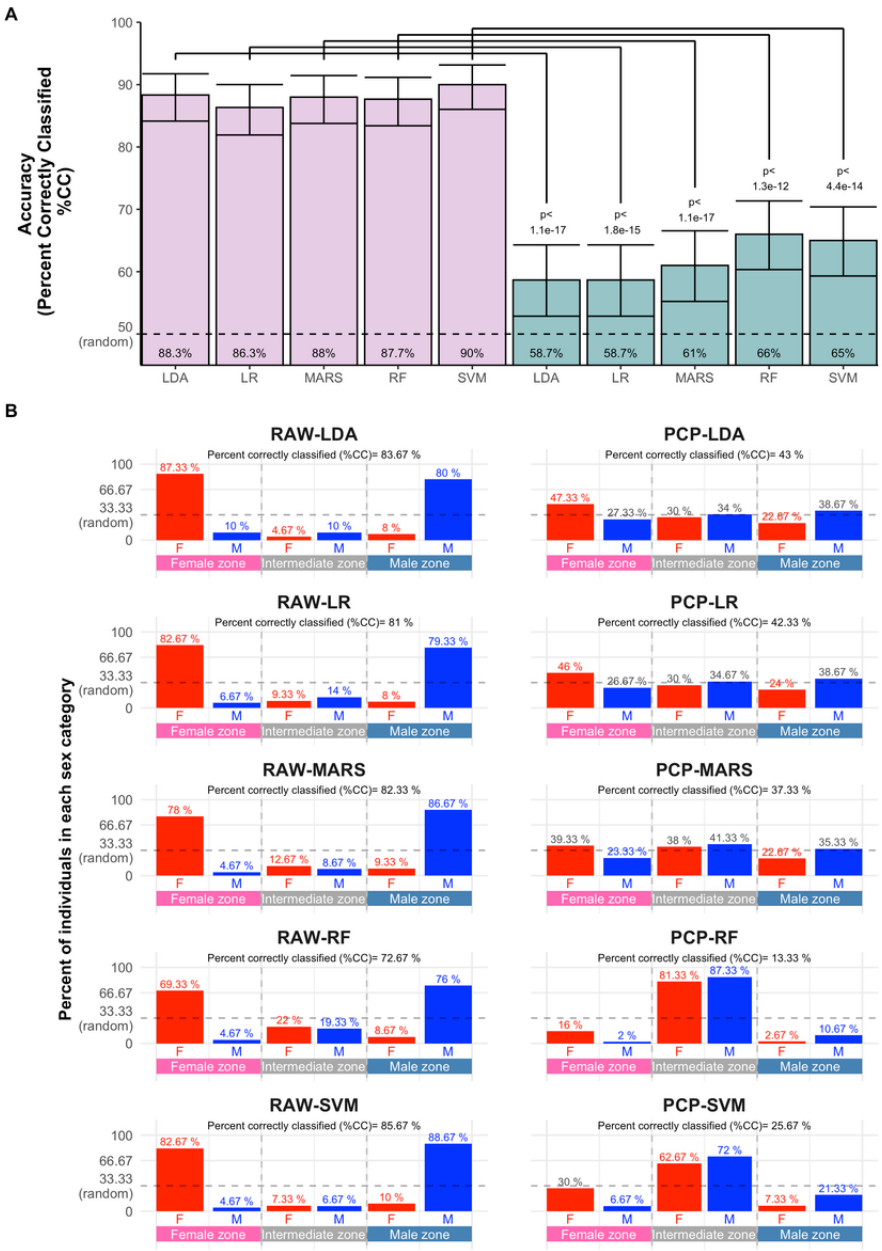

Figure 1

Sex classification accuracy. A) Dichotomous classification. 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. B) Trichotomous classification. Bars illustrate the percent of individuals in each sex category classified in the “female zone”, “intermediate zone”, and “male zone”, respectively. Numbers in gray denote accuracy scores that are not significantly different from chance levels. See Supplementary Tables 2A-2D and 3A-3F for a complete statistical output.

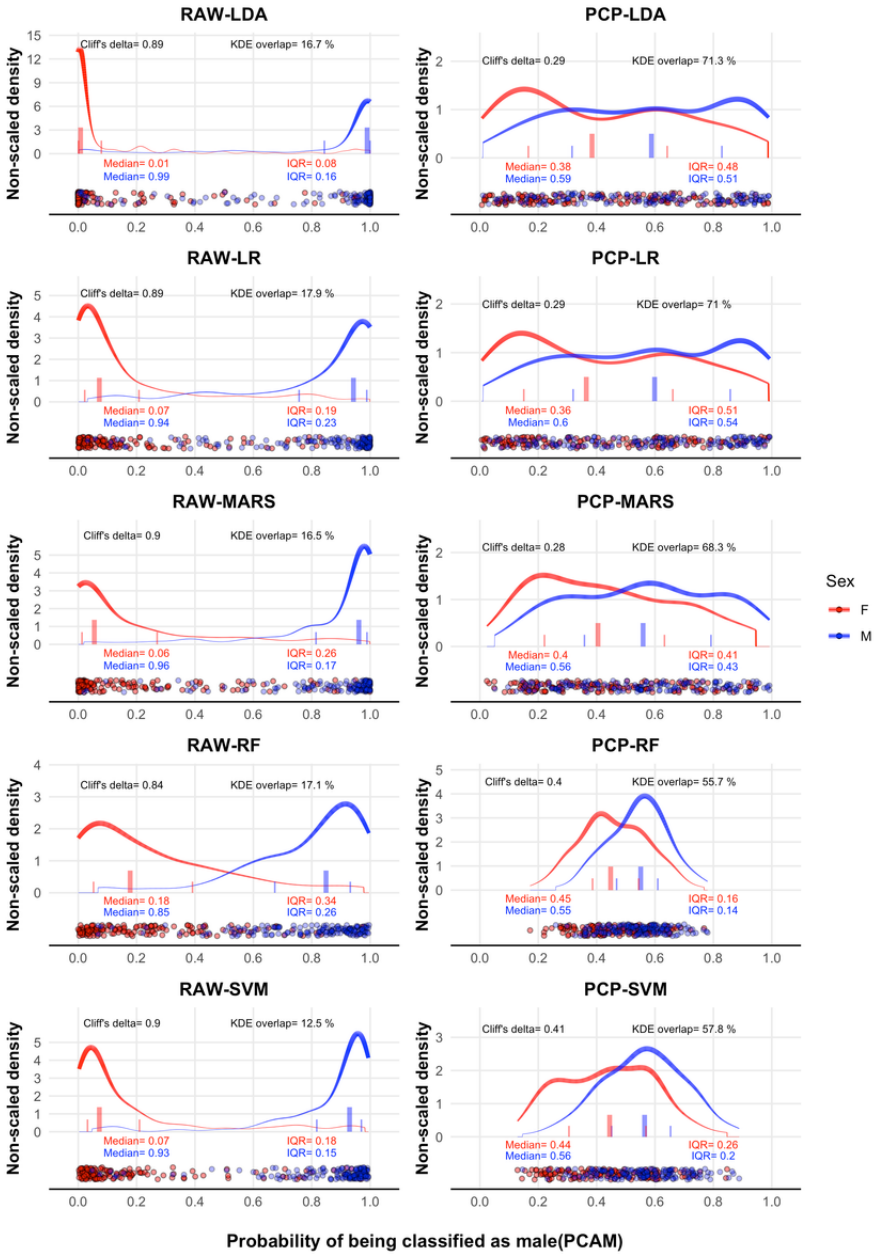

Figure 2

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 Tables 4A-4E and Supplementary Figure 1.

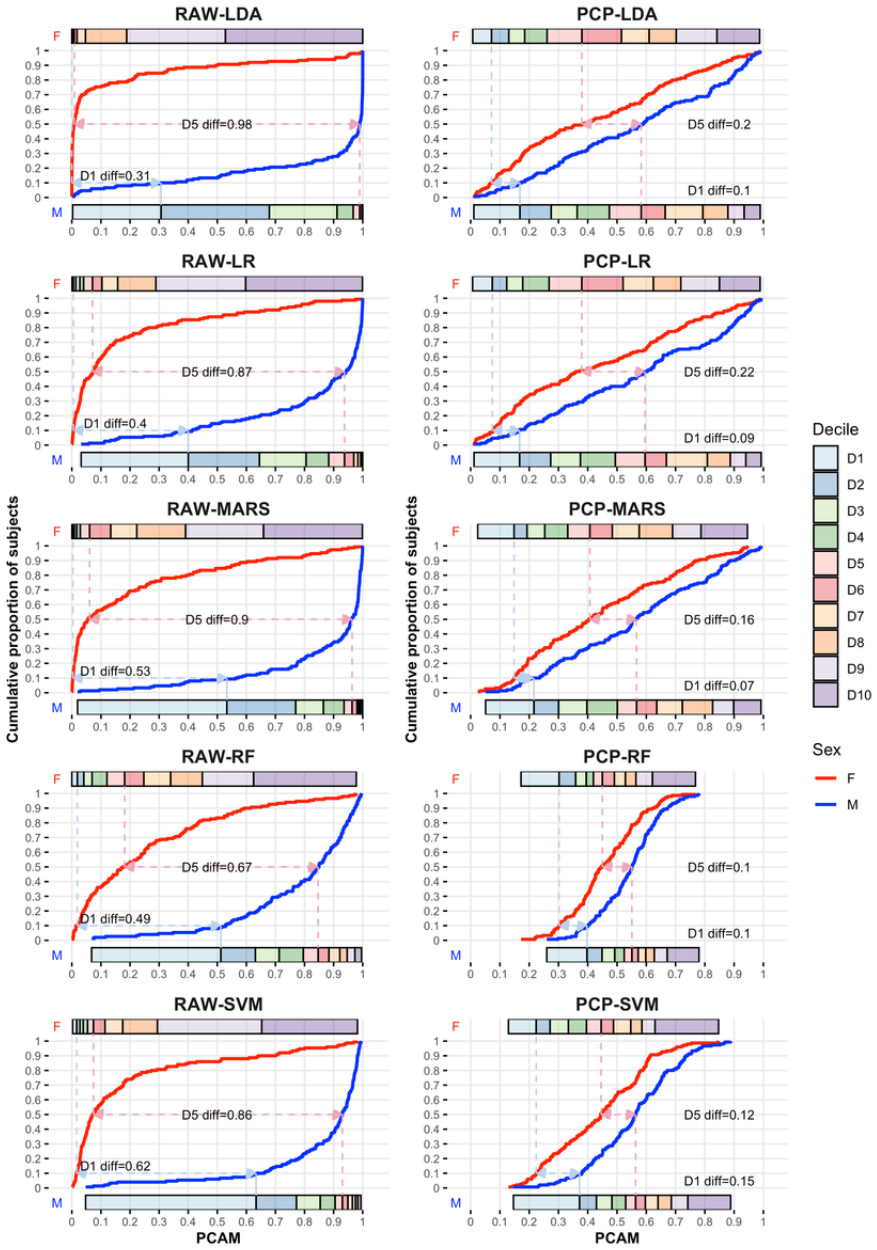

Figure 3

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 Tables 5A-5F and in the accompanying Supplementary Figures 2 and 3.

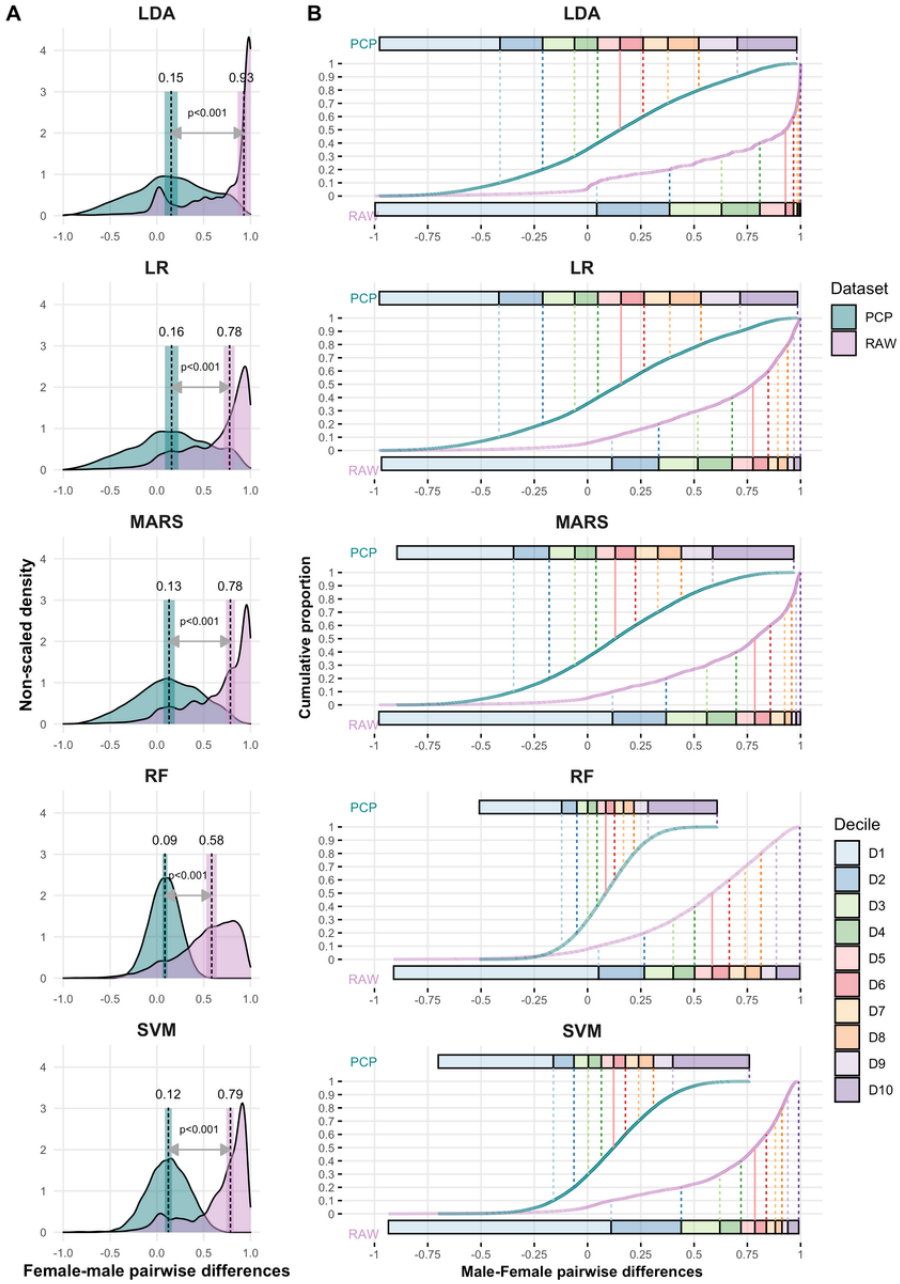

Figure 4

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 Tables 6A-6E and in the accompanying Supplementary Figures 4 and 5.

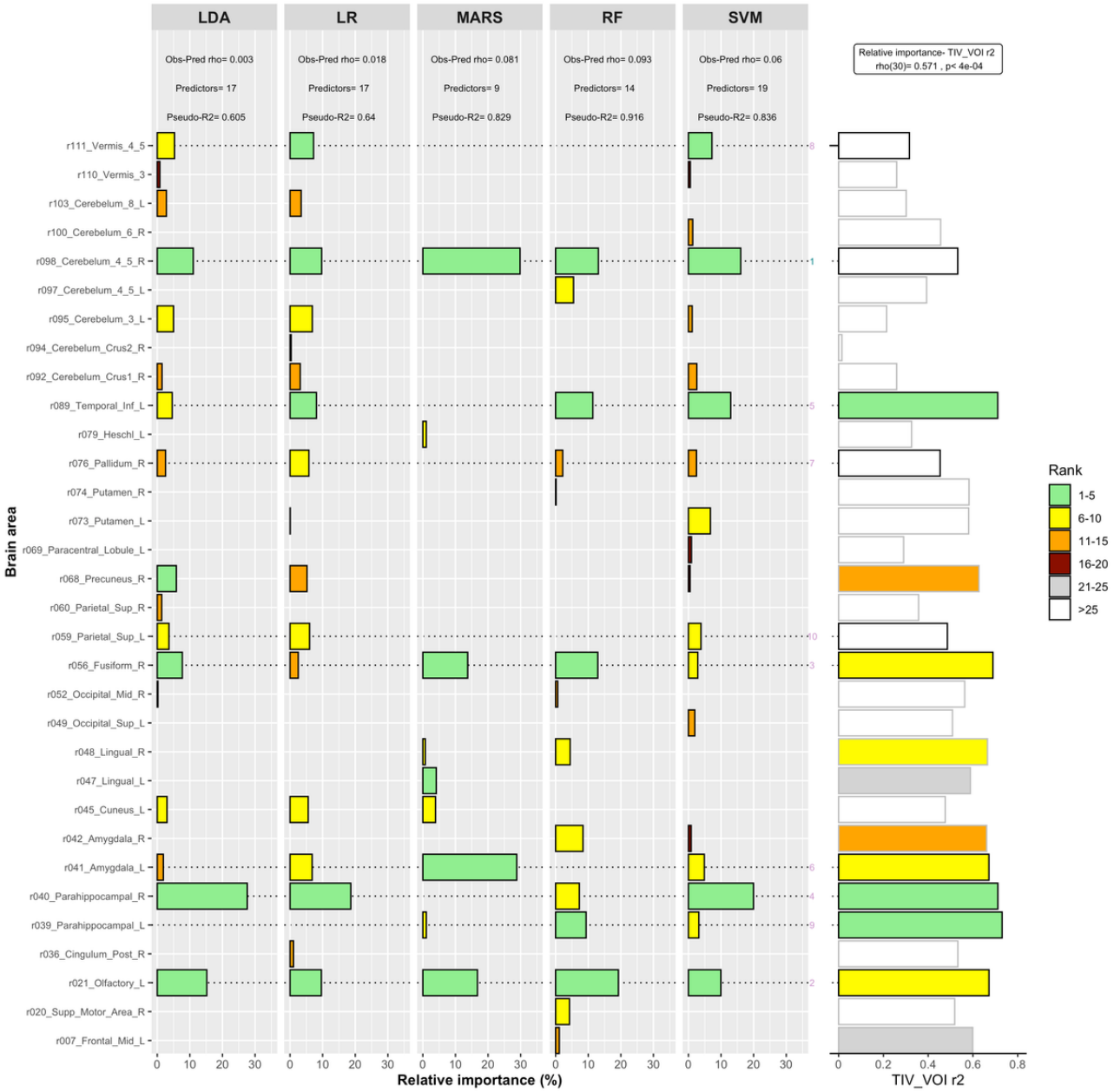

Figure 5

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 Tables 7A,7C, 7E-7G, 7H,7J, and 7L.

Figure 6

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 Tables 7B,7D, 7E-7G, 7I,7K, and 7M.

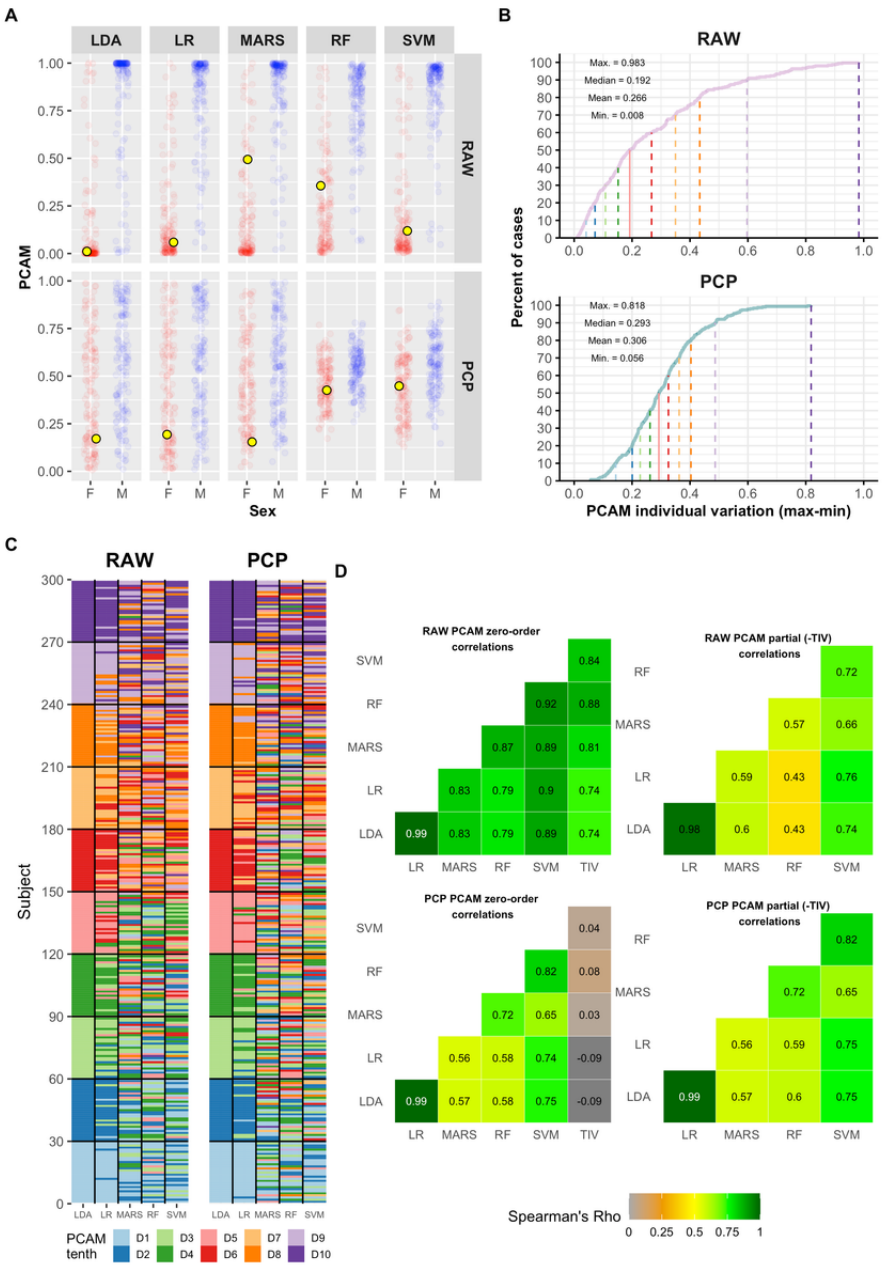

Figure 7

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 Tables 8A-8F and in the accompanying Supplementary Figure 6.

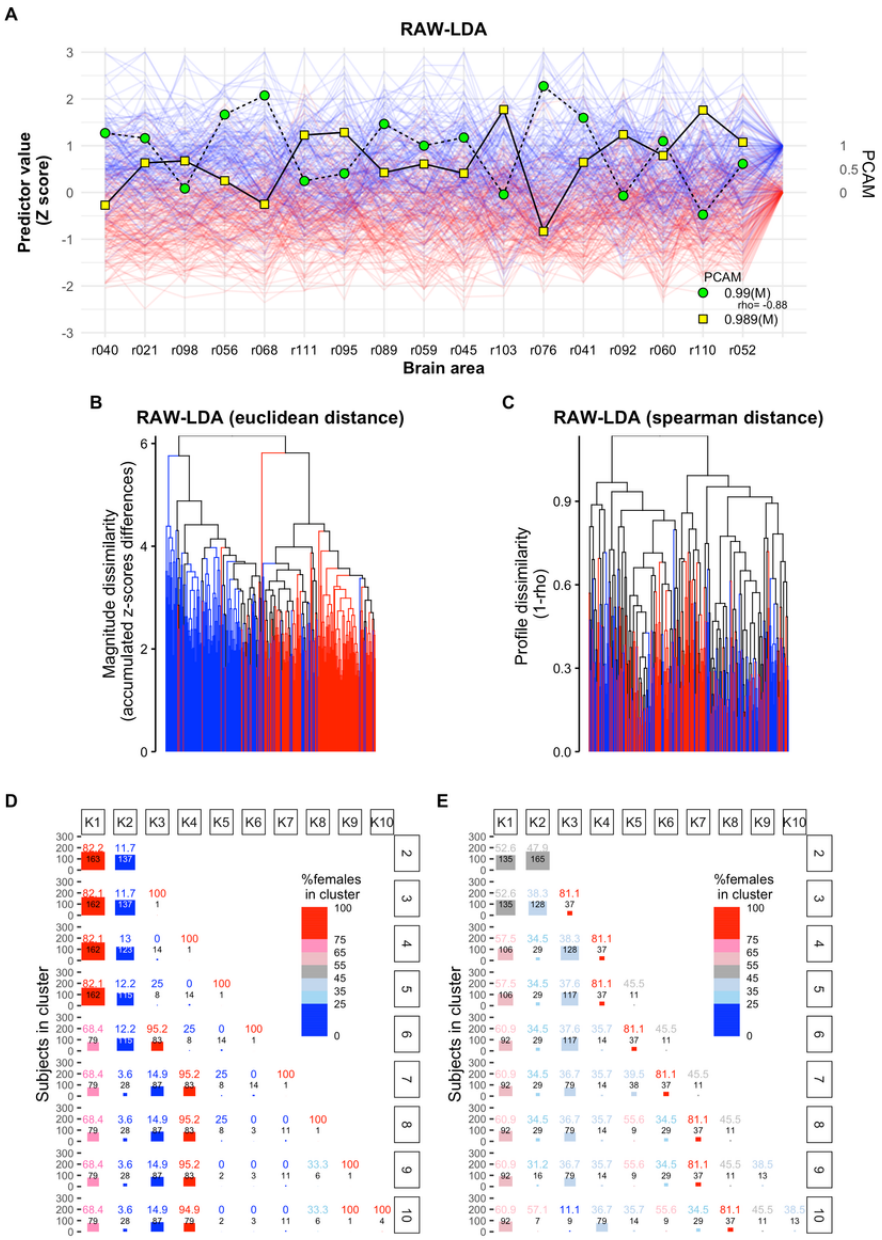

Figure 8

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 Figure 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 Figures 7-11).

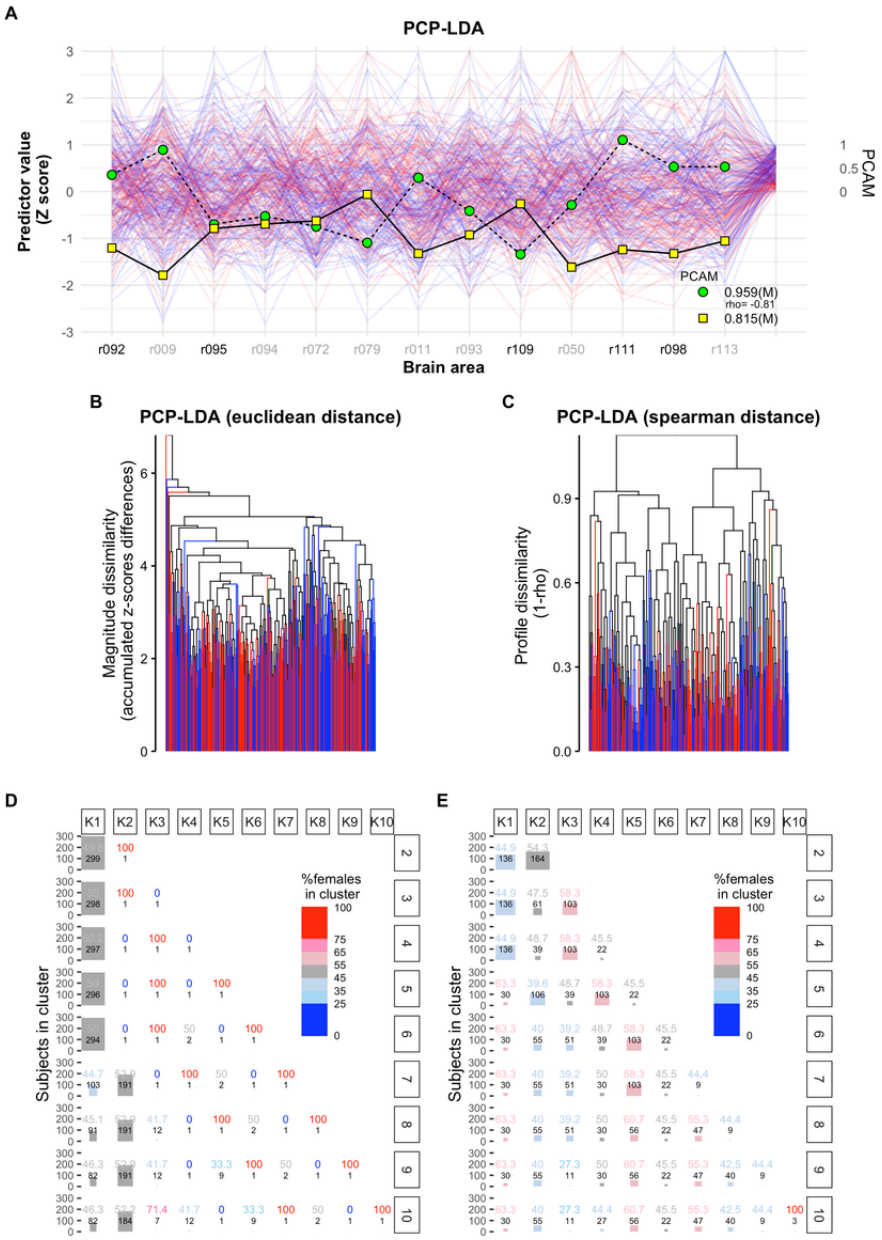

Figure 9

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 Figure 7; grey 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 Figures 8-12).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytablesfigs.xlsx