

Aus dem Institut für Systemische Neurowissenschaften der Heinrich-Heine-Universität
Düsseldorf

**Multivariate Statistical Approaches to investigate Sex Differences in Brain and
Cognition**

Dissertation

zur Erlangung des Grades eines “Doctor rerum medicarum”

(Dr. rer. med.)

der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Lisa Wiersch

(2024)

Teile dieser Arbeit wurden veröffentlicht

Wiersch, L., & Weis, S. (2021). Sex differences in the brain: More than just male or female. *Cognitive Neuroscience*, 12(3-4), 187-188. [1]

Jockwitz, C., Wiersch, L., Stumme, J., & Caspers, S. (2021). Cognitive profiles in older males and females. *Scientific Reports*, 11(1), 6524. [2]

Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K. R., ... & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, 45(6), e26683. [3]

Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., ... & Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*, 13(1), 13868. [4]

Zusammenfassung

Die Dekodierung von individueller Variabilität in Kognition und der Organisation des Gehirns ist essentiell, um unser Verständnis der menschlichen Vielfalt in Gehirn und Verhalten zu verbessern. Dabei ist die individuelle Variabilität oft mit Phänotypen verknüpft, wobei das Geschlecht ein wichtiger Phänotyp ist, der zur individuellen Variabilität beiträgt. Die Erforschung der geschlechtsspezifischen Variabilität verbessert daher nicht nur unser Verständnis von geschlechtsspezifischer Differenzierung in kognitiven Prozessen und der Organisation des Gehirns, sondern hilft darüber hinaus, die Heterogenität in neuropsychologischen Krankheiten, bei der das Geschlecht eine wichtige Rolle spielt, besser zu verstehen. Das Hauptziel dieser Arbeit ist es, multivariate statistische Verfahren als effektive Methodik zur Identifizierung von Mustern in komplexen Datensätzen wie Bildungsdaten des Gehirns oder kognitiven Daten zu präsentieren (Kommentar). Studie 1 zielte insbesondere auf die Untersuchung der geschlechtsspezifischen Differenzierung in neuropsychologischen Daten mit Hilfe von Strukturgleichungsmodellen ab. Die Ziele der weiteren Studien waren, die geschlechtsspezifische Variabilität in der funktionellen (Studie 2) und strukturellen (Studie 3) Gehirnorganisation mit Hilfe von maschinellem Lernen zu untersuchen. In beiden Studien wurden zusätzlich methodische Aspekte untersucht, wie z. B. der Einfluss des Trainings-Datensatzes auf die Generalisierungsperformanz (Studie 2) und der Einfluss konfundierender Variablen (Studie 3). Der Kommentar erläutert die Bedeutung neuer methodischer Ansätze wie multivariates statistisches Lernen, um unser Verständnis der komplexen Natur von Geschlechtsunterschieden zu verbessern. Die Ergebnisse von Studie 1 zeigten, dass es geschlechtsspezifische kognitive Profile gibt, die auf Geschlechtsunterschiede in der kognitiven Verarbeitung zurückzuführen sind. Die Ergebnisse von Studie 2 zeigten Geschlechtsunterschiede in der funktionellen Gehirnorganisation für bestimmte Hirnregionen; wobei allgemein die höchsten Generalisierungsperformanz erreicht wurde, wenn Modelle zur Geschlechtsklassifizierung auf einem großen und heterogenen Datensatz trainiert wurden, welcher die Daten mehrerer Datensätze umfasste. Zusätzlich zeigten die Ergebnisse von Studie 3 Geschlechtsunterschiede in der strukturellen Organisation des Gehirns durch eine akkurate Klassifizierung des Geschlechts mit Modellen, die durch Stratifikation der Gehirngrößen von Männern und Frauen um den konfundierenden Einfluss der Gehirngröße bereinigt wurden. Insgesamt zeigen die vorliegenden Studien, dass multivariate statistische Ansätze die geschlechtsspezifische Variabilität mit strukturellen und funktionellen Bildungsdaten dekodieren können, mit besonderer Berücksichtigung methodischer Aspekte.

Summary

Decoding individual variability in cognition and brain organization is essential to enhance our understanding of heterogeneity in the brain and behavior. Individual variability is often related to specific demographic phenotypes, with sex being a prominent phenotype contributing to individual variability. Examining how differences between males and females are reflected in cognitive and neuroimaging data advances the understanding of sex differences in cognitive processing, brain organization, and the heterogeneity of neuropsychological and mental diseases. To characterize common sources of variability such as sex, the present work aims to present multivariate statistical methods as powerful tools to identify patterns in complex datasets such as neuroimaging or cognitive data (commentary). By using multivariate statistical approaches, the present work examines sex differences in neuropsychological (study 1) and brain imaging data (study 2 & study 3). Specifically, study 1 examined sex-specific cognitive profiles derived from a battery of neuropsychological tests using structural equation modeling. Studies 2 and 3 supplement this investigation by examining sex-related variability in the functional (study 2) and structural (study 3) brain organization using machine learning (ML) approaches. Additionally, methodological considerations in ML were taken into account such as the influence of training samples on the generalization performance of ML models (study 2) and the influence of confounding variables (study 3).

The commentary highlighted the importance of new methodological approaches such as multivariate statistical learning to enhance our understanding of the complex nature of sex differences in rich data. Study 1 identified sex-specific cognitive profiles pertaining to sex differences in component solutions in cognitive processing strategies. Results of study 2 revealed sex differences in the functional brain organization for some, but not all brain regions, with the highest generalization performance when sex classification models were trained on a large and heterogeneous sample comprising the data of multiple datasets. Study 3 demonstrated sex differences in the structural brain organization by accurately classifying sex with ML models that were debiased for the confounding influence of brain size by matching males and females for brain size. In sum, the present studies demonstrated that multivariate statistical approaches can effectively decode sex-related variability in cognitive as well as structural and functional brain imaging data while incorporating important methodological considerations.

List of abbreviations

ADHD	Attention-deficit/hyperactivity disorder
BOLD	Blood Oxygenation Level Dependent
CFA	Confirmatory factor analysis
CM	Cisgender men
CSF	Cerebrospinal fluid
CV	Cross-validation
CW	Cisgender women
DMN	Default mode network
GM	Grey matter
GMV	Grey matter volumes
ML	Machine Learning
MRI	Magnetic resonance imaging
PCA	Principal component analysis
RS	Resting-state
RSFC	Resting-state functional connectivity
SEM	Structural Equation Modeling
SVM	Support Vector Machine
TIV	Total intracranial volume
TM	Transgender men
TW	Transgender women
WM	White matter

Table of Contents

1 INTRODUCTION.....	1
1.1 SEX DIFFERENCES IN COGNITION AND THE BRAIN.....	2
1.2 DIFFERENCES AND SIMILARITIES BETWEEN THE SEXES IN COGNITION AND THE BRAIN	4
1.3 SEX AND GENDER.....	5
1.4 MULTIVARIATE STATISTICAL ANALYSES.....	6
1.4.1 STRUCTURAL EQUATION MODELING.....	7
1.4.2 MACHINE LEARNING	9
1.4.3 METHODOLOGICAL CONSIDERATIONS IN MULTIVARIATE ANALYSES OF SEX DIFFERENCES...	10
1.5 AIMS OF THE THESIS	16
1.6 ETHICS VOTE.....	18
 2 WIERSCH, L., & WEIS, S. (2021). SEX DIFFERENCES IN THE BRAIN: MORE THAN JUST MALE OR FEMALE. COGNITIVE NEUROSCIENCE, 12(3-4), 187-188.....	 19
 3 JOCKWITZ, C., WIERSCH, L., STUMME, J., & CASPERS, S. (2021). COGNITIVE PROFILES IN OLDER MALES AND FEMALES. SCIENTIFIC REPORTS, 11(1), 6524. .	 20
 4 WIERSCH, L., FRIEDRICH, P., HAMDAN, S., KOMAYER, V., HOFFSTAEDTER, F., PATIL, K. R., ... & WEIS, S. (2024). SEX CLASSIFICATION FROM FUNCTIONAL BRAIN CONNECTIVITY: GENERALIZATION TO MULTIPLE DATASETS. HUMAN BRAIN MAPPING, 45(6), E26683.....	 21
 5 WIERSCH, L., HAMDAN, S., HOFFSTAEDTER, F., VOTINOV, M., HABEL, U., CLEMENS, B., ... & WEIS, S. (2023). ACCURATE SEX PREDICTION OF CISGENDER AND TRANSGENDER INDIVIDUALS WITHOUT BRAIN SIZE BIAS. SCIENTIFIC REPORTS, 13(1), 13868.....	 22
 6 DISCUSSION	 23
6.1 ADDRESSING METHODOLOGICAL CONSIDERATIONS IN MULTIVARIATE ANALYSES OF SEX DIFFERENCES	23
6.2 SEX DIFFERENCES IN COGNITION AND THE BRAIN: INSIGHTS FROM MULTIVARIATE ANALYSES.....	25
6.3 UNIVARIATE AND MULTIVARIATE STATISTICAL APPROACHES.....	27
6.4 FACTORS MODULATING SEX-RELATED VARIABILITY	28
6.5 OUTLOOK.....	30
6.6 CONCLUSION.....	30
 7 REFERENCES.....	 32

1 Introduction

A central aim in neuroscience and psychology is to decode the individual variability that drives human diversity in brain and behavior. Common sources of such variability include genetic, cultural, or developmental effects [5, 6], as well as specific phenotypic associations, e.g., the age or sex of a person [7-9]. Linking the individually exhibited variability in brain and behavior to these respective sources advances our understanding of brain organization and cognitive function [10].

A high amount of variability is introduced by the distinction between male and female individuals [8, 11, 12]. Examining how differences between males and females are reflected in cognitive and brain imaging data advances the understanding of sex-related variability in cognitive processing styles and brain organization. Further, there are distinct sex differences in the risk for neuropsychiatric and neurodegenerative disorders [13, 14] such as Alzheimer's disease, Depression, Anxiety, Attention-deficit/hyperactivity disorder (ADHD), Parkinson's disease, and Schizophrenia [15-17]. Understanding the nature of sex differences in brain and behavior can enhance our understanding of the sexual differentiation in a variety of disorders, potentially leading to the development of sex-specific treatments and prevention strategies [14].

There has been considerable debate in the literature regarding sex differences in cognitive abilities and brain organization. While sex differences in structural and functional brain organization, as well as cognition, have been widely studied [12, 18-35], some researchers have argued that the overall similarity between the sexes might be greater than the differences for particular brain regions and cognitive processing [8, 11, 36, 37]. However, much of the work to date has adopted a group differences approach - directly comparing males and females - and focused on specific brain regions or selected single cognitive tasks. We have still yet to understand the fundamental differences between males and females on a global level, beyond individual regions or functions that encompass the overall organization of the brain or cognition.

In contrast to the group differences that have been traditionally used, multivariate methods provide a powerful tool for identifying patterns in rich and complex datasets. Multivariate

analyses offer a way of articulating sex differences in brain and behavior at the global level by identifying differences across multiple variables and considering the interactions among those variables. Consequently, multivariate methods are increasingly applied to study the relationship between phenotypical and neuroimaging data [38-52]. The present work examines sex differences in neuropsychological variables as well as structural and functional brain organization using different multivariate methods, including structural equation modeling and machine learning (ML). Furthermore, the present work focuses on important methodological considerations when adopting multivariate methods to study sex differences to ensure the validity and interpretation of the results.

1.1 Sex differences in cognition and the brain

Males and females have been reported to differ across various cognitive domains. According to Mansouri et al. (2016 [31]), males and females exhibit different modulations of contextual control functions as well as overall executive control. Furthermore, sex differences have been found in the processing of language, showing a female advantage in most verbal tests [18]. This difference is reflected in a female behavioral advantage in linguistic flexibility, verbal fluency, speech articulation, and grammatical skills [18]. A second prominent example of sex differences in cognition occurs in the domain of visual-spatial attention [18, 23-25, 33], in which males tend to outperform females in a variety of visual-spatial, problem-solving, and mental rotation tests [18, 25, 34]. Moreover, sex differences have been observed in memory, particularly working memory [23, 24, 27, 28]. Specifically, sex differences were evident in the processing of the distinct working memory components [24], as males are reported to perform tasks of visuospatial processing more rapidly than females [23, 34]. Overall, several studies have reported sex differences in cognitive task performance across a variety of cognitive domains.

These sex differences in cognitive performance are accompanied by sex differences in performance-related functional magnetic resonance imaging (MRI) brain activation. Brain activation during cognitive task performance can be assessed using Blood Oxygenation Level Dependent (BOLD) imaging, which reflects changes in the blood flow and blood oxygenation, indicating which parts of the brain are activated [53]. For instance, sex differences have been reported in brain activation patterns in regions associated with language during a variety of language tasks [19-22]. Similarly, some studies have reported

that activation patterns of the brain during visual-spatial attention tasks differ between males and females [26, 54]. Moreover, sex differences in memory at the behavioral level are accompanied by differences in functional activity in working memory-related regions [29, 30, 35]. Taken together, there are sex-related differences in behavior and brain activation across various cognitive domains.

In addition to the sex differences in cognition, sex differences have also been reported for the intrinsic functional connectivity in the brain. Functional connectivity is defined as the temporal correlation in the BOLD signal between different regions in the brain [55]. Functional connectivity can be assessed while performing a specific cognitive task or during ‘rest’, meaning when a participant is left to think for themselves and not of anything specific [56]. Resting-state functional connectivity (RSFC) can be used to study large-scale brain networks associated with trait characteristics and behavioral performances [57]. A network of brain regions that exhibits decreased activity during tasks but not in resting-state (RS) is known as the default mode network (DMN [56]). The DMN encompasses brain regions such as the precuneus, posterior cingulate cortex, medial prefrontal cortex, as well as the medial, lateral, and inferior parietal cortices [58]. So far, several studies reported sex differences in the DMN, with females exhibiting a higher level of connectivity than males [58-62].

The basis of functional networks emerges from the structural architecture of the brain [63]; therefore, when examining sex differences in brain function, it is imperative to also consider sex differences in brain structure. The most prominent difference between male and female brains is the overall brain size, known as total intracranial volume (TIV), which is on average higher in males than in females [8, 12, 64]. Moreover, males and females are reported to differ in brain measures such as grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF), with males exhibiting larger volumes than females on average [8, 11, 12]. Several studies have also demonstrated that the GM volumes (GMV) in several brain regions differ between the sexes. For instance, females show higher GMV in medial and lateral prefrontal areas, the superior temporal sulcus, the posterior insula, and prefrontal cortex [32], as well as in the right frontal pole, inferior and middle frontal gyri, pars triangularis, and planum temporale [12]. In contrast, males exhibit higher GMV in subcortical temporal structures, including the amygdala and hippocampus, as well as the temporal pole, fusiform gyrus, visual primary cortex, posterior cingulate gyri, precuneus, putamen, and motor areas [12, 32]. It is important to note that these regional differences are not fully attributable to the

overall differences in TIV [11], which indicates that females and males differ not only globally but also in more localized regions of the brain.

1.2 Differences and Similarities between the sexes in cognition and the brain

While some studies emphasize the differences between males and females, other studies report more similarities than differences, indicating a contradiction in the literature. The contradiction usually refers to the fact that while a sex difference may be significant, the overlap in distributions representing males and females can be greater than the difference between them. Such an overlap was found for several structural brain measures such as brain volumes, GM and WM organization, cortical thickness, cortical surface area, gyrification, and subregional analyses [8, 11, 64]. Furthermore, there is also a high degree of overlap in brain regions related to specific functions, including language lateralization in brain structure, which demonstrated no consistent differences among males and females in terms of GM distributions [22]. Moreover, Weiss et al. (2003 [54]) reported comparable patterns of brain activation for visual-spatial tasks for males and females. Likewise, Voyer et al. (2017 [24]) found that the functioning of visual-spatial working memory is more similar than different between the sexes. Furthermore, Zell et al. (2015 [65]) conducted a metasynthesis demonstrating that sex differences fluctuated across several psychological domains (for example, cognitive variables, social and personality variables, and well-being) but remained largely constant across age, culture, and generations, with most of the effects being relatively small.

The findings of these studies corroborate a study by Hyde (2005 [36]) stating that males and females are more similar than different in most, but not all, psychological variables [36, 37] and that although differences between males and females may exist, the effect sizes of these differences are usually small. Hyde (2005 [36]) reviewed 46 meta-analyses covering different categories, including several cognitive variables, communication, social or personality variables, psychological well-being, and motor behaviors. In one-third of the cases, the effect sizes of the sex differences were small or close to zero, a few were moderate, and none were large [36, 37]. The finding of small effect sizes in the majority of outcomes (close to 80%) provides strong evidence that males and females are more similar than different [36].

Taken together, the literature presents a contradiction regarding the extent of differences and similarities between males and females in terms of brain organization and cognitive functioning. This contradiction may stem from analyses based on the assumption of a clear-cut sexual dimorphism of the brain. The term ‘sexual dimorphism’ refers to a well-defined distinction between males and females due to differences in terms of physical, biological, or behavioral characteristics [66]. Group comparisons contrasting males and females often rely on the assumption of a sexual dimorphism. However, the assumption of a sexual dimorphism of the brain has recently been called into question. According to Joel et al. (2015 [67]), the sexually dimorphic view of a ‘male brain’ and a ‘female brain’ can only hold true if differences in brain features are dimorphic in the sense of a high internal consistency of sex differences with minimal overlap between the distributions. The findings of Joel et al. (2015 [67]), however, demonstrated an extensive overlap between the distributions of males and females for multiple brain measurements, including GM, WM, and connectivity measurements. Further analyses of internal consistency revealed that brains with features that are consistently at one end of the ‘maleness-femaleness’ continuum are extremely rare. Instead, most brains consist of a unique ‘mosaic’ of features, some of which are more common in females compared to males, some more common in males compared to females, and some common in both [67, 68]. According to these results, Joel et al. (2015 [67]) concluded that, although there are sex differences in the human brain, human brains do not fall into the distinct categories of a ‘male brain’ or ‘female brain’. These findings contradict the sexually dimorphic view of the human brain and highlight the need for methods that consider the heterogeneity and variability of the human brain and behavior when investigating sex-related variability.

1.3 Sex and Gender

One important variable that is closely linked to the phenotype sex and its associated variability is the gender a person identifies with. Following the linguistic guidelines for submission of transgender healthcare materials [69], the term ‘sex’ refers to the sex assigned at birth based on anatomical features, while the term ‘gender (identity)’ refers to an individual’s subjective identification influenced by social and psychological factors [70]. The congruence of sex and gender is described as ‘cisgender’ for cisgender men (CM) and women (CW). In contrast, transsexualism refers to the incongruence of sex and gender and

is often associated with gender dysphoria, which describes the clinically significant stress that transgender individuals experience due to this incongruence [71]. To alleviate this distress, transgender individuals may opt for surgery and cross-sex hormone treatment. While ‘transsexualism’ is an umbrella term referring to a variety of gender identities incongruent with the sex assigned at birth, the present work only considers transgender individuals identifying either as men (TM) or women (TW). Moreover, samples of cisgender individuals are mostly based on the “presented sex” of a person (derived based on their name, outer appearance, or self-reported sex) without explicitly collecting information regarding the coherence of sex and gender. For individuals who do not describe themselves as transgender, their gender identity is assumed to be coherent with their sex assigned at birth. It is important to acknowledge that considering sex and gender as binary variables is still a reductionist approach that does not account for the nuanced gender categories of individuals identifying between, outside, or beyond the gender binary [72].

To date, only a few studies have investigated whether and how an incongruence between sex and gender identity might also be reflected in the brain. In general, transgender individuals have been reported to show an altered brain structure compared to cisgender individuals [73]. More specifically, regional differences were found in the putamen [74, 75], insula [76, 77], hypothalamus [78], cortical brain volumes, surface areas [79], and the third ventricle, leading to changes in the overall TIV [78], demonstrating structural brain differences between cis- and transgender individuals. Moreover, sex and gender are both important modifying variables in a variety of diseases, highlighting the need to also consider variability that is not only related to sex alone [16].

1.4 Multivariate statistical analyses

Considering the complexity of various factors associated with the phenotype sex, it is imperative to utilize methods that are capable of addressing this complexity. Specifically, it is necessary to determine the extent to which observed patterns are genuinely related to sex as opposed to other sources of variability within the studied variables. Thus, the question arises whether males and females differ fundamentally in their intrinsic brain organization and cognitive function or whether they are largely similar with only minor differences. To address this question, the group comparisons conducted so far using univariate analyses are not sufficient. Univariate analyses focus only on a single dependent variable but not on the

potential association and interaction of multiple variables [80], which may provide a more nuanced comprehension of sex-related differences between individuals.

Consequently, we should transition from univariate statistical methods to multivariate methods, which allow us to examine the relationships between multiple variables simultaneously. Multivariate analyses aim to identify and interpret the underlying factors of all the variables provided as input for the analyses [80]. Underlying factors are not directly observable variables that typically influence more than one observed measure and are the constructs researchers are most interested in [81-83]. An analysis of the underlying factors also allows us to evaluate whether there are fundamental differences in those factors that are associated with sex. Compared to univariate methods, multivariate analyses offer a more holistic approach by considering multiple variables simultaneously, thereby enabling the identification of potential sex differences in the underlying constructs of the respective variables. Thus, multivariate methods allow for a more accurate representation of the complex characteristics of the phenotype sex and further related factors in the brain and cognition.

Multivariate statistical analyses can incorporate ML techniques, which are a key aspect of statistical learning. In contrast to classical statistics, statistical learning approaches do not primarily focus on confirmatory but exploratory data analysis to infer a model from the presented data inductively [84]. Furthermore, classical statistics are focused on in-sample estimates, whereas statistical learning methods are focused on individual predictions of out-of-sample estimates [84].

1.4.1 Structural Equation Modeling

Multivariate data analysis can be accomplished using various statistical learning approaches, allowing researchers to model and estimate complex relationships among multiple variables [82] and also to determine whether males and females show differences in these models. One particular method to analyze multivariate data is Structural Equation Modeling (SEM [82]). The term ‘SEM’ is an umbrella term for the collection of statistical techniques that enable the examination of multiple variables in relation to one another [82, 85]. Specifically, SEM examines the relationship among underlying factors as measured by one or more indicator variables [82, 83]. These underlying factors influence one or more observed variables and account for the relationships among them [81]. An observed variable can be a participant’s

response to a questionnaire or any directly measured raw data, often referred to as items or indicators [82, 83]. The underlying factors are derived from the intercorrelations shared by multiple indicators [81]. The overall model specifying the relationships among factors, indicators, and their respective correlations are evaluated to determine whether it adequately reflects the actual data [86]. Using SEM can reveal sex differences on a broader level than univariate analysis by assessing not only a single variable as an indicator but the relationships between multiple indicators, their respective intercorrelations, and underlying factors. Such sex differences may be reflected in sex-specific model accuracies, which would indicate that sex-stratified analyses might better reflect the actual data. Whether such differences exist can be evaluated based on the fit of a model to the actual data.

Confirmatory Factor Analysis (CFA) is a useful method to evaluate the fit of a model. CFA is a specific form of SEM, which can be a part of or a precursor of SEM [81, 83] and focuses on measurement models assessing the relationship between indicators and their underlying factors ([81] Figure 1). The relationship is predefined as the model specifies the number of factors and the expected pattern of the indicator-factor relationship in advance based on theoretical grounds [81, 83, 85]. The central aim of CFA is to test this hypothesized structure [85] by assessing how well the variance-covariance matrix of the indicators (observed variables) is reproduced in the model solution of the proposed model [81], which is known as the model fit. The model's fit can be evaluated using a variety of fit estimates. Considering multiple fit indices provides a global summary of the model fit with a more conservative and reliable evaluation of the model fit [81]. Overall, CFA allows for a deeper understanding of how a set of observed indicators is organized based on their underlying factor structure. By evaluating the model fit, CFA allows one to determine whether a proposed factor solution is an adequate model fit for the present data. In the context of multivariate analyses of sex differences, CFA allows assessing whether a factor solution fits both males and females equally well or whether individual factor solutions are required to provide an adequate model fit. This multivariate approach allows us to examine whether there are fundamental differences between males and females in the underlying factor structure of a set of variables.

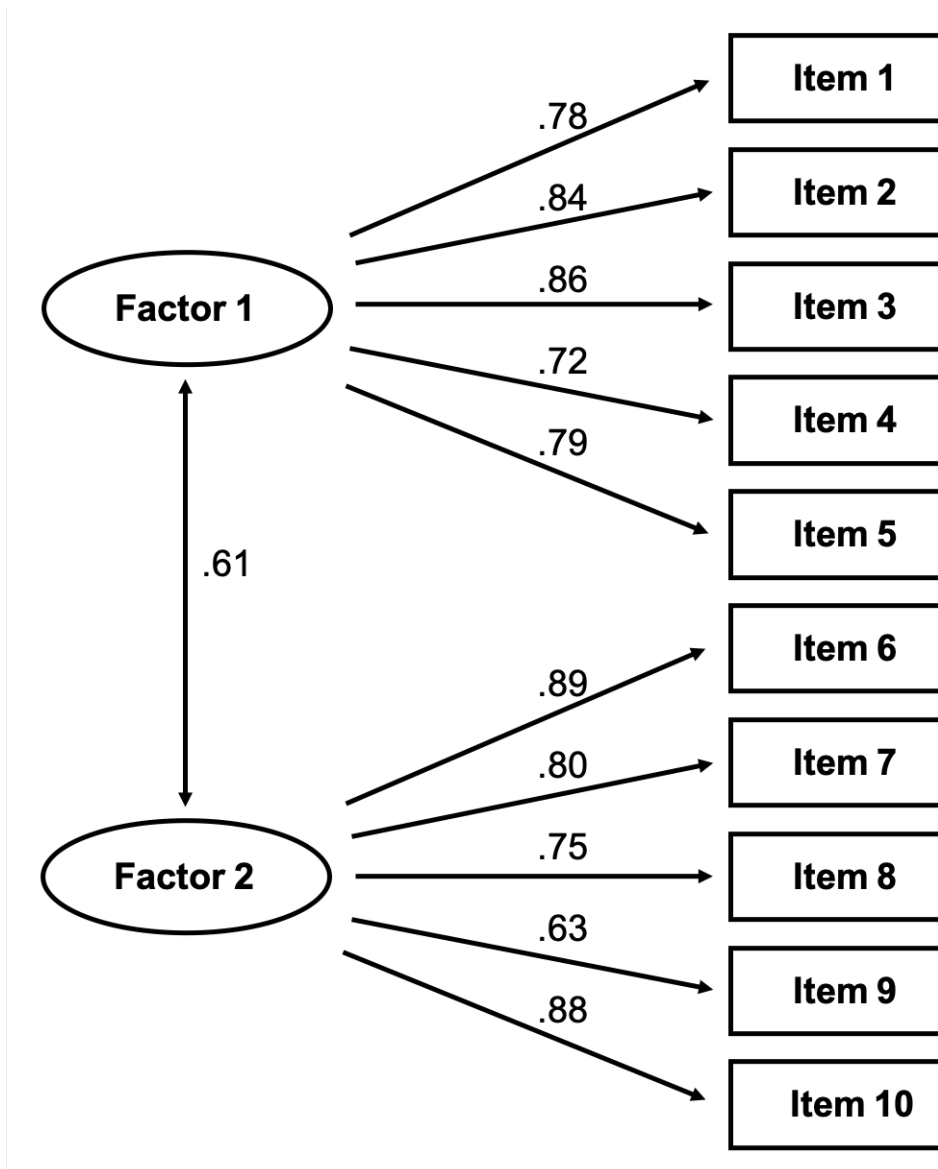


Figure 1. Two-factor confirmatory factor analysis (CFA) with factor loadings

1.4.2 Machine Learning

An alternative method for addressing the complex variability in multivariate data associated with, e.g., the phenotype sex is ML models [87]. ML models are designed to identify patterns in complex data, with the fundamental goal of generalizing these patterns to new and unseen data [88]. Recently, ML approaches have been increasingly applied to neuroimaging data [39, 40, 89-91] to advance our understanding of the brain and how it functions [41]. Combining ML approaches with neuroimaging data has expanded the field from population-based analyses to individualized biomarkers [40].

In the context of using ML to study sex differences, the sex of a person can be used as the outcome variable predicted by multiple input variables in a supervised ML approach. In supervised learning, labels are assigned to the outcome variable (target variable), e.g., the labels ‘male’ and ‘female’ if the sex of a person is the target variable. Within the training process of supervised ML, the given input data (features, e.g., structural or functional neuroimaging data) are related to the respective target variables of each person [92-95]. After learning the feature-target relationship during the training phase, the ML model should be able to generalize the learned patterns beyond the training data. As a result, the ML model is able to make predictions for the target variable in unknown data (out-of-sample predictions [42, 96]. The accuracy of out-of-sample predictions allows us to evaluate the model's ability to make personalized predictions of a person's phenotype, which is referred to as a model's generalization performance [88, 90, 97].

ML models that aim to predict sex by categorizing the target variable as ‘male’ or ‘female’ are referred to as sex classification models. Recent sex classification studies using neuroimaging data have utilized either structural or functional brain imaging features. Studies using RSFC data for sex classification have reported accuracies that range from 62% to 87% [51, 98-100]. Studies using structural brain imaging features for sex classification have reported even higher accuracies that range from 82% to 94% [76, 101-103]. Generally, a high sex classification accuracy indicates that the patterns within the given neuroimaging data differ sufficiently between males and females, allowing a classification model to derive individualized classifications of a person's sex based on these patterns. However, it is essential to determine whether these distinct patterns rely on actual sex differences or on potential methodological shortcomings within the ML models.

1.4.3 Methodological considerations in multivariate analyses of sex differences

Generalizability in Machine Learning

One important methodological consideration in ML is to ensure that the model can actually generalize the learned feature-target relationship to new and unknown datasets. In cases where the model does not capture the general feature-target pattern but rather specializes in

recognizing patterns specific to the training dataset, the ML model is prone to overfit. An overfitted model results in a high model accuracy for the training data but poor generalizability to other datasets [88]. It is essential to avoid overfitting as it hinders making individualized predictions in new data, which impedes the purpose of ML. Failure to address overfitting in sex classification models could lead to ungeneralizable models and an overinterpretation of potential sex differences in the, e.g., structural or functional neuroimaging features due to a high accuracy in the training sample. Therefore, it is crucial to evaluate and avoid overfitting so that the results of the obtained sex classification models truly reflect the predictability of the phenotype sex based on the respective features.

In order to ensure that an ML model does not overfit and is thus able to make generalizable predictions in new data, it is essential to choose a suitable training dataset. Such a dataset is characterized by a large sample size, which is advantageous with regard to the generalization performance [88, 93, 104-106]. Furthermore, most ML models assume that the training data have a similar distribution as the test data [97]. Consequently, a discrepancy in the distributions of training and test data may result in poor generalization performance due to overfitting. It is therefore imperative that the training data is also representative of the test data, so that the model accuracy will not decline [97]. Hence, employing a large and heterogeneous training sample is useful to better represent the heterogeneity of the population from which we can derive potential test samples [93]. Combining several datasets for training purposes is advantageous to achieve such heterogeneity [107-109]. In summary, the choice of the training dataset involves considering a number of factors to ensure accurate generalization performance. It is crucial to consider these factors when selecting an appropriate sample for training sex classification models when analyzing sex differences using multivariate methods such as ML. By doing so, the results from a sex classification model trained on a large, heterogeneous sample are more likely to not depend on the respective training sample. In the present work, study 2 examines the influence of the training sample on the generalization performance and whether a large, heterogeneous dataset can provide more generalizable predictions regarding differences between males and females.

Besides selecting a suitable training set, it is essential to evaluate the ML model within the training process to prevent the model from overfitting and ensure accurate and generalizable predictions. This is accomplished by utilizing a method called cross-validation (CV). Within

this process, the data is split into training and test data, and the test data is used to evaluate the model after it has been trained. In a 10-fold CV, the data is split into 90% training data and 10% test data. This is an iterative process, with each iteration involving testing on a different set of 10% data (Figure 2). The CV process involves adjusting the model parameters with each iteration, therefore increasing the likelihood of high generalization performance for out-of-sample predictions. It is thus possible to obtain a more reliable estimate of its accuracy and to make accurate out-of-sample predictions. Both sex classification studies (studies 2 & 3) in the present work incorporate the CV process into the analyses to ensure reliable and generalizable results.

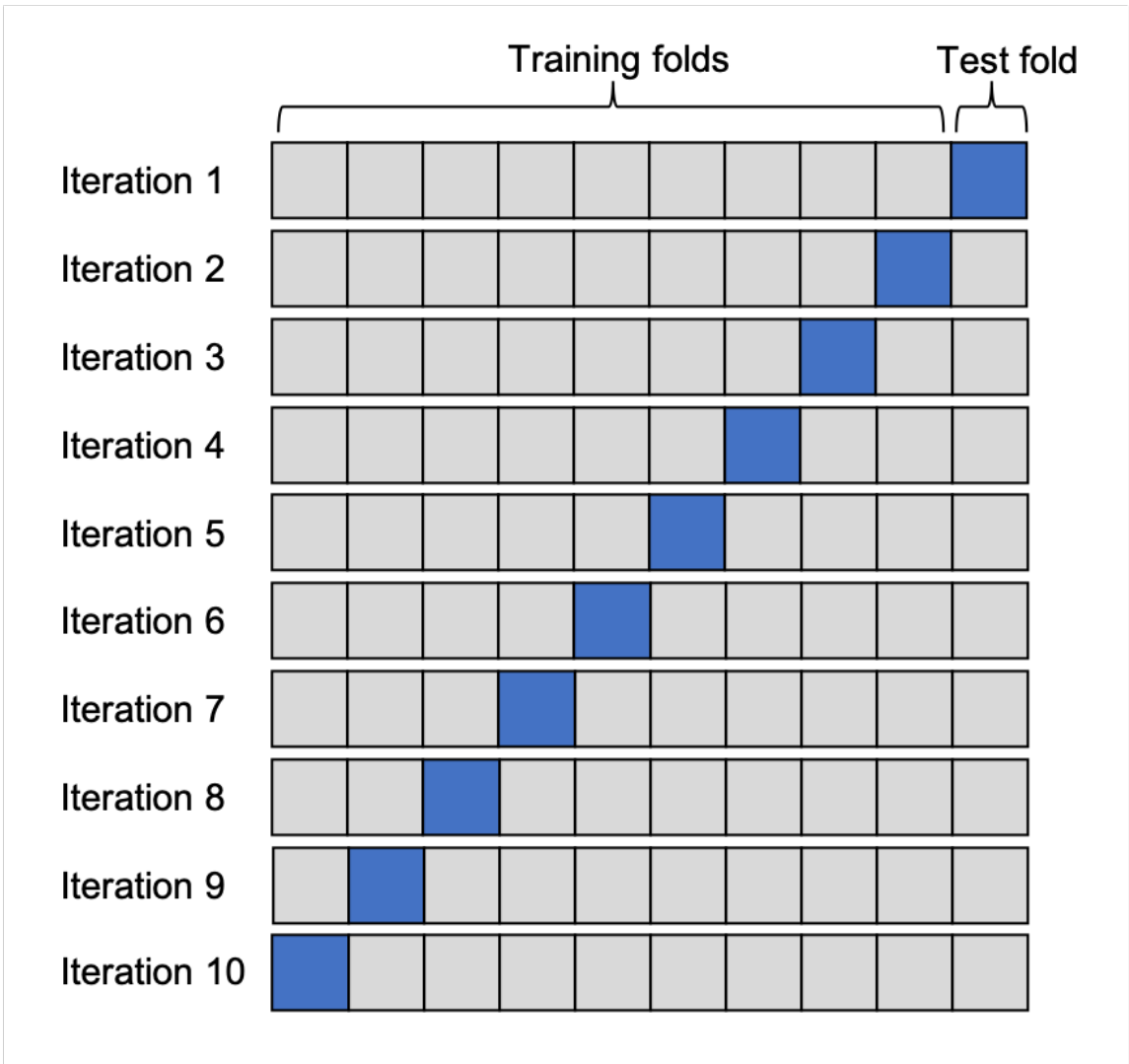


Figure 2. 10-fold cross-validation (CV).

Dimensionality

To ensure a high generalization performance for out-of-sample predictions, it is equally important to examine the feature input provided by the dataset. For instance, neuroimaging data frequently contains features with a high dimensionality, which presents a number of challenges known as the “curse of dimensionality” [93]. One of these challenges is that the higher the dimensionality of the features, the greater the risk that the model will overfit [110]. Similarly, the higher the dimensionality of the input feature space, the more data is required to ensure that the model actually learns the underlying feature-target associations in order to prevent the model from overfitting. For appropriate model training, feature dimensionality must be balanced with the number of available subject data, or, ideally, there should be more subject data than feature dimensionality [93, 111]. Overall, it is therefore essential to reduce highly dimensional data to achieve generalizable predictions in ML.

The principal component analysis (PCA) is a multivariate analysis technique that can be used for dimensionality reduction and to gain a better understanding and exploration of data [112]. PCA reduces complex multivariate datasets to their essential features [112] to provide an overview of complex multivariate datasets [112, 113] and is therefore frequently applied to reduce high dimensional data [114]. In the present work, PCA was applied in studies 1 and 3 to reduce the dimensionality of neuropsychological performance measures and structural neuroimaging data, respectively.

An alternative approach to reduce feature dimensionality when working with neuroimaging data is to use a parcellation of the brain. A brain parcellation is the delineation of distinct spatial partitions in the brain, referred to as brain parcels [115]. Brain parcellations are fundamental for decoding the human brain [115] as they provide insights into organizational principles of the brain along brain structure, function, or connectivity. In addition, they provide a biologically informed data reduction strategy to summarize the data of thousands of voxels into a manageable set of brain regions [115]. In the present work, study 2 followed a parcelwise approach to reduce the dimensionality of the RSFC in order to avoid the curse of dimensionality while enabling the assessment of spatially specific effects. Taken together, employing informed dimensionality reduction techniques is crucial to circumvent overfitting, thus establishing the groundwork for high generalization performances in multivariate analyses.

Choice of algorithm

To address the sex classification analyses in the present work, various ML algorithms could be applied to differentiate between the sexes. Previous work demonstrated successful generalization results for sex classification analyses based on structural [76] and functional [51] neuroimaging data using a support vector machine (SVM) algorithm. Following these approaches, the present ML studies also utilized SVM. Using the supervised SVM algorithm [116-118] for sex classification analyses involves classifying individuals into one of two classes - male or female - based on the labels of the input features. The goal of SVM is to separate the two classes with the widest gap possible, known as the hyperplane (Figure 3 [116, 117]). Upon successfully training the model, new data should be classified accordingly based on its position relative to the hyperplane, anticipating accurate generalization of the model.

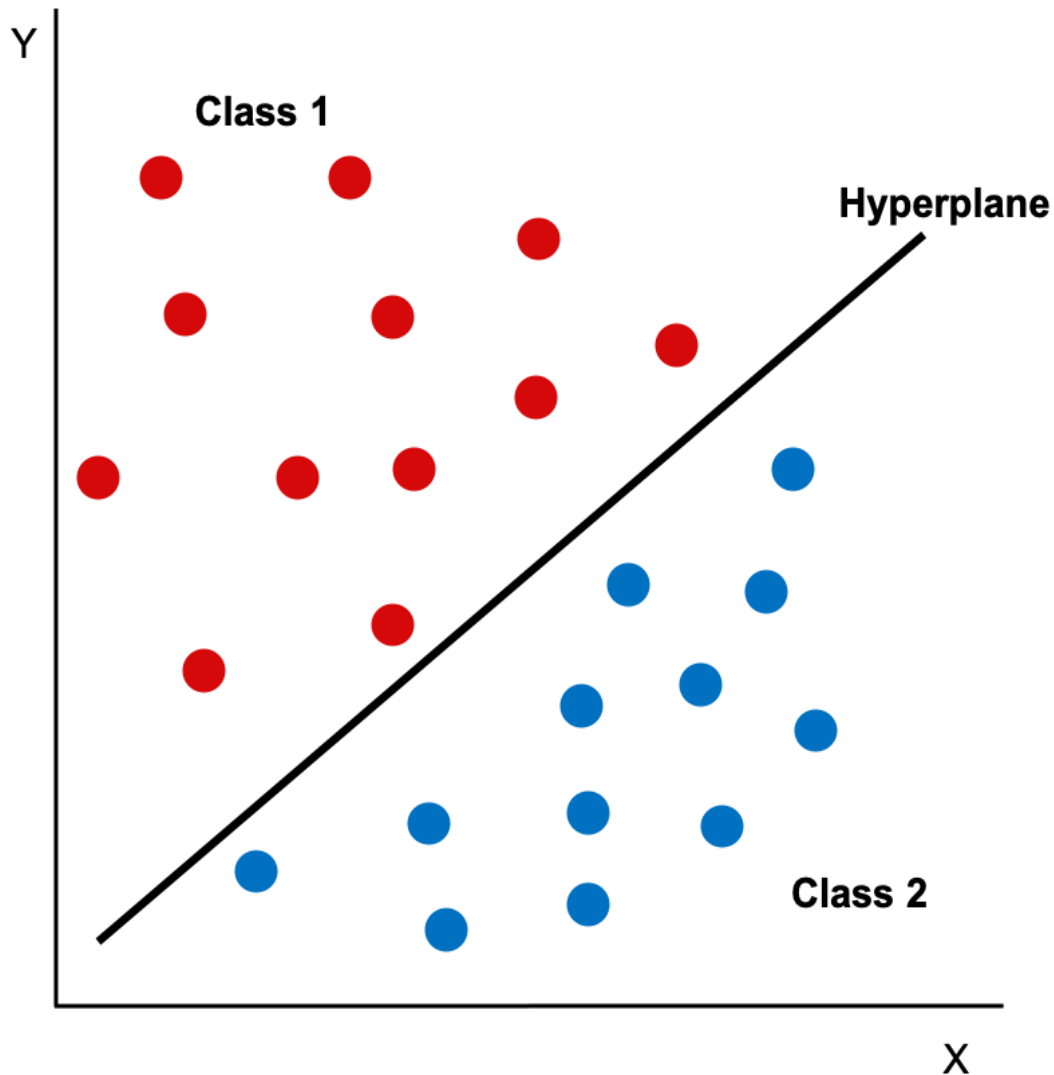


Figure 3. Support Vector Machine (SVM) Functionality

Confounding variables

While aiming for accurate and robust out-of-sample predictions, it is also essential to ensure that the target variable is actually predicted by the features of interest. A major concern in this context is the methodological issue of confounding variables. A confounding variable is a variable that is associated not only with the features of interest but also with the target variable [119]. The presence of a confounding variable within the features used as ML input might bias the predictions of the target variable. In a biased prediction, the causal feature-target effect can be either hidden by the confounder, or the confounder may suggest an effect where none exists [120]. As a result, confounding effects can lead to spurious conclusions or cause important associations to be missed [121]. Therefore, it is crucial to appropriately control for confounding variables in multivariate analyses to ensure accurate individualized predictions.

A prominent example of a confounding variable in examining sex differences in neuroimaging data is the overall brain size difference between males and females. Particularly in the context of structural brain imaging data, TIV is not of primary interest but is nonetheless embedded in the features derived from the structural data [122]. TIV can bias predictions when using e.g. GMV as structural brain imaging data for training a sex classification model. In this case, a TIV-biased model will rather learn to classify males and females based on the overall difference in TIV at the expense of learning sex-differentiating patterns in the GMV structures, which are the actual features of interest [122]. These biased predictions occur since TIV alone is an effective predictor of a person's sex [123].

Different methodologies to control for confounding variables exist, such as randomization, restriction, and matching for the respective confounding variable [119, 124]. However, most studies reported a reduced sex classification accuracy with TIV-controlled compared to TIV-uncontrolled structural brain information [123, 125, 126]. Thus, while it is critical to control for confounding variables in multivariate analyses, it is so far unclear whether it is also possible to maintain a high level of classification accuracy in spite of removing information such as TIV. In the present work, study 3 addresses this problem by evaluating two methods to control for confounding variables, using the example of TIV as a confounder in structural sex classification analyses.

1.5 Aims of the thesis

We have elaborated on the ongoing debate in the literature regarding the extent of sex differences in brain and cognition. Despite significant differences in specific cognitive domains and brain regions reported by studies using group comparisons, males and females may share more similarities than differences in the overall organization of the brain and cognition. In order to investigate sex differences more globally, it is essential to use statistical methods that allow to assess sex differences beyond one single variable to examine whether males and females differ in the overall organization of the brain and cognition. Multivariate analysis approaches, including e.g. SEM and ML, provide suitable statistical tools for addressing this complexity. Furthermore, multivariate statistical learning approaches utilizing sex classification analyses enable modeling complex relationships and generalizing to new data. For this purpose, it is crucial to ensure an appropriate learning of the feature-target association by carefully considering several methodological considerations.

The present work presents multivariate methods as powerful tools to disentangle the complex pattern of sex-related variations in cognition and brain organization (commentary). Specifically, the present studies examine sex-related variability in cognition using SEM (study 1) as well as functional (study 2) and structural (study 3) brain imaging data using ML models. In light of the methodological considerations inherent in sex classification analyses, study 2 examined the impact of the training sample on the generalization performance and study 3 examined the effective removal of confounding variables. Taken together, the findings of studies 2 and 3 aim to ensure a valid interpretation of sex differences in brain organization using sex classification analyses. Overall, all three studies utilize multivariate statistical methods based on data-driven approaches to examine sex differences on a holistic level, encompassing multiple variables.

Commentary

Wiersch, L., & Weis, S. (2021). Sex differences in the brain: More than just male or female. *Cognitive Neuroscience*, 12(3-4), 187-188.

The commentary in the present work emphasizes the significance of innovative methodological approaches such as statistical learning, as these multivariate techniques offer effective tools to disentangle the complex patterns of variability provided by multivariate data. The usage of these tools is particularly valuable in assessing how large and meaningful

the differences between the sexes are in relation to the overall variability in cognitive and neuroimaging data.

Study 1

Jockwitz, C., Wiersch, L., Stumme, J., & Caspers, S. (2021). Cognitive profiles in older males and females. *Scientific Reports*, 11(1), 6524.

Study 1 examined sex-related variability in cognition based on neuropsychological performance measures. Based on the assumption that males and females exhibit differences in their cognitive processing in neuropsychological tasks, we expected these differences to be reflected in the form of sex-specific cognitive profiles. Using a PCA, neuropsychological performance was first decomposed into cognitive components. The derived component solutions were then evaluated using CFA to assess whether males and females demonstrated fundamental differences in their component solutions.

Study 2

Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K. R., ... & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, 45(6), e26683.

In study 2, we systematically investigated the generalization performance in sex classification analyses based on functional neuroimaging data. Specifically, we investigated which kind of training sample is best suited to optimize the generalization performance in sex classification models. For this purpose, we compared sex classifiers trained on the parcelwise RSFC profile of either single samples or compound samples containing data from four different datasets with varying sample characteristics. In order to evaluate the generalization performance of each of the models derived according to the different training samples, the models were tested on multiple datasets.

Study 3

Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., ... & Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*, 13(1), 13868.

In study 3, we addressed the methodological consideration of confounding variables, particularly how to control for the influence of TIV bias in sex classification analyses. Specifically, two approaches for removing confounding information were evaluated:

featurewise confound removal of TIV information and stratifying training samples for TIV. Finally, we evaluated the efficacy of each approach with regard to successful debiasing for TIV while retaining accurate model performances.

1.6 Ethics vote

In study 1, data was obtained from the 1000Brains dataset [127]. All participants in this dataset gave written informed consent, and all experiments were performed in accordance with relevant guidelines and regulations. The 1000BRAINS study protocol was approved by the Ethics Committee of the University of Duisburg-Essen (reference number: 11-4678, 12-5199-BO).

The ethics committee of the medical faculty of the RWTH Aachen approved the acquisition of data for one sample used in study 3 (EK 088/09). Further usage of data analyzed in study 3 and 2 was approved by the Ethics Committee of the Medical Faculty of the Heinrich-Heine-University Düsseldorf (4039, 4096, 5139, 2018-317-RetroDEuA). All data were collected in research projects approved by a local Review Board, for which all participants provided written informed consent. All experiments were performed in accordance with relevant guidelines and regulations.

2 Wiersch, L., & Weis, S. (2021). Sex differences in the brain: More than just male or female. *Cognitive Neuroscience*, 12(3-4), 187-188.

COMMENTARY



Sex differences in the brain: More than just male or female

Lisa Wiersch^{a,b} and Susanne Weis^{a,b} 

^aInstitute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany; ^bInstitute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

ABSTRACT

Sex differences in the brain are widely studied, but results are often inconsistent and it is assumed that many negative findings are not even being reported. The lack of consistent findings might be based on the highly questionable assumption of a clear-cut sexual dimorphism in brain structure and function, that underlies commonly used group comparisons between males and females. Without having to rely on this assumption, state of the art statistical learning methods based on large neuroimaging data sets might offer the tools necessary to disentangle the complex pattern of sex-related variations in brain structure and organization.

ARTICLE HISTORY

Received 3 October 2020
Revised 7 November 2020
Published online xx xxx xxxx

Employing a qualitative review and an activation likelihood estimation (ALE) meta-analysis of eight functional magnetic resonance imaging (fMRI) papers, Spets and Slotnick claim to have identified compelling evidence for substantial sex differences in brain activity during long-term memory retrieval. Unfortunately, their methodological approach is questionable. In an era of very large-scale neuroimaging (e.g., >5000 subjects in (Ritchie et al., 2018)), all studies included in their meta-analysis comprise (much) less than 50 subjects. Small participant numbers have been shown to impact the reliability of cognitive neuroscience studies (Thirion et al., 2007), a problem that can, in principle, be circumvented by proper use of meta-analyses. However, the present ALE analysis ignores the clear recommendation – based on a simulation study by the authors of the ALE approach (Eickhoff et al., 2016) – to include at least 17 experiments. When even including less than ten experiments, ALE scores of a single experiment may already be close to significance relative to the overall null-distribution and results of the meta-analysis might simply reflect results of a single experiment (Eickhoff et al., 2009; Muller et al., 2018).

Furthermore, considering that sex differences as well as long-term memory are highly researched topics, the inclusion of only eight studies in the meta-analysis in itself might point towards a file drawer problem, where negative findings on differences between the sexes are simply not reported. This assumption is supported by results of a large meta-analysis involving 179 studies,

which indicates an excess of false positives as well as a strong publication bias in the sex differences literature (David et al., 2018).

Looking beyond the present study, the above considerations point to a much more fundamental problem in sex differences research in neuroimaging (and in general): The commonly adopted group comparison approach simply is not sufficient to capture the complex nature of sex differences in the brain. Still, the vast majority of sex differences research is based on the highly questionable assumption of a clear-cut sexual dimorphism in the brain, which would only be justified if male and female brain features could be assumed to cluster distinctively and consistently at opposite ends of a single-gender continuum (Rippon et al., 2014).

On the contrary, recent research (Joel et al., 2015; Weis et al., 2020) based on big data sets indicates that it is time to move away from considering sex differences in the brain as fixed, in-variant over time, or binary with sharply defined category boundaries (Rippon et al., 2014). Rather, most brain features appear to be highly overlapping between the sexes (Joel et al., 2015), indicating that sex differences in the brain are not defined by biological sex alone but rather modulated by a variety of factors, some of which might even be dynamically changing over relative short time frames (e.g., the female menstrual cycle).

In an era of very large-scale neuroimaging, new methodological approaches like statistical learning are

CONTACT Susanne Weis  S.Weis@fz-juelich.de  Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Commentary on: 'Are there sex differences in brain activity during long-term memory?' by Dylan Spets and Scott Slotnick.

© 2021 Informa UK Limited, trading as Taylor & Francis Group

needed to understand the complex nature of sex differences in the brain. In fact, the actual question that should be asked is not whether or not sex differences exist in the brain, but rather how large and meaningful such differences are in relation to variation within each sex as well as sex-independent inter- and intra-individual variance. While we agree with the authors in that we should 'question the widespread practice of collapsing across sex in the field of cognitive neuroscience,' much more detailed research is needed to actually understand sex differences in the brain to an extent that is transferable to real-life and clinical applications. Almost inevitably, such studies will have to rely on very large neuroimaging data sets (Ritchie et al., 2018) in combination with appropriate statistical learning approaches (Weis et al., 2020).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Susanne Weis  <http://orcid.org/0000-0001-7726-6470>

References

- David, S. P., Naudet, F., Laude, J., Radua, J., Fusar-Poli, P., Chu, I., Stefanick, M. L., Ioannidis, J. P. A. (2018). Potential Reporting Bias in Neuroimaging Studies of Sex Differences. *Scientific Reports*, 30(9), 6082. PMID: 29666377, PMCID: PMC5904173. <https://doi.org/10.1038/s41598-018-23976-1>
- Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox P. T., Bzdok, D., & Eickhoff, C.R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage*, 137, 70–85. PMID: 27179606, PMCID: PMC4981641. <https://doi.org/10.1016/j.neuroimage.2016.04.072>
- Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9), 2907–2926. <https://doi.org/10.1002/hbm.20718>
- Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., Shefi, N., Pool, J., Urchs, S., Margulies, D. S., Liem, F., Hänggi, J., Jäncke, L., & Assaf, Y. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences*, 112(50), 15468–15473. PMID: 26621705, PMCID: PMC4687544. <https://doi.org/10.1073/pnas.1509654112>
- Muller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., Tench, C. R., Yarkoni, T., Nichols, T. E., Turkeltaub, P. E., Wager, T. D., & Eickhoff, S. B. (2018). Ten simple rules for neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 84, 151–161. PMID: 29180258, PMCID: PMC5918306. <https://doi.org/10.1016/j.neubiorev.2017.11.012>
- 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. *Frontiers in Human Neuroscience*, 8, 650. PMID: 25221493, PMCID: PMC4147717. <https://doi.org/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. *Cerebral Cortex*, 28(8), 2959–2975. <https://doi.org/10.1093/cercor/bhy109>
- Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage*, 35(1), 105–120. <https://doi.org/10.1016/j.neuroimage.2006.11.054>
- Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., & Eickhoff, S. B. (2020). Sex Classification by Resting State Brain Connectivity. *Cerebral Cortex*, 30(2), 824–835. <https://doi.org/10.1093/cercor/bhz129>

3 Jockwitz, C., Wiersch, L., Stumme, J., & Caspers, S. (2021). Cognitive profiles in older males and females. *Scientific Reports*, 11(1), 6524.



OPEN

Cognitive profiles in older males and females

C. Jockwitz^{1,2}✉, L. Wiersch³, J. Stumme^{1,2} & S. Caspers^{1,2}

Males and females are subject to differences in cognitive processing strategies, i.e. the way males and females solve cognitive tasks. So far primarily reported for younger adults, this seems to be especially important in older adults, who also show sex differences in cognitive impairments. Therefore, the aim of the current study was to examine the older adult population with respect to cognitive profiles derived from a large variety of cognitive functions. Using an exploratory component analysis with consecutive confirmatory factor analysis in a sample of 676 older adults, neuropsychological performance data in a variety of cognitive domains was decomposed into cognitive components. A general cognitive profile based on the whole group fits unequally well on the two sexes. Importantly, cognitive profiles based on either males or females differ in terms of their composition of cognitive components, i.e. three components in males versus four components in females, with a generally better model fit in females. Thus, related to the established differences in processing styles between males and females the current study found a rather decomposed (or local) cognitive profile in females while males seem to show a holistic (or global) cognitive profile, with more interrelations between different cognitive functions.

There has been a longstanding debate about whether males and females differ in terms of cognitive abilities. Males are often supposed to outperform females in visual spatial tasks, while females outperform males in terms of verbal and episodic memory tasks^{1–6}. While these sex stereotypes are well accepted in our society⁷, there is a non-negligible amount of studies showing exactly the opposite, namely that men and women do not differ in most of the cognitive tasks, also referred to as the “Gender Similarity Hypothesis”^{8,9}. That is, cognitive performance differences on average show an effect size of $d = 0.22$ (range: 0.05–0.57) which is interpreted as rather small differences. Using a meta-synthesis approach, Zell et al.¹⁰, however, concluded that sex differences in terms of psychological and cognitive variables is rather small but stable across ages, generations and cultures.

Besides investigating sex differences in absolute cognitive performance outcome measures (i.e. females remember more words from a word list as compared to males), recent studies rather focussed on sex differences in cognitive processing styles, i.e. the way males and females solve a given cognitive task^{11,12}. For example, in spatial navigation tests, females were found to use local landmarks to find a specific route, while males rather construct cognitive maps of the environment^{11,13,14}. Interestingly, when males and females are instructed to actively choose a landmark-based style, females outperform males in this task¹³. Similarly, in a verbal fluency task, Weiss et al.¹⁵ as well as Lanting et al.¹⁶ showed that the males’ processing strategy is typically characterized by a systematic and extensive scan of the word space of a given category before moving to the next one (e.g. listing jobs, males would first list all jobs within a hospital, then within an office etc.). In contrast, females switch more often between different categories. Changing the instructions, i.e. inducing more switches between categories, led to superior performance of females¹². Thus, based on previous research investigating specific cognitive tasks, it has been established that males and females use different cognitive processing strategies: Males seem to use a rather holistic processing style with a focus on global aspects of the task (i.e. having in mind the whole map of a city when performing a spatial navigation task). Females instead use a decomposed processing style with a focus on more local aspects of the task (i.e. remember more details of a given word list). Similar sex differences in terms of a global versus local focus have been found for other tasks such as mental rotation tasks¹⁷, number-comparison-task¹⁸ and Navon paradigms¹⁹.

Although sex-related differences in cognitive processing styles do not necessarily result in differences in performance in everyday life, i.e. males and females perform equally good in an everyday multitasking paradigm²⁰, they give rise to the question of whether males and females do not only differ in single cognitive abilities. Rather, the two sexes might generally differ in the overall composition of their cognitive abilities. So far, studies mostly

¹Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany. ²Institute for Anatomy I, Medical Faculty & University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ³Institute of Neuroscience and Medicine (INM-7), Research Centre Jülich, Jülich, Germany. ✉email: c.jockwitz@fz-juelich.de

focus on cognitive profiles that are predefined based on specific cognitive theories or derived from data-driven approaches (e.g. principal component analyses)^{21–25}. For example, single cognitive abilities are often categorized into cognitive domains, such as attention, memory and executive functions, based on correlations between performance in the tasks administered²¹. Performance within the cognitive domains, then, together represent cognitive profiles. Typically, such approaches are based on an entire group including both, males and females. However, whether these cognitive components and profiles reflect the cognitive architecture equally well across the two sexes remains unclear. The relation between performance in distinct cognitive tasks might be differentially related to each other in males and females and therefore might form different sex-specific cognitive profiles.

Particularly interesting in this aspect is the older adult population, since sex differences in cognitive performance were found to persist until late adulthood and might even lead to differences in cognitive impairments during older age and disease^{2,22,23}. So far, the majority of studies investigate cognitive performance during aging while correcting for sex differences. Averaged over the two sexes, cognitive performance decline is well established during the aging process^{24–27} with a significant decline starting in the mid 50's²⁴, especially in the domains of executive functions, working memory and episodic memory. However, previous studies not only showed that sex-differences in cognitive performance persist until late adulthood^{2,22,23}, they also reported unbalanced prevalence in neurodegenerative diseases that are accompanied by different cognitive impairments, i.e. males rather suffer from MCI and Parkinson's disease, while females are more often affected by Alzheimer's disease^{28,29}. Potentially, different interrelations between cognitive functions might explain parts of these different age-related trajectories and therefore depict a promising research topic. To examine this, the current study took advantage of a large older adult population of males and females between 55 and 85 years from the 1000BRAINS cohort, matched for age and education, and examined the sex-specificity of cognitive profiles based on a large variety of neuropsychological functions. Using a data-driven approach, neuropsychological test performance was first decomposed into cognitive components. Afterwards the different component solutions were statistically compared between the two sexes. Based on the sex-specific strategies found when investigating specific cognitive tasks (i.e. global versus local processing strategies), we would expect these differences to be also reflected in sex-specific cognitive profiles.

Methods

Subjects. Subjects included in the current study were drawn from 1000BRAINS³¹, a population-based epidemiological cohort study, recruited from the Heinz-Nixdorf recall study that has been conducted in the Ruhr area in Germany³². Along the line of being population-based, exclusion from the study was based on eligibility for MR measurements for scientific purposes. From the initial cohort of 1314 subjects, 968 subjects being 55 years and older were selected to assess the older adult population. 20 subjects had to be excluded due to missing variables of interest for the current study (DemTect³³: $n = 18$; or information on education: $n = 2$). Furthermore, subjects missing more than three values of the neuropsychological assessment ($n = 2$; for all other subjects missing values (ranging from 0 to 2.1% depending on the test) were replaced by the median of the respective age- (<60; 60–64; 65–69; 70–74; 75–79; >79) and sex-group. Subjects representing outliers ($n = 83$; outliermax > mean + 3*SD; outliermin < mean – 3*SD) in at least one of the cognitive variables were removed from the dataset. To establish similar demographic conditions in the two sex groups, propensity score matching (method = "nearest", caliper = 0.25; implemented in R: matchit, version 3.0.3) was used to match males and females for age and education (measured by ISCED³⁰) which resulted in a final sample size of 676 subjects between 55 and 87 years of age: 338 males with a mean age of 66.9 years (SD = 6.7) and a mean ISCED score of 6.3 (SD = 1.74) and 338 females with a mean age of 66 years (SD = 6.5) and a mean ISCED score of 6.1 (SD = 1.86). All participants gave written informed consent before participating in 1000BRAINS. All experiments were performed in accordance with relevant named guidelines and regulations. The study protocol was approved by the local Ethics Committee of the University of Essen.

Neuropsychological assessment. All subjects underwent intensive neuropsychological testing during their participation in 1000BRAINS³¹. In total, 16 different cognitive functions, namely selective attention, processing speed, problem solving, concept shifting, susceptibility to interference, figural fluency, phonematic and semantic verbal fluency, vocabulary, verbal episodic memory, figural memory, visual-, visual-spatial- and verbal short-term/working memory were assessed. For cognitive functions and tests used, as well as raw mean scores for males and females, see Table 1.

Statistical analyses. First, sex differences in cognitive performance were examined for the different cognitive functions assessed in 16 different neuropsychological tests using Independent Sample T-Tests. Effect sizes were calculated using Cohen's d . Afterwards, we calculated z -scores for each variable followed by Pearson correlations between all neuropsychological variables included in the current analysis for the whole group, as well as for males and females separately.

The major research question in this study concerned whether males and females would show different cognitive profiles, i.e. different compositions of cognitive components. To investigate this, we divided our analyses in two parts (for an overview of analyses, see Fig. 1, part A and part B). In the first part (part A), we extracted cognitive components for both, the whole group ($n = 676$) including males and females, as it is commonly done in research investigating cognitive performance (e.g. see^{46–51}, as well as for males ($n = 338$) and females ($n = 338$) separately to identify commonalities as well as differences in cognitive profiles between the two sexes. For all the groups (whole, males and females) a two-step approach was applied:

Cognitive Function	Test description	Females: mean $\pm$ SD (Min; Max)	Males: mean $\pm$ SD (Min; Max)	T-value	p-value	Cohen's d
Age		65.99 $\pm$ 6.5 (55.2;85.4)	66.87 $\pm$ 6.65 (55.1;85.4)	-1.735	0.083	0.132
DemTect	DemTect ³³ : Global cognitive score	15.55 $\pm$ 2.22 (8;18)	14.17 $\pm$ 2.36 (8;18)	7.860	0.000	-0.587
ISCED97	International Classification ³⁰ : Education classification	6.1 $\pm$ 1.86 (3;10)	6.29 $\pm$ 1.74 (3;10)	-1.370	0.171	0.109
Problem solving	Leistungsprüfungssystem 50+ (Subtest 3) ³⁴ : Number of correctly identified non-matching figures among geometrical figures	20.39 $\pm$ 4.71 (8;35)	20.82 $\pm$ 5.13 (8;34)	-1.132	0.258	0.084
Visual STM	Block-Tapping-Test ³⁵ : Number of correctly repeated blocks, forwards	6.32 $\pm$ 1.76 (2;10)	6.57 $\pm$ 1.65 (2;10)	-1.937	0.053	0.154
Visual WM	Block-Tapping-Test ³⁵ : Number of correctly repeated blocks, backwards	4.69 $\pm$ 1.65 (1;10)	5.04 $\pm$ 1.7 (0;10)	-2.738	0.006	0.208
VisualSpatial STM	Visual pattern (Jülich version; similar to ³⁶): Number of memorized patterns presented in a grid of black and white squares	7.32 $\pm$ 1.7 (4;12)	8.06 $\pm$ 1.68 (4;12)	-5.711	0.000	0.443
Verbal STM	Zahlennachsprechen (from Nürnberger Alters-Inventar ³⁷): Number of correctly repeated digits, forwards	7.63 $\pm$ 1.84 (4;13)	7.66 $\pm$ 2.02 (4;13)	-0.179	0.858	0.013
Verbal WM	Zahlennachsprechen (from Nürnberger Alters-Inventar ³⁷): Number of correctly repeated digits, backwards	6.79 $\pm$ 1.65 (2;12)	6.87 $\pm$ 1.77 (2;12)	-0.653	0.514	0.049
Figural memory	Benton-Test ³⁸ : Number of errors during free recall of 20 remembered figures	-16.33 $\pm$ 7.57 (-40; -2)	-16.17 $\pm$ 7.56 (-36; -1)	-0.275	0.784	-0.021
Selective attention	Alters-Konzentrations-Test ³⁹ : Time(s) to recognize target figures among distractors	-33.54 $\pm$ 8.74 (-64.78; -17)	-33.66 $\pm$ 8.38 (-65.87; -18.22)	0.183	0.855	0.014
Interference	Farb-Wort-Interferenztest (Jülich version; similar to: Bäumlert ⁴⁰ ; Stroop ⁴¹): Time(s) to name ink color of words with color meaning but printed in a different color (subtracted by the time(s) to read color words)	-39.63 $\pm$ 16.64 (-110.6; -9.47)	-43.36 $\pm$ 17.58 (-109.97; -3.66)	2.833	0.005	0.212
Figural fluency	Fünf-Punkte-Test (Jülich version; similar to: Regard et al. ⁴²): Number of unique drawn patterns by connecting five points in 3 min	26.15 $\pm$ 6.89 (4;44)	26.38 $\pm$ 7.22 (11;49)	-0.425	0.671	0.032
Episodic memory	Verbaler Gedächtnistest ⁴³ : Number of free recalled words in five trials from a list containing 15 words	45.76 $\pm$ 10.05 (2;66)	38.61 $\pm$ 10.01 (6;65)	9.262	0.000	-0.714
Phonematic fluency	Regensburger Wortflüssigkeitstest ⁴⁴ : Number of produced words beginning with the letter "B"	19.32 $\pm$ 6.04 (4;37)	17.49 $\pm$ 5.93 (5;37)	3.992	0.000	-0.310
Semantic fluency	Regensburger Wortflüssigkeitstest ⁴⁴ : Number of produced words belonging to the category "jobs"	24.47 $\pm$ 6.19 (11;44)	23.39 $\pm$ 6.76 (6;43)	2.153	0.032	-0.159
Processing speed	Trail Making Test (taken from CERAD-Plus ⁴⁵): Time(sec.) to connect randomly arranged digits in ascending order	-38.62 $\pm$ 11.71 (-79.41; -16.06)	-40.22 $\pm$ 13 (-84.18; -16.13)	1.677	0.094	0.123
Concept shifting	Trail Making Test (taken from CERAD-Plus ⁴⁵): Time(sec.) to alternately connect letters and numbers in ascending order (TMT B), then calculating: TMT B-TMT A	-48.71 $\pm$ 28.33 (-183.44; -1.78)	-54.28 $\pm$ 32.46 (-166.6; 0.67)	2.375	0.018	0.171
Vocabulary	Wortschatztest ⁴⁶ : Number of correctly identified real words among five pseudo-words	30.96 $\pm$ 4.34 (16;40)	30.8 $\pm$ 4.17 (16;40)	0.493	0.622	-0.039

Table 1. Descriptives of neuropsychological variables including cognitive functions, tasks used, mean and standard deviation (SD) and Min; Max values, T value of group comparison with corresponding *p* value and effect size measured with Cohen's *d*. Values written in bold indicate significant differences between groups (*p* < .05). STM = short-term memory, WM = working memory.

- (1) *Data reduction* We reduced the cognitive performance data into independent cognitive components by using exploratory principal component analysis (ePCA) with Varimax rotation (implemented in the "psych" package, R Studio), as one of the most commonly used technique for data reduction⁵². The number of extracted components was based on the eigenvalue criterion (eigenvalue > 1). This resulted in three independent data-driven component solutions: whole ePCA based on the whole group, male ePCA within males only, females ePCA within females.
- (2) *Component solution validation within respective groups* To validate the obtained component solutions in their respective group (whole ePCA, male ePCA, females ePCA), a confirmatory factor analysis (CFA, implemented in the "lavaan" package, R Studio) was set up with Maximum Likelihood estimator with robust standard errors and a Satorra-Bentler scaled test statistic. In detail, each component solution represents a measurement model that is composed of a specific number of cognitive components, with each including a specific number of cognitive performance tests. In the current study, we based the measurement models on the component solutions obtained by ePCA and included all cognitive performance tests with a component loading of at least 0.4⁵³.

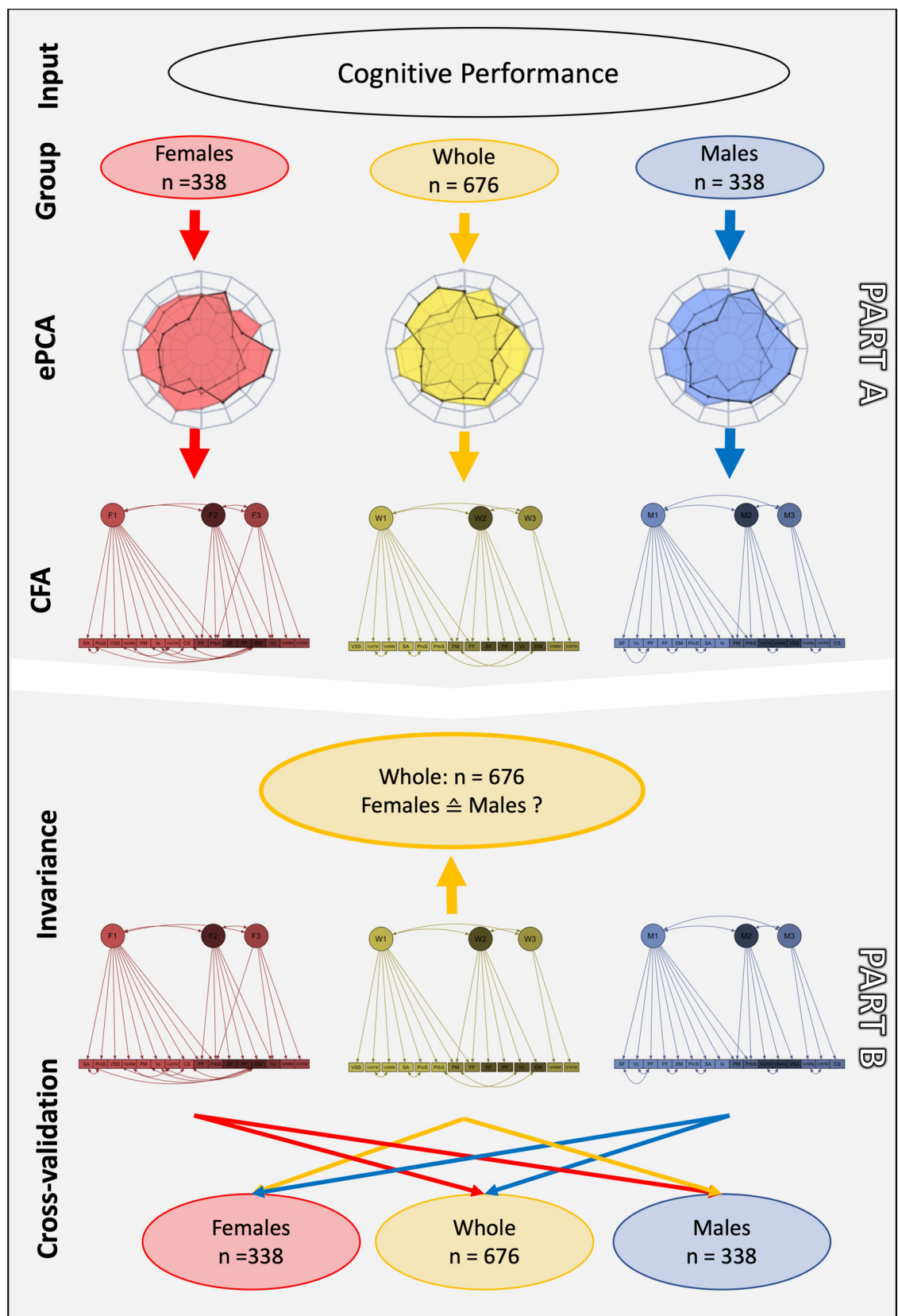


Figure 1. Flowchart presenting the study design.

To examine model fit of the respective ePCAs, we used comparative fit index (CFI), tucker lewis index (TLI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). Quality of model fit was assessed based on frequently reported fit indices indicating excellent model fit at $CFI > 0.95$, $TLI > 0.95$, $RMSEA \leq 0.06$, $SRMR < 0.09$ ^{54,55}. All initial models were subsequently refined to increase model fit: From the initial model, we first modelled residual covariances (included when residual covariances were > 0.1) between variables and components, and afterwards, removed non-significant variables from the model, if present.

After this measurement model configuration, we attempted to validate the established models across the two sexes (Fig. 1, part B). To do so, we first examined measurement invariance for all three models (whole CFA, male CFA, female CFA). Measurement invariance addresses the question whether a scale measures the same attribute in different groups of subjects. Hence, in the current study, measurement invariance would test whether the different cognitive component solutions, i.e. cognitive profiles would be the same across males and females. Measurement invariance was tested with the following aspects: (1) configural invariance: the measurement models derived from the CFA would fit equally well in males and females (same data structure across variables); (2) loading invariance: loadings of variables onto a cognitive component would be the same for males and females (groups have the same factor loadings); (3) intercept invariance: males and females would show the same intercept on the measured variables (groups have same intercepts of the observed variables); (4) mean invariance: males and females would show the same means on the measured variables (groups have the same means across the observed variables). In a second step, we applied a strict cross-validation by applying the sex-specific models to the other sex group only to test whether the male component solution would also obtain a good fit in females and vice versa. Model fit changes across the models were considered as significant with a change in $CFI > 0.01$ ⁵⁶ and a significant likelihood chi square difference test ($p < 0.05$).

Results

The current study assessed sex differences in cognitive profiles between older males and females based on a large battery of cognitive tests assessing attention, memory, executive and language functions. Differences in performance between males and females were already observed at single test level in several of the 16 neuropsychological tests used in the current study. For example, males performed significantly better in tasks requiring visual and visual-spatial abilities, e.g. visual-spatial memory, whereas females performed better in tasks requiring verbal abilities, such as episodic learning, phonematic and semantic fluency (see Table 1).

Investigating intercorrelations between cognitive performance scores of the different cognitive functions revealed a second interesting and important observation: While we overall found high intercorrelations between the assessed cognitive performances, these intercorrelations do not seem to be identical in males and females, already hinting at differences in cognitive profiles for the two sexes (for chord diagrams for the whole group, males and females separately as well differences between males and females, see Fig. 2, for Pearson correlation values between cognitive task, see Supplement, Tables S1–S3). Sex differences in cognitive performance correlations are shown in Fig. 2d. Noticeably, females show higher correlations between verbal and non-verbal test performance while males show higher correlations between verbal, non-verbal and executive functions (e.g. interference, concept shifting and problem solving).

Principal component analyses and confirmatory factor analyses for the whole group and males and females separately. Based on the correlations between cognitive performance tests, ePCA was applied to individual cognitive performance measurements of the whole group as well as males and females separately. Extraction of components was based on the eigenvalue criterion (eigenvalue > 1 , see Supplement, Table S4). Three components were extracted for the whole group (eigenvalues: 4.94, 1.48, 1.19) as well as when assessing males only (eigenvalues: 5.08, 1.58, 1.25). Regarding females only, the optimal component solution consisted of four cognitive components (eigenvalues: 4.97, 1.36, 1.09, 1.02). For all eigenvalues, see Supplement, Table S4.

For the whole group, the extracted components were dominated by the following functions: The first component covered a variety of non-verbal cognitive functions such as visual working memory, attention, executive functions and memory. The second component included fluency as well as memory. The third component was dominated by verbal functions, such as verbal working memory and vocabulary (Fig. 3a, for component loadings of all groups, see Supplement, Table S5). Afterwards, we extracted fit values for the PCA-derived three-component model using CFA. All variables were found to significantly contribute to the components (> 0.4). However, fit values of the initial model were not to be considered as of sufficient quality ($CFI = 0.894$; $TLI = 0.866$; $RMSEA = 0.07$, $SRMR = 0.053$). After model refinement via inclusion of residual covariances and exclusion of non-significant variables, the model improved significantly, but did not reach the threshold for being an excellent model in all fit indices ($CFI = 0.941$; $TLI = 0.921$; $RMSEA = 0.054$, $SRMR = 0.045$). The resulting model is shown in Fig. 3b (for results of the CFA for all groups, see Supplement, Table S6 and S7).

The male model (Fig. 3c,d), also a three-component model, consisted of a first component that included fluency, memory, attention and executive function, a second component that was dominated by visual working memory and executive functions and a third component including verbal working memory and executive functions. The initial male model revealed fit values not to be considered as of sufficient quality ($CFI = 0.883$, $TLI = 0.854$, $RMSEA = 0.071$, $SRMR = 0.06$). After additional refinement, the male model fitted on males revealed fit indices of: $CFI = 0.94$, $TLI = 0.923$, $RMSEA = 0.051$, $SRMR = 0.049$. This is a significant increase in model fit although it still does not reach the threshold for being an excellent model.

The female component solution (Fig. 3e,f) revealed one component dominated by visual memory and working memory, a second component including fluency and vocabulary and executive functions, a third component that consisted of executive functions and memory and a fourth component including verbal working memory

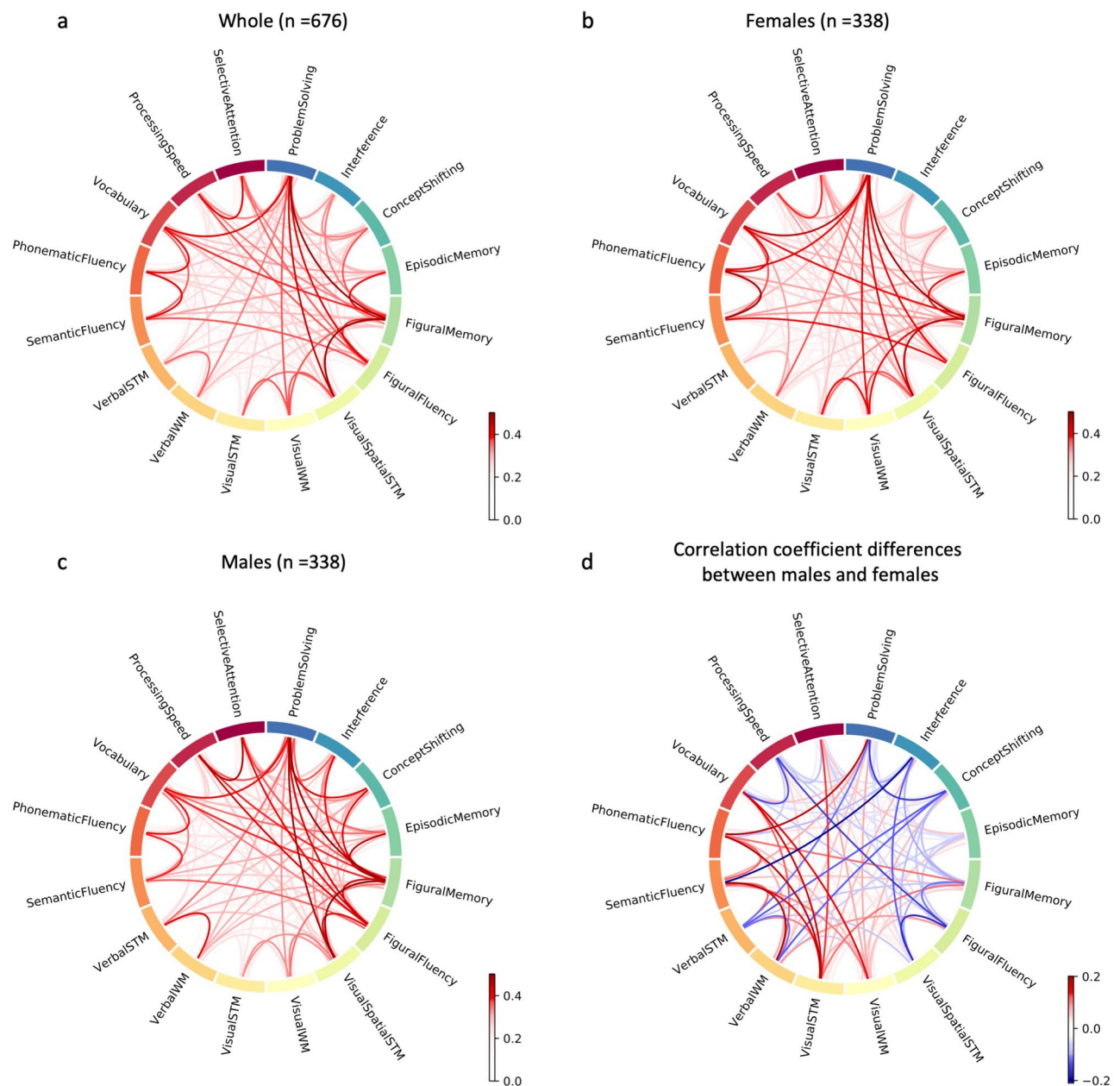


Figure 2. Chord diagrams of correlations between cognitive performance tests: (a) whole group, (b) females and (c) males, (d) differences in correlation coefficients between males and females: blue = males > females, red = females > males.

and vocabulary (for variable loading on the different components, see Supplement, Table S5). The initial female model fitted on females revealed fit indices of: CFI = 0.964, TLI = 0.953, RMSEA = 0.037, SRMR = 0.039. Although this model fulfilled the requirements for being an excellent model, we additionally refined the model by the same conditions we did before. This resulted in an additional significant increase in model fit (CFI = 0.984, TLI = 0.979, RMSEA = 0.025, SRMR = 0.034).

Taken together, the investigation of data-driven cognitive components in the three groups (whole group, males and females) hint at different compositions of cognitive components in older males and females (i.e. three versus four components, for an additional overview of three versus four component solutions for the whole group, males and females, see Supplement, Figure S1). Comparing these to the again slightly different component solution derived from the whole group (including both males and females) raises the question of whether these so far descriptively compared differences would be statistically meaningful, which was tested afterwards.

Measurement invariance and cross-validation. In the second part of the study (Fig. 1, part B), we addressed the distinctiveness of cognitive components between males and females by using measurement invariance and cross-validation. In detail, we started with the component solution that was derived from the whole group (including males and females) and tested whether this whole group cognitive component solution would

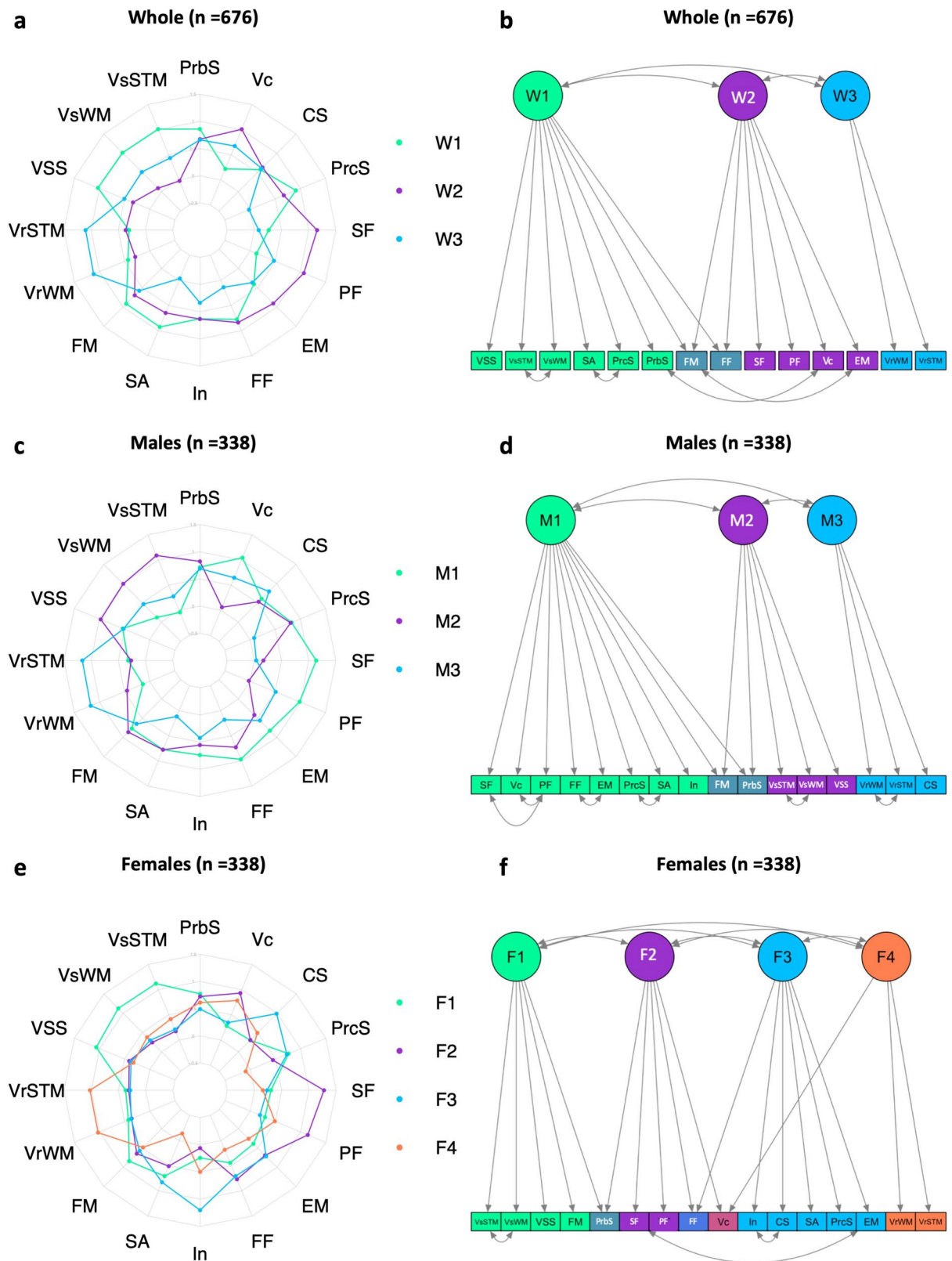


Figure 3. Exploratory Principal Component Analysis (ePCA) and Confirmatory Factor Analysis (CFA): (a,c,e): ePCA for the whole group, males and females. (b,d,f): CFA for the whole group, males and females. *PrbS* problem solving, *VsSTM* visual spatial short-term memory, *VsWM* visual spatial working memory, *VSS* visual working memory, *VrSTM* verbal short-term memory, *VrWM* verbal working memory, *FM* figural memory, *SA* selective attention, *In* interference, *FF* figural fluency, *EM* episodic memory, *PF* phonemic fluency, *ST* semantic fluency, *PrcS* processing speed, *CS* concept shifting, *Vc* vocabulary.

Model	Group	X2	CFI	TLI	RMSEA	SRMR
WHOLE	WHOLE	201.647	0.941	0.921	0.054	0.045
WHOLE	MALES	165.421	0.919	0.892	0.065	0.058
WHOLE	FEMALES	109.756	0.962	0.949	0.043	0.043
MALE	WHOLE	221.151	0.95	0.935	0.045	0.04
MALE	MALES	175.919	0.94	0.923	0.051	0.049
MALE	FEMALES	142.817	0.959	0.947	0.04	0.043
FEMALE	WHOLE	247.081	0.939	0.921	0.05	0.04
FEMALE	MALES	207.401	0.917	0.891	0.061	0.051
FEMALE	FEMALES	117.601	0.979	0.972	0.029	0.036

Table 2. Model fit indices for male and female refined models applied to the different groups. Values in bold reach the threshold for being an excellent model.

be statistically the same across males and females, i.e. invariant (for CFA model estimates, see Supplement, Table S8). Model fit indices did not reach the threshold for measurement invariance in terms of configural model (i.e. same data structure: CFI = 0.939, RMSEA = 0.055) and loading invariance (i.e. same factor loadings: CFI = 0.938, RMSEA = 0.053), it did even less so in the intercept (i.e. same intercept: CFI = 0.893, RMSEA = 0.067) and means invariance (i.e. same means: CFI = 0.871, RMSEA = 0.073). Thus, the cognitive component solution derived from the whole group, as it is often done in research investigating cognitive performance, does not seem to be completely generalizable over males and females. This, in turn, leads to the question which group (males or females) would fit better to the whole group component solution. While the model fit increased when the whole group model was applied to females only (whole group: CFI = 0.941; TLI = 0.921; RMSEA = 0.054, SRMR = 0.045; females: CFI = 0.962, TLI = 0.949, RMSEA = 0.043, SRMR = 0.043), it significantly decreased when investigating males only (CFI = 0.919, TLI = 0.892, RMSEA = 0.065, SRMR = 0.058). Thus, the current results indicate that the overall composition of cognitive components derived from the whole group is better suited for the female group as compared to the male group.

In a final cross validation, we applied the different cognitive component models obtained by either the whole group, males or females to the other groups, e.g. male component solution fitted onto the female group and vice versa (for fit indices, see Table 2). Applying the whole group model to males and females separately revealed an excellent model fit for the female group and a worse model fit for the male group. Applying the female cognitive component model to the male group reveals an overall insufficient model fit, which underpins the results obtained by the examination of measurement invariance. In turn, applying the male component model to the female group revealed a reasonable fit, with excellent fit indices. Thus, while males' cognitive performance does not seem to be sufficiently explained by the female model, female's cognitive performance can be sufficiently composed into both, male and female component solutions, with a slightly better fit of the female cognitive component model. Nevertheless, applying the male component solution to the female group revealed high covariances between the components (> 1), which indicates collinearity between the components. Thus, the validation of the component solutions indicate that separate cognitive component solutions might better describe a cognitive profile as compared to a common component solution.

Discussion

Using a data-driven approach, the current study examined sex-specific cognitive profiles based on a large variety of cognitive functions in older males and females. Our results show that a general model consisting of cognitive components that combine numerous cognitive tasks calculated based on the whole group (including both, males and females) fit unequally well on the two sexes. Males and females differ in terms of their composition of cognitive components, i.e. three components in males versus four components in females, with a generally better model fit in females. Thus, the current study found a rather decomposed (or local) cognitive profile in females while males seem to show a holistic (or global) cognitive profile, with more interrelations between different cognitive functions.

In a first step, we systematically examined sex differences in 16 different cognitive functions, namely selective attention, processing speed, problem solving, concept shifting, susceptibility to interference, figural fluency, phonematic and semantic verbal fluency, vocabulary, verbal episodic memory, figural memory, visual-, visual-spatial- and verbal short-term/working memory. We showed that older women perform better in verbal fluency, verbal episodic memory, processing speed and interference while older men significantly performed better on visual and visual-spatial working memory tasks. Importantly, these differences were rather small with only visual short-term memory and episodic memory showing medium effect sizes. Hence, the results are in line with a large amount of previous studies showing that males and females differ in some but not all cognitive functions and that these differences tend to be small^{5,9,10}. Thus, in normal older adults, we were able to show that those tasks requiring high verbal versus visuospatial processing show the largest sex differences.

Further, de Frias et al.² presented long-term sex differences in cognitive performance in a sample of adults with an age range from 35 to 80 years (at baseline). Over a period of ten years, women remained better in tasks assessing verbal episodic memory and verbal fluency, while men outperformed women in tasks assessing visuospatial functions. Additionally, and in line with Maitland et al.⁵⁷ and Pauls et al.⁵⁸ we showed that sex

differences, especially in the verbal versus spatial domains remain stable even in older ages. Thus, the current study adds to the notion that, even in later decades of life, sex differences in verbal versus visuospatial cognitive functions persist.

The observed sex differences in cognitive performance might be due to different processing styles to solve cognitive problems. Men usually inspect new scenes in a more 'global' way (e.g. having in mind the whole map in a spatial navigation task), while women usually prefer to inspect tasks more locally (i.e. remember more details of a given word list)^{18,19,59}. This might explain why men outperform women with respect to visual-spatial tasks and why women perform better in verbal episodic memory. Based on these task specific differences between the two sexes, the main goal was to investigate whether we could extend this global versus local phenomenon, to cognitive profiles in males and females, i.e. the relations between cognitive abilities. Using a data-driven ePCA we revealed a three-component solution for the whole including: (1) a non-verbal component composed of tasks including attention, executive functions and (working-) memory, (2) a mixed component including verbal and non-verbal fluency and memory functions and (3) a verbal short-term/working memory. This data-driven cognitive component solution shows the high complexity between cognitive functions, i.e. verbal fluency tasks require a large memory span and vice versa, an observation that has been found to be impaired in amnesic mild cognitive performance⁶⁰. It furthermore shows that cognitive components do not necessarily comply with the classical theory-driven cognitive domains of attention, executive functions, working and episodic memory and language functions, an observation that has already been described by Harvey²¹.

Noticeably, CFA was used to examine the model fit indices of this component analysis and whether this component solution fit equally well to males and females. The overall model shows an acceptable, although not excellent fit ($CFI > 0.95$)^{54,55} for the whole group (even after refinement of the model by including residual covariances between cognitive variables and exclusion of non-significant variables). When examining measurement invariance between the two sexes, thus whether a cognitive profile would fit equally well to males and females, we again found an acceptable but not excellent model fit already in the configural model (composition of the components), with further significant decreases when it comes to mean and intercept invariances. While some fit values do not differ from previous results obtained by Siedlecki et al.²², who interpreted a CFI value of 0.941 as being acceptable, they are low as compared to other studies investigating measurement invariance in cognitive or psychological profiles between, e.g. healthy adults and Alzheimer patients or using longitudinal models of sex differences over the whole adulthood^{61,62}. These differences in model fit to the aforementioned studies might be due to differences in neuropsychological tests used or differences in group characteristics. In the current sample, normal older adults were examined that were matched for age and sex, since both factors are well known to correlate with cognitive performance^{24,63}. Thus, the sex-specific effects found in the current study regarding cognitive profiles line up with previous studies showing that sex differences exist, and might be of special importance for our society, but are of rather small effect size¹⁰.

After stratifying the current sample for sex, we again performed an ePCA and obtained different component solutions for each group. While in the male group, three components were preferred (according to the Eigenvalue criterion), females' cognitive performance was best described by a four-component solution. More importantly, the extracted components differed in their composition, i.e. cognitive tasks involved in the different components. While for the whole group, the first component was composed of heterogeneous but consistently non-verbal functions, verbal as well as non-verbal functions belong to the same component in males, additionally including fluency, memory, attention and executive functions. The second male component contained visual working memory and executive functions and the third component consisted of verbal working memory and executive functions. Relating these results to the observations made regarding task specific differences in processing styles, i.e. global-local hypothesis of sex-differences¹¹, one could argue that males' holistic/global processing style to solve cognitive task, is in line with the current cognitive profile. Males show a quite holistic first component, including attention, executive functions, episodic memory and fluency tasks, hinting at higher interrelations between different cognitive abilities. Furthermore, since executive functions and/or attention depict essential parts in all three components it could be assumed that these functions serve as a higher order executive-attention system that monitors cognitive performance⁶³⁻⁶⁶. Thus, this would mean that males rely strongly on their attentional and executive functions, e.g. goal-directed planning, monitoring, mental flexibility, to process cognitive tasks belonging to different cognitive domains. In terms of a global way of cognitive processing, males potentially manage cognitive processing using one superordinate system that links different cognitive abilities. Likewise, if these functions decline, a decline of all other cognitive domains follow, as has been stated by theories, such as the frontal executive theory of aging⁶⁷. Investigating females only revealed a different picture compared to both the whole group or males only. Females' cognitive performance within the functions examined is best decomposed into four cognitive components. In contrast to the males' first component which was quite heterogeneous including fluency, executive functions and attention, in the females' cognitive profile visual-verbal fluency and executive functions—attention—built separate components; together with a component composed of visual (working) memory and executive functions and another component dominated by verbal functions including working memory and vocabulary. Thus, females' cognitive profiles consist of more subsystems as compared to males, with systems including different cognitive functions (i.e. [1] visual (working) memory/[2] fluency/[3] executive functions/[4] verbal (working) memory). Although these functions share covariances, they themselves represent distinct cognitive systems or modules. On the other hand, males might have a superordinate system, i.e. the attentional-executive-fluency-memory component, which includes several cognitive domains, thereby representing a stronger interplay of cognitive functions with a probably superordinate system (i.e. executive functions). Hence, this could be potentially related to a more global processing strategy during cognitive performance, meaning that irrespective of the task (e.g. memory or fluency), males might activate similar cognitive processing strategies. In contrast to that, females would rather choose different processing strategies, depending on the cognitive task, e.g. visual versus verbal working memory. Together, similarly to the global versus local

processing at single task level^{11,12,15,18,68}, cognitive profiles derived from either males or females seem to be differentially composed along the global vs. local processing dichotomy in the current study.

Furthermore, focussing on the cross-validation model fit values, an additional support for the existence of sex-related cognitive profiles in line with these processing strategies became evident. While applying the female component solution to the female group reveals excellent fit values, the male component solution only reveals acceptable fit values when applied to the male group. These lower fit values might arise from the stronger interconnectedness of different cognitive functions in the male group, which has been shown when comparing correlation strength between males and females (cf Fig. 2). For example, interference is correlated to both, verbal fluency as well as visual spatial short-term memory, which in turn is correlated with figural fluency. As a consequence, a clear division of cognitive functions into different (independent) cognitive domains, might not be possible in the male group. Thus, males' cognitive abilities seem to be not fully suitable for a modular cognitive structure as compared to females. This again, would be in accordance with global versus local processing styles.

Importantly, the current study investigated an older adult population to examine sex-specific cognitive profiles. This population is of special interest when examining sex differences in cognitive performance and cognitive profiles since previous research has shown that first, sex-differences in cognitive functions remain stable until older ages, and second, pathological conditions with cognitive impairments differ in prevalence between males and females^{28,57}. However, research so far, most often includes sex as a covariate of non-interest when assessing cognitive impairment.

Previous studies often showed steeper decline in general cognitive functions in males^{1,69}. Similarly, in pathological conditions, such as Parkinson's disease, males were reported to show a faster decline in cognitive functions²⁸. However, when it comes to Alzheimer's disease, females show a faster decline in memory scores as compared to males²⁹. This observation might be related to distinct cognitive profiles in older males and females. If, within the 'global' cognitive profile of males, the executive-attentional monitoring system breaks down this would lead to a global decline in cognitive functions. Especially for the aging process, theories such as the prefrontal-executive theory⁶⁷ as well as the processing speed theory⁷⁰ of aging, stating that decreasing executive functions and attention, respectively, predict cognitive decline in a diversity of cognitive functions belonging to different domains⁶⁴. Thus, in males these two theories that try to explain cognitive performance decline during the aging process, would be in line with the current results. On the other hand, if females' cognitive profiles are rather composed of different cognitive subsystems or modules (thus 'local' cognitive profile), impairments within the executive-attentional component would not necessarily lead to an impairment in other cognitive components. Hence, this would rather result in function-specific cognitive decline, e.g. executive impairment. These differences in cognitive profiles might thus serve as a possible explanation for why males show generally steeper decreases in overall cognitive abilities during aging⁶⁹.

Methodological considerations. The current study has several advantages and disadvantages. While we were able to show that cognitive profiles differ, when investigating males and females independent of each other, it is important to mention that the effects of sex differences are rather of smaller sizes, which becomes obvious when focussing on the differences in terms of intercorrelations between different cognitive tasks. Nevertheless, as stated by Zell et al.¹⁰, although effect sizes might be small, when investigating sex differences in cognitive performance, these differences might be important to understand cognitive performance differences.

In addition, it has to be mentioned that the current study investigated these cognitive profiles in a sample ranging from 55 to 85 years of age. It might be the case that with increasing age, cognitive profiles change, especially when cognitive impairments arise, e.g. due to pathological conditions. Future studies should investigate this topic, especially using longitudinal data, to show whether cognitive profiles change in the course of the aging process, potentially also with respect to pathological conditions.

Further, it has to be mentioned that the set-up of cognitive profiles is not straightforward. We used a Principal Component Analysis with Varimax rotation method for extracting cognitive components in the two groups and extracted four factors for females and three factors for males, based on the Eigenvalue criterion (cut-off for selection of components being an Eigenvalue > 1). Nevertheless, the fourth Eigenvalue is only slightly above one for females (1.02) and the fourth Eigenvalue is only slightly below 1 (0.97) for males, which both are very close to the cut-off value. Further, the model refinement highly depends on the input data (in this case the cognitive tasks used). Until now, there is no gold standard in this respect. More research is needed to address this important topic.

Finally, the question that arises when observing these differences is which factors might be responsible for the development of sex differences. From previous studies it is known that males and females differ in terms of brain structure and function, which might relate to differences in cognitive processing strategies^{71,72}. Furthermore, it has been shown that hormonal differences, but also genetic variations might be related to differences in cognitive and social behavior between the two sexes⁷³. Social factors, such as gender role models, significantly influence differences in cognitive performance, which is less pronounced in countries that promote gender equality⁷⁴. Further studies are warranted to examine this question.

Conclusion

Conclusively, males and females show not only differences in specific cognitive tasks but generally in cognitive profiles across cognitive domains. Males are likely to use a more holistic way of processing, by integrating different cognitive functions to solve specific tasks. This could be, for example, a higher executive control and memory function in a verbal fluency task, which in turn, would result in larger clusters of the same category. Females, on the other hand are likely to process cognitive tasks in smaller, rather domain-specific subsystems. The results showed that older males and females exhibit different cognitive profiles, that are likely to be related to

differences in cognitive decline across the aging process. Therefore, the current research stresses the importance to use sex-stratified analyses when assessing cognitive performance. Future research is warranted to extend the current results to pathological conditions, such as Alzheimer's disease. Furthermore, differences in cognitive profiles might not only be important in basic research but, might also impact clinical prevention programs, i.e. cognitive training.

Received: 14 September 2020; Accepted: 4 February 2021

Published online: 22 March 2021

References

- McCarrey, A. C., An, Y., Kitner-Triolo, M. H., Ferrucci, L. & Resnick, S. M. Sex differences in cognitive trajectories in clinically normal older adults. *Psychol. Aging* **31**, 166–175. <https://doi.org/10.1037/pag0000070> (2016).
- de Frias, C. M., Nilsson, L. G. & Herlitz, A. Sex differences in cognition are stable over a 10-year period in adulthood and old age. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* **13**, 574–587. <https://doi.org/10.1080/13825580600678418> (2006).
- Maitland, S. B., Herlitz, A., Nyberg, L., Bäckman, L. & Nilsson, L. G. Selective sex differences in declarative memory. *Mem. Cogn.* **32**, 1160–1169 (2004).
- Mansouri, F. A., Fehring, D. J., Gaillard, A., Jaberzadeh, S. & Parkinson, H. Sex dependency of inhibitory control functions. *Biol. Sex Differ.* **7**, 11. <https://doi.org/10.1186/s13293-016-0065-y> (2016).
- Halpern, D. F. *et al.* The science of sex differences in science and mathematics. *Psychol. Sci. Public Interest* **8**, 1–51. <https://doi.org/10.1111/j.1529-1006.2007.00032.x> (2007).
- Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W. & Delazer, M. Sex differences in cognitive functions. *Person. Individ. Differ.* **35**, 863–875. [https://doi.org/10.1016/s0264-3707\(03\)00061-9](https://doi.org/10.1016/s0264-3707(03)00061-9) (2003).
- Hirnst, M., Coloma Andrews, L. & Hausmann, M. Gender-stereotyping and cognitive sex differences in mixed- and same-sex groups. *Arch. Sex Behav.* **43**, 1663–1673. <https://doi.org/10.1007/s10508-014-0311-5> (2014).
- Hyde, J. S. The gender similarities hypothesis. *Am. Psychol.* **60**, 581–592. <https://doi.org/10.1037/0003-066X.60.6.581> (2005).
- Hyde, J. S. Sex and cognition: gender and cognitive functions. *Curr. Opin. Neurobiol.* **38**, 53–56. <https://doi.org/10.1016/j.conb.2016.02.007> (2016).
- Zell, E., Krizan, Z. & Teeter, S. R. Evaluating gender similarities and differences using metasynthesis. *Am. Psychol.* **70**, 10–20. <https://doi.org/10.1037/a0038208> (2015).
- Pletzer, B., Scheuringer, A. & Scherndl, T. Global-local processing relates to spatial and verbal processing: implications for sex differences in cognition. *Sci. Rep.* **7**, 10575. <https://doi.org/10.1038/s41598-017-11013-6> (2017).
- Scheuringer, A., Wittig, R. & Pletzer, B. Sex differences in verbal fluency: The role of strategies and instructions. *Cogn Process* **18**, 407–417. <https://doi.org/10.1007/s10339-017-0801-1> (2017).
- Saucier, D. M. *et al.* Are sex differences in navigation caused by sexually dimorphic strategies or by differences in the ability to use the strategies? *Behav. Neurosci.* **116**, 403–410. <https://doi.org/10.1037/0735-7044.116.3.403> (2002).
- Andersen, N. E., Dahmani, L., Konishi, K. & Bohbot, V. D. Eye tracking, strategies, and sex differences in virtual navigation. *Neurobiol. Learn. Mem.* **97**, 81–89. <https://doi.org/10.1016/j.nlm.2011.09.007> (2012).
- Weiss, E. M. *et al.* Sex differences in clustering and switching in verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **12**, 502–509. <https://doi.org/10.1017/s1355617706060656> (2006).
- Lanting, S., Haugrud, N. & Crossley, M. The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **15**, 196–204. <https://doi.org/10.1017/S1355617709090237> (2009).
- Peña, D., Contreras, M. J., Shih, P. C. & Santacreu, J. Solution strategies as possible explanations of individual and sex differences in a dynamic spatial task. *Acta Physiol. (Oxf.)* **128**, 1–14. <https://doi.org/10.1016/j.actpsy.2007.09.005> (2008).
- Pletzer, B. Sex-specific strategy use and global-local processing: A perspective toward integrating sex differences in cognition. *Front. Neurosci.* **8**, 425. <https://doi.org/10.3389/fnins.2014.00425> (2014).
- Roalf, D., Lowery, N. & Turetsky, B. I. Behavioral and physiological findings of gender differences in global-local visual processing. *Brain Cogn.* **60**, 32–42. <https://doi.org/10.1016/j.bandc.2005.09.008> (2006).
- Hirnst, M., Laroi, F. & Laloyaux, J. No sex difference in an everyday multitasking paradigm. *Psychol. Res.* **83**, 286–296. <https://doi.org/10.1007/s00426-018-1045-0> (2019).
- Harvey, P. D. Domains of cognition and their assessment. *Dialogues Clin. Neurosci.* **21**, 227–237. <https://doi.org/10.31887/DCNS.2019.21.3/pharvey> (2019).
- Siedlecki, K. L., Falzarano, F. & Salthouse, T. A. Examining gender differences in neurocognitive functioning across adulthood. *J. Int. Neuropsychol. Soc.* **25**, 1051–1060. <https://doi.org/10.1017/S1355617719000821> (2019).
- Munro, C. A. *et al.* Sex differences in cognition in healthy elderly individuals. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* **19**, 759–768. <https://doi.org/10.1080/13825585.2012.690366> (2012).
- Hedden, T. & Gabrieli, J. D. Insights into the ageing mind: A view from cognitive neuroscience. *Nat. Rev. Neurosci.* **5**, 87–96. <https://doi.org/10.1038/nrn1323> (2004).
- Schaie, K. W. When does age-related cognitive decline begin? Salthouse again reifies the “cross-sectional fallacy”. *Neurobiol. Aging* **30**, 528–529. <https://doi.org/10.1016/j.neurobiolaging.2008.12.012> (2009) (**discussion 530–533**).
- Schaie, K. W. & Willis, S. L. The Seattle longitudinal study of adult cognitive development. *ISSBD Bull.* **57**, 24–29 (2010).
- Habib, R., Nyberg, L. & Nilsson, L. G. Cognitive and non-cognitive factors contributing to the longitudinal identification of successful older adults in the betula study. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* **14**, 257–273. <https://doi.org/10.1080/13825580600582412> (2007).
- Cholerton, B. *et al.* Sex differences in progression to mild cognitive impairment and dementia in Parkinson's disease. *Parkinsonism Relat. Disord.* **50**, 29–36. <https://doi.org/10.1016/j.parkreldis.2018.02.007> (2018).
- Sohn, D. *et al.* Sex differences in cognitive decline in subjects with high likelihood of mild cognitive impairment due to Alzheimer's disease. *Sci. Rep.* **8**, 7490. <https://doi.org/10.1038/s41598-018-25377-w> (2018).
- Organisation for Economic Co-operation and Development. *Classifying educational programmes: manual for ISCED-97 implementation in OECD countries*. 1999 edn, (Organisation for Economic Co-operation and Development, 1999).
- Caspers, S. *et al.* Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Front. Aging Neurosci.* **6**, 149. <https://doi.org/10.3389/fnagi.2014.00149> (2014).
- Schmermund, A., Mohlenkamp, S., Stang, A., Gronemeyer, D. & Seibel, R., *et al.* Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL Study. *Am Heart J* **144**, 212–18 (2002).
- Kalbe, E. *et al.* DemTect: A new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *Int. J. Geriatr. Psychiatry* **19**, 136–143. <https://doi.org/10.1002/gps.1042> (2004).

34. Sturm, W., Willmes, K. & Horn, W. *Leistungsprüfungssystem für 50–90jährige (LPS 50+): Handanweisung* (Hogrefe, Verlag für Psychologie, 1993).
35. Schellig, D. *Block-tapping-test* (Swets Test Services Frankfurt, Frankfurt, 1997).
36. Della Sala, S., Gray, C., Baddeley, A. & Wilson, L. Visual patterns test: A test of short-term visual recall. *Thames Valley Test Company* **40** (1997).
37. Oswald, W. & Fleischmann, U. *The Nürnberger-Alters Inventory* (Hogrefe, Göttingen, 1997).
38. Benton, A. L., Sivan, A., Spreen, O. & Der Steck, P. *Benton-Test Huber* (Hogrefe, Göttingen, 2009).
39. Der Gatterer, G. *Alters-Konzentrations-Test* 2nd edn. (Hogrefe, Göttingen, 2008).
40. Bäuml, G. & Stroop, J. *Farbe-Wort-Interferenztest Nach JR Stroop (FWIT)* (Hogrefe, Verlag für Psychologie, 1985).
41. Stroop, J. R. Studies of interference in serial verbal reactions. *J. Exp. Psychol.* **18**, 643 (1935).
42. Regard, M., Strauss, E. & Knapp, P. Children's production on verbal and non-verbal fluency tasks. *Percept. Mot. Skills* **55**, 839–844 (1982).
43. Lux, S., Hartje, W., Reich, C. & Nagel, C. *VGT: Verbaler Gedächtnistest: Bielefelder Kategoriale Wortlisten* (Verlag Hans Huber, Göttingen, 2012).
44. Aschenbrenner, S., Tucha, O. & Lange, K. *Regensburger Wortflüssigkeits-Test (RWT)* (Hogrefe, Göttingen, 2000).
45. Morris, J. *et al.* The consortium to establish a registry for Alzheimer's disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **39**, 1159–1159 (1989).
46. Schmidt, K. & Metzler, P. *WST-Wortschatztest. Gött Beltz Test* (1992).
47. Finkel, D., Andel, R., Gatz, M. & Pedersen, N. L. The role of occupational complexity in trajectories of cognitive aging before and after retirement. *Psychol. Aging* **24**, 563–573. <https://doi.org/10.1037/a0015511> (2009).
48. Jockwitz, C. *et al.* Influence of age and cognitive performance on resting-state brain networks of older adults in a population-based cohort. *Cortex* **89**, 28–44 (2017).
49. Stumme, J., Jockwitz, C., Hoffstaedter, F., Amunts, K. & Caspers, S. Functional network reorganization in older adults: Graph-theoretical analyses of age, cognition and sex. *NeuroImage* **214**, 116756 (2020).
50. Heaton, R. K. *et al.* Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *J. Int. Neuropsychol. Soc.* **20**, 588–598. <https://doi.org/10.1017/S1355617714000241> (2014).
51. Gross, A. L. *et al.* Effects of education and race on cognitive decline: An integrative study of generalizability versus study-specific results. *Psychol. Aging* **30**, 863–880. <https://doi.org/10.1037/pag0000032> (2015).
52. Wilhalm, H. *et al.* A comparison of theoretical and statistically derived indices for predicting cognitive decline. *Alzheimers Dement (Amst)* **6**, 171–181. <https://doi.org/10.1016/j.dadm.2016.10.002> (2017).
53. Stevens, J. P. *Applied Multivariate Statistics for the Social Sciences* (Routledge, Abingdon, 2012).
54. Fournet, N. *et al.* Multigroup confirmatory factor analysis and structural invariance with age of the behavior rating inventory of executive function (BRIEF)–French version. *Child Neuropsychol.* **21**, 379–398. <https://doi.org/10.1080/09297049.2014.906569> (2015).
55. Hu, L. T. & Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscipl. J.* **6**, 1–55 (1999).
56. Chen, F. F. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model. Multidiscipl. J.* **14**, 464–504. <https://doi.org/10.1080/10705510701301834> (2007).
57. Maitland, S. B., Intrieri, R. C., Schaie, W. K. & Willis, S. L. Gender differences and changes in cognitive abilities across the adult life span. *Aging Neuropsychol. Cognit.* **7**, 32–53. <https://doi.org/10.1076/anec.7.1.32.807> (2010).
58. Pauls, F., Petermann, F. & Lepach, A. C. Gender differences in episodic memory and visual working memory including the effects of age. *Memory* **21**, 857–874. <https://doi.org/10.1080/09658211.2013.765892> (2013).
59. Kimchi, R., Amishav, R. & Sulitzeanu-Kenan, A. Gender differences in global-local perception? Evidence from orientation and shape judgments. *Acta Psychol (Amst)* **130**, 64–71. <https://doi.org/10.1016/j.actpsy.2008.10.002> (2009).
60. Mueller, K. D. *et al.* Verbal fluency and early memory decline: results from the wisconsin registry for Alzheimer's prevention. *Arch. Clin. Neuropsychol.* **30**, 448–457. <https://doi.org/10.1093/arclin/acr030> (2015).
61. Maitland, S. B., Herlitz, A., Nyberg, L., Backman, L. & Nilsson, L. G. Selective sex differences in declarative memory. *Mem. Cognit.* **32**, 1160–1169. <https://doi.org/10.3758/bf03196889> (2004).
62. Johnson, D. K., Storandt, M., Morris, J. C., Langford, Z. D. & Galvin, J. E. Cognitive profiles in dementia: Alzheimer disease vs healthy brain aging. *Neurology* **71**, 1783–1789. <https://doi.org/10.1212/01.wnl.0000335972.35970.70> (2008).
63. Scarmeas, N., Albert, S. M., Manly, J. J. & Stern, Y. Education and rates of cognitive decline in incident Alzheimer's disease. *J. Neurol. Neurosurg. Psychiatry* **77**, 308–316. <https://doi.org/10.1136/jnnp.2005.072306> (2006).
64. Albinet, C. T., Boucard, G., Bouquet, C. A. & Audiffren, M. Processing speed and executive functions in cognitive aging: How to disentangle their mutual relationship?. *Brain Cogn.* **79**, 1–11. <https://doi.org/10.1016/j.bandc.2012.02.001> (2012).
65. Cahn-Weiner, D. A., Malloy, P. F., Boyle, P. A., Marran, M. & Salloway, S. Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. *Clin. Neuropsychol.* **14**, 187–195. [https://doi.org/10.1076/1385-4046\(200005\)14;2:1-Z;FT187](https://doi.org/10.1076/1385-4046(200005)14;2:1-Z;FT187) (2000).
66. Mitchell, M. & Miller, L. S. Prediction of functional status in older adults: the ecological validity of four Delis–Kaplan executive function system tests. *J. Clin. Exp. Neuropsychol.* **30**, 683–690. <https://doi.org/10.1080/13803390701679893> (2008).
67. West, R. L. An application of prefrontal cortex function theory to cognitive aging. *Psychol. Bull.* **120**, 272–292. <https://doi.org/10.1037/0033-2909.120.2.272> (1996).
68. Scheuringer, A. & Pletzer, B. Sex differences in the Kimchi–Palmer task revisited: Global reaction times, but not number of global choices differ between adult men and women. *Physiol. Behav.* **165**, 159–165. <https://doi.org/10.1016/j.physbeh.2016.07.012> (2016).
69. Laws, K. R., Irvine, K. & Gale, T. M. Sex differences in cognitive impairment in Alzheimer's disease. *World J. Psychiatry* **6**, 54–65. <https://doi.org/10.5498/wjp.v6.i1.54> (2016).
70. Salthouse, T. A. The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* **103**, 403–428. <https://doi.org/10.1037/0033-295x.103.3.403> (1996).
71. Jancke, L., Sele, S., Liem, F., Oschwald, J. & Merillat, S. Brain aging and psychometric intelligence: A longitudinal study. *Brain Struct. Funct.* **225**, 519–536. <https://doi.org/10.1007/s00429-019-02005-5> (2020).
72. Young, K. D., Bellgowan, P. S. F., Bodurka, J. & Drevets, W. C. Functional neuroimaging of sex differences in autobiographical memory recall. *Hum. Brain Mapp.* **34**, 3320–3332. <https://doi.org/10.1002/hbm.22144> (2013).
73. Ristori, J. *et al.* Brain sex differences related to gender identity development: Genes or hormones?. *Int. J. Mol. Sci.* <https://doi.org/10.3390/ijms21062123> (2020).
74. Bonsang, E., Skirbekk, V. & Staudinger, U. M. As you sow, so shall you reap: Gender-role attitudes and late-life cognition. *Psychol. Sci.* **28**, 1201–1213. <https://doi.org/10.1177/0956797617708634> (2017).

Acknowledgements

This project was partially funded by the German National Cohort and the 1000BRAINS-Study of the Institute of Neuroscience and Medicine, Research Centre Jülich, Germany. We thank the Heinz Nixdorf Foundation (Germany) for the generous support of the Heinz Nixdorf Study. We thank the investigative group and the study

staff of the Heinz Nixdorf Recall Study and 1000BRAINS. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3; SC) as well as from the Initiative and Networking Fund of the Helmholtz Association (SC).

Author contributions

C.J. Investigation, Methodology, Formal analysis, Visualization, Writing—original draft, Writing—review and editing. L.W. Methodology, Formal analysis, Writing—review and editing. J.S. Data curation, Formal analysis; Writing—review and editing. S.C. Conceptualization, Supervision, Resources, Funding acquisition, Writing—review and editing.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-84134-8>.

Correspondence and requests for materials should be addressed to C.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

4 Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K. R., ... & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, 45(6), e26683.

RESEARCH ARTICLE

Sex classification from functional brain connectivity: Generalization to multiple datasets

Lisa Wiersch^{1,2}  | Patrick Friedrich^{1,2} | Sami Hamdan^{1,2} | Vera Komeyer^{1,2,3} | Felix Hoffstaedter^{1,2} | Kaustubh R. Patil^{1,2} | Simon B. Eickhoff^{1,2} | Susanne Weis^{1,2}

¹Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

²Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany

³Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Correspondence

Susanne Weis, Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
Email: s.weis@fz-juelich.de

Funding information

Deutsche Forschungsgemeinschaft (DFG), Grant/Award Numbers: 491111487, 431549029; National Institute of Mental Health, Grant/Award Number: R01-MH074457; the Helmholtz Portfolio Theme “Supercomputing and Modeling for the Human Brain”; European Union's Horizon 2020 Research and Innovation Programme, Grant/Award Number: 945539

Abstract

Machine learning (ML) approaches are increasingly being applied to neuroimaging data. Studies in neuroscience typically have to rely on a limited set of training data which may impair the generalizability of ML models. However, it is still unclear which kind of training sample is best suited to optimize generalization performance. In the present study, we systematically investigated the generalization performance of sex classification models trained on the parcelwise connectivity profile of either single samples or compound samples of two different sizes. Generalization performance was quantified in terms of mean across-sample classification accuracy and spatial consistency of accurately classifying parcels. Our results indicate that the generalization performance of parcelwise classifiers (pwCs) trained on single dataset samples is dependent on the specific test samples. Certain datasets seem to “match” in the sense that classifiers trained on a sample from one dataset achieved a high accuracy when tested on the respected other one and vice versa. The pwCs trained on the compound samples demonstrated overall highest generalization performance for all test samples, including one derived from a dataset not included in building the training samples. Thus, our results indicate that both a large sample size and a heterogeneous data composition of a training sample have a central role in achieving generalizable results.

KEYWORDS

big data, generalizability, machine learning, neuroimaging, resting-state functional connectivity, sex classification

1 | INTRODUCTION

Machine learning (ML) is a powerful tool to relate neuroimaging data to behavior and phenotypes (Genon et al., 2022; Varoquaux & Thirion, 2014) and is therefore increasingly being employed in neuroscience applications (Buch et al., 2018; Jollans et al., 2019; Kohoutova et al., 2020; Varoquaux, 2018). Successful applications of ML approaches include the decoding of mental states (Haynes &

Rees, 2006), classification of mental disorders (Chen et al., 2020; Zhang et al., 2021), as well as the prediction of demographic and behavioral phenotypes (More et al., 2023; Nostro et al., 2018; Pläschke et al., 2020; Smith et al., 2015; Varikuti et al., 2018).

ML models learn the feature-target relationship given a training sample. Subsequently, the model is applied to make predictions on previously unseen data (Dhamala et al., 2023) and successful generalization to independent data samples is the central goal in ML

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

(Domingos, 2012; Varoquaux, 2018; Chung, 2018). For example, a recent study (Weis et al., 2020) demonstrated successful generalization of sex prediction models based on regionally specific functional brain connectivity patterns, which were trained on the data of the Human Connectome Project (HCP, Van Essen et al., 2012; Van Essen et al., 2013). For this spatially specific approach, independent classifiers were trained on the functional brain connectivity patterns of parcels covering the whole brain. In this case, assessing generalization performance should not only consider the averaged across-sample accuracy. Rather, if the classifiers generalize well, the same parcels should achieve high classification accuracies during cross-validation (CV) and across-sample testing.

Further sex classification studies (Menon & Krishnamurthy, 2019; Smith, Vidaurre, et al., 2013; Zhang et al., 2018), as well as other applications of ML models employed the HCP dataset to predict phenotypes such as task activation (Cohen et al., 2020), and individual behavioral and demographic scores (Cui & Gong, 2018; Smith et al., 2015) like age (Sanford et al., 2022). The HCP dataset is characterized by high-quality multi-modal imaging data acquired from a large group of healthy young adults. However, both the high quality of the brain imaging data as well as the narrow age range is not typical of other datasets, especially when dealing with clinical data (Arslan, 2018; Jansma et al., 2020; Rutten & Ramsey, 2010). This raises the question whether results based on the HCP data can be generalized to other datasets with different characteristics. Weis et al. (2020) demonstrated that sex classifiers trained on the HCP data generalized well to an independent subset of the HCP dataset as well as to the 1000Brains dataset (Caspers et al., 2014). Additional evidence from the application of such classifiers to data from datasets with diverse characteristics would provide even stronger evidence of model generalization.

Especially in neuroimaging, differences between datasets may result from several different sources. On the one hand, participants may differ with respect to demographic characteristics, such as age, education, or economic status. On the other hand, data samples likely differ with regard to the MRI acquisition parameters and data processing. Considering these differences, it is so far unresolved what kind of training sample leads to good generalization performance across multiple test samples.

Various characteristics of the training data can influence the generalization performance of ML models (Dhamala et al., 2023). For instance, larger sample size is beneficial for generalization performance (Cui & Gong, 2018; Domingos, 2012). Ensuring that the training data is representative of the target sample is another crucial factor for achieving good generalization performance (Ishida, 2019; Yang et al., 2020). Furthermore, data from different acquisition sites are likely heterogeneous with respect to demographic characteristics, data acquisition, and processing parameters. Due to the variability across different datasets and sites, a ML model trained on a compound of such data is more likely to capture the shared biological variability in all datasets while disregarding the variability resulting from differences between the datasets. This distinction supports models

focusing solely on the biological variability independent of specific dataset characteristics. Hence, such models are less likely to overfit and more likely to generalize to new data. Thus, aggregating data from multiple sites should be beneficial for improving generalization performance. Indeed, this has been partially shown by studies concerning clinical applications of ML approaches (Chang et al., 2018; Nielsen et al., 2020; Willemink et al., 2020). These results suggest that training ML models on diverse datasets covering a wide range of characteristics may improve the overall generalization performance.

In the present study, we aimed to evaluate the generalization performance of multiple sets of sex classification models derived from different training samples. The different training samples were created from four different datasets with varying demographic characteristics. In addition, sex classifiers were trained on compound samples combining data from all datasets to obtain training samples with heterogeneous sample characteristics. Both compound samples comprise the same ratios of datasets, sex and age distributions, but differ in sample size to additionally assess the effect of training sample size. Following the parcelwise approach by Weis et al. (2020), we trained independent sex classifiers on the resting state (RS) connectivity patterns of 436 parcels covering the whole brain. For each parcel, a sex classification model was built based on the individual connectivity profile, resulting in one classification accuracy value per parcel. This was done for each of the six training samples, resulting in six sets of parcelwise classifiers (pwCs). These pwCs were applied to test samples from the four original datasets and one dataset which was not part of the training samples. Then, accuracy maps, representing the spatial distribution of classification accuracies for each parcel were generated for CV (within-sample accuracy) and for application of the pwCs to the different test samples (across-sample accuracy). The comparison of these accuracy maps enabled us to evaluate generalization performance of classifiers by (i) examining the mean accuracy of all pwCs across the 10% best classifying parcels and (ii) comparing the spatial location of highly classifying parcels between CV and across-sample test. Good generalization performance with regard to spatial consistency is characterized by identical parcels performing well in CV and across-sample testing. We hypothesized that the pwC trained on the compound sample with a smaller sample size should outperform pwCs trained on single samples due to the heterogeneous data composition, while the compound sample with a higher sample size should achieve the overall best generalization performance (Chang et al., 2018; Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Nielsen et al., 2020; Willemink et al., 2020).

2 | MATERIALS AND METHODS

2.1 | Data

We employed RS functional magnetic resonance imaging (fMRI) data of subsets of four large datasets to train and test sex classification models. For all datasets, we only included healthy subjects aged



20 years or older. Within each training sample, we matched females and males for age and included a similar number of women and men. The first sample, taken from the HCP dataset (900 subjects data release; Van Essen et al., 2012; Van Essen, 2013), comprised 878 subjects with a mean age of 28.49 years (range: 22–37 years). The second sample, taken from the Brain Genomics Superstructure Project (GSP; Holmes et al., 2015) comprised 854 subjects with a mean age of 22.92 years (range: 21–35 years). The third sample was a subset from the Rockland Sample of the Enhanced Nathan Klein Institute (eNKI; Nooner et al., 2012), comprising 190 subjects with a mean age of 46.02 years (range: 20–83 years). The fourth sample, taken from the 1000Brains dataset (Caspers et al., 2014), comprised 1000 subjects with a mean age of 61.18 years (range: 21–85 years). This sample was included to examine generalization performance to an older sample. A fifth sample (“compound854”) was constructed by combining subsamples of the HCP, GSP, eNKI and 1000Brains samples, with a mean age of 40.05 (range: 20–85), resulting in a sample size of 854 subjects. This sample size is equal to the GSP sample, but larger than the eNKI and lower than the HCP and 1000Brains samples, therefore representing an intermediate sample size compared to the other data samples. Another sixth sample (“compound2190”) was constructed by combining 75% of the HCP, GSP, eNKI and 1000Brains samples resulting in a sample size of 2190 subjects in total. The compound 2190 sample comprised a mean age of 40.10 years (range: 20–85 years). Thus, both compound samples display a large difference in sample size but ratios of dataset representation, sex and age distribution have been maintained. This allows us to evaluate the influence of data composition compared to the sample size of a training sample on the generalization performance of sex classification models.

RS fMRI data from an additional dataset was included to evaluate classifiers on an additional independent sample. This sample comprised 370 subjects (214 females) with a mean age of 22.50 years (range 20–26 years) from the AOMIC dataset (Snoek et al., 2021). It was not additionally balanced for sex to maintain the maximum number of participants for evaluation. Data usage of the included datasets was approved by the Ethics Committee of the Medical Faculty of the Heinrich-Heine University Düsseldorf (4039, 5193, 2018-317-RetroDEuA). All data was collected in research projects approved by a local Review Board, for which all participants provided written informed consent. All experiments were performed in accordance with relevant guidelines and regulations.

2.2 | Data acquisition

2.2.1 | HCP

The RS fMRI data of the HCP dataset were acquired on a Siemens Skyra 3 T MRI scanner with multiband echo-planar imaging with a duration of 873 s and the following parameters: 72 slices; voxel size, $2 \times 2 \times 2 \text{ mm}^3$; field of view (FOV), $208 \times 180 \text{ mm}^2$; matrix, 104×90 ; TR, 720 ms; TE, 33 ms; flip angle, 52° (https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf).

Participants were instructed to lie in the scanner with eyes open, with a “relaxed” fixation on a white cross on a dark background and think of nothing in particular, and to not fall asleep (Smith, Beckmann, et al., 2013).

2.2.2 | GSP

RS data were acquired on a 3 T Tim Trio Scanner with a duration of 372 s and the following parameters: 47 slices; voxel size, $3 \times 3 \times 3 \text{ mm}^3$; FOV, 216 mm; TR, 3 s; TE, 30 ms; flip angle, 85° . During data acquisition, participants were instructed to lay still, stay awake, and keep eyes open while blinking normally (https://static1.squarespace.com/static/5b58b6da7106992fb15f7d50/t/5b68650d8a922db3bb807a90/1533568270847/GSP_README_140630.pdf, Holmes et al., 2015).

2.2.3 | eNKI

Participants in the eNKI dataset were underwent RS scanning for 650 s in a Siemens Magnetom Trio Tim syngo MR scanner with the following parameters: 38 slices; voxel size, $3 \times 3 \times 3 \text{ mm}^3$, FOV, $256 \times 200 \text{ mm}^2$; TR, 2500 ms; TE, 30 ms; flip angle, 80° . Participants were instructed to keep their eyes closed, relax their minds and not to move (Betz et al., 2014).

2.2.4 | 1000Brains

Subjects were scanned for 660 s on a Siemens TRIO 3 T MRI scanner with the following parameters: 36 slices; voxel size, $3.1 \times 3.1 \times 3.1 \text{ mm}^3$; FOV, $200 \times 200 \text{ mm}^2$; matrix, 64×64 , TR = 2.2 s; TE = 30 ms; flip angle, 90° . During RS data acquisition, participants were instructed to keep their eyes closed and let the mind wander without thinking of anything in particular (Caspers et al., 2014).

2.2.5 | AOMIC

The AOMIC dataset includes two subsamples, PIOP1 and PIOP2, comprising data of healthy university students scanned on a Philips 3 T scanner. Participants were instructed to keep their gaze fixated on a fixation cross on the screen and let their thoughts run freely (Snoek et al., 2021). Both samples were acquired with a voxel size of $3 \times 3 \times 3 \text{ mm}^3$ and a matrix size of 80×80 . While PIOP1 was acquired for 360 s with multi-slice acceleration, 480 volumes and a 0.75 TR, PIOP2 was acquired for 480 s without multi-slice acceleration, 240 volumes and a 2 s TR (further details in <https://www.nature.com/articles/s41597-021-00870-6/tables/10>).



2.3 | Data preprocessing

2.3.1 | HCP

RS data from the 'HCP S1200 Release' analyzed here was fully preprocessed and denoised via the Connectome Workbench software. In short, data were corrected for spatial distortions, head motion, B_0 distortions and were registered to the T1-weighted structural image (Smith, Beckmann, et al., 2013). Concatenating these transformations with the structural-to-MNI nonlinear warp field resulted in a single warp per time point, which was applied to the timeseries to achieve a single resampling in the 2 mm isotropic MNI space. Afterwards, global intensity normalization was applied and voxels that were not part of the brain were masked out. Locally noisy voxels as measured by the coefficient of variation were excluded and all the data were regularized with 2 mm Full width half maximum (FWHM) surface smoothing (Glasser et al., 2013; Smith, Beckmann, et al., 2013). The temporal preprocessing included corrections and removal of physiological and movement artifacts by an independent component analysis (ICA) of the FMRIB's X-noisifier (FIX, Salimi-Khorshidi et al., 2014). This method decomposes data into independent components and identifies noise components based on a variety of spatial and temporal features through pattern classification.

2.3.2 | GSP, eNKI, 1000Brains

RS data of the GSP, eNKI and 1000Brains samples were preprocessed in the same way. Initially, FSL was used for the removal of noise and motion artifacts by applying the FIX-denoising procedure (Jenkinson et al., 2012; Salimi-Khorshidi et al., 2014) using the appropriate pre-trained dataset for noise classification. As FIX does not include normalization to MNI space, denoised data were further preprocessed with SPM12 (SPM12 v6685, Wellcome Centre for Human Neuroimaging, 2018; <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>) using Matlab R2014a (Mathworks, Natick, MA). For each subject, the first four echo-planar-imaging (EPI) volumes were discarded and the remaining ones were corrected for head movement by an affine registration with two steps: First, the images were aligned to the first image. Second, the images were aligned to the mean of all volumes. The mean EPI image was spatially normalized to the MNI152 template (Holmes et al., 1998) using the "unified segmentation" approach (Ashburner & Friston, 2005) and the resulting deformation was applied to the FIX-denoised images and resampled to 2 mm³.

2.3.3 | AOMIC

Fully preprocessed data was used provided via OpenNeuro, where it was preprocessed using Fmriprep version 1.4.1 (Esteban et al., 2019; Esteban et al., 2020), a Nipype based tool for reproducible preprocessing in neuroimaging data (Gorgolewski et al., 2011). Data were motion corrected using mcflirt (FSLv5.0.9, (Jenkinson et al., 2002)) followed by distortion correction by co-registering the functional image

to the respective T1 weighted image with inverted intensity (Huntenberg, 2014; Wang et al., 2017) with six degrees of freedom, using bbrregister (FreeSurfer v6.0.1). In a following step, motion correction transformations, field distortion correction warp, BOLD-to-T1-weighted transformation and the warp from T1-weighted to MNI were concatenated and applied using antsApplyTransforms (ANTs v2.1.0.) using Lanczos interpolation (Snoek et al., 2021).

2.4 | Connectome extraction

Following the parcelwise approach by Weis et al. (2020), individual RS connectomes were extracted based on 400 cortical parcels of the Schaefer Atlas (Schaefer et al., 2018), and 36 subcortical parcels of the Brainnetome Atlas (Fan et al., 2016). Each parcel's time series was cleaned by excluding variance that could be explained by mean white matter and cerebrospinal fluid signal (Satterthwaite et al., 2013). Data was not further cleaned for motion related variance as this variance was already removed during FIX preprocessing. For each of the 436 parcels, the activation time series was computed as the mean of all voxel time courses within that parcel. Then, for each parcel, pairwise Pearson correlations were computed between the parcel's time series and those of all other 435 remaining parcels, representing the individual RS functional connectivity (RSFC) profile of the parcel.

2.5 | Parcelwise sex classification

Sex classification models were trained based on the individual multivariate RSFC profile of each parcel. Specifically, the connectivity values between each parcel and the 435 remaining parcels were used as features to train a sex classification model per parcel, resulting in a set of 436 pwC (Weis et al., 2020). Since each model provides one final accuracy value, one pwC provides an accuracy map covering the entire brain. Training sex classification models based on the connectivity profile of each parcel allows for a reduction of the feature dimensionality for each model (1×436) as compared to training one model based on the overall connectivity profile (436×436). Furthermore, using this parcelwise approach allows us to identify the highest classifying brain regions. In the following steps, we evaluated generalization performance in terms of classification accuracies and spatial consistency of highly classifying parcels across the entire brain.

All models were built using support vector machine (SVM) classifiers. SVM is a supervised ML method that separates the data into distinct classes with the widest possible gap between these classes (Boser et al., 1992; Rafi & Shaikh, 2013; Vapnik, 1998; Zhang et al., 2021). Based on its operational principles regarding a supervised binary classification task and successful applications in previous sex classification studies (Flint et al., 2020; Weis et al., 2020; Wiersch et al., 2023), SVM is a suitable method for the present task. SVM models were built in Julearn (Hamdan et al., 2023; <https://juaml.github.io/julearn/main/index.html>) including a hyperparameter search nested within a 10-fold CV with five repetitions. The parameter



search included choice of kernel (linear vs. radial basis function (rbf) kernel) as well as the hyperparameters gamma and C, which is used to set the strength of regularization (https://scikit-learn.org/stable/auto_examples/svm/plot_svm_scale_c.html). The SVM algorithm used in the present study incorporates a squared L2 regularization. The regularization parameter controls the trade-off between the model fit to the training data and generalizable predictions beyond the training data in order to avoid overfitting and to optimize model performance and generalizability (<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>).

Confounding effects of age were regressed out in a CV-consistent manner by removing age-related variance before training the classifiers. By estimating confound regression models only for training subsets and applying them to training and test sets, leakage of information from test to training data within the CV-process can be avoided (More et al., 2021). The best performing combination of hyperparameters was used for the final model for each individual parcel. Within-sample classification accuracy for each individual parcel was determined by averaging accuracies over CV folds and repetitions.

For a cross-sample classification, single dataset pwCs were tested on the respective other three samples, while pwC compound 854 and pwC compound 2190 were tested on the remaining 25% of the HCP ($n = 220$, mean age: 29.68, age range: 22–36), GSP ($n = 214$, mean age: 22–72, age range: 21–31), eNKI ($n = 48$, mean age: 47.52, age range: 20–75) and 1000Brains ($n = 250$, mean age: 52.08, age range: 22–80) sample. Here, for computing time reasons, we restricted the choice of the SVM kernel to rbf (see Weis et al., 2020). Finally, generalization performance of all six pwCs was assessed on the AOMIC sample. All reported accuracies are balanced accuracies.

2.6 | Statistical analyses

2.6.1 | Across-sample classification accuracy

To statistically compare the classification accuracies of pwCs across the different test samples, we employed independent *t*-tests between the different across-sample accuracies over the respectively 10% highest classifying parcels. Additional analyses using all 436 parcels are reported in the supplements (Table S3 and below).

Significance levels were Bonferroni-corrected according to the number of dependent tests (15 dependent tests for comparing across-sample accuracies of all six pwCs on the AOMIC test sample, 10 dependent tests for comparing the across-sample accuracy of both compound pwCs for the five test samples and for comparing pwC performances against each other for each of the five test samples; six dependent tests for all other comparisons).

2.6.2 | Consistency of highly classifying brain regions

Previous studies have demonstrated that sex classification accuracies for models trained on parcelwise RSFC patterns do not achieve

uniformly high performance across the whole brain (Weis et al., 2020; Zhang et al., 2018). Thus, we assessed generalization performance of the different pwCs by examining the consistency of highly classifying brain regions during CV and across-sample testing. Consistency was assessed by computing Dice coefficients (DSC) to evaluate the similarity in spatial distribution of parcels achieving certain accuracies in both CV and across-sample testing. This consistency was evaluated for different accuracy thresholds above chance (0.5–0.7 at 0.02 steps). For each threshold, Dice coefficients were computed as the number of common parcels achieving within- and across-sample accuracies above or equal to that threshold (p_{com}) multiplied by 2 and divided by the total number of parcels achieving a within (p_{tr}) or across-sample (p_{te}) accuracy above or equal to that accuracy level in CV (Dice, 1945; Sorensen, 1948).

$$DSC = \frac{2 * p_{com}}{p_{tr} + p_{te}}$$

To facilitate comparison of the dice score distributions between the different pwCs and test samples, we summarized each contribution into one score by computing a weighted mean (wmDice) as the average of each dice coefficient weighted by the accuracy threshold for which the respective dice coefficient was calculated.

3 | RESULTS

The generalization performance of pwCs trained on each of the single dataset samples (HCP, GSP, eNKI, & 1000Brains) and on both compound samples were compared with respect to mean across-sample accuracy averaged across the best 10% classifying parcels. Additionally, we evaluated the consistency of the spatial distribution of accurately classifying parcels between CV and across-sample testing to determine whether pwCs trained on compound samples exhibit more generalizable results in contrast to pwCs trained on single samples.

3.1 | Training and test classification accuracies

For the single samples pwCs, the mean within-sample performance across the top 10% classifying parcels was at a similar level for pwC GSP (66.8%), pwC eNKI (66.9%) and pwC 1000Brains (66.3%) and ranged up to 73.5% for pwC HCP. The mean across-sample accuracies averaged for the top 10% classifying parcels ranged between 58.4% (for pwC HCP tested on AOMIC and pwC eNKI tested on 1000Brains) and 65.8% (for pwC GSP tested on eNKI). Details for within- and across-sample performance are reported in Table S1 and Figure 1 and Figure S1. Parcelwise within- and across-sample accuracies are displayed as accuracy maps in Figure 1a and the distribution of test accuracies is shown in Figure 3 (red boxplots). Here, accuracy maps represent the spatial distribution of classification accuracies resulting from the 436 individual ML models trained on the respective multivariate RSFC profile of each parcel.



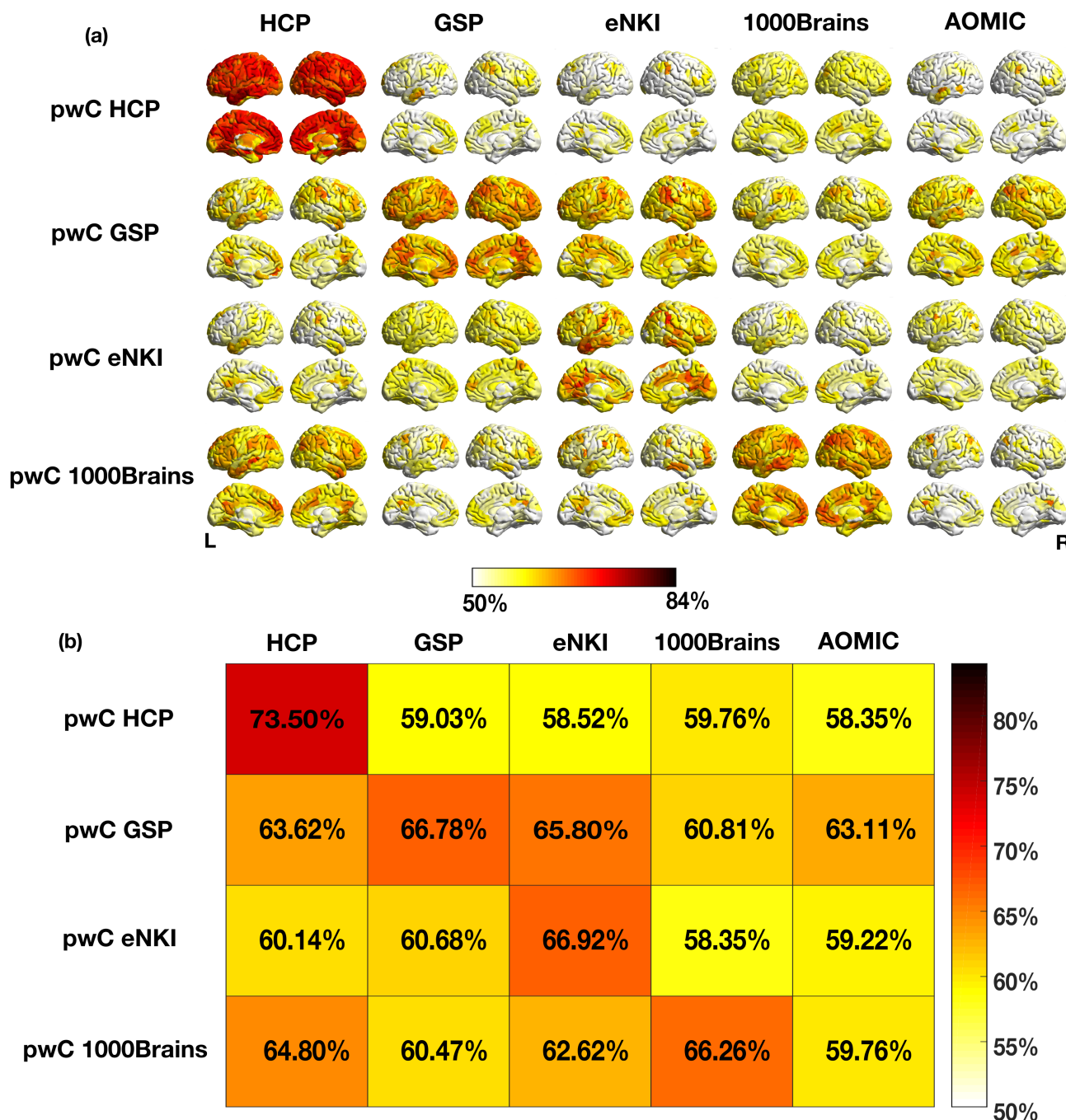


FIGURE 1 Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for parcelwise classifiers (pwCs) trained on single samples. (a) Spatial distribution of parcelwise sex classification accuracies across the brain. Within-sample accuracies are depicted on and across-sample accuracies off the diagonal. Only parcels with an accuracy of 0.5 or higher are displayed. (b) Mean accuracies averaged across the top 10% classifying parcels for each cross-validation (CV) and across-sample prediction.

Accuracy maps for the different combinations of training and test samples were compared using independent t-tests across the top 10% classifying parcels in each prediction (details in Table S2). First, we analyzed differences in classification accuracies between test samples for each pwC (horizontal comparisons, Figure 1): For pwC HCP, testing on 1000Brains achieved the highest mean classification accuracy (59.8%). The averaged accuracy for this test sample

was descriptively higher than for the GSP and significantly higher than for the eNKI and AOMIC test samples. PwC GSP achieved significantly higher accuracies for the eNKI test sample (65.8%) than for any other test sample, while pwC eNKI showed highest accuracies for the GSP test sample (60.7%). This across-sample prediction showed descriptively higher accuracies than pwC eNKI did for the HCP test sample and significantly higher accuracies than for the

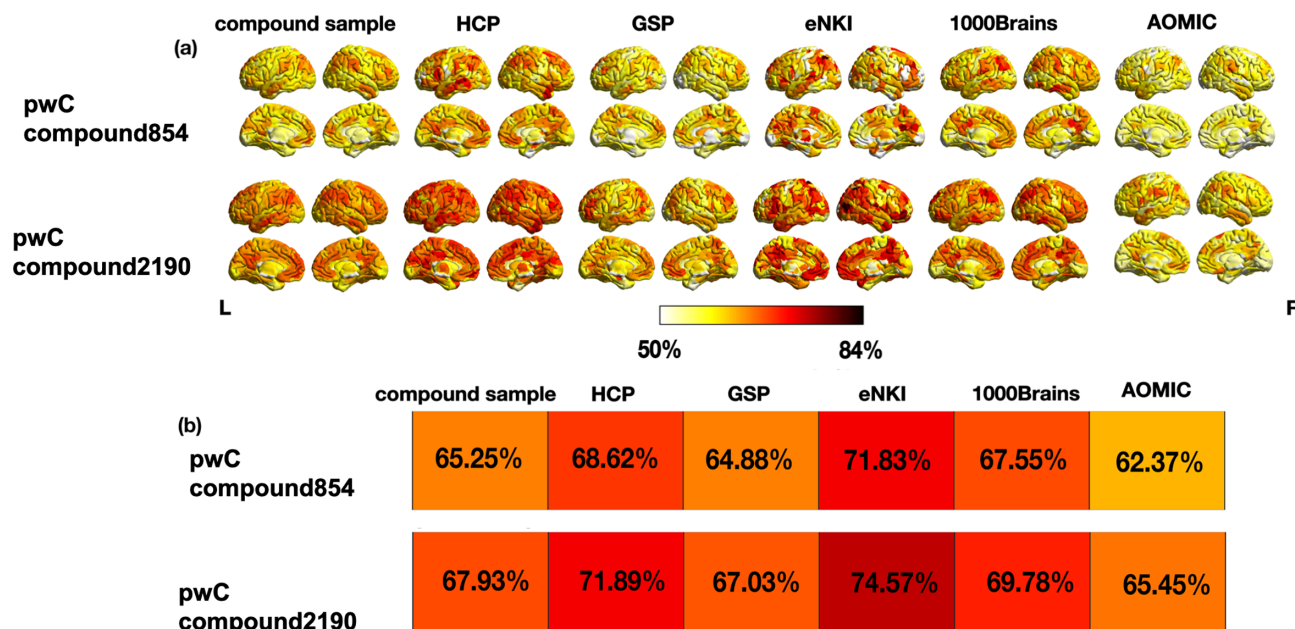


FIGURE 2 Accuracy maps and tile plots of mean accuracies of top 10% classifying parcels for parcelwise classifiers (pwC) compound 854 and pwC compound 2190. (a) Spatial distribution of parcelwise sex classification accuracies across the brain. Only parcels with an accuracy of 0.5 or higher are displayed. (b) Mean accuracies averaged across the top 10% classifying parcels for the respective cross-validation (CV)- (first column) and across-sample predictions.

AOMIC and 1000Brains samples. For pwC 1000Brains, testing on the HCP showed significantly higher accuracies (64.8%) than testing on the eNKI, GSP and AOMIC sample. Details of all statistical comparisons are given in Table S2.

PwC compound854 achieved a mean within-sample accuracy of 65.3% for the top 10% classifying parcels, while mean across-sample accuracies of the highest classifying parcels ranged between 62.4% (pwC compound854 tested on AOMIC) and 71.8% (pwC compound854 tested on eNKI, further details in Table S1 and Figure 2, Figure S2). PwC compound2190 achieved a mean within-sample accuracy of 67.9% within the top 10% classifying parcels. The mean across-sample accuracies averaged across the top 10% classifying parcels ranged between 65.5% (pwC compound2190 tested on AOMIC) and 74.6% (pwC compound2190 tested on eNKI, details in Table S1 and Figure 2, Figure S2).

Contrasting the top 10% classifying parcels in the accuracy maps of pwC compound854 and pwC compound2190 displayed peaks in accuracies for the eNKI test sample (71.8% and 74.6%) resulting in significantly higher accuracies than for the remaining test samples, respectively (Figure 2 and Table S2). We also contrasted how the six pwCs performed on each test sample by employing independent t-tests: pwC compound 854 outperformed all pwCs trained on single samples for all test samples, except for the AOMIC test sample, where pwC GSP achieved higher accuracies within the best 10% classifying parcels (Table S2). PwC compound 2190 outperformed all other pwCs for the HCP, GSP, eNKI and AOMIC test sample with regards to the top 10% classifying parcels in each across-sample prediction (Figure 2). Details for all statistical comparisons are shown in Table S2.

3.2 | Consistency of correctly classifying parcels

To evaluate the spatial consistency of accurately classifying parcels, we calculated the dice coefficient between thresholded within- and across-sample accuracy maps at different levels of accuracy. Here, a high dice coefficient indicates a high overlap in highly classifying parcels between within and across-sample predictions at a given accuracy level. The results are depicted in the blue bar plots in Figure 3. Regarding spatial consistency within a given pwC (horizontal comparison in Figure 3), pwC HCP overall demonstrated relatively low spatial consistency while it was highest for 1000Brains (wmDice = 0.1765, all other wmDice < 0.1112). Spatial consistency for pwC GSP was highest for the eNKI sample (wm = 0.3103) and lowest for 1000Brains (wmDice = 0.1810) with spatial consistency for HCP (wmDice = 0.2407) and AOMIC (wmDice = 0.2607) test samples ranging in between. PwC eNKI showed overall low spatial consistency for the HCP, 1000Brains and AOMIC sample (wmDice: 0.1244–0.1523) and highest for the GSP sample (wmDice = 0.2072). Spatial consistency of pwC 1000Brains was lower for the GSP, eNKI and AOMIC test sample (wmDice: 0.1201–0.1853) but considerably higher for the HCP test sample (wmDice = 0.3159). Spatial consistency of pwC compound854 ranged between 0.2865–0.3221 for the HCP, GSP, eNKI and 1000Brains sample and achieved 0.2546 for the AOMIC sample. PwC compound2190 demonstrated a relatively similar spatial consistency for HCP, GSP, eNKI and 1000Brains (wmDice: 0.3641–0.4168) and lower spatial consistency with the AOMIC sample (wmDice = 0.2960). Concerning the comparisons within each test sample (vertical comparisons in Figure 3) pwC compound854 demonstrated higher spatial consistency than single sample

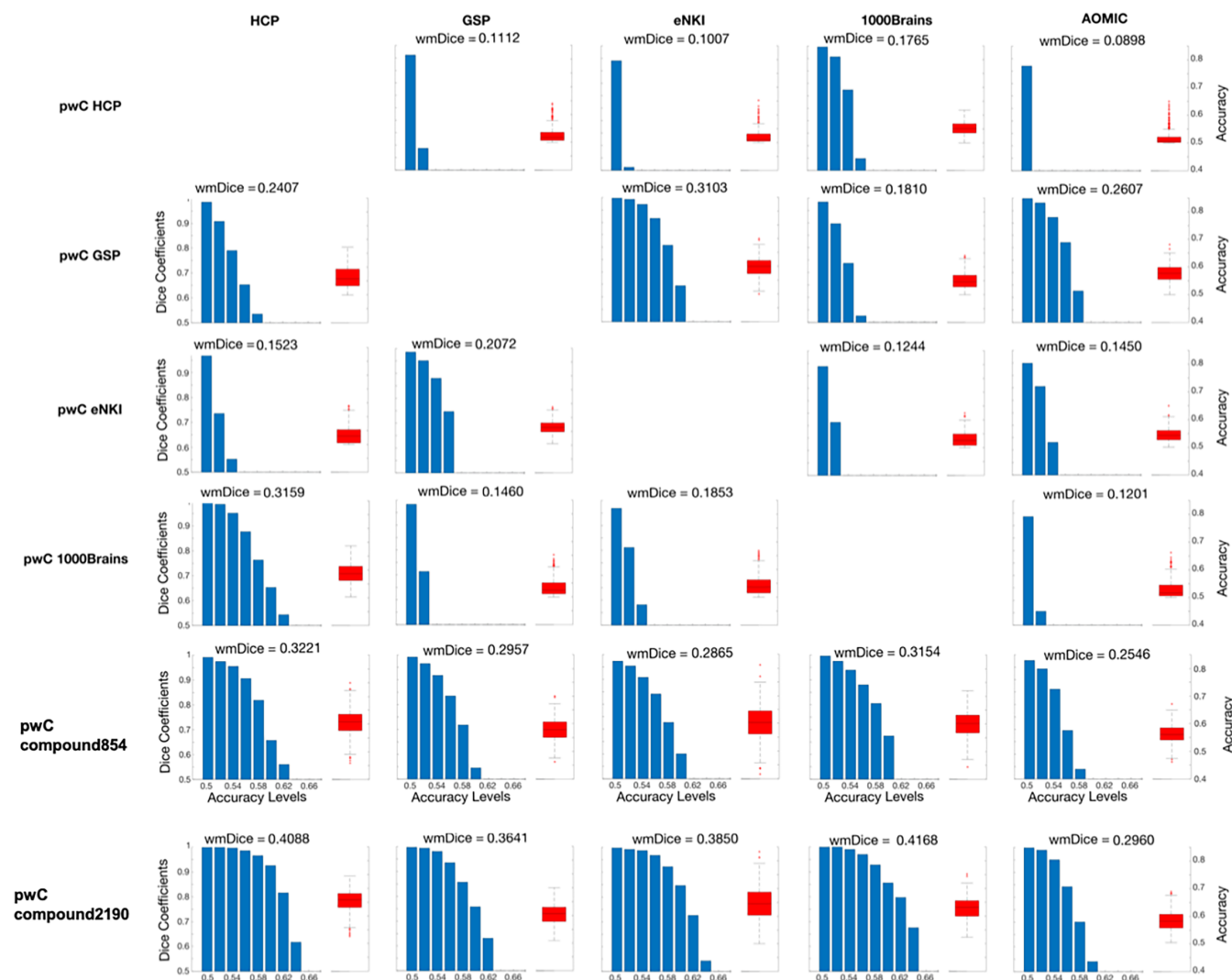


FIGURE 3 Spatial consistency of all parcelwise classifiers (pwCs). For each combination of training (rows) and test sample (columns), the right side of each subplot (red boxplot) depicts the distribution of accuracies across all parcels (right y-axis). The left side of each subplot (blue barplot) shows the dice coefficients (left y-axis), representing the overlap of accuracy maps between cross-validation (CV) and test predictions at different accuracy levels (x-axis). For each accuracy-threshold, the respective dice coefficient was calculated as the number of similar parcels classifying above a certain accuracy-threshold in both, respective CV and test prediction, in relation to the total number of parcels of both predictions classifying at this level. For each combination of pwC and test sample, the weighted mean of the dice coefficients (wmDice) across accuracy levels is displayed above the subplot to allow for a straightforward comparison between the distributions of dice coefficients.

pwCs for the HCP, GSP and 1000Brains test samples and pwC compound2190 demonstrated higher spatial consistency than the other six pwCs. Dice coefficients for the top 10% classifying parcels are reported in Figure S3.

4 | DISCUSSION

In the present study, we examined the generalization performance of parcelwise sex classification models trained on different samples. Here, we operationalized generalization performance in terms of both mean classification accuracy of best classifying parcels during across-sample testing as well as spatial consistency in highly classifying

parcels between CV and across-sample testing. Since not all parcels are expected to achieve high classification accuracies (Weis et al., 2020; Zhang et al., 2018), we mainly focused on the top 10% classifying parcels. Overall, our results showed that classifiers trained on single dataset samples generalized well only for certain test samples. In contrast, classifiers trained on the compound samples tend to outperform classifiers trained on single dataset samples both in terms of accuracy and consistency of accurately classifying parcels.

To evaluate generalization performance with respect to mean classification accuracies of the top 10% classifying parcels, for each pwC, we compared across-sample classification accuracies between the different test samples. Results indicate that certain datasets seem to “match” in the sense that classifiers trained on a sample from one

of the datasets achieved a high accuracy when tested on the respective other one and vice versa. This was the case for HCP and 1000Brains as well as for GSP and eNKI with the former matching the results of a previous study (Weis et al., 2020). Based on the good across-sample performance of sex classifiers trained on an HCP sample on a subsample of the 1000Brains, Weis et al. (2020) suggested that parcelwise sex classification generalizes well between different samples. No additional samples from other datasets were considered in Weis et al. (2020). The present results extend the findings of the previous study by showing that good generalization performance of the HCP classifiers appears to be specific to the 1000Brains sample. Generalization to samples from other datasets (GSP, eNKI and AOMIC) is, however, rather poor. Thus, our study demonstrates that the generalizability of pwCs trained on single dataset samples depends on the train-test data combination, which is in line with a previous study that employed sex classification based on regional homogeneity of RS time series (Huf et al., 2014). The limited generalization performance of the pwCs trained on single dataset samples to the majority of test samples from other datasets might be attributed to the homogeneity of each single dataset training sample arising due to demographic factors such as the age range (Damoiseaux, 2017; Damoiseaux et al., 2008; Scheinost et al., 2015) as well as technical details such as fMRI acquisition parameters (Brown et al., 2011; Yu et al., 2018). Homogeneous data characteristics within each dataset will result in a homogeneity of the feature space on which ML models are trained. Such homogeneous features might lead the ML model to learn dataset-specific characteristics that are predictive of the target variable, which might not translate to other test samples, resulting in inaccurate across-sample predictions (Huf et al., 2014). Thus, training ML models on a single, homogenous sample may not be ideal to achieve a good generalization performance on diverse test samples (Belur Nagaraj et al., 2020; Di Tanna et al., 2020; Huf et al., 2014; Janssen et al., 2018). In contrast, training classifiers on a combination of multiple datasets (pwC compound854 and pwC compound2190) achieved significantly higher accuracies for all test samples, including the sample from a dataset which was not included in the compound training sample. We contrasted performances of both pwCs trained on compound samples to evaluate potential sample size effects. Here, pwC compound854 demonstrated higher accuracies and spatial consistency in the majority of across-sample predictions compared to single sample pwCs, but did not outperform pwC compound2190. These results suggest that the sample size of the training sample is an important factor in determining the generalization performance of ML analyses. These results align with the findings of several other studies highlighting the importance of the sample size in ensuring accurate ML results (Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Ishida, 2019; Yang et al., 2020). However, pwC compound854 still predominantly demonstrated a higher generalization performance compared to single sample pwCs with a similar or even higher sample size. Thus, it is evident that the composition of a training sample is crucial in ensuring generalizable ML results, as reported by previous studies (Chang et al., 2018; Huf et al., 2014; Willemink et al., 2020). While a high sample size is beneficial to assure reliable and accurate

ML predictions (Dhamala et al., 2023), the heterogeneity and representativeness of a composite sample led to significantly better results than single sample pwCs with a higher sample size in the present ML analyses. Thus, the high generalization performance of both compound samples is not only attributable to the sample size but also to the heterogeneity of data characteristics included in a training sample created from various datasets. This heterogeneity likely enables the model to learn patterns that do not rely on specific sample characteristics, but actually capture the underlying relationship between features and target, enabling the model to generalize better, even to data from datasets that were not included in training. Therefore, the heterogeneity of a composite training sample is essential for generalizable ML outcomes and may also serve to minimize sample-specific biases (Li et al., 2022). Thus, training on a compound sample comprising the variability of multiple datasets is preferable to training on single dataset samples in order to achieve high generalization performances (Chang et al., 2018; Huf et al., 2014; Willemink et al., 2020).

While undesirable sources of variability, e.g. due to scanner differences, may be accounted for by using data harmonization (Fortin et al., 2017; Yu et al., 2018), in the present study we intentionally refrained from using harmonization techniques. Here, we evaluated the generalization performances of differently trained pwCs in order to determine which may generalize best to unseen data. Harmonization techniques such as ComBat are not suitable for this purpose because they require a sufficient amount of data from each sample and site (Orlhac et al., 2022).

The parcelwise classification approach allowed us to investigate generalization performance not only in terms of accuracy but also with respect to the spatial distribution of accurately classifying parcels. To quantify the overlap of accurately classifying parcels between CV and across-sample testing, we computed dice coefficients between within- and across sample accuracy maps at different accuracy thresholds. We observed a pattern similar to the one found for classification accuracies, with the train-test pairing of HCP and 1000Brains and GSP and eNKI, respectively, showing highest spatial consistency, relative to other combinations. Thus, also when considering spatial consistency, generalization performance depended on the specific pairing of training and test datasets. For pwCs trained on single samples, training sample characteristics appeared to be the most important factor in driving generalization performance across test samples. In contrast, pwC compound854 achieved superior spatial consistency in most test samples and pwC compound2190 in all test samples, as compared to pwCs trained on single samples. Thus, the classifiers trained on the compound samples achieved both higher classification accuracies as well as more consistency in accurately classifying parcels as opposed to the classifiers trained on single dataset samples. Altogether, the high generalization performance for pwC compound854 and pwC compound2190 can likely be attributed to the data heterogeneity in the respective training samples which was achieved by combining multiple samples for training. These findings match results of previous studies (Chang et al., 2018; Huf et al., 2014; Nielsen et al., 2020; Willemink et al., 2020).



Overall, the aggregation of multiple samples in pwC compound854 and pwC compound2190 for training sex classifiers resulted in superior generalization performance compared to pwCs trained on single samples. Firstly, the classification accuracies were comparable between CV and the different across-sample test classifications. Secondly, highly classifying parcels overlapped to a large degree between training and test. The overall high generalization performance of pwC compound2190 across all test samples could be attributed to several possible explanations: first, the compound2190 sample is more than twice as large as compared to any of the single dataset samples. Such high sample size has been shown to be beneficial for generalization (Cui & Gong, 2018; Domingos, 2012; Ishida, 2019; Yang et al., 2020). However, sample size alone is likely not sufficient to explain the high generalization performance. For instance, the eNKL sample consists of only 190 participants, but the classifiers trained on this sample achieved better generalization performance than those trained on the HCP sample, which included 878 participants. In addition, analyses with pwC compound854 also demonstrated a superior generalization performance with respect to classification accuracies as well as spatial consistency compared to single sample pwCs, despite the smaller sample size. A second explanation for the good performance of both compound pwCs may lie in the heterogeneous nature of its training sample as discussed above. Having the different samples represented within the compound sample may have allowed the classifiers to classify sex based on sample-unspecific information. Another potential explanation is that the training samples of pwC compound854 and pwC compound2190 partially consist of data from datasets on which we evaluated the test performance. In general, training on data that is representative of the test data typically results in an increased generalization performance (Chung et al., 2018). Here, both training samples for the compound pwCs composed data from four different datasets. Although each dataset had a different sample size and thus a different share in the respective compound training sample, the model applications to the eNKL test sample showed highest accuracies for the best 10% classifying parcels. This result stems from few parcels classifying at a high level for the eNKL test data (up to 83%), resulting in such a high mean accuracy for the top 10% parcels (Figure 2). Furthermore, the mean accuracy averaged across all 436 parcels confirms that there are only few parcels responsible for the high accuracy in the top 10% parcels, as the eNKL dataset did not exhibit the overall highest mean accuracy across all parcels.

In contrast to both compound pwCs, CV and across sample test performances differed considerably for pwCs trained on single dataset samples. This lack of generalization performance was especially apparent for pwC HCP which showed a rather high performance during CV in combination with the lowest generalization performance both with respect to accuracy and spatial consistency. While homogeneity of a data sample has been argued to lead to high CV classification accuracy (Huf et al., 2014), sample characteristics such as the age range were comparable between HCP and the GSP sample, with the latter outperforming HCP in generalization performance. Thus, the comparably poor performance of classifiers trained on the HCP sample may be

partially attributed to sample homogeneity but also to other factors such as the differences in preprocessing pipelines. For the HCP sample, connectome extraction was based on the FIX denoised preprocessed version of the data. The eNKL, GSP and 1000Brains samples were preprocessed using the same pipeline in FSL/SPM12 also including FIX-denoising, while the AOMIC sample was preprocessed using fMRIPrep without FIX. Given that comparative performance evaluation of fMRI data is sensitive to preprocessing decisions (Bhagwat et al., 2021), it is likely that this difference in preprocessing may contribute to the poor generalization performance of pwC HCP when tested on the other single samples. Furthermore, the high within-sample accuracy coupled with the lack of generalization performance may also indicate an overfitting effect of pwC HCP during training (Cui & Gong, 2018; Domingos, 2012).

The present study, however, does not primarily aim to build a classifier attaining highest sex classification accuracies but rather to evaluate the impact of the training sample in ML models, particularly the size and composition of the training sample.

Altogether, our results highlight the importance of the sample size and also a heterogeneous, diverse, and representative data composition for training ML models (Cui & Gong, 2018; Dhamala et al., 2023; Domingos, 2012; Gong et al., 2019; Li et al., 2022), which can be achieved by combining data from multiple sites and datasets (Chang et al., 2018; Nielsen et al., 2020; Willemink et al., 2020). By minimizing sample-specific biases, we can aim for maximizing the generalizability of ML models.

4.1 | Limitations

The present results consistently demonstrated the superior generalizability of sex classifiers trained on compound samples as compared to those trained on single dataset samples, but they come with some limitations. First of all, the high spatial consistency of pwC compound2190 might partially be attributed to the generally higher accuracy of the across-sample predictions. Dice coefficients across the top 10% classifying parcels showed a more differentiated pattern. Here, pwC compound2190 did not always outperform pwCs trained on single samples. Overall, the predominantly higher generalization performance of pwC compound2190 can be attributed to the sample size and sample composition of its training sample. However, an additional systematic study would be required to determine the exact degree to which each factor contributes to high generalization performance.

Another limitation in the present study is that, while we accounted for age as a potential confound during training of the classifiers, there might be other confounds that were not considered. For example, we did not control for structural variables such as brain size, which have been reported to influence brain functions (Batista-Garcia-Ramo & Fernandez-Verdecia, 2018) and RS brain connectivity in particular (Zhang et al., 2018). Thus, in principle, different distributions of brain size within the different samples might have influenced the present results. However, Weis et al. (2020) demonstrated that at



least with their training sample, classification based on RS connectivity was not systematically influenced by brain size. Still, there might be other demographic variables which differ between samples and might influence classification accuracies (Li et al., 2022; Mehrabi et al., 2021; Sripada et al., 2021).

A further limitation of the present study is the potential impact of different preprocessing approaches which may affect the outcomes in ML analyses. In neuroimaging data, there can be various sources of noise and artifacts. Prior to data analysis, it is necessary to preprocess the data to mitigate these issues and enhance the data quality. However, the impact of preprocessing steps on the outcomes of fMRI analyses has been well documented. For instance, conceptually similar preprocessing packages such as AFNI, FSL, or SPM can produce differences in fMRI results (Bowring et al., 2019). Differences on the level of preprocessing steps may also produce dissimilarities (Carp, 2012). Even differences in the order of preprocessing steps can lead to differences in the graph theoretical outcomes derived from RS functional connectivity (Gargouri et al., 2018). Thus, it is plausible that discrepancies in preprocessing pipelines may lead to differences in classification outcomes. Indeed, one study that compared ML results for patient and healthy control classification across different preprocessing pipelines indicated differences in the classification accuracy (Vergara et al., 2017). Overall, while different preprocessing approaches may lead to differences in the fMRI and ML results, in the present study these differences represent an additional source of variance that may occur when using data of various datasets. Despite various potential sources of variance within the training samples of the compound pwCs, pwC compound854 and pwC compound2190 demonstrate a comparatively good performance compared to the single sample pwCs. While it is reasonable to anticipate that aligned preprocessing approaches may improve predictions; however, conducting a systematic evaluation on the effect of preprocessing pipelines is beyond the scope of the present study and remains an important open question for future research.

Another factor which has not been considered in the present analyses are fluctuating sex hormones, which have been shown to influence functional brain connectivity in RS (Arélin et al., 2015; Haraguchi et al., 2021; Weis et al., 2019). These dynamic changes in female and male connectivity patterns (Coenjaerts et al., 2023; Kogler et al., 2016; McEwen & Milner, 2017) will likely influence overall sex classification accuracies. However, unfortunately, most publicly available datasets do not provide information on hormone levels, making it impossible to consider these variations in the analyses. Future large-scale studies should include hormone levels in data acquisition, enabling model training on a combination of multiple independent datasets with well characterized phenotypes to achieve most accurate results.

5 | CONCLUSION

The present results show that parcelwise sex classification models generalize best when trained on a compound sample including data

with different demographic and data acquisition characteristics. Our results demonstrate that a large and heterogeneous training sample including multiple datasets is best suited to achieve accurate generalization performance. This observation carries practical implications for future neuroimaging studies employing ML models for generalizable predictions.

ACKNOWLEDGMENTS

We would like to thank Dr. Federico Raimondo and Dr. Georgios Antonopoulos and all people contributing in the “ML hour” initiative from INM-7 for the insightful discussions. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The work was supported by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) – Project-ID 431549029 - Collaborative Research Centre CRC1451 on motor performance project B05, the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme “Supercomputing and Modeling for the Human Brain”, and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3). Open access publication funded by the DFG – 491111487.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The datasets HCP, GSP, eNKI and AOMIC are publicly available and free to download: <https://www.humanconnectome.org/study/hcp-young-adult/data-releases> <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/25833> <https://openneuro.org/datasets/ds001021/versions/1.0.0> <https://nilab-uva.github.io/AOMIC.github.io/>. Data of the 1000Brains are available upon request from the responsible Principal Investigator (Caspers et al., 2014). The code for preprocessing, data preparation, model training and computation of further analyses is available on Github: https://jugit.fz-juelich.de/l.wiersch/functional_sex_classification_code https://jugit.fz-juelich.de/f.hoffstaedter/bids_pipelines/-/tree/master/func.

ORCID

Lisa Wiersch  <https://orcid.org/0000-0001-8006-8678>

REFERENCES

- Arélin, K., Mueller, K., Barth, C., Rekkas, P. V., Kratzsch, J., Burmann, I., Villringer, A., & Sacher, J. (2015). Progesterone mediates brain functional connectivity changes during the menstrual cycle—a pilot resting state MRI study. *Frontiers in Neuroscience*, 9, 44. <https://doi.org/10.3389/fnins.2015.00044>
- Arslan, A. (2018). Application of neuroimaging in the diagnosis and treatment of depression. *Understanding Depression: Volume 2. Clinical Manifestations, Diagnosis and Treatment*, 2, 69–81.
- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3), 839–851. <https://doi.org/10.1016/j.neuroimage.2005.02.018>



- Batista-Garcia-Ramo, K., & Fernandez-Verdecia, C. I. (2018). What we know about the brain structure-function relationship. *Behavioral Sciences*, 8(4), 1–14. <https://doi.org/10.3390/bs8040039>
- Belur Nagaraj, S., Pena, M. J., Ju, W., Heerspink, H. L., & BEAT-DKD Consortium. (2020). Machine-learning-based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. *Diabetes, Obesity and Metabolism*, 22(12), 2479–2486.
- Betzel, R. F., Byrge, L., He, Y., Goni, J., Zuo, X. N., & Sporns, O. (2014). Changes in structural and functional connectivity among resting-state networks across the human lifespan. *NeuroImage*, 102(Pt 2), 345–357. <https://doi.org/10.1016/j.neuroimage.2014.07.067>
- Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C. M. T., Misić, B., Kennedy, D. N., & Poline, J. B. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience*, 10(1), 1–13. <https://doi.org/10.1093/gigascience/giaa155>
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). Association for Computing Machinery.
- Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, 40(11), 3362–3384.
- Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., McCarthy, G., Voyvodic, J., Glover, G., Diaz, M., Yetter, E., Ozyurt, I. B., Jorgensen, K. W., Wible, C. G., Turner, J. A., Thompson, W. K., Potkin, S. G., & Function Biomedical Informatics Research Network. (2011). Multisite reliability of cognitive BOLD data. *NeuroImage*, 54(3), 2163–2175. <https://doi.org/10.1016/j.neuroimage.2010.09.076>
- Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *The British Journal of General Practice*, 68(668), 143–144. <https://doi.org/10.3399/bjgp18X695213>
- Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience*, 6, 149.
- Caspers, S., Moebus, S., Lux, S., Pundt, N., Schutz, H., Muhleisen, T. W., Gras, V., Eickhoff, S. B., Romanzetti, S., Stöcker, T., Stirnberg, R., Kirlangic, M. E., Minnerop, M., Pieperhoff, P., Mödder, U., Das, S., Evans, A. C., Jöckel, K. H., Erbel, R., ... Amunts, K. (2014). Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Frontiers in Aging Neuroscience*, 6, 149. <https://doi.org/10.3389/fnagi.2014.00149>
- Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., Rosen, B., Rubin, D. L., & Kalpathy-Cramer, J. (2018). Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8), 945–954. <https://doi.org/10.1093/jamia/ocy017>
- Chen, J., Patil, K. R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., Aleman, A., Sommer, I. E., Liemburg, E. J., Hoffstaedter, F., Habel, U., Derntl, B., Liu, X., Fischer, J. M., Kogler, L., Regenbogen, C., Diwadkar, V. A., Stanley, J. A., Riedl, V., ... Pharmacotherapy Monitoring and Outcome Survey (PHAMOUS) Investigators. (2020). Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biological Psychiatry*, 87(3), 282–293. <https://doi.org/10.1016/j.biopsych.2019.08.031>
- Chung, Y., Haas, P. J., Upfal, E., & Kraska, T. (2018). Unknown examples & machine learning model generalization. *arXiv Preprint arXiv:1808.08294*.
- Coenjaerts, M., Adrovic, B., Trimborn, I., Philipsen, A., Hurlemann, R., & Scheele, D. (2023). Effects of exogenous oxytocin and estradiol on resting-state functional connectivity in women and men. *Scientific Reports*, 13(1), 3113. <https://doi.org/10.1038/s41598-023-29754-y>
- Cohen, A. D., Chen, Z., Parker Jones, O., Niu, C., & Wang, Y. (2020). Regression-based machine-learning approaches to predict task activation using resting-state fMRI. *Human Brain Mapping*, 41(3), 815–826. <https://doi.org/10.1002/hbm.24841>
- Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, 178, 622–637. <https://doi.org/10.1016/j.neuroimage.2018.06.001>
- Damoiseaux, J. S. (2017). Effects of aging on functional and structural brain connectivity. *NeuroImage*, 160, 32–40. <https://doi.org/10.1016/j.neuroimage.2017.01.077>
- Damoiseaux, J. S., Beckmann, C. F., Arigita, E. J., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Rombouts, S. A. (2008). Reduced resting-state brain activity in the "default network" in normal aging. *Cerebral Cortex*, 18(8), 1856–1864. <https://doi.org/10.1093/cercor/bhm207>
- Dhamala, E., Yeo, B. T. T., & Holmes, A. J. (2023). One size does not fit all: Methodological considerations for brain-based predictive modeling in psychiatry. *Biological Psychiatry*, 93(8), 717–728. <https://doi.org/10.1016/j.biopsych.2022.09.024>
- Di Tanna, G. L., Wirtz, H., Burrows, K. L., & Globe, G. (2020). Correction: Evaluating risk prediction models for adults with heart failure: A systematic literature review. *PLoS One*, 15(7), e0235970. <https://doi.org/10.1371/journal.pone.0235970>
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.
- Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., Kent, J. D., Goncalves, M., DuPre, E., Gomez, D. E. P., Ye, Z., Salo, T., Valabregue, R., Amlen, I. K., Liem, F., Jacoby, N., Stojic, H., Cieslak, M., Urchs, S., ... Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature Protocols*, 15(7), 2186–2202. <https://doi.org/10.1038/s41596-020-0327-3>
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1), 111–116. <https://doi.org/10.1038/s41592-018-0235-4>
- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The human Brainnetome atlas: A new brain atlas based on connective architecture. *Cerebral Cortex*, 26(8), 3508–3526. <https://doi.org/10.1093/cercor/bhw157>
- Flint, C., Förster, K., Koser, S. A., Konrad, C., Zwitterlood, P., Berger, K., Hermesdorf, M., Kircher, T., Nenadic, I., Krug, A., Baune, B. T., Dohm, K., Redlich, R., Opel, N., Arolt, V., Hahn, T., Jiang, X., Dannlowski, U., & Grotegerd, D. (2020). Biological sex classification with structural MRI data shows increased misclassification in transgender women. *Neuropsychopharmacology*, 45(10), 1758–1765. <https://doi.org/10.1038/s41386-020-0666-3>
- Fortin, J. P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161, 149–170. <https://doi.org/10.1016/j.neuroimage.2017.08.047>
- Gargouri, F., Kallel, F., Delphine, S., Ben Hamida, A., Lehericy, S., & Valabregue, R. (2018). The influence of preprocessing steps on graph theory measures derived from resting state fMRI. *Frontiers in Computational Neuroscience*, 12, 8.
- Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews. Neuroscience*, 23(5), 307–318. <https://doi.org/10.1038/s41583-022-00584-7>

- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80, 105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127>
- Gong, Z., Zhong, P., & Hu, W. (2019). Diversity in machine learning. *IEEE Access*, 7, 64323–64350.
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 5, 13. <https://doi.org/10.3389/fninf.2011.00013>
- Hamdan, S., More, S., Sasse, L., Komeyer, V., Patil, K. R., & Raimondo, F. (2023). Julearn: An easy-to-use library for leakage-free evaluation and inspection of ML models. *arXiv Preprint arXiv:2310.12568*.
- Haraguchi, R., Hoshi, H., Ichikawa, S., Hanyu, M., Nakamura, K., Fukasawa, K., Poza, J., Rodríguez-González, V., Gómez, C., & Shigihara, Y. (2021). The menstrual cycle alters resting-state cortical activity: A magnetoencephalography study. *Frontiers in Human Neuroscience*, 15, 652789.
- Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, 7(7), 523–534. <https://doi.org/10.1038/nrn1931>
- Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., Fischl, B., Rosen, B. R., Mair, R. W., Roffman, J. L., Smoller, J. W., & Buckner, R. L. (2015). Brain genomics Superstruct project initial data release with structural, functional, and behavioral measures. *Scientific Data*, 2, 150031. <https://doi.org/10.1038/sdata.2015.31>
- Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., & Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography*, 22(2), 324–333.
- Huf, W., Kalcher, K., Boubela, R. N., Rath, G., Vecsei, A., Filzmoser, P., & Moser, E. (2014). On the generalizability of resting-state fMRI machine learning classifiers. *Frontiers in Human Neuroscience*, 8, 502.
- Huntenberg, J. M. (2014). Evaluating nonlinear coregistration of BOLD EPI and T1w images. (Doctoral dissertation, Freie Universität Berlin).
- Ishida, E. E. (2019). Machine learning and the future of supernova cosmology. *Nature Astronomy*, 3(8), 680–682.
- Jansma, J. M., Rutten, G. J., Ramsey, L. E., Snijders, T. J., Bizzi, A., Rosengarth, K., Dodoo-Schittko, F., Hattingen, E., de la Peña, M. J., von Campe, G., Jehna, M., & Ramsey, N. F. (2020). Correction to: Automatic identification of atypical clinical fMRI results. *Neuroradiology*, 62(12), 1723. <https://doi.org/10.1007/s00234-020-02565-y>
- Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(9), 798–808. <https://doi.org/10.1016/j.bpsc.2018.04.004>
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841. [https://doi.org/10.1016/s1053-8119\(02\)91132-8](https://doi.org/10.1016/s1053-8119(02)91132-8)
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, 62(2), 782–790.
- Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., Martinot, J. L., Paus, T., Smolka, M. N., Walter, H., Schumann, G., Garavan, H., & Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, 199, 351–365. <https://doi.org/10.1016/j.neuroimage.2019.05.082>
- Kogler, L., Muller, V. I., Seidel, E. M., Boubela, R., Kalcher, K., Moser, E., Habel, U., Gur, R. C., Eickhoff, S. B., & Derntl, B. (2016). Sex differences in the functional connectivity of the amygdalae in association with cortisol. *NeuroImage*, 134, 410–423. <https://doi.org/10.1016/j.neuroimage.2016.03.064>
- Kohoutova, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C. W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, 15(4), 1399–1435. <https://doi.org/10.1038/s41596-019-0289-5>
- Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, 8(11), eabj1812. <https://doi.org/10.1126/sciadv.abj1812>
- McEwen, B. S., & Milner, T. A. (2017). Understanding the broad influence of sex hormones and sex differences in the brain. *Journal of Neuroscience Research*, 95(1–2), 24–39. <https://doi.org/10.1002/jnr.23809>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.
- Menon, S. S., & Krishnamurthy, K. (2019). A comparison of static and dynamic functional Connectivities for identifying subjects and biological sex using intrinsic individual brain connectivity. *Scientific Reports*, 9(1), 5729. <https://doi.org/10.1038/s41598-019-42090-4>
- More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S. B., Patil, K. R., & Alzheimer's Disease Neuroimaging Initiative. (2023). Brain-age prediction: A systematic comparison of machine learning workflows. *NeuroImage*, 270, 119947. <https://doi.org/10.1016/j.neuroimage.2023.119947>
- More, S., Eickhoff, S. B., Caspers, J., & Patil, K. R. (2021). Confound removal and normalization in practice: A neuroimaging based sex prediction case study. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 3–18.
- Nielsen, A. N., Barch, D. M., Petersen, S. E., Schlaggar, B. L., & Greene, D. J. (2020). Machine learning with neuroimaging: Evaluating its applications in psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(8), 791–798. <https://doi.org/10.1016/j.bpsc.2019.11.007>
- Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R. T., Kamiel, S. M., Anwar, A. R., Hinz, C. M., Kaplan, M. S., Rachlin, A. B., ... Milham, M. P. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6, 152. <https://doi.org/10.3389/fnins.2012.00152>
- Nostro, A. D., Müller, V. I., Varikuti, D. P., Pläschke, R. N., Hoffstaedter, F., Langner, R., Patil, K. R., & Eickhoff, S. B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain Structure & Function*, 223(6), 2699–2719. <https://doi.org/10.1007/s00429-018-1651-z>
- Orlhac, F., Eertink, J. J., Cottureau, A. S., Zijlstra, J. M., Thieblemont, C., Meignan, M., Boellaard, R., & Buvat, I. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, 63(2), 172–179.
- Pläschke, R. N., Patil, K. R., Cieslik, E. C., Nostro, A. D., Varikuti, D. P., Plächti, A., Lösche, P., Hoffstaedter, F., Kalenscher, T., Langner, R., & Eickhoff, S. B. (2020). Age differences in predicting working memory performance from network-based functional connectivity. *Cortex*, 132, 441–459. <https://doi.org/10.1016/j.cortex.2020.08.012>
- Rafi, M., & Shaikh, M. S. (2013). A comparison of SVM and RVM for document classification. *arXiv Preprint arXiv:1301.2785*.
- Rutten, G. J., & Ramsey, N. F. (2010). The role of functional magnetic resonance imaging in brain surgery. *Neurosurgical Focus*, 28(2), E4. <https://doi.org/10.3171/2009.12.FOCUS09251>
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith, S. M. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90, 449–468. <https://doi.org/10.1016/j.neuroimage.2013.11.046>

- Sanford, N., Ge, R., Antoniadis, M., Modabbernia, A., Haas, S. S., Whalley, H. C., Galea, L., Popescu, S. G., Cole, J. H., & Frangou, S. (2022). Sex differences in predictors and regional patterns of brain age gap estimates. *Human Brain Mapping*, 43(15), 4689–4698. <https://doi.org/10.1002/hbm.25983>
- Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64, 240–256. <https://doi.org/10.1016/j.neuroimage.2012.08.052>
- Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global Parcelation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9), 3095–3114. <https://doi.org/10.1093/cercor/bhx179>
- Scheinost, D., Finn, E. S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M., & Constable, R. T. (2015). Sex differences in normal age trajectories of functional brain networks. *Human Brain Mapping*, 36(4), 1524–1535. <https://doi.org/10.1002/hbm.22720>
- Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Glasser, M. F. (2013). Resting-state fMRI in the human connectome project. *NeuroImage*, 80, 144–168. <https://doi.org/10.1016/j.neuroimage.2013.05.039>
- Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11), 1565–1567. <https://doi.org/10.1038/nn.4125>
- Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Ugurbil, K., & Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, 17(12), 666–682. <https://doi.org/10.1016/j.tics.2013.09.016>
- Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., & Steven Scholte, H. (2021). The Amsterdam open MRI collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8(1), 85. <https://doi.org/10.1038/s41597-021-00870-6>
- Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *I kommission hos E. Munksgaard*, 5, 1–34.
- Sripada, C., Angstadt, M., Taxali, A., Clark, D. A., Greathouse, T., Rutherford, S., Dickens, J. R., Shedden, K., Gard, A. M., Hyde, L. W., Weigard, A., & Heitzeg, M. (2021). Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Translational Psychiatry*, 11(1), 571. <https://doi.org/10.1038/s41398-021-01704-0>
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, 80, 62–79. <https://doi.org/10.1016/j.neuroimage.2013.05.041>
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., ... WU-Minn HCP Consortium. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4), 2222–2231. <https://doi.org/10.1016/j.neuroimage.2012.02.018>
- Vapnik, V. (1998). *Statistical learning theory* (p. 2). Wiley.
- Varikuti, D. P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K. R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K., Davatzikos, C., & Eickhoff, S. B. (2018). Evaluation of non-negative matrix factorization of grey matter in age prediction. *NeuroImage*, 173, 394–410. <https://doi.org/10.1016/j.neuroimage.2018.03.007>
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180, 68–77. <https://doi.org/10.1016/j.neuroimage.2017.06.061>
- Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1), 1–7.
- Vergara, V. M., Mayer, A. R., Damaraju, E., & Calhoun, V. D. (2017). The effect of preprocessing in dynamic functional network connectivity used to classify mild traumatic brain injury. *Brain and Behavior*, 7(10), e00809.
- Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. *Frontiers in Neuroinformatics*, 11, 17. <https://doi.org/10.3389/fninf.2017.00017>
- Weis, S., Hodgetts, S., & Hausmann, M. (2019). Sex differences and menstrual cycle effects in cognitive and sensory resting state networks. *Brain and Cognition*, 131, 66–73. <https://doi.org/10.1016/j.bandc.2017.09.003>
- Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., & Eickhoff, S. B. (2020). Sex classification by resting state brain connectivity. *Cerebral Cortex*, 30(2), 824–835. <https://doi.org/10.1093/cercor/bhz129>
- Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., ... Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*, 13(1), 13868.
- Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4–15. <https://doi.org/10.1148/radiol.2020192224>
- Yang, F., Wanik, D. W., Cerrai, D., Bhuiyan, M. A. E., & Anagnostou, E. N. (2020). Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration. *Sustainability*, 12(4), 1525.
- Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, 39(11), 4213–4227. <https://doi.org/10.1002/hbm.24241>
- 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. *Human Brain Mapping*, 39(4), 1765–1776. <https://doi.org/10.1002/hbm.23950>
- Zhang, Z., Li, G., Xu, Y., & Tang, X. (2021). Application of artificial intelligence in the MRI classification task of human brain neurological and psychiatric diseases: A scoping review. *Diagnostics (Basel)*, 11(8), 1402. <https://doi.org/10.3390/diagnostics11081402>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K. R., Eickhoff, S. B., & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, 45(6), e26683. <https://doi.org/10.1002/hbm.26683>

5 Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., ... & Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*, 13(1), 13868



OPEN

Accurate sex prediction of cisgender and transgender individuals without brain size bias

Lisa Wiersch^{1,2}, Sami Hamdan^{1,2}, Felix Hoffstaedter^{1,2}, Mikhail Votinov^{3,4}, Ute Habel^{3,4}, Benjamin Clemens^{3,4}, Birgit Derntl^{5,6}, Simon B. Eickhoff^{1,2}, Kaustubh R. Patil^{1,2,7}✉ & Susanne Weis^{1,2,7}✉

The increasing use of machine learning approaches on neuroimaging data comes with the important concern of confounding variables which might lead to biased predictions and in turn spurious conclusions about the relationship between the features and the target. A prominent example is the brain size difference between women and men. This difference in total intracranial volume (TIV) can cause bias when employing machine learning approaches for the investigation of sex differences in brain morphology. A TIV-biased model will not capture qualitative sex differences in brain organization but rather learn to classify an individual's sex based on brain size differences, thus leading to spurious and misleading conclusions, for example when comparing brain morphology between cisgender- and transgender individuals. In this study, TIV bias in sex classification models applied to cis- and transgender individuals was systematically investigated by controlling for TIV either through featurewise confound removal or by matching the training samples for TIV. Our results provide strong evidence that models not biased by TIV can classify the sex of both cis- and transgender individuals with high accuracy, highlighting the importance of appropriate modeling to avoid bias in automated decision making.

Machine Learning (ML) approaches have become increasingly popular in medical imaging, especially for neuroimaging data^{1–3}. Previous studies applying ML approaches to neuroimaging data coming from individuals with mental and neurodegenerative disorders have provided valuable insights into the complex mechanisms underlying psychopathology^{4–6}. The ability of ML models to make predictions about previously unseen individual subjects has expanded the field from population-based analyses to investigation of individualized biomarkers^{5,6}. However, it is important to ensure that predictions are not confounded by variables that are not part of the causal pathway of interest, but are associated with both the features the model was trained on and the target^{6,7}, as results from confounded analyses might potentially lead to inaccurate and spurious conclusions^{8,9}. Using brain size bias in sex classification as an example, the present study examines which confound removal strategy is most suitable to achieve high classification accuracy while effectively removing brain size bias^{8–10}.

ML approaches have been successfully applied to the study of sex differences in the brain by training a classifier to predict sex based on features derived from structural brain imaging data, e.g. regional grey matter volume (GMV). Such a sex classifier is expected to capture multivariate brain organizational patterns that differ between the sexes. High classification accuracies on out-of-sample data^{11,12} are then taken as evidence for qualitative sex differences in the brain^{13,14}. So far, studies using sex classification approaches based on structural brain imaging data achieved classification accuracies ranging from 82 up to 94%^{11,12,15–17}. However, a sex classifier biased by brain size (measured as total intracranial volume, TIV^{18,19}) will result in predictions that are driven by TIV differences rather than actual sex differences in brain structure^{9,10,20}. As a result, a TIV-biased model will classify

¹Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ²Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany. ³Department of Psychiatry, Psychotherapy and Psychosomatics, Faculty of Medicine, RWTH Aachen University, Aachen, Germany. ⁴Institute of Neuroscience and Medicine (INM-10: Brain Structure-Function Relationships), Research Centre Jülich, Jülich, Germany. ⁵Department of Psychiatry and Psychotherapy, Tübingen Center for Mental Health, University of Tübingen, Tübingen, Germany. ⁶LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany. ⁷These authors contributed equally: Kaustubh R. Patil and Susanne Weis. ✉email: k.patil@fz-juelich.de; s.weis@fz-juelich.de

individuals with higher TIV as males and individuals with lower TIV as females, while making more mistakes for individuals with intermediate TIV.

The use of such a TIV-biased sex classifier is particularly problematic when analyzing data of individuals for whom local and global brain structural alterations have been reported, such as those with "gender incongruence," where a person's sex and gender identity differ²¹. In the present paper, following the linguistic guidelines provided by the Professional Association of Transgender Health²², the term "sex" is used to refer to the sex that a person was assigned at birth based on their anatomical sexual characteristics, whereas the term "gender (identity)" is used to denote the subjective identification of an individual as female, male, or one of the other gender identities which might be also fluid or non-binary. While the coherence of sex and gender is termed cisgender for cisgender men and women (CM, CW), gender incongruent individuals are denoted as transgender men and women (TM, TW,²¹).

To date, it is not yet fully understood if and to which extent local and global brain organization of transgender individuals is driven by factors matching their gender identity on top of those matching their sex. So far, studies contrasting groups of cisgender and transgender individuals reported regional GMV differences in the putamen²³, insula¹⁶ as well as in surface areas, cortical and subcortical brain volumes²⁴. Additionally, transgender individuals undergoing cross-sex hormone treatment (CHT) were reported to show structural alterations in the hypothalamus and the third ventricle²⁵. Thus, there is some evidence indicating that transgender individuals display local brain volume differences^{24,26–28}. Extending the results of group studies contrasting cisgender and transgender individuals, sex classification approaches—building a classifier on cisgender individuals' data and then applying it to transgender individuals—have reported reduced sex classification accuracies for transgender compared to cisgender samples (76.2% vs. 82.6%¹⁷; 61.5% vs. 93.2–94.9%¹⁶). Higher rates of misclassification of sex in transgender as opposed to cisgender individuals have been taken to indicate that transgender brains might differ from those typical for their sex, implying an interaction between sex and gender at the neuroanatomical level^{16,17,29}. However, before such conclusions can be drawn, biases that can influence a sex classifier must be taken into account, particularly those related to TIV^{18,19}. It is crucial to be aware of the impact of local and global structural brain alterations that can lead to increases or decreases of TIV resulting in the TIV of transgender individuals falling between TIV of cisgender women and men²⁵. Consequently, the predictions of a TIV-biased classifier might erroneously be interpreted as evidence for transgender brain organization to align with gender identity as has been reported before^{16,29}.

Here, we investigate the impact of TIV bias by examining two approaches to control for confounding effects of TIV¹⁰ in sex classification to evaluate which approach is most suited to account for TIV bias in the present sex classification analysis. We compare two statistically different approaches of controlling for TIV bias in comparison to a baseline model that does not account for the influence of TIV. For the first approach, we built debiased models through featurewise confound control by removing confounding effects of TIV during training (Fig. 1, ^{20,30}). In the second approach, we trained models on a stratified sample where women and men were matched for TIV. Model performance and TIV bias were assessed on hold-out samples of cisgender individuals to compare performance of the biased to the debiased models. We hypothesized that a TIV-biased model should achieve high performance but also exhibit a biased output pattern. In contrast, a model not biased by TIV will likely exhibit a drop in classification accuracy. However, importantly, misclassifications of such a model should be largely independent of TIV. In the final step, the debiased models were applied to application samples comprising both cisgender and transgender individuals to examine whether models without a TIV bias provide any evidence for an interaction of sex and gender influences on structural brain organization, as previously suggested¹⁷.

Results

Classifiers employing Support Vector Machine (SVM) models with radial basis function kernel (rbf) were trained on whole-brain voxelwise GMV data of two large, non-overlapping cisgender samples to classify sex assigned at birth. In the first sample, women and men were matched for age (AM sample) to create a sample with a natural occurring TIV-distribution (Fig. S1 and Table S1). As a baseline, we trained the first model on this sample without any control for TIV bias (AM model), following the methodology of a previous study¹⁶. We then compared the baseline model to other models, which integrated two different approaches for confound control in order to assess which approach successfully removes TIV bias while accurately classifying sex. For the first approach, a ML model was also trained on the AM sample, but additionally controlled for TIV bias by featurewise confound removal (AM+cr model), while the third model comprised stratification for TIV by training the model on a sample of women and men who were matched for both age and TIV (ATM; see Fig. S1 and Table S1 for demographic details and TIV distribution of the samples). While the third model was trained on the ATM sample without additional TIV-control (ATM model) to evaluate stratification in itself, the fourth model employed a combination of both approaches to assess whether the addition of featurewise confound removal might further improve results (AM+cr model, Fig. 1). Subsequently, all models were calibrated to ensure that the prediction probabilities of the models match the respective class label (Figs. S2 and S3, Supplementary Results, <https://scikit-learn.org/stable/modules/calibration.html#calibration>). To evaluate model performance on hold-out data, each sample (AM and ATM) was split into a training sample (80%) and a hold-out sample (20%). As the two approaches—featurewise confound removal and stratification by matching—might exhibit differences in model performance since they are based on different statistical processes⁸, all four models were evaluated on both AM and ATM hold-out samples. This allowed for a thorough understanding of model behavior and evaluation of whether both approaches successfully remove TIV bias. Assessing model performance on the first sample (AM hold-out sample), which exhibits a naturally occurring TIV-distribution among women and men, enables a realistic evaluation of the model's effectiveness in broader populations beyond those included in the present study. In turn, the ATM hold-out sample enables a more in-depth evaluation of the model performance, as it displays

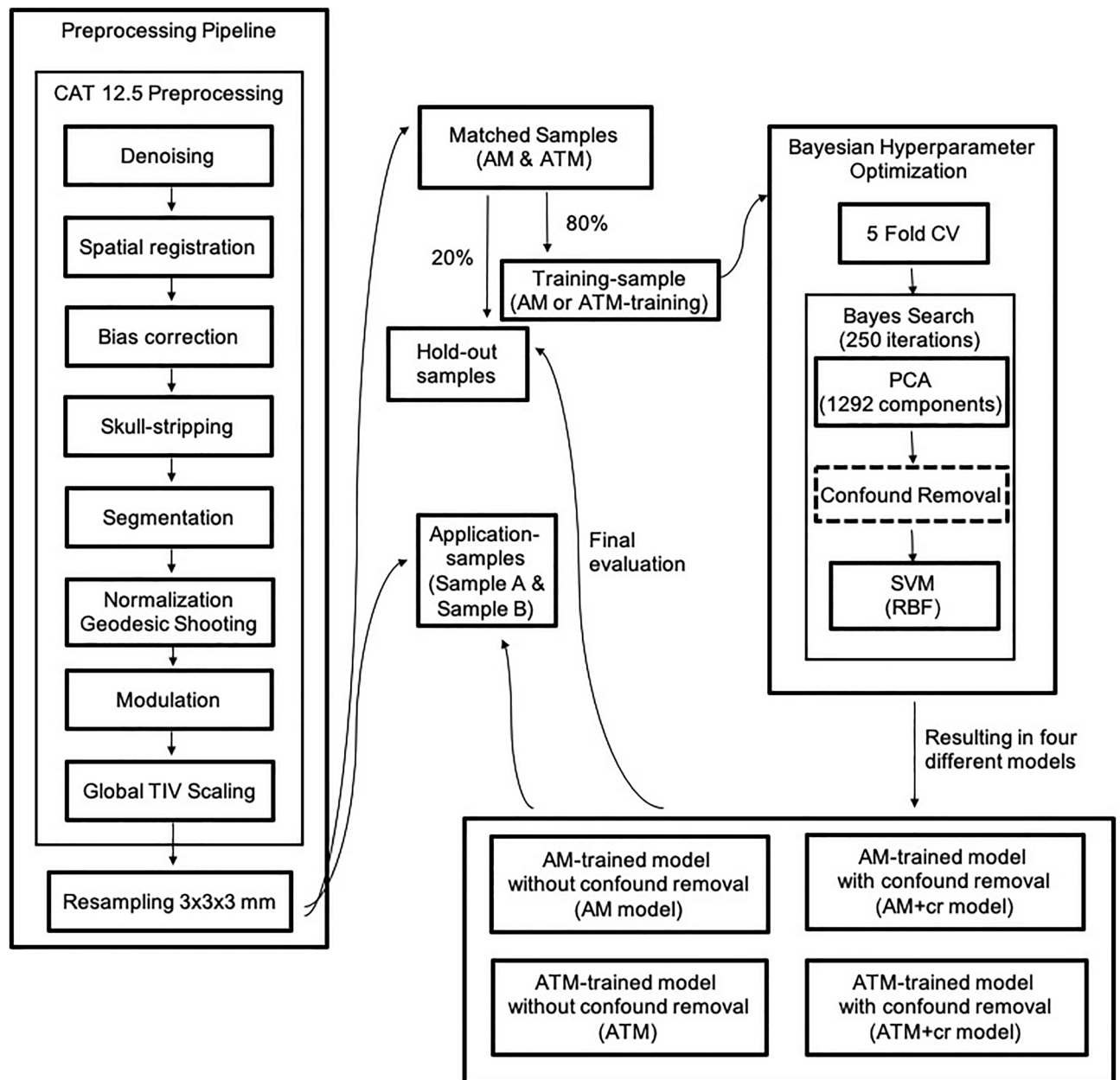


Figure 1. Analysis pipeline. Workflow of the sex classification analysis.

no significant difference in TIV between women and men. Consequently, an accurate model performance for the ATM hold-out sample indicates a non-TIV-biased model behavior as the model classifies a person's sex based on other features than TIV, providing a “confound-free accuracy”³¹. Additionally, the models were tested on two independent application samples comprising transgender and cisgender individuals (sample A, sample B, see Fig. S1 and Table S1 for demographic details and TIV distribution of the samples).

Evidence for TIV bias in the AM model. The application of the AM model to the AM hold-out sample resulted in a high classification accuracy of 96.89% (Table 1, Table S2, and Fig. 2). Accordingly, the assigned probability of being classified as male (prediction probability) was higher for men than for women (Fig. 3a). The comparison of TIV distributions revealed that men who were classified congruently with their sex as male had a significantly higher TIV than incongruently classified men (Fig. 3b). Similarly, women classified incongruently with their sex as male on average had a higher TIV than congruently classified women, even though this difference was not significant (details in Table 2).

When applied to the ATM hold-out sample, the AM model resulted in a much lower classification accuracy of 79.19% (Tables 1 and S2), presumably as the model could not rely on TIV for classifying in the ATM sample. Still, we observed a similar pattern as above, with men having a higher prediction probability than women (Fig. 3c), significantly higher TIV in sex congruently as opposed to incongruently classified men, and significantly lower

	AM model	AM+cr model	ATM model	AM+cr model
Model performance for the AM hold-out sample				
Recall:	0.9503	0.7329	0.8820	0.8571
Specificity:	0.9876	0.5031	0.8509	0.8571
F1:	0.9684	0.6574	0.8685	0.8571
BA*:	0.9689	0.6180	0.8665	0.8571
Model performance for the ATM hold-out sample				
Recall:	0.7453	0.8323	0.9255	0.9193
Specificity:	0.8385	0.6273	0.9255	0.9317
F1:	0.7818	0.7549	0.9255	0.9250
BA*:	0.7919	0.7298	0.9255	0.9255
Model performance for sample A				
Recall:	0.9474	0.7895	1	0.9474
Specificity:	0.8276	0.7241	0.8276	0.8448
F1:	0.8926	0.7627	0.9194	0.9000
BA*:	0.8875	0.7568	0.9138	0.8961
Model performance for sample B				
Recall:	0.8889	0.8333	0.9722	0.8889
Specificity:	0.9608	0.5882	0.9020	0.9020
F1:	0.9143	0.6897	0.9211	0.8767
BA*:	0.9248	0.7108	0.9371	0.8954

Table 1. Model performance of all models applied to the hold-out and application samples (* Balanced Accuracy). Model performance of all models applied to the hold-out and application samples.

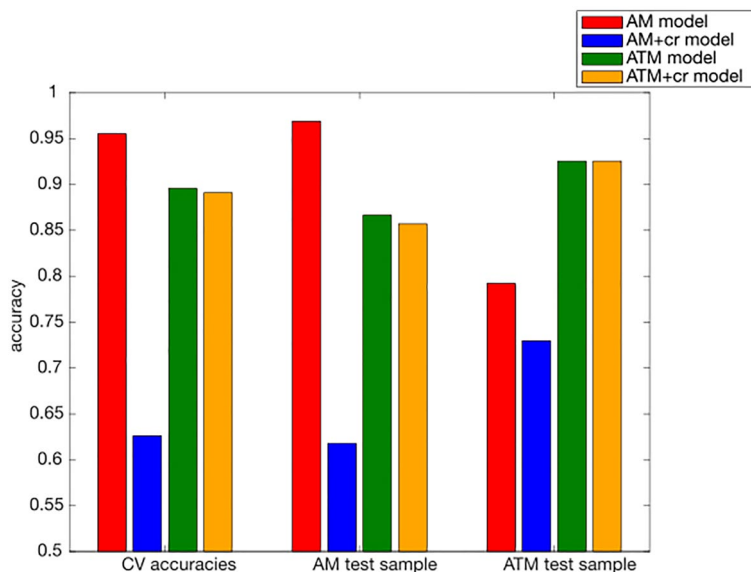


Figure 2. Sex classification accuracy. Accuracy values of the four different models for the cross validation (CV)-folds and applied to the AM and ATM hold-out sample.

TIV in sex congruently as opposed to incongruently classified women (Fig. 3d and Table 2). Altogether, across both hold-out samples, this model tended to classify subjects with higher TIV as male and those with lower TIV as female, clearly indicating a brain size bias inherent in this model.

Reducing TIV bias by confound removal. Featurewise control for TIV in the AM+cr model resulted in decreased classification accuracies both for the AM (61.80%) and the ATM (72.98%; further details in Fig. 2, Table 1 and Table S2) hold-out samples. In comparison to the AM model with no TIV control (Fig. 3a) prediction probability displayed a much larger overlap between women and men (Fig. 3e, g). Further evaluation did not reveal any evidence for a TIV bias—i.e. neither did sex congruently classified men show higher TIV than

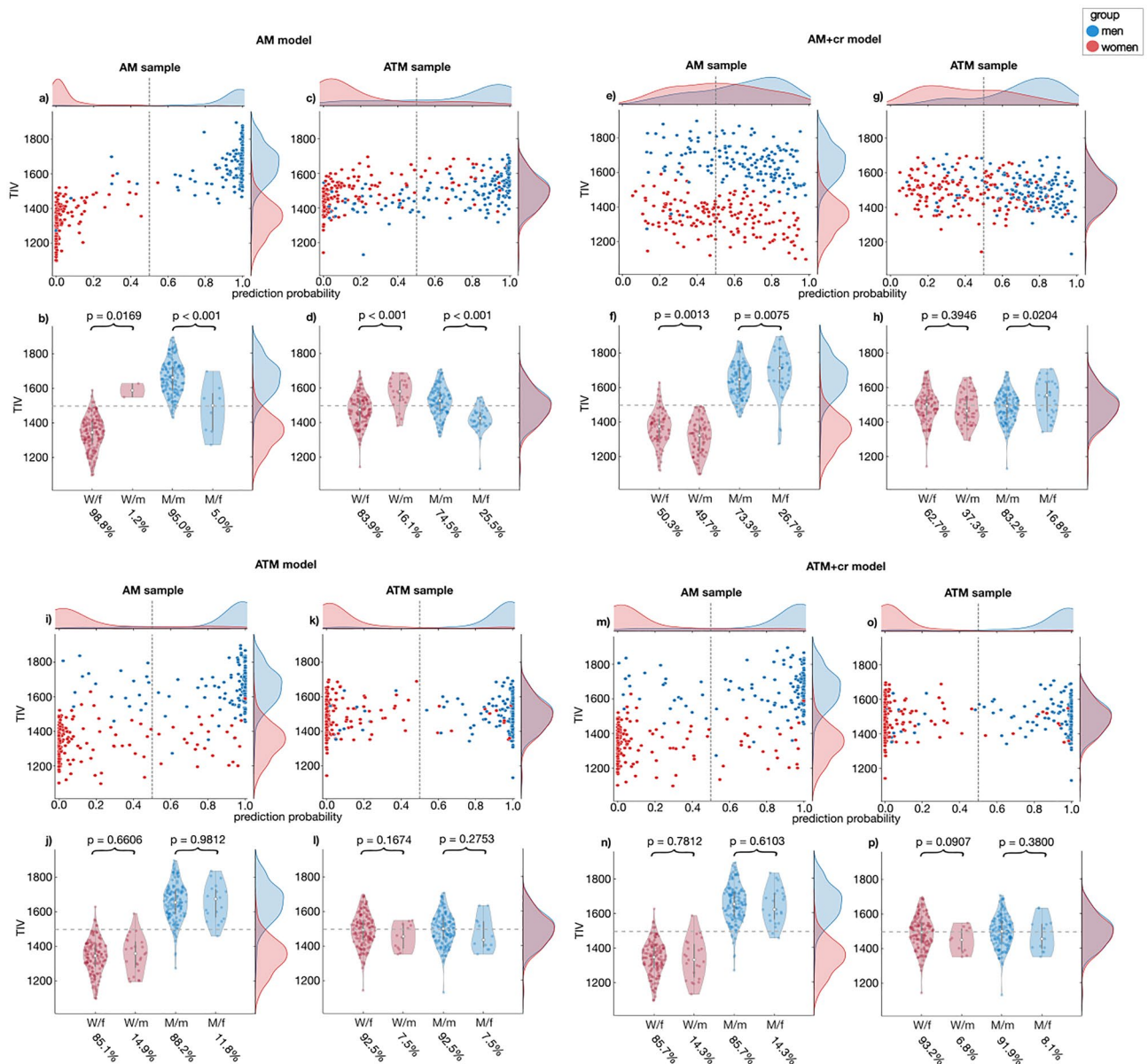


Figure 3. Association between prediction probability and TIV. Prediction probability (a, c, e, g, i, k, m, o) and TIV distribution (b, d, f, h, j, l, n, p) of sex congruently and incongruently classified women (red) and men (blue) of all four models applied to the AM and ATM hold-out sample. (W/f: women classified as female; W/m: women classified as male; M/m: men classified as male; M/f: men classified as female).

	TIV women classified as female versus classified as male	TIV men classified as male versus classified as female
AM hold-out sample		
AM model	$T = 12,722, z = -2.3885, p = 0.0169, \eta^2 = 0.0354$	$T = 12,829, z = 3.3879, p < 0.001, \eta^2 = 0.0713$
AM+cr model	$T = 7514, z = 3.2204, p = 0.0013, \eta^2 = 0.0644$	$T = 8858, z = -2.6727, p = 0.0075, \eta^2 = 0.0444$
ATM model	$T = 11,004, z = -0.4390, p = 0.6606, \eta^2 = 0.0012$	$T = 11,507, z = 0.0236, p = 0.9812, \eta^2 < 0.001$
AM+cr model	$T = 11,236, z = 0.2778, p = 0.7812, \eta^2 < 0.001$	$T = 11,284, z = 0.5097, p = 0.6103, \eta^2 = 0.0016$
ATM hold-out sample		
AM model	$T = 9908, z = -4.7156, p < 0.001, \eta^2 = 0.1381$	$T = 11,325, z = 6.2257, p < 0.001, \eta^2 = 0.2407$
AM+cr model	$T = 8425, z = 0.8513, p = 0.3946, \eta^2 = 0.0045$	$T = 10,341, z = -2.3190, p = 0.0204, \eta^2 = 0.0334$
ATM model	$T = 12,284, z = 1.3806, p = 0.1674, \eta^2 = 0.0118$	$T = 12,239, z = 1.0910, p = 0.2753, \eta^2 = 0.0074$
AM+cr model	$T = 12,403, z = 1.6918, p = 0.0907, \eta^2 = 0.0178$	$T = 12,130, z = 0.8780, p = 0.3800, \eta^2 = 0.0048$

Table 2. Wilcoxon rank sum tests of the hold-out samples. Comparison of individuals classified as female versus male (Wilcoxon rank sum tests) for the AM and ATM sample.

incongruently classified men nor did sex congruently classified women show lower TIV than incongruently classified women in both the AM (Fig. 3f) and the ATM (Fig. 3h and Table 2) hold-out samples.

Reducing bias by matching the training sample for TIV. The application of the two models built using TIV matched data with and without featurewise TIV control (ATM and ATM+cr model, respectively) to the AM hold-out sample resulted in similarly high classification accuracy (86.65% for ATM, 85.71% for ATM+cr model, details in Tables 1 and S2), performing between accuracies achieved by the AM and the AM+cr model. Thus, for the ATM models, additional featurewise TIV control did not result in decreased model performance. This is further reflected in similar prediction probability distributions (Fig. 3i, m), which were higher for men than for women. Likewise, the TIV of sex congruently and incongruently classified individuals did not differ significantly from each other both for women and for men (Fig. 3j, n and Table 2). Application of these models to the ATM hold-out sample (details in Tables 1 and S2), displayed better performance (92.55%) than for the AM hold-out sample. Furthermore, prediction probability distributions showed a comparable (Fig. 3k, o) but more pronounced pattern for the ATM hold-out sample. Again, when testing on the ATM hold-out sample, there was no difference between TIV of sex congruently and incongruently classified individuals both for the model without (Fig. 3l and Table 2) and with additional confound removal (Fig. 3p and Table 2).

Overall, the AM model achieved highest classification accuracy, but evaluation of the model output identified clear evidence for a TIV bias of the model. Reducing TIV-related variance by featurewise confound removal in the AM+cr model resulted in a less biased model, which also displayed a pronounced decrease in model performance, especially for the AM hold-out sample. Both models trained on the TIV balanced sample (ATM, ATM+cr model) did not show evidence of a TIV bias while still retaining high classification performance and appropriate calibration curves (Figs. S2 and S3), indicating that—at least for the present classification problem—training on a matched sample is more appropriate than featurewise confound removal. Thus, in the following, we will focus on comparing the performance of the biased AM model and the nonbiased ATM model on cisgender and transgender individuals in the application samples (sample A, sample B). Results for the AM+cr and ATM+cr models are provided in the Supplementary Results and Fig. S4.

Biased performance of the AM model for cisgender and transgender individuals. The application of the TIV-biased AM model resulted in an overall high performance of 88.70% for sample A, with an accuracy of 81.63% for cisgender and 93.43% for transgender individuals (detailed measures in Tables 1 and S3). Likewise, for sample B, the model achieved high overall accuracy of 93.10% (Tables 1 and S3) with an accuracy of 90.24% for cisgender individuals and 95.65% for transgender individuals. Matching the high accuracies, the prediction probability showed a sex congruent pattern with higher prediction probabilities for CM and TW (assigned male at birth) than for CW and TM (assigned female at birth) in both sample A (Fig. 4a, c) and sample B (Fig. 4e, g). A comparison of probability distributions of cis- and transgender individuals with the same sex revealed a trend for higher prediction probability for CW than for TM in sample A ($t = 1.98$, $p = 0.0527$, Cohen's $d = 0.53$), which was significant in sample B ($t = 3.58$, $p < 0.001$, Cohen's $d = 1.01$), matching the TIV-distributions showing higher TIV for CW than TM (Fig. S1).

The comparison of prediction probabilities for CM versus TW was not significant in both samples (Sample A: $t = -0.55$, $p = 0.5820$, Cohen's $d = -0.15$; Sample B: $t = 1.07$, $p = 0.2922$, Cohen's $d = 0.36$), while the effect size indicated a trend of lower prediction probability for TW than CM. While TIV-distributions for sex congruently and incongruently classified individuals did not differ significantly (Table 3), sex congruently classified CW and TM had a lower TIV than those classified in a sex incongruent manner. Sex congruently classified CM and TW had a higher TIV than those classified sex incongruently (Fig. 4b, d, f, h), indicating a similar bias of this model for both cisgender and transgender individuals.

Nonbiased ATM model: similar performances for cisgender and transgender individuals. The application of the ATM model to sample A displayed a high overall sex classification accuracy of 91.30% (91.84% for cisgender and 90.01% for transgender individuals). This model also performed accurately on sample B with an overall accuracy of 93.10% (92.68% for cisgender and 93.48% for transgender individuals, details in Table 1 and S3). In both samples, the ATM model yielded sex congruent prediction probabilities for all four groups (Fig. 4i, k, m, o). As opposed to the biased model, here, TM showed a trend of higher prediction probability than CW in Sample B (CW vs TM: $t = -1.27$, $p = 0.2093$, Cohen's $d = -0.36$; Sample A: $t = 0.47$, $p = 0.6425$, Cohen's $d = -0.12$). This gender congruent trend was not observed for TW (CM vs. TW: Sample A: $t = 0.31$, $p = 0.7577$, Cohen's $d = 0.08$; Sample B: $t = -2.02$, $p = 0.0510$, Cohen's $d = -0.68$). The comparison of TIV distributions between sex congruently and incongruently classified individuals (Fig. 4j, l, n, p) did not reveal any significant differences (Table 3), neither for cisgender nor for transgender individuals, thus displaying no evidence for a TIV bias of this model.

Discussion

In this work, we systematically compared two confound removal approaches, featurewise confound removal and sample stratification, with the aim to train accurate sex classification models without a TIV bias. In order to directly compare our findings to those of a previous study, we implemented a ML pipeline that has demonstrated high levels of sex classification accuracy¹⁶. This pipeline consisted of principal component analysis (PCA) for dimensionality reduction, followed by an SVM model with rbf kernel for learning, but did not report any consideration of the confounding effects of TIV.

Consistent with previous results, the baseline AM model which does not consider confounding effects of TIV achieved near-perfect classification accuracy on the AM hold-out sample by accurately classifying men with

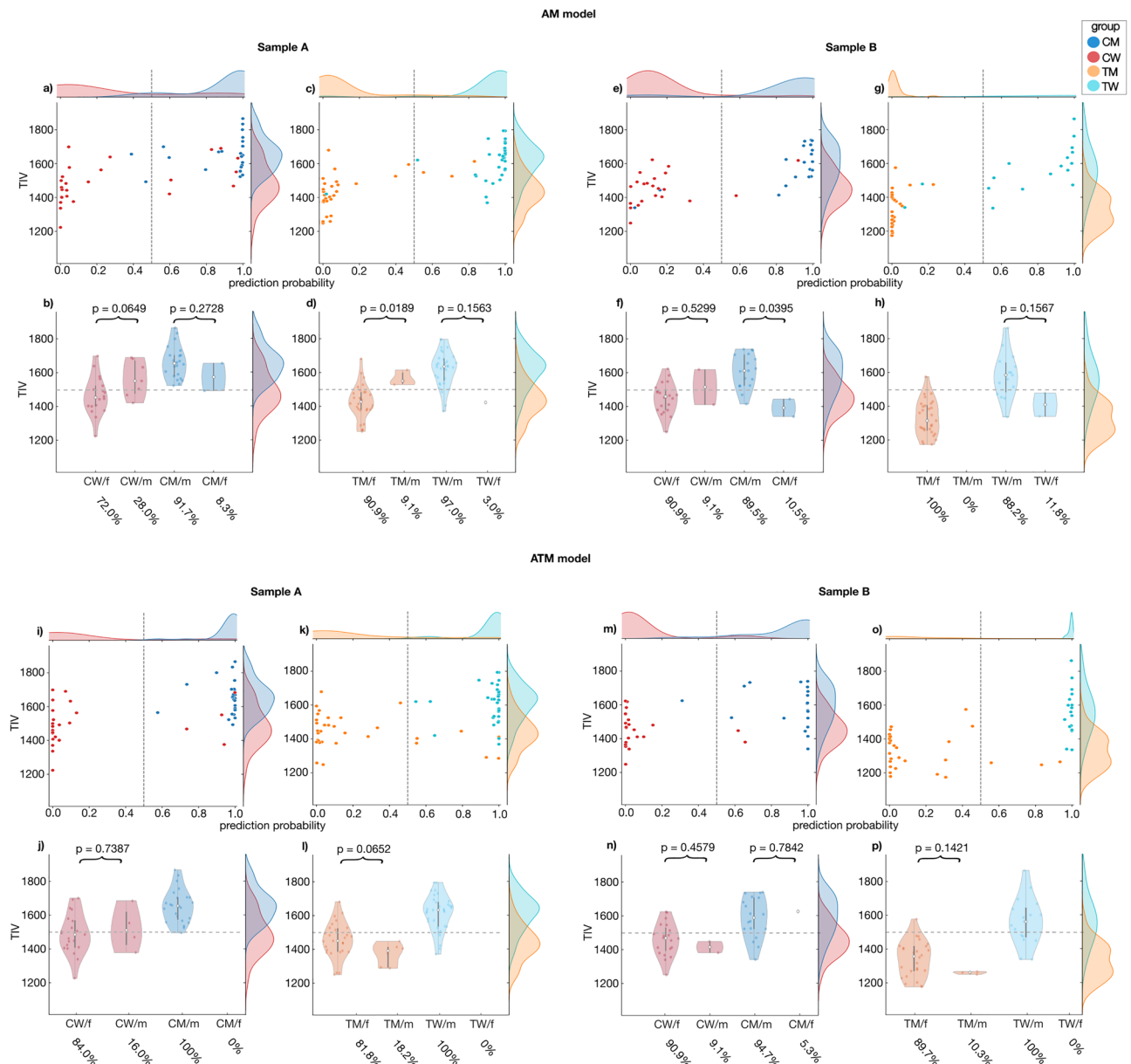


Figure 4. Association between prediction probability and TIV for the AM and ATM models in the two application samples. The upper row (a–h) shows the prediction probability (a, c, e, g) and TIV distribution (b, d, f, h) of sex congruently and incongruently classified CM, CW, TM and TW in the AM model in sample A and B. The bottom row (i–p) shows the prediction probability (i, k, m, o) and TIV distribution (j, l, n, p) of sex congruently and incongruently classified CM, CW, TM and TW in the ATM model in sample A and B. (CW/f: CW classified as female; CW/m: CW classified as male; CM/m: CM classified as male; CM/f: CM classified as female; TM/f: TM classified as female; TM/m: TM classified as male; TW/m: TW classified as male; TW/f: TW classified as female).

high TIV as male and women with low TIV as female^{11,12,16,17}, but relied on TIV as a proxy for sex, indicating a pronounced TIV bias (Fig. 3b). The TIV bias was even more pronounced when the model was applied on the ATM hold-out sample presumably as the AM model was more likely to make mistakes for men with relatively lower TIV and women with relatively higher TIV. The pronounced TIV bias observed here is especially interesting, since the GMV data had already been scaled for TIV during preprocessing. Thus, our results align with previous claims that while the absolute amount of tissue is corrected for individual TIV, such scaling does not fully remove TIV-related variance^(32, <http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf>).

For the AM+cr model, where a featurewise removal of TIV was performed on the AM data, the misclassifications of both women and men were not systematically related to TIV differences, indicating that this model was not biased by TIV. This suggests that the AM+cr model based its classifications on different information than the AM model did. Our results match the findings of previous studies^{20,30,33,34}, reporting a decrease in accuracy for sex classification models controlling for TIV in contrast to TIV-biased models. This decrease is likely related to the

a)	TIV CW classified as female versus classified as male	TIV CM classified as male versus classified as female
AM model	$T = 203, z = -1.8459, p = 0.0649, \eta^2 = 0.1363$	$T = 286, z = 1.0967, p = 0.2728, \eta^2 = 0.0501$
AM+cr model	$T = 249, z = 0.8776, p = 0.3802, \eta^2 = 0.0308$	$T = 236, z = -1.0457, p = 0.2957, \eta^2 = 0.0456$
ATM model	$T = 268, z = -0.3336, p = 0.7387, \eta^2 = 0.0045$	<i>no CM classified as female</i>
AM+cr model	$T = 268, z = -0.3336, p = 0.7387, \eta^2 = 0.0045$	$T = 294, z = 0.8668, p = 0.3861, \eta^2 = 0.0313$
	TIV TM classified as female versus classified as male	TIV TW classified as male versus classified as female
AM model	$T = 472, z = -2.3483, p = 0.0189, \eta^2 = 0.1671$	$T = 558, z = 1.4178, p = 0.1563, \eta^2 = 0.0609$
AM+cr model	$T = 477, z = 2.7689, p = 0.0056, \eta^2 = 0.2323$	$T = 442, z = 0.6931, p = 0.4882, \eta^2 = 0.0146$
ATM model	$T = 499, z = 1.8437, p = 0.0652, \eta^2 = 0.1030$	<i>no TW classified as female</i>
AM+cr model	$T = 506, z = 1.4812, p = 0.1386, \eta^2 = 0.0665$	$T = 532, z = 0.3395, p = 0.7342, \eta^2 = 0.0035$
b)	TIV CW classified as female versus classified as male	TIV CM classified as male versus classified as female
AM model	$T = 224, z = -0.6281, p = 0.5299, \eta^2 = 0.0179$	$T = 186, z = 2.0591, p = 0.0395, \eta^2 = 0.2231$
AM+cr model	$T = 199, z = 1.8328, p = 0.0668, \eta^2 = 0.1527$	$T = 159, z = -1.3948, p = 0.1631, \eta^2 = 0.1024$
ATM model	$T = 237, z = 0.7424, p = 0.4579, \eta^2 = 0.0250$	$T = 178, z = -0.2739, p = 0.7842, \eta^2 = 0.0039$
AM+cr model	$T = 237, z = 0.7424, p = 0.4579, \eta^2 = 0.0250$	$T = 138, z = -1.1500, p = 0.2501, \eta^2 = 0.0696$
	TIV TM classified as female versus classified as male	TIV TW classified as male versus classified as female
AM model	<i>no TM classified as male</i>	$T = 145, z = 1.4162, p = 0.1567, \eta^2 = 0.1180$
AM+cr model	$T = 289, z = 2.7714, p = 0.0056, \eta^2 = 0.2648$	$T = 115, z = -0.1698, p = 0.8651, \eta^2 = 0.0017$
ATM model	$T = 411, z = 1.4680, p = 0.1421, \eta^2 = 0.0743$	<i>no TW classified as female</i>
AM+cr model	$T = 411, z = 1.4680, p = 0.1421, \eta^2 = 0.0743$	<i>no TW classified as female</i>

Table 3. Wilcoxon rank sum tests of the application samples. Comparison of individuals classified as female versus male (Wilcoxon rank sum tests) for application sample A (a) and sample (b).

removal of TIV-related variance during featurewise confound removal, which might have decreased the overall amount of information available for the AM+cr model in contrast to the AM model^{20,30,33,34}. This observation is in line with the results of a previous study suggesting that TIV alone contains enough information to classify sex at a similar level of accuracy as TIV-uncorrected GMV³⁴. Considering that features in the AM sample can be assumed to contain more TIV-related variance than the ATM sample presumably explains why the drop in accuracy between the AM and the ATM+cr is less pronounced for the ATM hold-out sample than for the AM sample. Altogether, featurewise confound removal reduced TIV bias at the cost of classification accuracy. While a lack of bias in a model is desirable, so is high accuracy, suggesting that featurewise confound removal might not be the ideal approach to reduce TIV bias in structural sex classification.

In contrast to the models trained on the AM sample, both ATM trained models resulted in high and unbiased model performance for the AM as well as the ATM hold-out samples. The slightly higher accuracy for the ATM hold-out sample is likely due to the ATM hold-out sample better matching the characteristics of the ATM training sample, in particular with respect to TIV distribution, which is highly related to the target variable sex³⁰. The better performance of the ATM and ATM+cr model on the ATM hold-out samples also supports the relevance of stratifying training and hold-out samples with respect to relevant variables that may interact with the target^{35,36}.

The comparison of TIV of sex congruently and incongruently classified women and men did not indicate a TIV bias, which is in line with a study proposing beforehand matching to be a more efficient approach than feature-wise confound removal in the statistical analysis⁹. However, another study argued against the matching of data, arguing that matching for specific characteristics creates a sample that is not representative of the whole population²⁰. While we agree that the ATM sample does not strictly represent the TIV distribution of the population by rather comprising men with relatively low and women with relatively high TIV, the ensuing models achieved high classification accuracies, even when applied to the AM hold-out sample which reflects the natural TIV distribution. This indicates that the models themselves are not biased by training sample characteristics, especially the restricted TIV range. In fact, the models appear to correctly capture sex differences in a generalizable manner as exemplified by their performance on the two hold-out samples. However, we would like to emphasize that both confound removal approaches employed in the present study rely on different statistical operations which are anticipated to result in different outcomes and model performances⁸. Thus, high model performance of one approach does not imply the other one to behave in a similar manner. For this reason, testing which approach is most suited for an individual ML-problem is crucial. The present results demonstrated that matching women and men for TIV in the training sample provides an appropriate approach for creating unbiased and accurate sex classification models.

In contrast to previous studies^{16,17}, we observed similarly high classification accuracies for cis- and transgender individuals regardless of whether the models were debiased or not. This discrepancy may partly be explained by the fact that TIV of the transgender individuals in the present samples matched TIV of cisgender subjects of the same sex rather than aligning with gender identity (Fig. S1). Thus, even a biased classifier could accurately classify transgender individuals. However, in samples where the TIV values for transgender individuals indeed fall in-between those of cisgender men and women, as reported previously²⁵ TIV-biased models would misclassify transgender individuals in accordance with their gender identity, which could explain prior findings¹⁶. Future

studies should apply TIV-debiased models to additional datasets to help disentangle the complex interaction of sex, gender and the brain. It would be particularly interesting to apply our debiased models, which are available to other researchers (https://github.com/juaml/sex_prediction_vbm) to those datasets for which a reduction of sex classification accuracy for transgender participants has previously been reported^{16,29}. Another explanation for the discrepancy between present and previous results^{16,29}, might be that our classifiers learnt fundamentally different models, e.g. employing different feature weights than those in previous studies, which in turn might be caused by differences in characteristics of the training samples and in turn different parameters learnt during model optimization. Beside the differences due to different training samples, other factors affecting ML models and respective results might relate to differences in age-distribution. Here, we not only balanced for sex but also employed an exact matching of men and women with regards to age which might have reduced variance in comparison to the training-samples of other studies^{16,29} leading to differences in the fundamental model and results. In addition to age in the training sample, the age distribution of the application sample could also play a role, due to age-related GMV decline. Thus, older TW could be misclassified due to age-related GMV changes.

The present models were trained on a diverse collection of samples, ensuring a heterogeneity in several variables, such as age, scanning characteristics, and nationality. Likewise, as application samples we used two completely independent datasets comprising TW and TM. To our knowledge, previous studies have focused on test samples only comprising TW when applying a sex classifier trained on structural data of cisgender individuals to transgender individuals^{16,29}, limiting conclusions to TW rather than transgender individuals in general. Notably, one study employing data of both TW and TM did not report significantly lower classification accuracy for transgender data¹⁷, which is in line with the present results. While we did not observe decreased sex classification accuracy for transgender individuals, this cannot be taken as a proof of absence of such structural brain differences, which might be revealed by the investigation of different sets of brain features or different analysis approaches.

Future studies can benefit by incorporating confound control approaches within interpretable ML pipelines that can provide insight into how many and which brain regions are most relevant for sex differences. Those insights can shed further light on which features are more common in men, women or both, thereby carrying implications for hypotheses as the mosaic of the human brain³⁷, which exceeds the scope of the current study design. Methodologically sound studies, including both sex and gender aspects, are needed to improve our understanding of sex and gender-related differences in behavior and prevalence rates of mental disorders to advance development of sex-specific treatments^{38,39}. Viewing patients through the lens of sex and gender is an essential step towards personalized care and individualized medicine^{6,40}. Therefore, to achieve the ultimate goal of neuroimaging-based precision medicine, the present study takes a first step towards exploring appropriate confound removal in ML-based sex classification⁴¹. Although each ML analysis must consider confounds specific to the research question at hand, TIV is an important confound to consider in neuroimaging data in general, as also shown by others^{9,18,33,34,42}. In addition to its application in sex classification analyses, as demonstrated here, appropriate confound control should also be considered for other ML applications. We, therefore, recommend that researchers should investigate which confound removal method is appropriate for their ML analysis.

Conclusion

Our findings demonstrate that stratification via TIV-matching effectively eliminates TIV bias while achieving high levels of classification accuracy in a sex classification analysis using structural brain imaging features. Contrary to previous results¹⁶, our sex classification model demonstrated comparable levels of classification accuracy for both cisgender and transgender individuals. Our study emphasizes the importance of removing TIV bias appropriately in sex classification tasks to prevent incorrect interpretations. In general, confounding is a common issue in many ML-based modeling tasks, albeit with varying confounds and levels of confounding effects. Therefore, future studies utilizing ML approaches on brain imaging data should diligently examine for biases and implement appropriate confound control measures.

Materials and methods

Data. *Data pool for model training and evaluation.* To ensure a heterogeneous sample for training the classifiers, we combined data from 10 large cohorts into one data pool of structural magnetic resonance imaging (MRI) images from subjects differing in nationality, imaging parameters and age range. Supplementary Table S4 gives further details on the composition of the data pool, and details of the MRI data acquisition parameters can be found in the Supplementary Material. We only included subjects aged between 18 and 65 years with no indication of any psychiatric disorder, resulting in a total N of 5557 subjects. It is important to note, that the majority of large datasets, which have been employed for sex classification studies so far, likely report sex based on “presented sex”, i.e. the name and outer appearance of participants or on self-reported sex without explicitly collecting information on gender identity. We assume that among subjects not describing themselves as transgender, self-reported gender identity is equivalent to sex assigned at birth, while acknowledging that this match may neither be perfect nor binary.

Sixteen subjects whose TIV values differed more than three standard deviations from the mean TIV of the data pool were excluded as outliers. Then, two non-overlapping samples were extracted from the data pool. In the first sample (AM), women and men were matched for age to control for age-related GMV decline^{43–46}. In the second sample (ATM), women and men were additionally matched for TIV. Possible differences between samples and sites in scanning acquisition were controlled for by including similar numbers of subjects from the different samples in the AM and ATM-sample respectively. Both the AM and ATM sample comprised 276 subjects from 1000 Brains, 146 subjects from Cam-CAN, 168 subjects from CoRR, 50 subjects from DLBS, 94 subjects from

eNKI, 192 subjects from GOBS, 396 subjects from HCP, 96 subjects from IXI, 76 subjects from OASIS3, and 120 subjects from PNC. Each sample was split into a training (80%) and a hold-out sample (20%).

Age-matched (AM) sample. For the AM sample ($N=1614$, 807 women), women and men were matched for age within each site (including multiple sites within one sample) by including a male counterpart from the same site whose age differed by no more than one year for each female subject. The age range in this sample was 18–65 years ($M=37.96$, $SD=15.28$). Further detailed information can be found in Table S1, and a plot of the TIV distribution of women and men is displayed in Fig. S1. There was no significant difference in age between women and men ($t=0.01$, $p=0.99$); however, the sexes differed significantly with respect to TIV ($t=-61.06$, $p<0.001$). Splitting the sample into training (80%) and hold-out samples (20%) resulted in 1292 subjects (646 women) for training and 322 subjects (161 women) for testing. The training and hold-out samples did not differ with respect to age ($t=0.98$, $p=0.33$) or TIV ($t=-0.11$, $p=0.91$). The age difference between sexes remained nonsignificant within both the training ($t=-0.00$, $p=0.99$) and the hold-out sample ($t=0.03$, $p=0.97$), whereas the TIV difference was significant for both samples (training: $t=-54.79$, $p<0.001$, hold-out: $t=-26.90$, $p<0.001$).

Age-TIV-matched (ATM) sample. For the ATM sample ($N=1614$, 807 women), women and men were matched for age and TIV within each site. For each female subject, a male counterpart was included whose age differed by no more than one year and whose TIV differed by no more than 3%. The age range in this sample comprised 18–65 years ($M=38.15$, $SD=15.35$). More detailed information is displayed in Table S1, and the distribution of TIV for women and men in this sample is shown in Fig. S1. In this sample, women and men did not differ significantly in age ($t=0.01$, $p=0.99$), or in TIV ($t=-1.25$, $p=0.21$). The ATM sample was also divided into 80% for training and 20% hold-out for testing, again resulting in 1292 subjects (646 women) for training and 322 subjects (161 women) for testing. The training and hold-out samples did not differ with respect to age ($t=0.02$, $p=0.98$) or TIV ($t=-0.53$, $p=0.60$). Additionally, there was no significant difference between women and men in age or TIV in the training (age: $t=0.01$, $p=0.99$; TIV: $t=-0.99$, $p=0.32$) or hold-out sample (age: $t=-0.01$, $p=0.99$; TIV: $t=-0.83$, $p=0.41$).

Application samples. The first application sample (Sample A) was acquired in Aachen (Germany). This data set consisted of 115 individuals (24 CM, 25 CW, 33 TM, 33 TW). All cisgender participants were recruited via a public announcement around Aachen, whereas TM and TW were recruited in self-help groups and at the Department of Gynaecological Endocrinology and Reproductive Medicine of the RWTH Aachen University Hospital, Germany. All cisgender and transgender subjects in this sample reported no presence of neurological disorders, other medical conditions affecting the brain metabolism or first-degree relatives with a history of mental disorders. The Ethics Committee of the Medical Faculty of the RWTH Aachen University approved the study (EK 088/09²³). At the time of MRI measurement, 15 TM and 16 TW each were receiving hormone treatment. The age of the participants ranged from 18 to 61 years ($M=30.38$, $SD=11.03$). More detailed demographic information can be found in Table S1 and Fig. S1.

The second application sample (Sample B) consisted of an open-source dataset acquired in Barcelona, available via (<https://data.mendeley.com/datasets/hjmfvr6vmg/2>,^{47–49}). The data set contained 87 subjects (19 CM, 22 CW, 29 TM, 17 TW) with an age range of 17 to 39 years ($M=22.23$, $SD=4.97$). More detailed information related to age and TIV in all four groups can be found in Table S1 and Fig. S1, though no information were available regarding the status of potential hormone treatment.

Model applications were evaluated on both application samples separately to further understand the model behavior on samples with differing characteristics (Table S1).

The data usage of the second application sample as well as the data for the AM and ATM-sample was approved by the Ethics Committee of the Medical Faculty of the Heinrich-Heine University Düsseldorf (2018-317, 4039, 4096, 5193). All subjects were participants in research projects approved by a local Institutional Review Board and provided written informed consent and all experiments were performed in accordance with relevant guidelines and regulations.

Preprocessing of structural data. Structural T1-weighted MR images of all datasets were preprocessed using the Computational Anatomy Toolbox (CAT12.5 r1363, <http://www.neuro.uni-jena.de/cat12/>) in SPM (r6685) running under Matlab 9.0. After initial denoising (spatial-adaptive Non-Local Means), the pipeline included spatial registration, bias-correction, skull-stripping and segmentation by an adaptive maximum a posteriori approach⁵⁰ with using a partial volume model⁵¹. Subsequently, an optimized version of the Geodesic Shooting Algorithm⁵² was applied for normalization to MNI space and the resulting Jacobians were used for non-linear only modulation of grey matter segments, before final resampling to a $3 \times 3 \times 3$ mm resolution via FSL. The non-linear only modulated images (m0wp1) were globally scaled for TIV internally with an approximation of TIV, i.e. every voxel was scaled by the relative linear transformation to the MNI152 template. Consequently, while TIV-related variance was likely not fully removed from the data, the GMV data included in the analyses were not fully TIV-naïve.

Predictive modelling. Whole-brain voxelwise GMV were used as features for training the classifiers, resulting in 77779 brain features (voxels) per subject. For each of the AM and the ATM training samples, classifiers were trained to predict sex with and without featurewise removal of TIV-related variance, resulting in the four different models: AM, AM+cr, ATM and AM+cr model (Fig. 1). For all four models, we employed a SVM classifier with rbf kernel⁵³ using Julearn (<https://juaml.github.io/julearn>). Before training the classifier, PCA was performed to reduce the dimensionality of the data¹⁶. The maximum number of components ($n=1292$, num-

ber of subjects in the training sample) was retained. Where applicable, for featurewise TIV control TIV-related variance was removed after dimensionality reduction by subtracting the fitted values of each feature in a cross-validation (CV)-consistent manner to avoid data leakage^{20,30}. Stratified tenfold CV was performed to assess generalization performance. The two hyperparameters, C ($1 - 1e^8$, log-uniform) and gamma ($1e^{-7} - 1$, log-uniform), were tuned via Bayesian Hyperparameter Optimization with 250 iterations within a fivefold CV inner loop following the analysis employed in a previous study¹⁶. The best performing combination of hyperparameters from the Bayesian Hyperparameter Optimization was used to train the final model on the full sample (details depicted in Supplementary Material).

The four final models were used to obtain predictions for the AM and ATM hold-out samples and both application samples (Fig. 1). Before application of the models to the hold-out samples, we ensured that the models were calibrated (<https://scikit-learn.org/stable/modules/calibration.html#calibration>) by assessing probabilities of classifying an individual into a respective class in relation to the actual labels of the individuals (Supplementary Figs. S2 and S3, Supplementary Results). These calibrations allow for checking whether the models gave accurate estimates of class probabilities and support probability predictions. To distinguish between the predicted and actual label of the sex a person identifies with, we refer to the terms “male” and “female” as predicted labels of an ML model whereas we refer to “men” and “women” as actual (true) label of an individual.

To further explore model behaviour, we compared the TIV-distributions of individuals classified in accordance with their sex and those who were not, by use of violin plots⁵⁴ and by Wilcoxon rank sum tests. Due to the amount of comparisons conducted here, we chose a conservative significance level of $\alpha = 0.005$ with effect sizes estimated accordingly⁵⁵. To examine whether models were confounded by total GMV, we first tested whether GMV differed between the sexes in the two samples. In the AM sample, similarly to TIV, sexes exhibited significant differences in total GMV (two-sample t-test; $t = -31.21$, $p < 0.001$). However, matching for TIV in the ATM sample also resulted in a non-significant difference in total GMV ($t = 0.85$, $p = 0.40$), indicating that matching on TIV was effective also for GMV. We then compared the GMV distributions of individuals classified correctly in accordance with their sex and those who were misclassified (Tables S5 and S6) with the same conservative significance level as for TIV-differences of $\alpha = 0.005$. Further details can be found in the Supplementary Results and Tables S5 and S6. To assess potential differences between cis- and transgender individuals in prediction probabilities, we statistically compared probabilities of CM and TW as well as CW and TM. A power-analysis for these comparisons was conducted using G*Power to compute sample size required for effect sizes as found in previous work with a α -level of 0.05 and power-level of 0.8^{29,56,57}.

Data availability

The data used in the study are available via open-source datasets, for which access information is provided in the supplementary information files together with the structural scanning parameter. Code is available on GitHub: https://github.com/juaml/sex_prediction_vbm.

Received: 13 December 2022; Accepted: 22 June 2023

Published online: 24 August 2023

References

- Willeminck, M. J. *et al.* Preparing medical imaging data for machine learning. *Radiology* **295**(1), 4–15 (2020).
- Buch, V. H., Ahmed, I. & Maruthappu, M. Artificial intelligence in medicine: Current trends and future possibilities. *Br. J. Gen. Pract.* **68**(668), 143–144 (2018).
- Chang, K. *et al.* Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* **25**(8), 945–954 (2018).
- Jollans, L. *et al.* Quantifying performance of machine learning methods for neuroimaging data. *Neuroimage* **199**, 351–365 (2019).
- Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652–656 (2019).
- Nielsen, A. N. *et al.* Machine learning with neuroimaging: Evaluating Its applications in psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**(8), 791–798 (2020).
- Kahlert, J. *et al.* Control of confounding in the analysis phase—an overview for clinicians. *Clin. Epidemiol.* **9**, 195–204 (2017).
- Pourhoseingholi, M. A., Baghestani, A. R. & Vahedi, M. How to control confounding effects by statistical analysis. *Gastroenterol. Hepatol. Bed Bench* **5**(2), 79 (2012).
- Sedgwick, P. Analysing case-control studies: Adjusting for confounding. *BMJ* **346**, f25 (2013).
- McNamee, R. Regression modelling and other methods to control confounding. *Occup. Environ. Med.* **62**(7), 500–506 (2005).
- Feis, D.-L. *et al.* Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *Neuroimage* **70**, 250–257 (2013).
- Chekroud, A. M. *et al.* Patterns in the human brain mosaic discriminate males from females. *Proc. Natl. Acad. Sci. U.S.A.* **113**(14), E1968 (2016).
- Bzdok, D. Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* **11**, 543 (2017).
- Weis, S. *et al.* Sex classification by resting state brain connectivity. *Cereb. Cortex* **30**(2), 824–835 (2020).
- Wang, L. *et al.* Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain: An MVPA approach. *Neuroimage* **61**(4), 931–940 (2012).
- Flint, C. *et al.* Biological sex classification with structural MRI data shows increased misclassification in transgender women. *Neuropsychopharmacology* **45**, 1758–1765 (2020).
- Baldinger-Melich, P. *et al.* Sex matters: A multivariate pattern analysis of sex- and gender-related neuroanatomical differences in cis- and transgender individuals using structural magnetic resonance imaging. *Cereb. Cortex* **30**(3), 1345–1356 (2020).
- Eliot, L. *et al.* Dump the “dimorphism”: Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neurosci. Biobehav. Rev.* **125**, 667–697 (2021).
- Kaczurkin, A. N., Raznahan, A. & Satterthwaite, T. D. Sex differences in the developing brain: Insights from multimodal neuroimaging. *Neuropsychopharmacology* **44**(1), 71–85 (2019).
- Snoek, L., Miletic, S. & Scholte, H. S. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage* **184**, 741–760 (2019).

21. Smith, E. *et al.* Gender incongruence and the brain - Behavioral and neural correlates of voice gender perception in transgender people. *Horm. Behav.* **105**, 11–21 (2018).
22. Bouman, W. P. *et al.* Language and trans health. *Int. J. Transgenderism* **18**(1), 1–6 (2017).
23. Clemens, B. *et al.* Replication of previous findings? Comparing gray matter volumes in transgender individuals with gender incongruence and cisgender individuals. *J. Clin. Med.* **10**(7), 1454 (2021).
24. Mueller, S. C. *et al.* The neuroanatomy of transgender identity: Mega-analytic findings from the ENIGMA transgender persons working group. *J. Sex Med.* **18**(6), 1122–1129 (2021).
25. Pol, H. E. H. *et al.* Changing your sex changes your brain: Influences of testosterone and estrogen on adult human brain structure. *Eur. J. Endocrinol.* **155**, S107–S114 (2006).
26. Spizzirri, G. *et al.* Grey and white matter volumes either in treatment-naïve or hormone-treated transgender women: A voxel-based morphometry study. *Sci. Rep.* **8**(1), 1–10 (2018).
27. Zubiaurre-Elorza, L., Junque, C., Gómez-Gil, E. & Guillamon, A. Effects of cross-sex hormone treatment on cortical thickness in transsexual individuals. *J. Sex. Med.* **11**(5), 1248–1261 (2014).
28. Fukao, T., Ohi, K. & Shioiri, T. Gray matter volume differences between transgender men and cisgender women: A voxel-based morphometry study. *Aust. N. Z. J. Psychiatry* **56**(5), 535–541 (2022).
29. Kurth, F. *et al.* Brain sex in transgender women is shifted towards gender identity. *J. Clin. Med.* **11**(6), 1582 (2022).
30. More, S., Eickhoff, S. B., Caspers, J., & Patil, K. R. Confound removal and normalization in practice: A neuroimaging based sex prediction case study in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 3–18 (2021)
31. Chyzyk, D., Varoquaux, G., Milham, M. & Thirion, B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* **11**, giac014 (2022).
32. Malone, I. B. *et al.* Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. *Neuroimage* **104**, 366–372 (2015).
33. Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á. J., Félix, S. & Forn, C. Beyond “sex prediction”: Estimating and interpreting multivariate sex differences and similarities in the brain. *NeuroImage* **257**, 119343 (2022).
34. Sanchis-Segura, C. *et al.* Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Sci. Rep.* **10**(1), 1–15 (2020).
35. Farias, F., Ludermir, T., & Bastos-Filho, C. Similarity Based Stratified Splitting: An approach to train better classifiers. arXiv Preprint at <https://arxiv.org/abs/2010.06099> (2020)
36. Uçar, M. K., Nour, M., Sindi, H. & Polat, K. The effect of training and testing process on machine learning in biomedical datasets. *Math. Probl. Eng.* <https://doi.org/10.1155/2020/2836236> (2020).
37. Joel, D. *et al.* Sex beyond the genitalia: The human brain mosaic. *Proc. Natl. Acad. Sci.* **112**(50), 15468–15473 (2015).
38. Bao, A. M. & Swaab, D. F. Sex differences in the brain, behavior, and neuropsychiatric disorders. *Neuroscientist* **16**(5), 550–565 (2010).
39. Bao, A. M. & Swaab, D. F. Sexual differentiation of the human brain: Relation to gender identity, sexual orientation and neuropsychiatric disorders. *Front. Neuroendocrinol.* **32**(2), 214–226 (2011).
40. Miller, V. M., Rocca, W. A. & Faubion, S. S. Sex differences research, precision medicine, and the future of women's health. *J. Womens Health (Larchmt)* **24**(12), 969–971 (2015).
41. Ruiz-Serra, V., Buslón, N., Philippe, O. R., Saby, D., Morales, M., Pontes, C., Andirkó, A.M., Holliday, G.L., Jené, A., Moldes, M., Rambla, J., . Cirillo, D. Addressing sex bias in biological databases worldwide. <https://biohackrxiv.org/n9dkg/> (2023)
42. Weber, K. A. *et al.* Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction. *Front. Neurol.* **13**, 960760 (2022).
43. Resnick, S. M. *et al.* One-year age changes in MRI brain volumes in older adults. *Cereb. Cortex* **10**(5), 464–472 (2000).
44. Good, C. D. *et al.* A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* **14**(1 Pt 1), 21–36 (2001).
45. Resnick, S. M. *et al.* Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain. *J. Neurosci.* **23**(8), 3295–3301 (2003).
46. Taki, Y. *et al.* Correlations among brain gray matter volumes, age, gender, and hemisphere in healthy individuals. *PLoS One* **6**(7), e22734 (2011).
47. Uribe, C. Original data of a functional MRI study in transgender individual. Mendeley Data, V2, doi: <https://doi.org/10.17632/hjmfvr6vmg> (2020)
48. Uribe, C. *et al.* Data for functional MRI connectivity in transgender people with gender incongruence and cisgender individuals. *Data Brief* **31**, 105691 (2020).
49. Uribe, C. *et al.* Brain network interactions in transgender individuals with gender incongruence. *Neuroimage* **211**, 116613 (2020).
50. Rajapakse, J. C., Giedd, J. N. & Rapoport, J. L. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Trans. Med. Imaging* **16**(2), 176–186 (1997).
51. Tohka, J., Zijdenbos, A. & Evans, A. Fast and robust parameter estimation for statistical partial volume models in brain MRI. *Neuroimage* **23**(1), 84–97 (2004).
52. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26**(3), 839–851 (2005).
53. Boser, B.E., Guyon, I. M., & Vapnik, V. N., A training algorithm for optimal margin classifiers in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152 (1992).
54. Bechtold, B. Violin Plots for Matlab, Github Project <https://github.com/bastibe/Violinplot-Matlab>, Doi: <https://doi.org/10.5281/zenodo.4559847> (2016).
55. Fritz, C.O., Morris, P.E., Richler, J.J. "Effect size estimates: Current use, calculations, and interpretation": Correction to Fritz et al. (2011). (2012).
56. Faul, F. *et al.* G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**(2), 175–191 (2007).
57. Faul, F. *et al.* Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **41**(4), 1149–1160 (2009).

Acknowledgements

The work was supported by: Deutsche Forschungsgemeinschaft (DFG, including DE 2319/2-2, /2-3, /2-4 and HA 3202/7-2, /7-3, /7-4). National Institute of Mental Health (R01-MH074457). Helmholtz Portfolio Theme “Supercomputing and Modeling for the Human Brain”. European Union’s Horizon 2020 Research. Innovation Programme under Grant Agreement No. 945539 (HBP SGA3). Open access publication funded by the DFG – 491111487.

Author contributions

K.R.P developed the idea of the study. K.R.P., S.W., S.H and L.W. conceptualized the study. M.V., U.H., B.C. and B.D. contributed sample A, F.H. preprocessed all data. M.V., F.H., L.W. preprocessed sample A and B, L.W.

prepared data for the ML-analysis, which was conducted by S.H. and K.R.P., L.W. prepared the results, including figures and tables, L.W. drafted the manuscript together with S.W., S.B.E. and all other authors commented and contributed to the final manuscript. K.R.P. and S.W. contributed equally to the manuscript as corresponding authors. This work has been done in partial fulfilment of the requirements for a PhD thesis.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

B.C. serves as scientific advisor for Dionysus Digital Health, Inc. and holds shares of this company. All other authors, L.W., S.H., F.H., M.V., U.H., B.D., S.B.E., K.R.P., S.W., declare they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-37508-z>.

Correspondence and requests for materials should be addressed to K.R.P. or S.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

6 Discussion

The present work highlights multivariate statistical approaches as useful tools to offer new and holistic insights into the complex characteristics introduced by the phenotype sex in brain and cognition. The commentary outlines the overall importance of multivariate statistical methods in investigating complex patterns of sex-related variability. In subsequent studies, sex-related variability in cognitive and brain imaging data was examined utilizing different multivariate statistical methods. Using a CFA, study 1 demonstrated sex differences in cognitive profiles as captured by different component solutions for males and females. The studies 2 and 3 examined sex differences in functional and structural brain organization using ML models while also focusing on methodological considerations in sex classification analyses. The former focused on the influence of the training sample on the generalization performance and the latter demonstrated the relevance of appropriate control of confounding variables.

6.1 Addressing methodological considerations in multivariate analyses of sex differences

Choice of dataset

One important factor that affects the generalization performance in terms of exhibiting differences between models is the choice of the sample on which the ML models are trained. In study 2, we systematically compared four different samples based on single datasets and two compound samples containing data from four different datasets with varying sample characteristics. All samples were used for training sex classification models, followed by an evaluation of the generalization performance when tested across multiple test samples. The results of study 2 demonstrated that models trained on single samples did not generalize well across all test samples, which is in line with the findings of a previous study [107]. Compared to models trained on single samples, both sets of models trained on compound samples demonstrated superior generalization performance. These findings indicate that training ML models on a compound sample may better generalize, as a compound training sample is more heterogeneous and more representative of the test data, as proposed by other studies [93, 97, 104, 128]. Furthermore, the larger compound sample outperformed the smaller compound sample, demonstrating that sample size affects generalization performance, which is

consistent with the results of previous studies [88, 93, 104-106]. Overall, the results of study 2 demonstrated that both the sample size and data composition of the training sample are crucial factors in achieving generalizable results when investigating sex differences using multivariate statistical analyses.

In line with these insights, a large and heterogeneous dataset was used in study 1 to obtain a representative sample that is also more likely to provide generalizable results. This study utilized a sample derived from the 1000Brains dataset, which is based on a population-based epidemiological cohort study. The 1000Brains dataset comprises more than 1000 subjects, a sample size comparable to previous large-scale neuroimaging studies acquired in Europe and the US [127, 129-131]. The 1000Brains dataset aims to represent the variability within the aging process of the human brain and therefore acquired a large cohort of elderly subjects from the general population [127]. As it was the objective of study 1 to study cognitive profiles in an older population, the 1000Brains sample was ideal for the research question at hand. Considering that this sample is representative of the general population, it is very likely that the results of study 1 could be replicated in other samples. However, future studies will be valuable to determine whether these results can actually be replicated with regard to the different component solutions for males and females in their cognitive profiles.

The findings of study 2 were also integrated into the study design of study 3 in selecting an appropriate training sample for the ML analysis. A sample with a high sample size and heterogeneous sample characteristics was constructed by combining data from 10 large cohorts, including subjects who differ in nationality, age range, and neuroimaging acquisition. Using a large and heterogeneous sample enabled study 3 to achieve accurate and generalizable results.

Taken together, all three studies incorporate the methodological consideration of the choice of dataset: For study 1, a large-scale sample was selected that is based on the general population. In study 2, the effect of the training sample was directly evaluated, demonstrating the combination of a large and heterogeneous sample leading to accurate and generalizable results. The results of study 3 confirmed this finding by showing accurate and generalizable sex classification results by training on a combination of large-scale datasets.

Confounding variables

It is crucial to ensure that accurate and generalizable results can actually be attributed to the feature-target relationship and are not a consequence of other confounding variables. Study

3 examined the example of TIV as a confounding variable in structural sex classification analyses based on GMV features. Here, the model that did not control for TIV as a confounder demonstrated a very high classification accuracy (up to 97%). However, study 3 also demonstrated that this high classification accuracy relates to a bias in the model, which classifies subjects with a high TIV as males and subjects with a low TIV as females (Figure 3 in study 3). When controlling for TIV by featurewise confound removal during the ML analysis, classification accuracies decreased to 62-73%. This finding aligns with previous studies reporting decreased sex classification accuracies when controlling for TIV [123, 125, 126]. In addition to featurewise confound removal, study 3 also investigated the effects of stratifying for TIV by matching as an approach to remove TIV-confounding information. The TIV-stratified models showed both an unbiased model behavior and a high level of accuracy (86-93%). Consequently, the matching approach is a more favorable approach to appropriately analyze for TIV bias in sex classification analyses based on GMV. In summary, study 3 demonstrates the importance of appropriately controlling for confounding variables in order to avoid biased predictions and misleading conclusions, as well as ensuring accurate interpretations about feature-target relationships.

The methodological consideration of confounding variables was also addressed in studies 1 and 2: For study 2, all samples were matched for sex and age. As a result, a similar number of males and females were enrolled in the study and the potential confounding influence of age on the RSFC was controlled for, as several studies reported age-related changes in functional connectivity [132-134]. In turn, study 1 controlled for the effects of age and education by matching males and females for these two variables, as both variables are reported to show effects on cognitive performance [135-138]. Overall, it is important to consider which variables may have a confounding influence in a given study design and how to appropriately control for these variables. All three studies in the present work included a control for potentially confounding variables in order to obtain accurate and unbiased results.

6.2 Sex differences in cognition and the brain: Insights from multivariate analyses

Beyond the methodological aspects presented in the studies 2 and 3, the results of both studies also demonstrated differences between males and females through the respective sex classification accuracies based on the functional and structural brain organization. Study 3

demonstrated high structural sex classification accuracies based on GMV (62-93%) that are consistent with the results found in previous sex classification studies [76, 101-103]. Thus, the sex differences in structural brain organization are as pronounced as to allow for accurate predictions of a person's sex in sex classification analyses. In addition, study 2 employed functional brain imaging data following a parcelwise approach to address topographic interference in RSFC by identifying spatially specific effects resulting from the highest classifying parcels. The results showed varying accuracies depending on the location and the dataset used in training the model ranging from 45-83%. Previous studies have also reported varying accuracy levels of up to 87% [51, 98-100]. In light of the results of study 2, a subsequent step would be to determine which parcels are capable of classifying on a high level, independent of the dataset used to train the sex classification models. It will be interesting to examine whether these parcels correspond to functional sex classification studies [51, 100] that reported highly classifying parcels located within the DMN [56, 139], aligning with other studies reporting sex differences in the DMN [60, 61, 140, 141]. Consistently highly classifying parcels across different datasets may indicate parcels that reflect differences in these functional connectivity patterns between males and females, independent of the respective dataset used to train the sex classification models.

By investigating differences between males and females in cognition, study 1 provides complementary insights to the findings of sex differences in structural and functional brain organization provided by studies 2 and 3. Specifically, study 1 assessed sex differences in cognitive profiles of older males and females, including attention, memory, language, and executive functions. The results of study 1 demonstrated sex differences in single domains, i.e. males to perform better in tasks related to visual-spatial attention, and females to perform better in tasks related to verbal abilities. These results match previous findings in the literature reporting sex differences in domains such as language, visual-spatial abilities, and memory [18, 23-25, 27, 28]. Moreover, study 1 demonstrated sex differences in the cognitive processing styles, indicating a rather decomposed (local) cognitive profile in females and a more holistic (global) profile in males, which is consistent with findings in previous literature [142-145].

Collectively, the studies in the present work highlight the insights provided by multivariate approaches that do not rely on the previously defined assumption of a sexual dimorphism. Instead, the three studies and the commentary demonstrate the advantages of using

multivariate approaches to study sex differences in brain and cognition on a holistic level covering multiple variables rather than focusing on single variables alone. In the literature, there has been a controversy regarding differences versus similarities between males and females that cannot be resolved without incorporating the wide range of variation between individuals across multiple variables. Ultimately, the question is whether, despite the general variability that can be attributed to various factors, sex also contributes a significant amount to this variability. The present data-driven approaches allowed us to examine sex-related variability in relation to the overall variability rather than to confirm or refute the hypothesis of a sexual dimorphism with regard to a single variable. In study 1, a PCA and CFA allowed us to identify whether males and females exhibit different factor structures in their cognitive profiles. In studies 2 and 3, the accuracy of the sex classification analyses allowed us to assess the extent of information available to classify the sex of a person based on functional or structural neuroimaging data. The results of the present work demonstrated distinct cognitive processing styles for males and females (study 1) and that sex differences in the functional (study 2) and structural (study 3) brain organization were so pronounced to allow for predictions of a person's sex according to these features. Thus, the phenotype sex contributes to a fundamental part of variability between individuals, at least to the extent that we can detect sex differences in cognition and brain organization. Overall, based on the multivariate approaches and the respective methodological considerations presented here, it is possible to evaluate the variability introduced by sex and also other phenotypes, allowing for appropriate conclusions regarding differences and similarities between the sexes.

6.3 Univariate and Multivariate statistical approaches

Taken together, the three studies in the present work demonstrate the value of multivariate statistical approaches to investigate sex differences in cognition and the brain. Single univariate group comparisons may indicate a significant sex difference in the behavioral and functional processing for a specific task, such as language [18-22], visual-spatial attention [18, 23-26, 33, 34] memory [23, 24, 27-30, 34, 35], and executive functioning [31], as well as single regions in the brain [12, 32]. However, several group comparisons do not necessarily reflect whether there are fundamental differences in the cognitive profile or brain organization. As opposed to univariate group comparisons contrasting males and females with regard to a specific variable, multivariate methods allow us to identify potential sex differences while taking into account multiple variables. Consequently, it is possible to

examine whether there are fundamental sex differences in the overall cognitive profile (study 1), and the functional (study 2) or structural (study 3) brain organization rather than identifying differences in individual cognitive functions or brain regions. In addition, localized approaches, such as parcelwise approaches, can be used to identify the specific brain regions that exhibit the greatest differences and thus the highest predictive variance. By using targeted group comparisons to analyze single variables, it is possible to further examine the direction and extent of differences identified by multivariate approaches. The combined use of multivariate and univariate approaches can therefore be complementary to one another to address questions on different levels.

Altogether, multivariate statistical methods provide a more holistic approach to analyze variability associated with the phenotype sex by considering multiple variables, enabling the identification of sex differences in the underlying constructs of the respective variables and features [80]. As such, multivariate approaches serve as a valuable complement to univariate analyses, which may address different types of questions, respectively. Overall, the findings of the present studies demonstrate the relevance of using multivariate statistical methods to investigate sex-related variability. In combination with large neuroimaging datasets, we can study parts of the complex nature of sex differences in brain and behavior.

6.4 Factors modulating sex-related variability

In general, the sex of a person is a prominent phenotype causing variability between individuals. Nonetheless, the variability in brain and cognition is not only modulated by sex alone, but several other factors might influence or even modulate the sex-related variability. Gender identity is one such modulating factor, as previous research has demonstrated an interaction of sex and gender indicated by differences in the brain structure when contrasting cis- and transgender individuals [73-78, 146]. Also, studies using multivariate methods reported indications for an interaction of sex and gender [76, 147]. In the present work, gender identity was examined in a multivariate approach in study 3, using structural sex classification models trained to classify sex assigned at birth to data from both cis- and transgender individuals. There have been previous studies that have also applied this methodology to transgender data, which reported a reduction in classification accuracy, indicating differences in the neuroanatomical structure between cis- and transgender individuals [76, 147]. Contrary to previous studies, the results in study 3 did not show such

a reduction in accuracy for transgender individuals compared to cisgender individuals assigned to the same sex at birth. However, these findings do not indicate that there are no differences between cis- and transgender individuals in the structural brain organization at all. The discrepancy with the previous studies may be explained by different TIV-distributions in the samples or by fundamentally different models being trained due to variations in feature weights, sample characteristics, and types of confound removal. Nevertheless, it is important to note that study 3 provides an accurate framework to detect and control for a TIV-biased model behavior, enabling the classifiers to accurately identify the sex assigned at birth based on GMV features for both cisgender and transgender individuals. Overall, study 3 provides the methodological framework for future investigations that will allow us to examine the interaction between sex and gender both in terms of structural and functional brain organization without relying on brain size alone to complement previous research [76, 147-149]. Future studies that incorporate both cis- and transgender data may provide more insights into the interaction of the entwined variables of sex and gender [150] and how both variables contribute to different clinical manifestations and outcomes of diseases [16].

Furthermore, dynamic changes in sex hormones may also influence sex differences in the brain [151-153]. In particular, hormonal levels in women fluctuating throughout the menstrual cycle [154-160], reproductive stage [161], or in response to oral contraceptives [162-164] may modulate sex-related variability in the brain. To gain a better understanding of the relationship between sex hormones and sex differences in the brain, large-scale and longitudinal studies should acquire and consider hormone levels in conjunction with well-characterized phenotypes.

Another important variable that may influence sex-related variability and modulate sex differences in the brain is age. Numerous studies have demonstrated an interaction of age for sex differences in cognition [165, 166], as well as structural and functional brain imaging data [61, 166-168]. In addition, other variables relating to socio-demographics might also influence sex-related variability, such as the level of education, ethnicity, and socio-economic status [169-171]. Consequently, it is necessary to consider other variables that may interact with sex-related variability to avoid confounding effects and ensure generalizable results of the present analyses [172].

6.5 Outlook

Future studies will benefit from using multivariate statistical approaches to further investigate individual variability in the brain and behavior that is related to a person's sex and other phenotypes. By combining the insights of all three studies in the present work, future studies may focus on integrating multiple modalities to examine sex differences in cognition, brain structure, and function using multivariate analysis approaches. Linking the three modalities can help to decode the complex interplay of variability in the brain and behavior on multiple levels. Generally, further investigations of individual variability, and especially sex-related variability, are needed to improve our understanding of sex differences with respect to different neuropsychological and neuropsychiatric disorders. The sex of a person is a crucial variable in the heterogeneity of several diseases, such as Depression, Anxiety, ADHD, Parkinson's, and Alzheimer's disease [15-17]. Thus, it is essential to include sex-related aspects in the investigation of mental disorders. Decoding sex-related variability in health- and disease-related conditions may help examining sex differences in the prevalence of neuropsychiatric disorders [13, 14]. Developing sex-specific treatments and prevention strategies is a crucial step towards a personalized medicine approach that aims to adapt therapeutic methods to meet individual needs. For this reason, multivariate approaches are important for future investigations as they provide crucial insights into the underlying biological mechanisms of diseases and disorders, thereby facilitating our understanding of their complexity [40, 42, 87].

The use of multivariate analyses has the potential to decode not only the variance associated with sex, but also the variance associated with phenotypes such as age, gender identity, and other demographic characteristics [46, 49, 50, 149]. There are several domains of sociodemographic characteristics [172] that should be considered when conducting a personalized analysis, as well as the interactions between each phenotype. Overall, integrating phenotypic information into multivariate analyses is essential for moving toward precision medicine.

6.6 Conclusion

In sum, the results of the present studies demonstrate the value of multivariate methods for understanding complex patterns in the variability of cognition and brain organization. In contrast to univariate group comparisons, multivariate approaches allow for studying sex-

related variability beyond single cognitive tasks or single brain regions. Instead, multivariate statistical methods offer a holistic framework to study differences between males and females from a broader perspective. Using SEM and ML, the present work highlights applications of multivariate approaches to study differences between males and females in cognitive processing as well as in structural and functional brain organization. Moreover, the present studies provide methodological recommendations to ensure accurate and generalizable results in sex classification analyses. Future studies may extend the findings of the present work to advance the understanding of how sex differences in brain and behavior translate into daily life and clinical applications. The present findings and methodological considerations are not limited to sex-related variability but can be extended to other variables that may also interact with sex, some of which may even dynamically change over time (e.g., aging and hormonal effects). In general, multivariate statistical approaches provide a comprehensive and holistic approach to study a large number of variables, enabling the modeling of complex relationships and the generalization of predictions to unknown data.

7 References

1. Wiersch, L. & Weis, S. (2021). Sex differences in the brain: More than just male or female. *Cognitive Neuroscience*, 12(3-4), 187-188.
2. Jockwitz, C., Wiersch, L., Stumme, J. & Caspers, S. (2021). Cognitive profiles in older males and females. *Scientific reports*, 11(1), 6524.
3. Wiersch, L., Friedrich, P., Hamdan, S., Komeyer, V., Hoffstaedter, F., Patil, K.R., . . . & Weis, S. (2024). Sex classification from functional brain connectivity: Generalization to multiple datasets. *Human Brain Mapping*, 45(6), e26683.
4. Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., . . . & Weis, S. (2023). Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific reports*, 13(1), 13868.
5. Van Horn, J.D., Grafton, S.T. & Miller, M.B. (2008). Individual Variability in Brain Activity: A Nuisance or an Opportunity? *Brain imaging and behavior*, 2, 327-334.
6. Radonjic, N.V., Hess, J.L., Rovira, P., Andreassen, O., Buitelaar, J.K., Ching, C.R.K., . . . & Faraone, S.V. (2021). Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders. *Molecular psychiatry*, 26(6), 2101-2110.
7. Niu, X., Zhang, F., Kounios, J. & Liang, H. (2020). Improved prediction of brain age using multimodal neuroimaging data. *Human Brain Mapping*, 41(6), 1626-1643.
8. Kaczurkin, A.N., Raznahan, A. & Satterthwaite, T.D. (2019). Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology*, 44(1), 71-85.
9. de Lacy, N., McCauley, E., Kutz, J.N. & Calhoun, V.D. (2019). Sex-related differences in intrinsic brain dynamism and their neurocognitive correlates. *Neuroimage*, 202, 116116.
10. Seghier, M.L. & Price, C.J. (2018). Interpreting and Utilising Intersubject Variability in Brain Function. *Trends in cognitive sciences*, 22(6), 517-530.
11. Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., . . . & Deary, I.J. (2018). Sex Differences in the Adult Human Brain: Evidence from 5216 UK Biobank Participants. *Cerebral cortex*, 28(8), 2959-2975.
12. Ruigrok, A.N., Salimi-Khorshidi, G., Lai, M.C., Baron-Cohen, S., Lombardo, M.V., Tait, R.J. & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39, 34-50.
13. Bao, A.M. & Swaab, D.F. (2010). Sex differences in the brain, behavior, and neuropsychiatric disorders. *The Neuroscientist*, 16(5), 550-65.
14. Bao, A.M. & Swaab, D.F. (2011). Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorders. *Frontiers in neuroendocrinology*, 32(2), 214-26.
15. Ferretti, M.T., Iulita, M.F., Cavedo, E., Chiesa, P.A., Schumacher Dimech, A., Santucci Chadha, A., . . . & Women's Brain Project and the Alzheimer Precision Medicine Initiative. (2018). Sex differences in Alzheimer disease - the gateway to precision medicine. *Nature Reviews Neurology*, 14(8), 457-469.
16. Mauvais-Jarvis, F., Bairey Merz, N., Barnes, P.J., Brinton, R.D., Carrero, J.J., DeMeo, D.L., . . . & Suzuki, A. (2020). Sex and gender: modifiers of health, disease, and medicine. *The Lancet*, 396(10250), 565-582.

17. Pinares-Garcia, P., Stratikopoulos, M., Zagato, A., Loke, H. & Lee, J. (2018). Sex: A Significant Risk Factor for Neurodevelopmental and Neurodegenerative Disorders. *Brain sciences*, 8(8).
18. Weiss, E.M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and individual differences*, 35(4), 863-875.
19. Harrington, G.S. & Farias, S.T. (2008). Sex differences in language processing: functional MRI methodological considerations. *Journal of Magnetic Resonance Imaging*, 27(6), 1221-1228.
20. Baxter, L.C., Saykin, A.J., Flashman, L.A., Johnson, S.C., Guerin, S.J., Babcock, D.R. & Wishart, H.A. (2003). Sex differences in semantic language processing: A functional MRI study. *Brain and Language*, 84(2), 264–272.
21. Xu, M., Liang, X., Ou, J., Li, H., Luo, Y.J. & Tan, L.H. (2020). Sex Differences in Functional Brain Networks for Language. *Cerebral Cortex*, 30(3), 1528-1537.
22. Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: a critical review. *Brain and Language*, 108(3), 175-183.
23. Loring-Meier, S. & Halpern, D.F. (1999). Sex differences in visuospatial working memory: components of cognitive processing. *Psychonomic Bulletin & Review*, 6, 464-71.
24. Voyer, D., Voyer, S.D. & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review*, 24, 307-334.
25. Joseph, R. (2000). The evolution of sex differences in language, sexuality, and visual-spatial skills. *Archives of Sexual Behavior*, 29, 35-66.
26. Rubia, K., Hyde, Z., Halari, R., Giampietro, V. & Smith, A. (2010). Effects of age and sex on developmental neural networks of visual-spatial attention allocation. *Neuroimage*, 51(2), 817-827.
27. Saylik, R., Raman, E. & Szameitat, A.J. (2018). Sex Differences in Emotion Recognition and Working Memory Tasks. *Frontiers in psychology*, 9, 1072.
28. Solianik, R., Brazaitis, M. & Skurvydas, A. (2016). Sex-related differences in attention and memory. *Medicina*, 52(6), 372–377.
29. Reed, J.L., Gallagher, N.M., Sullivan, M., Callicott, J.H. & Green, A.E. (2017). Sex differences in verbal working memory performance emerge at very high loads of common neuroimaging tasks. *Brain and Cognition*, 113, 56-64.
30. Speck, O., Ernst, T., Braun, J., Koch, C., Miller, E., & Chang, L. (2000). Gender differences in the functional organization of the brain for working memory. *Neuroreport*, 11(11), 2581–2585.
31. Mansouri, F.A., Fehring, D.J., Gaillard, A., Jaberzadeh, S. & Parkinson, H. (2016). Sex dependency of inhibitory control functions. *Biology of sex Differences*, 7, 11-13.
32. Lotze, M., Domin, M., Gerlach, F.H., Gaser, C., Lueders, E., Schmidt, C.O. & Neumann, N. (2019). Novel findings from 2,838 Adult Brains on Sex Differences in Gray Matter Brain Volume. *Scientific reports*, 9(1), 1671.
33. McCarrey, A.C., An, Y., Kitner-Triolo, M.H., Ferrucci, L. & Resnick, S.M. (2016). Sex differences in cognitive trajectories in clinically normal older adults. *Psychology and aging*, 31(2), 166.
34. De Frias, C.M., Nilsson, L.G. & Herlitz, A. (2006). Sex differences in cognition are stable over a 10-year period in adulthood and old age. *Aging, Neuropsychology, and Cognition*, 13(3-4), 574-587.

35. Zilles, D., Lewandowski, M., Vicker, H., Henseler, I., Diekhof, E., Melcher, T., . . . & Gruber, O. (2016). Gender Differences in Verbal and Visuospatial Working Memory Performance and Networks. *Neuropsychobiology*, 73(1), 52-63.
36. Hyde, J.S. (2005). The gender similarities hypothesis. *American psychologist*, 60(6), 581.
37. Hyde, J.S. (2014). Gender similarities and differences. *Annual review of psychology*, 65(1), 373-398.
38. Gong, W., Beckmann, C.F. & Smith, S.M. (2021). Phenotype discovery from population brain imaging. *Medical image analysis*, 71, 102050.
39. Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., . . . & Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *Neuroimage*, 199, 351-365.
40. Davatzikos, C. (2019). Machine learning in neuroimaging: Progress and challenges. *Neuroimage*, 197, 652-656.
41. Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1), 2047-217X.
42. Mateos-Perez, J.M., Dadar, M., Lacalle-Auriolles, M., Iturria-Medina, Y., Zeighami, Y. & Evans, A.C. (2018). Structural neuroimaging as clinical predictor: A review of machine learning applications. *Neuroimage: Clinical*, 20, 506-522.
43. Haynes, J.D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7), 523-534.
44. Zhang, Z., Li, G., Xu, Y. & Tang, X. (2021). Application of Artificial Intelligence in the MRI Classification Task of Human Brain Neurological and Psychiatric Diseases: A Scoping Review. *Diagnostics*, 11(8), 1402.
45. Chen, J., Patil, K.R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., . . . & Visser, E. (2020). Neurobiological Divergence of the Positive and Negative Schizophrenia Subtypes Identified on a New Factor Structure of Psychopathology Using Non-negative Factorization: An International Machine Learning Study. *Biological Psychiatry*, 87(3), 282-293.
46. Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E., Glasser, M.F., . . . & Miller, K.L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11), 1565-1567.
47. Nostro, A.D., Müller, V.I., Varikuti, D.P., Pläschke, R.N., Hoffstaedter, F., Langner, R., . . . & Eickhoff, S.B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain Structure and Function*, 223, 2699-2719.
48. Pläschke, R.N., Patil, K.R., Cieslik, E.C., Nostro, A.D., Varikuti, D.P., Plachti, A., . . . & Eickhoff, S.B. (2020). Age differences in predicting working memory performance from network-based functional connectivity. *Cortex*, 132, 441-459.
49. Varikuti, D.P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K.R., . . . & Eickhoff, S.B. (2018). Evaluation of non-negative matrix factorization of grey matter in age prediction. *Neuroimage*, 173, 394-410.
50. More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B., Patil, K.R. & Alzheimer's Disease Neuroimaging, I. (2023). Brain-age prediction: A systematic comparison of machine learning workflows. *Neuroimage*, 270, 119947.
51. Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T.T. & Eickhoff, S.B. (2020). Sex Classification by Resting State Brain Connectivity. *Cerebral Cortex*, 30(2), 824-835.

52. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. (2020). Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological psychiatry*, 88(11), 818-828.
53. Siero, J.C., Bhogal, A. & Jansma, J.M. (2013). Blood Oxygenation Level-dependent/Functional Magnetic Resonance Imaging: Underpinnings, Practice, and Perspectives. *PET clinics*, 8(3), 329-44.
54. Weiss, E., Siedentopf, C.M., Hofer, A., Deisenhammer, E.A., Hoptman, M.J., Kremser, C., . . . & Delazer, M. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: a functional magnetic resonance imaging study in healthy volunteers. *Neuroscience letters*, 344(3), 169-172.
55. Friston, K.J., Frith, C.D., Liddle, P.F. & Frackowiak, R.S. (1993). Functional connectivity: the principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1), 5-14.
56. Buckner, R.L., Andrews-Hanna, J.R. & Schacter, D.L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1), 1-38.
57. Fox, M.D. & Raichle, M.E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9), 700-711.
58. Mak, L.E., Minuzzi, L., MacQueen, G., Hall, G., Kennedy, S.H. & Milev, R. (2017). The Default Mode Network in Healthy Individuals: A Systematic Review and Meta-Analysis. *Brain connectivity*, 7(1), 25-33.
59. Tomasi, D. & Volkow, N.D. (2012). Gender differences in brain functional connectivity density. *Human brain mapping*, 33(4), 849-860.
60. Bluhm, R.L., Osuch, E.A., Lanius, R.A., Boksman, K., Neufeld, R.W., Theberge, J. & Williamson, P. (2008). Default mode network connectivity: effects of age, sex, and analytic approach. *Neuroreport*, 19(8), 887-891.
61. Scheinost, D., Finn, E.S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M. & Constable, R.T. (2015). Sex differences in normal age trajectories of functional brain networks. *Human brain mapping*, 36(4), 1524-1535.
62. Zhang, X., Liang, M., Qin, W., Wan, B., Yu, C. & Ming, D. (2020). Gender Differences Are Encoded Differently in the Structure and Function of the Human Brain Revealed by Multimodal MRI. *Frontiers in Human Neuroscience*, 14, 244.
63. Park, H.J. & Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science*, 342(6158), 1238411.
64. 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. *Neuroscience & Biobehavioral Reviews*, 125, 667-697.
65. Zell, E., Krizan, Z. & Teeter, S.R. (2015). Evaluating gender similarities and differences using metasyntesis. *American psychologist*, 70(1), 10.
66. Cabo, L.L., Brewster, C. P., & Luengo Azpiazu, J. (2012). Sexual dimorphism: interpreting sex markers. *A companion to forensic anthropology*, 248-286.
67. Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., . . . & Assaf, Y. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences*, 112(50), 15468-15473.
68. Joel, D. & Fausto-Sterling, A. (2016). Beyond sex differences: new approaches for thinking about variation in brain structure and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1688), 20150451.

69. Bouman, W.P., Schwend, A. S., Motmans, J., Smiley, A., Safer, J. D., Deutsch, M. B., ... & Winter, S. (2017). Language and trans health. *International Journal of Transgenderism*, 18(1), 1-6.
70. Diamond, M. (2002). Sex and Gender are Different: Sexual Identity and Gender Identity are Different. *Clinical child psychology and psychiatry*, 7(3), 320–334.
71. Kaltiala-Heino, R., Työlajärvi, M. & Lindberg, N. (2019). Gender dysphoria in adolescent population: A 5-year replication study. *Clinical child psychology and psychiatry*, 24(2), 379-387.
72. Motmans, J., Nieder, T.O. & Bouman, W.P. (2019). Transforming the paradigm of nonbinary transgender health: A field in transition. *International Journal of Transgenderism*, 20(2-3), 119-125.
73. Fukao, T., Ohi, K. & Shioiri, T. (2022). Gray matter volume differences between transgender men and cisgender women: A voxel-based morphometry study. *Australian & New Zealand Journal of Psychiatry*, 56(5), 535-541.
74. Clemens, B., Votinov, M., Puiu, A.A., Schuppen, A., Hüpen, P., Neulen, J., . . . & Habel, U. (2021). Replication of Previous Findings? Comparing Gray Matter Volumes in Transgender Individuals with Gender Incongruence and Cisgender Individuals. *Journal of Clinical Medicine*, 10(7), 1454.
75. Luders, E., Sánchez, F.J., Gaser, C., Toga, A.W., Narr, K.L., Hamilton, L.S. & Vilain, E. (2009). Regional gray matter variation in male-to-female transsexualism. *Neuroimage*, 46(4), 904-7.
76. Flint, C., Forster, K., Koser, S.A., Konrad, C., Zwitterlood, P., Berger, K., . . . & Grotegerd, D. (2020). Biological sex classification with structural MRI data shows increased misclassification in transgender women. *Neuropsychopharmacology*, 45(10), 1758-1765.
77. Spizzirri, G., Duran, F.L.S., Chaim-Avancini, T.M., Serpa, M.H., Cavallet, M., Pereira, C.M.A., . . . & Abdo, C.H.N. (2018). Grey and white matter volumes either in treatment-naïve or hormone-treated transgender women: a voxel-based morphometry study. *Scientific Reports*, 8(1), 736.
78. Pol, H.E.H., Cohen-Kettenis, P. T., Van Haren, N. E., Peper, J. S., Brans, R. G., Cahn, W., ... & Kahn, R. S. (2006). Changing your sex changes your brain: influences of testosterone and estrogen on adult human brain structure. *European Journal of Endocrinology*, 155(suppl_1), S107-S114.
79. Mueller, S.C., Guillamon, A., Zubiaurre-Elorza, L., Junque, C., Gomez-Gil, E., Uribe, C., . . . & Luders, E. (2021). The Neuroanatomy of Transgender Identity: Mega-Analytic Findings From the ENIGMA Transgender Persons Working Group. *The Journal of Sexual Medicine*, 18(6), 1122-1129.
80. Huberty, C.J., & Morris, J. D. (1992). Multivariate analysis versus multiple univariate analyses.
81. Brown, T.A., & Moore, M. T. (2012). Confirmatory factor analysis. *Handbook of structural equation modeling*, 361, 379.
82. Hair Jr, J.F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., Ray, S., ... & Ray, S. (2021). An introduction to structural equation modeling. *Partial least squares structural equation modeling (PLS-SEM) using R: a workbook*, 1-29.
83. Harrington, D. (2009). Confirmatory factor analysis. *Oxford university press*.
84. Bzdok, D. (2017). Classical Statistics and Statistical Learning in Imaging Neuroscience. *Frontiers in neuroscience*, 11, 543.
85. Jodie, B.U., & Ullman, B. (2006). Structural equation modeling: reviewing the basics and moving forward. *Journal of personality assessment*, 87(1), 35-50.

86. Savalei, V., & Bentler, P. M. (2006). Structural equation modeling. *The handbook of marketing research: Uses, misuses, and future advances*, 330, 36.
87. Nielsen, A.N., Barch, D.M., Petersen, S.E., Schlaggar, B.L. & Greene, D.J. (2020). Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(8), 791-798.
88. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.
89. Buch, V.H., Ahmed, I. & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice*, 68(668), 143-144.
90. Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180, 68-77.
91. Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T.D. & Woo, C.W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature protocols*, 15(4), 1399-1435.
92. Baştanlar, Y. & Özuysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, 105-28.
93. Dhamala, E., Yeo, B.T.T. & Holmes, A.J. (2023). One Size Does Not Fit All: Methodological Considerations for Brain-Based Predictive Modeling in Psychiatry. *Biological Psychiatry*, 93(8), 717-728.
94. Morales, E.F., & Escalante, H. J. (2022). A brief introduction to supervised, unsupervised, and reinforcement learning. In *Biosignal processing and classification using computational learning and intelligence*, Academic Press., (pp. 111-129).
95. Simeone, O. (2018). A brief introduction to machine learning for engineers. *Foundations and Trends® in Signal Processing*, 12(3-4), 200-431.
96. Kotsiantis, S.B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3-24.
97. Chung, Y., Haas, P. J., Upfal, E., & Kraska, T. (2018). Unknown examples & machine learning model generalization. *arXiv preprint arXiv:1808.08294*.
98. Casanova, R., Whitlow, C.T., Wagner, B., Espeland, M.A. & Maldjian, J.A. (2012). Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *The open neuroimaging journal*, 6, 1.
99. Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., . . . & Van Essen, D.C. (2013). Functional connectomics from resting-state fMRI. *Trends in cognitive sciences*, 17(12), 666-682.
100. 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. *Human brain mapping*, 39(4), 1765–1776.
101. 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(4), 931-940.
102. Feis, D.-L., Brodersen, K.H., Cramon, D.Y.v., Luders, E. & Tittgemeyer, M. (2013). Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *NeuroImage*, 70, 250–257.
103. Chekroud, A.M., Ward, E.J., Rosenberg, M.D. & Holmes, A.J. (2016). Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences of the United States of America*, 113(14), E1968.

104. Cui, Z. & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage*, 178, 622-637.
105. Ishida, E.E. (2019). Machine learning and the future of supernova cosmology. *Nature Astronomy*, 3(8), 680-682.
106. Yang, F., Wanik, D. W., Cerrai, D., Bhuiyan, M. A. E., & Anagnostou, E. N. (2020). Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration. *Sustainability*, 12(4), 1525.
107. Huf, W., Kalcher, K., Boubela, R. N., Rath, G., Vecsei, A., Filzmoser, P., & Moser, E. (2014). On the generalizability of resting-state fMRI machine learning classifiers. *Frontiers in human neuroscience*, 8, 502.
108. Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., . . . & Kalpathy-Cramer, J. (2018). Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8), 945-954.
109. Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., . . . & Lungren, M.P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, 295(1), 4-15.
110. Altman, N. & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nat Methods*, 15(6), 399-400.
111. Janssen, R.J., Mourão-Miranda, J. & Schnack, H.G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(9), 798-808.
112. Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1), 100.
113. Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9), 2812-2831.
114. Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, 26(3), 303-304.
115. Eickhoff, S.B., Yeo, B.T.T. & Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11), 672-686.
116. Rafi, M., & Shaikh, M. S (2013). A comparison of SVM and RVM for Document Classification. *arXiv preprint arXiv:1301.2785*.
117. Bradley, P.S., & Mangasarian, O. L (2000). Massive data discrimination via linear support vector machines. *Optimization methods and software*, 13(1), 1-10.
118. Boser, B.E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.
119. Pourhoseingholi, M.A., Baghestani, A. R., & Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*, 5(2), 79.
120. McNamee, R. (2005). Regression modelling and other methods to control confounding. *Occupational and environmental medicine*, 62(7), 500-506.
121. Sedgwick, P. (2013). Analysing case-control studies: adjusting for confounding. *Bmj*, 346.
122. Snoek, L., Miletić, S. & Scholte, H.S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage*, 184, 741-760.
123. Sanchis-Segura, C., Ibañez-Gual, M.V., Aguirre, N., Cruz-Gómez, Á.J. & Forn, C. (2020). Effects of different intracranial volume correction methods on univariate

- sex differences in grey matter volume and multivariate sex prediction. *Scientific Reports*, 10(1), 1-15.
124. Kahlert, J., Gribsholt, S.B., Gammelager, H., Dekkers, O.M. & Luta, G. (2017). Control of confounding in the analysis phase - an overview for clinicians. *Clinical epidemiology*, 195-204.
 125. Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á. J., Félix, S., & Forn, C. (2022). Beyond “Sex Prediction”: Estimating and Interpreting Multivariate Sex Differences and Similarities in the Brain. *NeuroImage*, 257, 119343.
 126. More, S., Eickhoff, S. B., Caspers, J., & Patil, K. R. (2021). Confound removal and normalization in practice: A neuroimaging based sex prediction case study. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, Springer International Publishing.*, 3-18.
 127. Caspers, S., Moebus, S., Lux, S., Pundt, N., Schutz, H., Muhleisen, T.W., . . . & Amunts, K. (2014). Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Frontiers in aging neuroscience*, 6, 149.
 128. Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y. & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*, 145, 166-179.
 129. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., . . . & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4), 869.
 130. Ikram, M.A., van der Lugt, A., Niessen, W.J., Krestin, G.P., Koudstaal, P.J., Hofman, A., . . . & Vernooij, M.W. (2011). The Rotterdam Scan Study: design and update up to 2012. *European journal of epidemiology*, 26, 811-824.
 131. Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., . . . & WU-Minn HCP Consortium. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4), 2222-2231.
 132. Geerligs, L., Renken, R.J., Saliasi, E., Maurits, N.M. & Lorist, M.M. (2015). A Brain-Wide Study of Age-Related Changes in Functional Connectivity. *Cerebral Cortex*, 25(7), 1987-1999.
 133. Song, J., Birn, R.M., Boly, M., Meier, T.B., Nair, V.A., Meyerand, M.E. & Prabhakaran, V. (2014). Age-related reorganizational changes in modularity and functional connectivity of human brain networks. *Brain connectivity*, 4(9), 662-676.
 134. Farras-Permanyer, L., Mancho-Fora, N., Montalà-Flaquer, M., Bartrés-Faz, D., Vaqué-Alcázar, L., Però-Cebollero, M. & Guàrdia-Olmos, J. (2019). Age-related changes in resting-state functional connectivity in older adults. *Neural regeneration research*, 14(9), 1544-1555.
 135. Hedden, T. & Gabrieli, J.D. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nature reviews neuroscience*, 5(2), 87-96.
 136. Schaie, K.W. & Willis, S.L. (2010). The Seattle Longitudinal Study of Adult Cognitive Development. *ISSBD Bull*, 57(1), 24-29.
 137. Habib, R., Nyberg, L. & Nilsson, L.G. (2007). Cognitive and non-cognitive factors contributing to the longitudinal identification of successful older adults in the betula study. *Aging, Neuropsychology, and Cognition*, 14(3), 257-273.

138. Le Carret, N., Lafont, S., Mayo, W. & Fabrigoule, C. (2003). The effect of education on cognitive performances and its implication for the constitution of the cognitive reserve. *Developmental neuropsychology*, 23(3), 317-337.
139. Biswal, B.B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., . . . & Milham, M.P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10), 4734–4739.
140. Allen, E.A., Erhardt, E.B., Damaraju, E., Gruner, W., Segall, J.M., Silva, R.F., . . . & Calhoun, V.D. (2011). A baseline for the multivariate comparison of resting-state networks. *Frontiers in systems neuroscience*, 5, 2.
141. Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G. & Rocca, M.A. (2013). The organization of intrinsic brain activity differs between genders: a resting-state fMRI study in a large cohort of young healthy subjects. *Human Brain Mapping*, 34(6), 1330-43.
142. Peña, D., Contreras, M. J., Shih, P. C., & Santacreu, J. (2008). Solution strategies as possible explanations of individual and sex differences in a dynamic spatial task. *Acta Psychologica*, 128(1), 1-14.
143. Roalf, D., Lowery, N. & Turetsky, B.I. (2006). Behavioral and physiological findings of gender differences in global-local visual processing. *Brain and cognition*, 60(1), 32-42.
144. Pletzer, B., Scheuringer, A. & Scherndl, T. (2017). Global-local processing relates to spatial and verbal processing: implications for sex differences in cognition. *Scientific reports*, 7(1), 10575.
145. Pletzer, B. (2014). Sex-specific strategy use and global-local processing: a perspective toward integrating sex differences in cognition. *Frontiers in neuroscience*, 8, 425.
146. Zubiaurre-Elorza, L., Junque, C., Gómez-Gil, E., & Guillamon, A. (2014). Effects of cross-sex hormone treatment on cortical thickness in transsexual individuals. *The journal of sexual medicine*, 11(5), 1248-1261.
147. Baldinger-Melich, P., Urquijo Castro, M.F., Seiger, R., Ruef, A., Dwyer, D.B., Kranz, G.S., . . . & Koutsouleris, N. (2020). Sex Matters: A Multivariate Pattern Analysis of Sex- and Gender-Related Neuroanatomical Differences in Cis- and Transgender Individuals Using Structural Magnetic Resonance Imaging. *Cerebral Cortex*, 30(3), 1345-1356.
148. Dhamala, E., Bassett, D.S., Yeo, B.T. & Holmes, A.J. (2024). Functional brain networks are associated with both sex and gender in children. *Science Advances*, 10(28), eadn4202.
149. Clemens, B., Derntl, B., Smith, E., Junger, J., Neulen, J., Mingoia, G., . . . & Habel, U. (2020). Predictive Pattern Classification Can Distinguish Gender Identity Subtypes from Behavior and Brain Imaging. *Cerebral Cortex*, 30(5), 2755-2765.
150. 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. *Frontiers in human neuroscience*, 8, 650.
151. McEwen, B.S. & Milner, T.A. (2017). Understanding the broad influence of sex hormones and sex differences in the brain. *Journal of neuroscience research*, 95(1-2), 24-39.
152. Coenjaerts, M., Adrovic, B., Trimborn, I., Philipsen, A., Hurlemann, R. & Scheele, D. (2023). Effects of exogenous oxytocin and estradiol on resting-state functional connectivity in women and men. *Scientific reports*, 13(1), 3113.

153. Kogler, L., Müller, V.I., Seidel, E.M., Boubela, R., Kalcher, K., Moser, E., . . . & Derntl, B. (2016). Sex differences in the functional connectivity of the amygdalae in association with cortisol. *Neuroimage*, 134, 410-423.
154. Zsido, R.G., Williams, A. N., Barth, C., Serio, B., Kurth, L., Mildner, T., ... & Sacher, J. (2023). Ultra-high-field 7T MRI reveals changes in human medial temporal lobe volume in female adults during menstrual cycle. *Nature Mental Health*, 1(10), 761-771.
155. Haraguchi, R., Hoshi, H., Ichikawa, S., Hanyu, M., Nakamura, K., Fukasawa, K., ... & Shigihara, Y. (2021). The menstrual cycle alters resting-state cortical activity: a magnetoencephalography study. *Frontiers in human neuroscience*, 15, 652789.
156. Weis, S., Hodgetts, S. & Hausmann, M. (2019). Sex differences and menstrual cycle effects in cognitive and sensory resting state networks. *Brain and cognition*, 131, 66-73.
157. Arélin, K., Mueller, K., Barth, C., Rekkas, P.V., Kratzsch, J., Burmann, I., . . . & Sacher, J. (2015). Progesterone mediates brain functional connectivity changes during the menstrual cycle-a pilot resting state MRI study. *Frontiers in neuroscience*, 9, 44.
158. Avila-Varela, D.S., Hidalgo-Lopez, E., Dagnino, P. C., Acero-Pousa, I., del Agua, E., Deco, G., ... & Escrichs, A. (2024). Whole-brain dynamics across the menstrual cycle: the role of hormonal fluctuations and age in healthy women. *npj Women's Health*, 2(1), 8.
159. Schmalenberger, K.M., Tauseef, H.A., Barone, J.C., Owens, S.A., Lieberman, L., Jarczok, M.N., . . . & Eisenlohr-Moul, T.A. (2021). How to study the menstrual cycle: Practical tools and recommendations. *Psychoneuroendocrinology*, 123, 104895.
160. Pritschet, L., Taylor, C.M., Santander, T. & Jacobs, E.G. (2021). Applying dense-sampling methods to reveal dynamic endocrine modulation of the nervous system. *Current opinion in behavioral sciences*, 40, 72-78.
161. Crestol, A., Rajagopal, S., Lissaman, R., LaPlume, A., Pasvanis, S., Olsen, R., ... & Rajah, M. N. (2023). The influence of chronological age and menopause status on the functional neural correlates of spatial context memory in middle-aged females. *bioRxiv*.
162. Pletzer, B., Kronbichler, M., Aichhorn, M., Bergmann, J., Ladurner, G. & Kerschbaum, H.H. (2010). Menstrual cycle and hormonal contraceptive use modulate human brain structure. *Brain research*, 1348, 55-62.
163. Heller, C., Kimmig, A.S., Kubicki, M.R., Derntl, B. & Kikinis, Z. (2022). Imaging the human brain on oral contraceptives: A review of structural imaging methods and implications for future research goals. *Frontiers in Neuroendocrinology*, 67, 101031.
164. Taylor, C.M., Pritschet, L., Olsen, R.K., Layher, E., Santander, T., Grafton, S.T. & Jacobs, E.G. (2020). Progesterone shapes medial temporal lobe volume across the human menstrual cycle. *Neuroimage*, 220, 117125.
165. Gurvich, C., Thomas, N. & Kulkarni, J. (2020). Sex differences in cognition and aging and the influence of sex hormones. *Handbook of clinical neurology*, 175, 103-115.
166. Gur, R.E. & Gur, R.C. (2002). Gender differences in aging: cognition, emotions, and neuroimaging studies. *Dialogues in clinical neuroscience*, 4(2), 197-210.
167. Gur, R.C., Mozley, P.D., Resnick, S.M., Gottlieb, G.L., Kohn, M., Zimmerman, R., . . . & Berretta, D. (1991). Gender differences in age effect on brain atrophy

- measured by magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 88(7), 2845-2849.
168. Costantino, S. & Paneni, F. (2020). Sex-related differences in the ageing brain: time for precision medicine? *Cardiovascular Research*, 116(7), 1246-1248.
 169. Rippon, G., Jordan-Young, R., Kaiser, A., Joel, D. & Fine, C. (2017). Journal of neuroscience research policy on addressing sex as a biological variable: Comments, clarifications, and elaborations. *Journal of neuroscience research*, 95(7), 1357-1359.
 170. Takeuchi, H., Taki, Y., Nouchi, R., Yokoyama, R., Kotozaki, Y., Nakagawa, S., . . . & Kawashima, R. (2018). The Effects of Family Socioeconomic Status on Psychological and Neural Mechanisms as Well as Their Sex Differences. *Frontiers in human neuroscience*, 12, 543.
 171. Liu, S.Z., Tahmasebi, G., Sheng, Y., Dinov, I.D., Tsilimingras, D. & Liu, X. (2023). Sex difference in the associations of socioeconomic status, cognitive function and brain volume with dementia in old adults: Findings from the OASIS study. *medRxiv*.
 172. Stadler, G., Chesaniuk, M., Haering, S., Roseman, J., Straßburger, V. M., Martina, S., ... & Mine, W. (2023). Diversified innovations in the health sciences: Proposal for a Diversity Minimal Item Set (DiMIS). *Sustainable chemistry and pharmacy*, 33, 101072.

Acknowledgements

My deepest gratitude goes to PD Dr. Susanne Weis for her support and supervision over the years. Working with her has been a pleasure, and I have gained a wealth of knowledge and insights, which have facilitated my growth. She empowered me to think independently and anticipate challenges. Her support has been crucial on my journey to making me the scientist I am today.

I would also like to thank Prof. Dr. Christian Bellebaum, as I started my journey in neuroscience in his department with my bachelor's thesis and since then, he has guided me through both my master's and doctoral theses as my co-supervisor, consistently providing invaluable feedback and support.

Special thanks are extended to Prof. Dr. Simon Eickhoff for introducing me to the Institute and guiding my scientific journey since my Master's degree. Working in the INM-7 provided me with an incredible amount of resources and opportunities for scientific exchange, profoundly shaping my research path. He is an inspirational leader who is always willing to help, provide feedback, and offer support, for which I am truly grateful.

I would also like to thank all members of the Brain Variability working group and those at the INM-7. The scientific exchange and feedback helped to develop my work, and the laughter at retreats and social meetings was filled with so much joy. The expertise coming from different areas encouraged me to view my work from multiple perspectives and also to take a step back to see the bigger picture. In addition to the scientific exchange, the personal exchange with other PhD students was very special for me. I would like to thank all of them who shared this journey with me, especially Gianna Kuhles, Lisa Mochalski, Vera Komeyer, Shammi More, Lya Paas Oliveros, Nevena Kraljević, Lisa Hahn, Ann-Christin Kimmig, and Kevin Wischnewski. Thank you for so many joyful moments during our time together at the Institute and conferences.

I am truly thankful to my colleagues in my new working group - Elvisha Dhamala, Katharina Brosch, and Erynn Christensen - for their constant support and understanding during the challenging transition period.

Lastly, I would like to thank my family - my parents, siblings, and grandparents – I am deeply grateful for their endless support, belief in my potential, and pride in my journey. Their encouragement has helped me reach places I never imagined I was capable of. Last but not least, I want to thank Dominik for simply everything, but above all for his unconditional support and understanding in every aspect throughout all these years.