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

Short Title: Brain size bias in sex classification

Authors

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]*, and Susanne Weis[1,2]*

Affiliations

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

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

[3]Department of Psychiatry, Psychotherapy and Psychosomatics, Faculty of Medicine, RWTH Aachen University, Aachen, Germany.

[4]Institute of Neuroscience and Medicine (INM-10: Decoding the human brain at systematic levels), Research Centre Jülich, Jülich, Germany.

[5]Department of Psychiatry and Psychotherapy, Tübingen Center for Mental Health, University of Tübingen, Tübingen, Germany.

[6]LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany.

*Equal contribution, corresponding authors: {s.weis, k.patil}@fz-juelich.de

Abstract

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.

Introduction

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 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 differ [21]. In the present paper, following the linguistic guidelines provided by the Professional Association of Transgender Health [22], 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 putamen [23], insula [16] as well as in surface areas, cortical and subcortical brain volumes [24]. Additionally, transgender individuals undergoing cross-sex hormone treatment (CHT) were reported to show structural alterations in the hypothalamus and the third ventricle [25]. 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 men [25]. 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 TIV [10] 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 (Figure 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 suggested [17].

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 (Figure S1, 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 study [16]. 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 Figure 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 (ATM+cr model, Figure 1). Subsequently, all models were calibrated to ensure that the prediction probabilities of the models match the respective class label (Figure S2-3, 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 processes [8], 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 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 Figure S1 and Table S1 for demographic details and TIV distribution of the samples).

***insert Figure 1 about here ***

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, Figure 2). Accordingly, the assigned probability of being classified as male (prediction probability) was higher for men than for women (Figure 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 (Figure 3b). Similarly, women classified incongruently with their sex as male had 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% (Table 1, Table 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 (Figure 3c), significantly higher TIV in sex congruently as opposed to incongruently classified men, and significantly lower TIV in sex congruently as opposed to incongruently classified women (Figure 3d, 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 Figure 2, Table 1 and Table S2) hold-out samples. In comparison to the AM model with no TIV control (Figure 3a) prediction probability displayed a much larger overlap between women and men (Figure 3e and 3g). Further evaluation did not reveal any evidence for a TIV bias — i.e. neither did sex congruently classified men show higher TIV than incongruently classified men nor did sex congruently classified women show lower TIV than incongruently classified women in both the AM (Figure 3f) and the ATM (Figure 3h, 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 Table 1 and Table 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 (Figure 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 (Figure 3j, n, Table 2). Application of these models to the ATM hold-out sample (details in Table 1, Table S2), displayed better performance (92.55%) than for the AM hold-out sample. Furthermore, prediction probability distributions showed a comparable (Figure 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 (Figure 3l, Table 2) and with additional confound removal (Figure 3p, 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 (Figure S2, 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 Figure S4.

***insert Figure 2 about here ***

***insert Figure 3 about here ***

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 Table 1 and S3). Likewise, for sample B, the model achieved high overall accuracy of 93.10% (Table 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 (Figure 4a, c) and sample B (Figure 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 (Figure S1).

The comparison of prediction probabilities for CM vs. 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 classified had a higher TIV than those classified sex incongruently (Figure 4b, d, f, h), indicating a similar bias of this model for both cisgender and transgender individuals.

***insert Figure 4 about here ***

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 (Figure 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 -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 (Figure 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 accuracy [16]. 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 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 (Figure 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 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 GMV [34]. 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 sex [30]. 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 analysis [9]. 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 population [20]. 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 performances [8]. 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 (Figure 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 previously [25] TIV-biased models would misclassify transgender individuals in accordance with their gender identity, which could explain prior findings [16]. 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 data [17], 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 brain [37], 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 classification [41]. 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 results [16], 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 1000Brains, 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 Figure 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 Figure 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 088/09, [23]). 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 Figure S1.

The second application sample (Sample B) consisted of an open-source dataset acquired in Barcelona, available via (https://data.mendeley.com/datasets/hjmfrv6vmg/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 Figure 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-striping and segmentation by an adaptive maximum a posteriori approach [50] with using a partial volume model [51]. Subsequently, an optimized version of the Geodesic Shooting Algorithm [52] 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 3x3x3 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-naive.

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 ATM+cr model (Figure 1). For all four models, we employed a SVM classifier with rbf kernel [53] using Julearn (https://juaml.github.io/julearn). Before training the classifier, PCA was performed to reduce the dimensionality of the data [16]. The maximum number of components (n = 1292, number 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 10-fold 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 5-fold CV inner loop following the analysis employed in a previous study [16]. 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 (Figure 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 Figure S2-3, 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 plots [54] 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 accordingly [55]. 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 (Table S5, 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].

References

1. Willemink, M.J., et al., *Preparing Medical Imaging Data for Machine Learning.* Radiology, 2020. **295**(1): p. 4-15.
2. Buch, V.H., I. Ahmed, and M. Maruthappu, *Artificial intelligence in medicine: current trends and future possibilities.* Br J Gen Pract, 2018. **68**(668): p. 143-144.
3. Chang, K., et al., *Distributed deep learning networks among institutions for medical imaging.* J Am Med Inform Assoc, 2018. **25**(8): p. 945-954.
4. Jollans, L., et al., *Quantifying performance of machine learning methods for neuroimaging data.* Neuroimage, 2019. **199**: p. 351-365.
5. Davatzikos, C., *Machine learning in neuroimaging: Progress and challenges.* Neuroimage, 2019. **197**: p. 652-656.
6. Nielsen, A.N., et al., *Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry.* Biol Psychiatry Cogn Neurosci Neuroimaging, 2020. **5**(8): p. 791-798.
7. Kahlert, J., et al., *Control of confounding in the analysis phase - an overview for clinicians.* Clin Epidemiol, 2017. **9**: p. 195-204.

8. Pourhoseingholi, M.A., Baghestani, A. R., & Vahedi, M., *How to control confounding effects by statistical analysis.* Gastroenterology and hepatology from bed to bench, 2012. **5**(2): p. 79.
9. Sedgwick, P., *Analysing case-control studies: adjusting for confounding.* Bmj, 2013. **346**.
10. McNamee, R., *Regression modelling and other methods to control confounding.* Occup Environ Med, 2005. **62**(7): p. 500-506.
11. Feis, D.-L., et al., *Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data.* NeuroImage, 2013. **70**: p. 250–257.
12. Chekroud, A.M., et al., *Patterns in the human brain mosaic discriminate males from females.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(14): p. E1968.
13. Bzdok, D., *Classical Statistics and Statistical Learning in Imaging Neuroscience.* Front Neurosci, 2017. **11**: p. 543.
14. Weis, S., et al., *Sex Classification by Resting State Brain Connectivity.* Cereb Cortex, 2020. **30**(2): p. 824-835.
15. 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, 2012. **61**(4): p. 931-940.
16. Flint, C., et al., *Biological sex classification with structural MRI data shows increased misclassification in transgender women.* Neuropsychopharmacology, 2020.
17. 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, 2020. **30**(3): p. 1345-1356.

18. Eliot, L., et al., *Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size.* Neurosci Biobehav Rev, 2021: p. 667-697.
19. Kaczkurkin, A.N., A. Raznahan, and T.D. Satterthwaite, *Sex differences in the developing brain: insights from multimodal neuroimaging.* Neuropsychopharmacology, 2019. **44**(1): p. 71-85.
20. Snoek, L., S. Miletic, and H.S. Scholte, *How to control for confounds in decoding analyses of neuroimaging data.* Neuroimage, 2019. **184**: p. 741-760.
21. Smith, E., et al., *Gender incongruence and the brain - Behavioral and neural correlates of voice gender perception in transgender people.* Horm Behav, 2018. **105**: p. 11-21.
22. Bouman, W.P., Schwend, A. S., Motmans, J., Smiley, A., Safer, J. D., Deutsch, M. B., ... & Winter, S., *Language and trans health.* International Journal of Transgenderism, 2017. **18**(1): p. 1-6.
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, 2021. **10**(7): p. 1454.
24. Mueller, S.C., et al., *The Neuroanatomy of Transgender Identity: Mega-Analytic Findings From the ENIGMA Transgender Persons Working Group.* J Sex Med, 2021. **18**(6): p. 1122-1129.
25. Pol, H.E.H., Cohen-Kettenis, P. T., Van Haren, N. E., Peper, J. S., Brans, R. G., Cahn, W., ... & Kahn, R. S., *Changing your sex changes your brain: influences of testosterone and estrogen on adult human brain structure.* European Journal of Endocrinology, 2006. **155**(suppl_1): p. S107-S114.

26. Spizzirri, G., et al., *Grey and white matter volumes either in treatment-naive or hormone-treated transgender women: a voxel-based morphometry study.* Sci Rep, 2018. **8**(1): p. 1-10.
27. Zubiaurre-Elorza, L., Junque, C., Gómez-Gil, E., & Guillamon, A. , *Effects of cross-sex hormone treatment on cortical thickness in transsexual individuals.* The journal of sexual medicine, 2014. **11**(5): p. 1248-1261.
28. Fukao, T., K. Ohi, and T. Shioiri, *Gray matter volume differences between transgender men and cisgender women: A voxel-based morphometry study.* Aust N Z J Psychiatry, 2022. **56**(5): p. 535-541.
29. Kurth, F., et al., *Brain Sex in Transgender Women Is Shifted towards Gender Identity.* J Clin Med, 2022. **11**(6): p. 1582.
30. More, S., Eickhoff, S. B., Caspers, J., & Patil, K. R., *Confound removal and normalization in practice: A neuroimaging based sex prediction case study.* Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2021: p. 3-18.
31. Chyzhyk, D., Varoquaux, G., Milham, M., & Thirion, B., *How to remove or control confounds in predictive models, with applications to brain biomarkers.* GigaScience, 2022. **11**.
32. Malone, I.B., et al., *Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance.* Neuroimage, 2015. **104**: p. 366-372.
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, 2022. **119343**.
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, 2020. **10**(1): p. 1-15.

35. Farias, F., Ludermir, T., & Bastos-Filho, C., *Similarity Based Stratified Splitting: an approach to train better classifiers.* arXiv preprint arXiv: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.* Mathematical Problems in Engineering, 2020. **2020**.
37. Joel, D., et al., *Sex beyond the genitalia: The human brain mosaic.* Proceedings of the National Academy of Sciences, 2015. **112**(50): p. 15468–15473.
38. Bao, A.M. and D.F. Swaab, *Sex differences in the brain, behavior, and neuropsychiatric disorders.* The Neuroscientist, 2010. **16**(5): p. 550-65.
39. Bao, A.M. and D.F. Swaab, *Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorders.* Front Neuroendocrinol, 2011. **32**(2): p. 214-26.
40. Miller, V.M., W.A. Rocca, and S.S. Faubion, *Sex Differences Research, Precision Medicine, and the Future of Women's Health.* J Womens Health (Larchmt), 2015. **24**(12): p. 969-71.
41. Ruiz-Serra, V., Buslón, N., Philippe, O. R., Saby, D., Morales, M., Pontes, C., ... & Cirillo, D., *Addressing sex bias in biological databases worldwide.* https://biohackrxiv.org/n9dkg/, 2023.
42. Weber, K.A., Teplin, Z. M., Wager, T. D., Law, C. S., Prabhakar, N. K., Ashar, Y. K., ... & Mackey, S., *Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction.* Frontiers in Neurology, 2022. **13**.
43. Resnick, S.M., et al., *One-year age changes in MRI brain volumes in older adults.* Cereb Cortex, 2000. **10**(5): p. 464-472.
44. Good, C.D., et al., *A voxel-based morphometric study of ageing in 465 normal adult human brains.* Neuroimage, 2001. **14**(1 Pt 1): p. 21-36.

45. Resnick, S.M., et al., *Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain.* J Neurosci, 2003. **23**(8): p. 3295-3301.
46. Taki, Y., et al., *Correlations among Brain Gray Matter Volumes, Age, Gender, and Hemisphere in Healthy Individuals.* Plos One, 2011. **6**(7): p. e22734.
47. Uribe, C., *original data of a functional MRI study in transgender individual, Mendeley Data, V2, doi: 10.17632/hjmfrv6vmg.* 2020.
48. Uribe, C., et al., *Data for functional MRI connectivity in transgender people with gender incongruence and cisgender individuals.* Data Brief, 2020. **31**: p. 105691.
49. Uribe, C., et al., *Brain network interactions in transgender individuals with gender incongruence.* Neuroimage, 2020. **211**: p. 116613.
50. Rajapakse, J.C., Giedd, J. N., & Rapoport, J. L., *Statistical approach to segmentation of single-channel cerebral MR images.* IEEE transactions on medical imaging, 1997. **16**(2): p. 176-186.
51. Tohka, J., A. Zijdenbos, and A. Evans, *Fast and robust parameter estimation for statistical partial volume models in brain MRI.* Neuroimage, 2004. **23**(1): p. 84-97.
52. Ashburner, J. and K.J. Friston, *Unified segmentation.* NeuroImage, 2005. **26**(3): p. 839–851.
53. Boser, B.E., Guyon, I. M., & Vapnik, V. N., *A training algorithm for optimal margin classifiers.* Proceedings of the fifth annual workshop on Computational learning theory, 1992: p. 144-152.
54. Bechtold, B., *Violin Plots for Matlab, Github Project https://github.com/bastibe/Violinplot-Matlab, DOI: 10.5281/zenodo.4559847.* 2016.
55. Fritz, C.O., P.E. Morris, and J.J. Richler, *" 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, 2007. **39**(2): p. 175-191.
57. Faul, F., et al., *Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses.* Behav Res Methods, 2009. **41**(4): p. 1149-1160.

Acknowledgments

Funding: 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.

Data and materials 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

Additional information

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.

Figure legends

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

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

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)

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)

Table 1. Performance of models.

	Model performance for the AM hold-out sample			
	AM model	**AM+cr model**	**ATM model**	**ATM+cr model**
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			
	AM model	**AM+cr model**	**ATM model**	**ATM+cr model**
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			
	AM model	**AM+cr model**	**ATM model**	**ATM+cr model**
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			
	AM model	**AM+cr model**	**ATM model**	**ATM+cr model**
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

Model performance of all models applied to the hold-out and application samples.

Table 2. Wilcoxon rank sum tests of the hold-out samples.

	TIV women classified as female vs. classified as male	TIV men classified as male vs. classified as female
	AM hold-out sample	
AM model	$T = 12722$, z = -2.3885, $p = 0.0169$, $\eta^2 = 0.0354$	$T = 12829$, 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	$T = 11004$, z = -0.4390, $p = 0.6606$, $\eta^2 = 0.0012$	$T = 11507$, z = 0.0236, $p = 0.9812$, $\eta^2 < 0.001$
ATM+cr model	$T = 11236$, z = 0.2778, $p = 0.7812$, $\eta^2 < 0.001$	$T = 11284$, z = 0.5097, $p = 0.6103$, $\eta^2 = 0.0016$
	ATM hold-out sample	
AM model	$T = 9908$, z = -4.7156, $p < 0.001$, $\eta^2 = 0.1381$	$T = 11325$, z = 6.2257, $p < 0.001$, $\eta^2 = 0.2407$
AM+cr model	$T = 8425$, z = 0.8513, p = 0.3946, $\eta 2 = 0.0045$	$T = 10341$, z = -2.3190, $p = 0.0204$, $\eta^2 = 0.0334$
ATM model	$T = 12284$, z = 1.3806, $p = 0.1674$, $\eta^2 = 0.0118$	$T = 12239$, z = 1.0910, $p = 0.2753$, $\eta^2 = 0.0074$
ATM+cr model	$T = 12403$, z = 1.6918, $p = 0.0907$, $\eta^2 = 0.0178$	$T = 12130$, z = 0.8780, $p = 0.3800$, $\eta^2 = 0.0048$

Comparison of individuals classified as female vs. male (Wilcoxon rank sum tests) for the AM and ATM sample.

Table 3. Wilcoxon rank sum tests of the application samples.

a)	TIV CW classified as female vs. classified as male	TIV CM classified as male vs. 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, $p = 0.7387$, $\eta^2 = 0.0045$	*no CM classified as female*
ATM+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 vs. classified as male	**TIV TW classified as male vs. 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, $p = 0.0652$, $\eta^2 = 0.1030$	*no TW classified as female*
ATM+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 vs. classified as male	TIV CM classified as male vs. 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, $p = 0.7842$, $\eta^2 = 0.0039$
ATM+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 vs. classified as male	**TIV TW classified as male vs. classified as female**
AM model	*no TM classified as male*	$T = 145$, 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, $p = 0.1421$, $\eta^2 = 0.0743$	*no TW classified as female*
ATM+cr model	$T = 411$, z = 1.4680, $p = 0.1421$, $\eta^2 = 0.0743$	*no TW classified as female*

Comparison of individuals classified as female vs. male (Wilcoxon rank sum tests) for application sample A (a) and sample (b).