

Supplementary Materials for
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*}, and
Susanne Weis^{1,2*}

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

This file includes:

Supplementary Text
Figs. S1 to S4
Tables S1 to S6

Supplementary Results

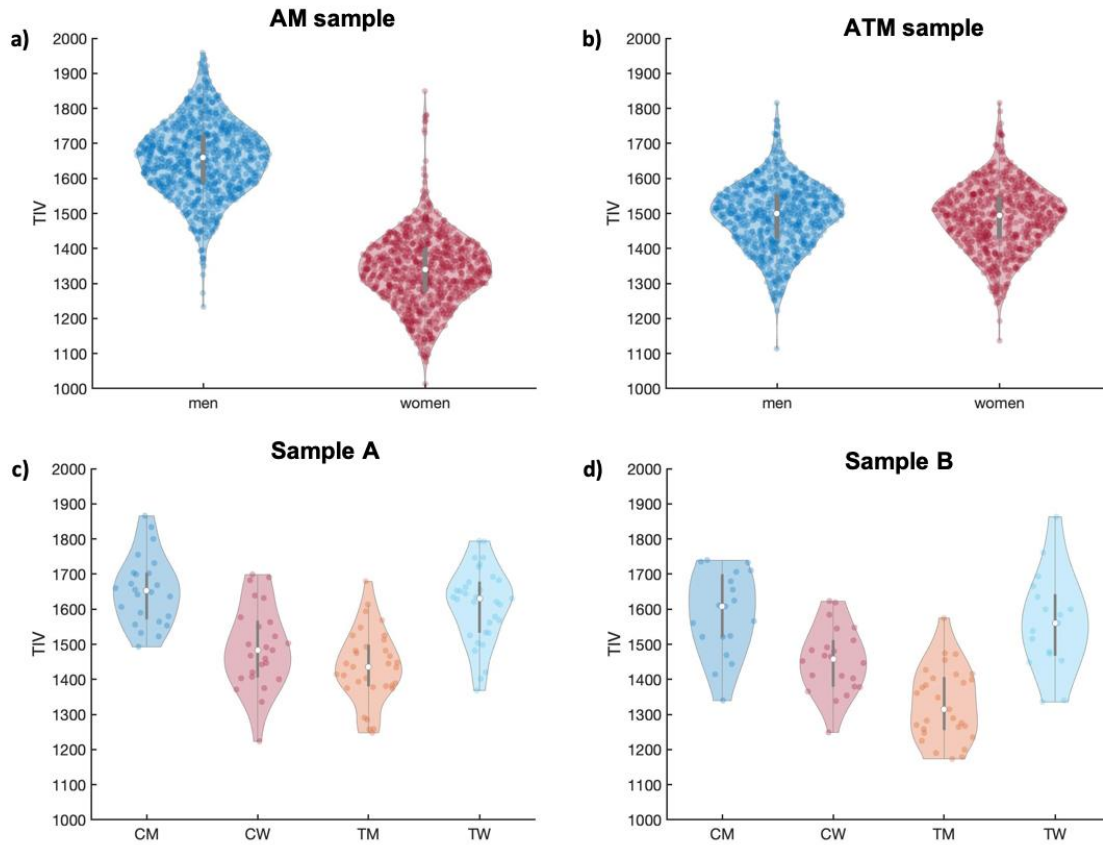


Figure. S1. TIV distribution of all samples.

- a) AM sample
- b) ATM sample
- c) Sample B
- d) Sample A

Shades of red and orange indicate samples comprising individuals with a female sex, while blue shades denote samples comprising individuals with a male sex.

Calibration results

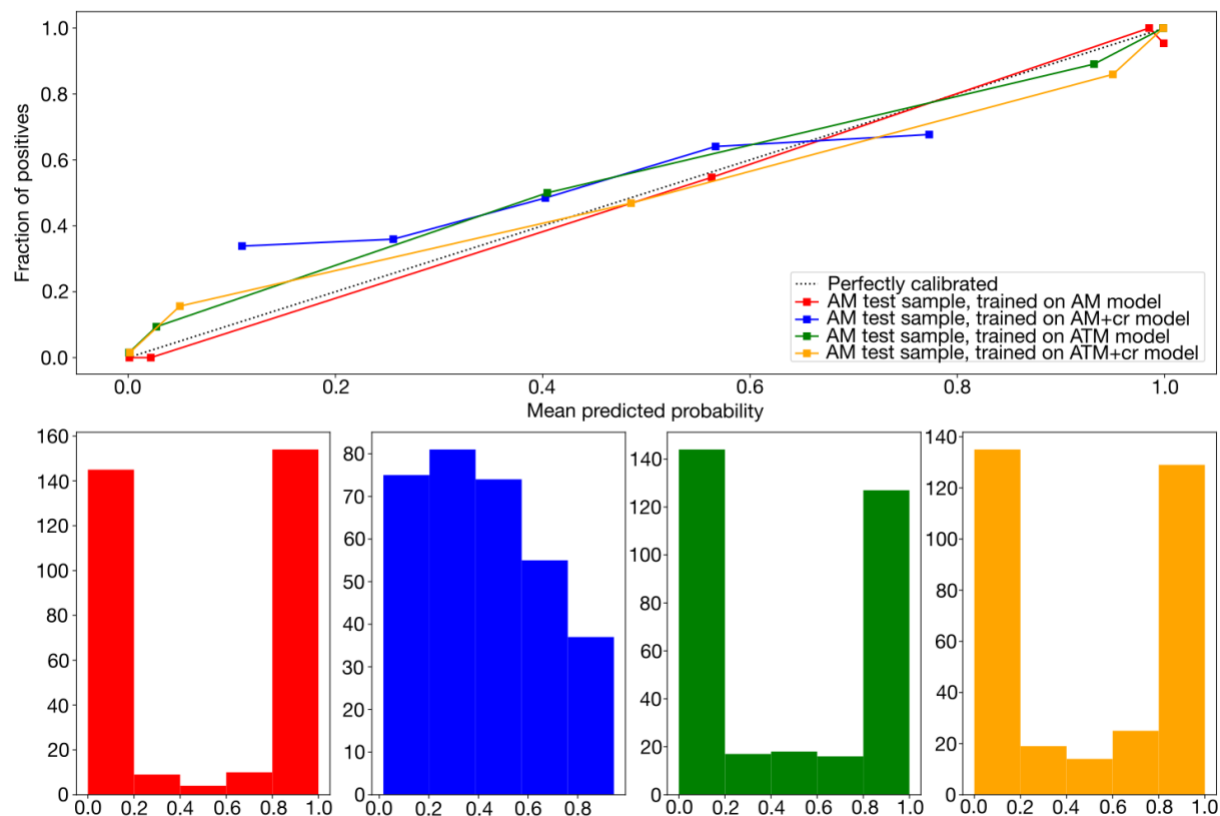


Figure S2. Calibration curves for all four models applied to the AM test sample. The plot in the upper row depicts the true frequency of the positive labels in relation to the respective predicted probability. The x-axis represents the average predicted probability and the y-axis represents the fraction of positives, meaning the proportion of samples whose class is the positive class.

The plots in the bottom line provide closer insight into the behavior of each classifier by showing the number of samples in each predicted probability bin (<https://scikit-learn.org/stable/modules/calibration.html#calibration>). The AM model (red) returned close to perfectly calibrated predictions for the AM test sample with respectively very high or very low probabilities for being classified as male when classifying men and women. In contrast, the AM+cr model (blue) showed a distribution of probabilities that neither started close to 0 nor ended close to 1, but was rather located somewhere in the middle of the probability spectrum,

indicating that across the sample (independent of sex) probabilities of being classified as male was neither close to 0 or 1. However, both ATM models (green and yellow) also showed a close to perfect calibration with two peaks at the respective end of the probability continuum, but showed more counts in the middle of the spectrum than the AM model, meaning that in both ATM models, most of the individuals were classified with a very high or low probability, but some were also assigned with a probability score between 0.2 and 0.8.

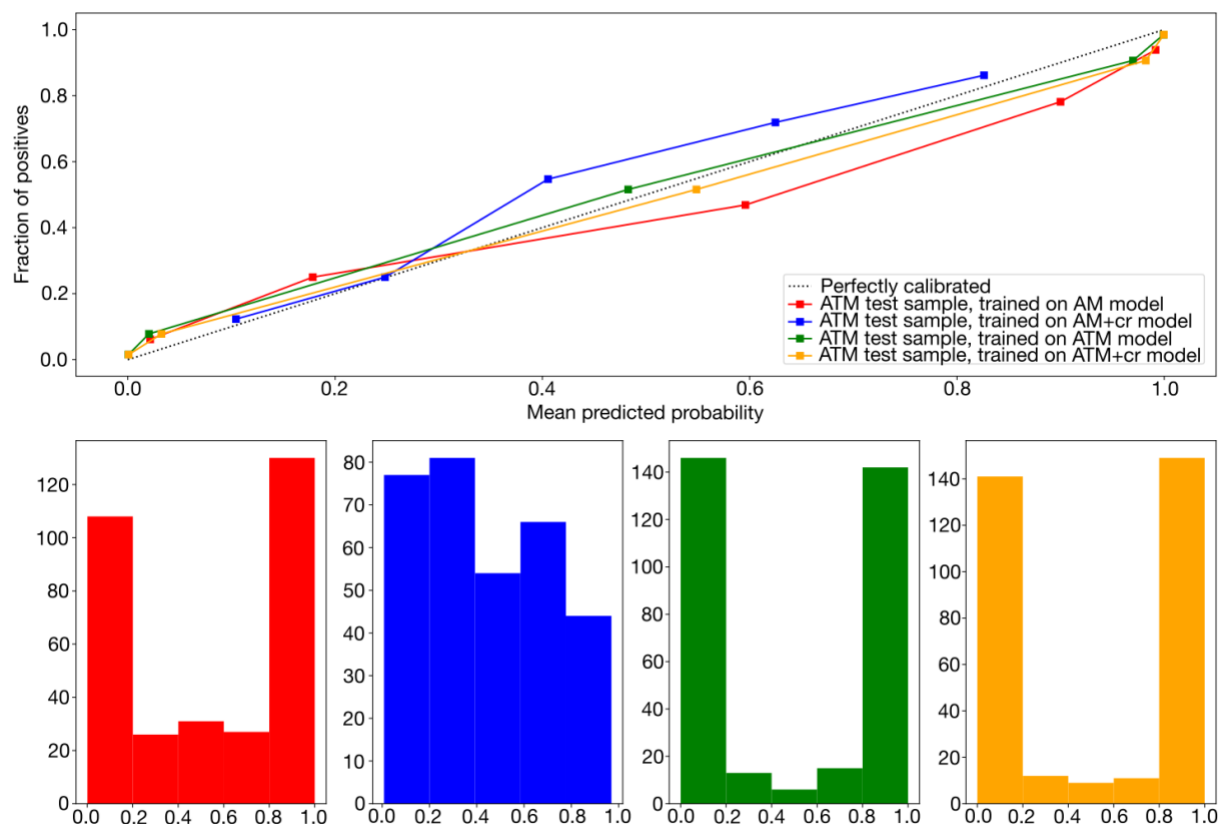


Figure S3. Calibration curves for all four models applied to the ATM test sample. The plot in the upper row depicts the true frequency of the positive labels in relation to the respective predicted probability. The x-axis represents the average predicted probability and the y-axis represents the fraction of positives, meaning the proportion of samples whose class is the positive class. The plots in the bottom line give closer insight into the behavior of each classifier by showing the number of samples in each predicted probability bin.

For the ATM test sample, the AM model (red) was also well calibrated and again showed two peaks at the probability close to 0.0 and 1.0, but in contrast to application to the AM sample (Figure S2), here, the model showed more counts distributed across the probability spectrum.

The AM+cr model (blue) showed a similar distribution in counts in the bottom line, but the calibration curve indicated a better calibration than for the AM sample.

Again, both ATM models (green and yellow) showed a close to perfect calibration with more pronounced peaks at the respective ends of the probability continuum. This means that the application of both ATM models to the ATM test sample resulted in a model behavior that

classified women with a very low probability and men with a very high probability.

We strongly encourage future studies to also inspect the behavior of a model according to the probabilities in calibration curves to gain confidence in the prediction. The mean predicted probability should correspond to the amount of positive predicted subjects.

To ensure that total GMV did not bias our results, we conducted Wilcoxon rank sum tests to compare the GMV distributions of individuals classified in accordance with their sex to those who were not for all four models. A significant outcome would suggest biased predictions with respect to GMV, similar to the tests done for TIV (Table 2-3). While we did observe a TIV-bias in the results of the AM model as men were classified in congruence with their sex as male had a significantly higher TIV than incongruently classified men and vice versa for women (Table 2), we did not observe a biased output pattern for GMV (all Wilcoxon rank sum tests were not significant at $\alpha = 0.005$ (Table S5, S6)).

Hyperparameter tuning

For each of the four models, a Bayesian hyperparameter search identified an optimal combination of hyperparameters: For the AM model $C = 113.49$ and $\gamma = 5.79$ were selected. These parameters changed only slightly when including TIV as a confound in the AM+cr model, resulting in $C = 116.40$ and $\gamma = 3.22$. Higher C values were selected for the ATM-trained models (ATM: $C = 1318.09$ and $\gamma = 1.64$; ATM+cr: $C = 6155.03$ and $\gamma = 3.23$).

Effects of TIV confound removal in application data

The application of the AM+cr model to sample A resulted in an overall accuracy of 75.65% with 77.55% for cisgender and 74.24% for transgender individuals. A similar pattern was found for sample B with an accuracy of 68.97%, with an accuracy of 78.05% for cisgender and 60.87% for transgender individuals (details in Table 1 and S3). In both samples, the prediction probabilities demonstrated a large overlap for cis- and transgender individuals (Figure S4a, c, e, g). We did

not observe any significant differences in probability distributions between CM and TW (Sample A: $t = 0.01$, $p = 0.9927$, Cohen's $d = 0.0025$; Sample B: $t = 1.34$, $p = 0.1886$, Cohen's $d = 0.45$) or between CW and TM (Sample A: $t = -1.18$, $p = 0.2447$, Cohen's $d = -0.31$; Sample B: $t = -0.98$, $p = 0.3335$, Cohen's $d = -0.28$), whereas TM displayed in both samples a gender congruent trend of a higher prediction accuracy with medium effect sizes [1]. The AM+cr model showed no indication of a TIV-bias, neither statistically (Table 3) nor visually (Figure S4b, d, f, h).

In contrast to the AM+cr model, the ATM+cr model resulted in a higher model performance with an overall accuracy of 89.57% with 89.80% for cisgender and 89.39% for transgender individuals. Similar model performance was achieved in sample B with an overall accuracy of 89.66% (85.37% for cisgender and 93.48% for transgender individuals, detailed information in Table 1 and S3). In accordance with the high sex classification accuracies, prediction probabilities showed a sex congruent distribution (Figure S4 i, k, m, o) with no significant differences in prediction probabilities between CW and TM (Sample A: $t = -0.38$, $p = 0.7050$, Cohen's $d = -0.10$; Sample B: $t = -1.28$, $p = 0.2073$, Cohen's $d = -0.36$), whereas we observed a gender congruent trend of TM having a higher prediction probability than CM with a large effect size [1]. This gender congruent trend was not observed for CM vs. TW (Sample A: $t = 0.15$, $p = 0.8818$, Cohen's $d = 0.04$; Sample B: $t = -2.16$, $p = 0.0380$, Cohen's $d = -0.72$). Similarly, as for the AM+cr model, the ATM+cr model indicated no evidence of TIV-biased model behavior (Table 3, Figure S4 j, l, n, p)

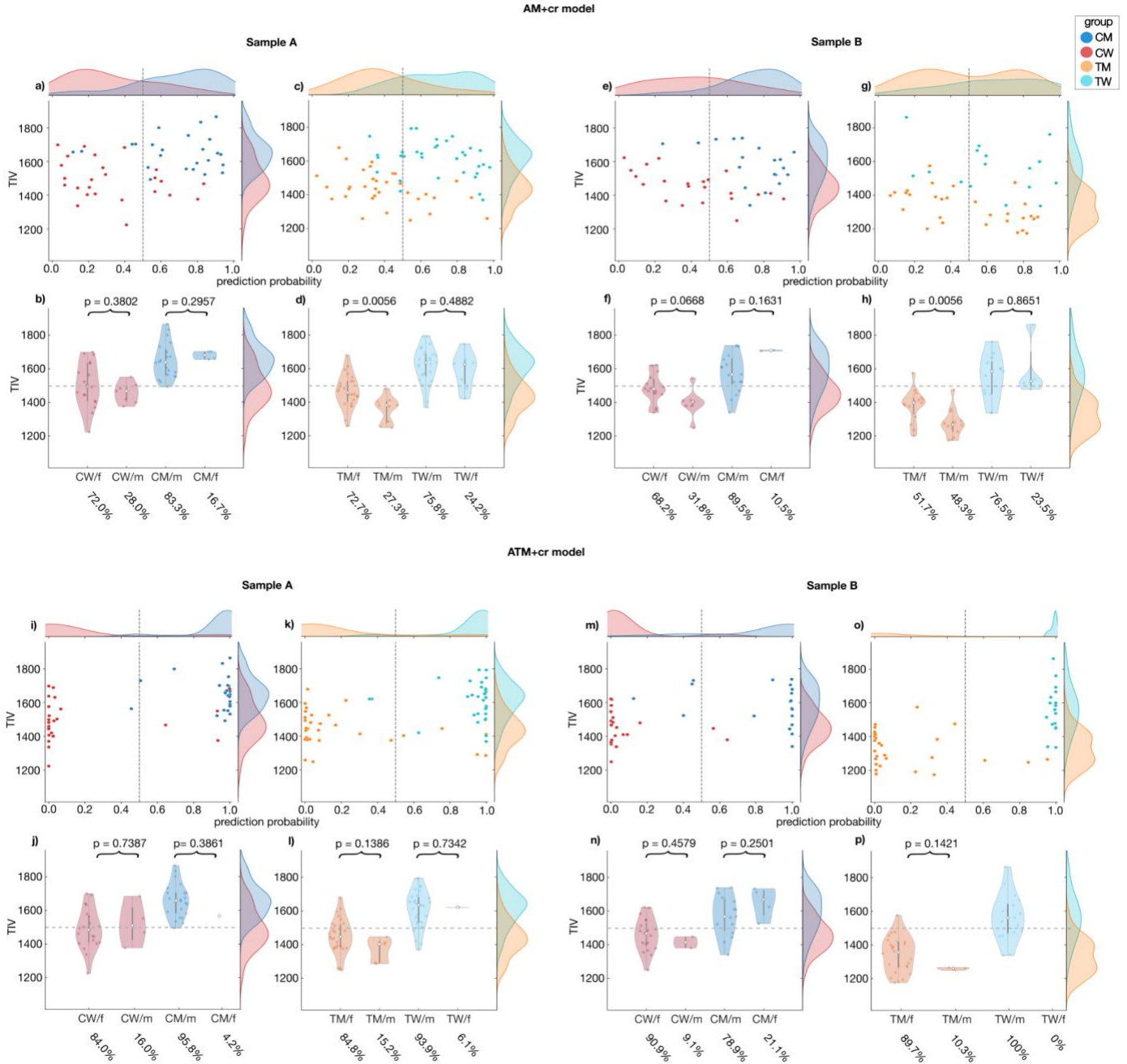


Figure S4. Association between prediction probability and TIV for the AM+cr and ATM+cr models in the two application samples. The upper row (a-h) shows the prediction probability (a, c, e, g) and TIV distribution (b, d, f, h) of sex congruently and incongruently classified CM, CW, TM and TW in the AM+cr 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+cr 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)

Supplementary Material and Methods

Structural scanning parameter

1000Brains

The population-based 1000Brains [2] investigated the variability of the human brain in a German cohort that also covered a wide age range. The subjects were measured in a Siemens TRIO 3 Tesla MRI scanner with the following parameters: 176 slices, TR = 2250 ms, TE = 3.03 ms, TI = 900 ms, FOV = 256 x 256 mm², flip angle = 9°, voxel resolution = 1 x 1 x 1 mm³. Data of 1000Brains are available upon request from the responsible Principal Investigator [2].

CamCAN

The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository provides data from of a population-based sample. All MRI datasets of this study were collected with a 3T Siemens TIM Trio scanner with a 32-channel head coil. Structural T1-weighted images were collected in an MPRAGE sequence with TR = 2250 ms, TE = 2.99 ms, flip angle = 9°, FOV = 256 x 240 x 192 mm³, and voxel size = 1 x 1 x1 mm³ [3].

CoRR

The Consortium for Reliability and Reproducibility (CoRR) addresses the challenge of reliability characterizing interindividual differences in human brain function, wherefore participants from several sites were scanned. Individual scanning parameters for each site are linked elsewhere [4].

DLBS

The Dallas lifespan brain study (DLBS) collected, among other data, anatomical MRI data to address research regarding the cognitive neuroscience of aging. The T1-weighted images were acquired in a Philips Achieva 3T scanner with the following parameters: TR = 8.135 ms, TE = 3.7 ms, matrix = 256 x 256, FOV = 204 x 256, slice thickness = 1 mm ([5]; www.nitrc.org/fcon_1000/htdocs/indi/retro/dlbs_content/dlbs_scan_params_anat.pdf).

eNKI

The Rockland Sample of the enhanced Nathan Kline Institute is a large-scale community sample of participants across the lifespan. T1-weighted images were acquired with an MPRAGE in a Siemens Trio Tim 3.0 T MRI scanner with the following parameters: TR = 2500 ms, TE = 30 ms, inversion time = 1200 ms, flip angle = 8°, FOV = 256 x 256mm², voxel size = 1 x 1 x 1 mm³ and number of slices = 192. T1-weighted images were used for spatial normalization and group-specific template generation [6, 7].

GOBS

The GOBS-sample provides a cohort of subjects that are of Mexican-American ancestry, parts of a large family and live in the region of San Antonio [8]. Diffusion imaging was performed at the University of Texas Health Science Center San Antonio (UTHSCSA) and Yale University with a 3T Trio Scanner (Siemens) with a spatial resolution of 1.7 x 1.7 x 3.0mm³, FOV = 200 mm, TR = 8000 ms, and TE = 87 ms [9].

HCP

The Human Connectome Project (HCP) contains a large cohort of healthy adults [10]. The parameters for the acquisition of T1 images as follows: TR = 2400 ms, TE = 2.14, TI = 1000, flip angle = 8°, FOV = 224 x 224 mm², voxel size of 0.7 mm (https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf).

IXI

The open source dataset for Information eXtraction from images (IXI) collected data from healthy participants at different hospitals in London. The data from the Hammersmith Hospital (HH) acquired the structural MRI data with a 3T Philips Medical Systems Scanner using the following parameters: TR = 9.6 ms, TE = 4.60 ms, 208 phase encoding steps, acquisition matrix = 208 x 208, flip angle = 8.0° (<http://brain-development.org/scanner-philips-medical-systems-intera-3t/>). The data in the Guy's Hospital were acquired using a 1.5T Philips scanner with a TR of 9.8 ms, TE of 4.6 ms, 192 steps of phase encoding and a flip angle of 8°.

OASIS-3

T1-weighted MRI images of the OASIS-3 study [11] were acquired once in a 1.5 T Magnetom Vision Siemens scanner with a 16-channel head coil with the following parameters: TR = 9.7 ms, TE = 4.0 ms, flip angle = 10°, number of slices = 128, FOV = 256 x 256, voxel size = 1.0 x 1.0 x 1.25 [12, 13]. Otherwise, participants were acquired with a 3T TIM Trio Siemens scanner with a 20-channel head coil with the following parameters: TR = 2400 ms, TE = 3.2 ms, flip angle = 8°,

voxel size = 1.0 x 1.0 x 1.0, FOV = 176 x 256 x 256 [14]. Data were provided by OASIS-3

Principal Investigators: T. Benzinger, D. Marcus, J. Morris.

PNC

The participants of the Philadelphia Neurodevelopmental Cohort (PNC) were scanned in a 3T Siemens TIM Trio scanner with a 32-channel head coil. The structural MRI data of T1-weighted images were assessed with an MPRAGE with the following parameters: TR = 1810 ms, TE = 3.5 ms, FOV = 180 mm³, 160 slices, flip angle = 9° [15, 16].

Table S1. Demographic variables of all samples.

	AM sample			
	women		men	
N	807		807	
mean age (SD)	37.96 (15.28)		37.95 (15.28)	
mean TIV (SD)	1338.2 (103.56)		1655.8 (105.42)	
	ATM sample			
	women		men	
N	807		807	
mean age (SD)	38.16 (15.33)		38.15 (15.38)	
mean TIV (SD)	1487.4 (94.01)		1493.2 (93.27)	
	Sample A			
	CW	CM	TM	TW
N	25	24	33	33
mean age (SD)	31.84 (11.25)	33.42 (10.53)	24.42 (6.88)	33.03 (12.62)
mean TIV (SD)	1491.4 (118.33)	1653.0 (97.86)	1436.8 (103.49)	1612.3 (106.08)
	Sample B			
	CW	CM	TM	TW
N	22	19	29	17
mean age (SD)	19.64 (2.42)	22.21 (4.35)	24.72 (6.19)	21.35 (3.94)
mean TIV (SD)	1454.8 (94.39)	1587.5 (118.02)	1330.0 (102.69)	1561.3 (139.12)

Age (years) and TIV (ml) for AM- and ATM-sample as well as for both application samples.

Table S2. Confusion matrices of all four models applied to both test samples.

AM test sample				ATM test sample		
AM model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	153	8		Male	120	41
Female	2	159		Female	26	135
AM+cr model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	118	43		Male	134	27
Female	80	81		Female	60	101
ATM model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	142	19		Male	149	12
Female	24	137		Female	12	149
ATM+cr model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	138	23		Male	148	13
Female	23	138		Female	11	150

Table S3. Confusion matrices of all four models applied to both application samples.

Sample A				Sample B		
AM model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	54	3		Male	32	4
Female	10	48		Female	2	49
AM+cr model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	45	12		Male	30	6
Female	16	42		Female	21	30
ATM model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	57	0		Male	35	1
Female	10	48		Female	5	46
ATM+cr model						
	Predicted Label:				Predicted Label:	
True Label:	Male	Female		True Label:	Male	Female
Male	54	3		Male	32	4
Female	9	49		Female	5	46

Table S4. Numbers of exact allocation of subjects per sample before and after matching in AM and ATM sample.

Name Sample/Site	Number of subjects before matching	Number of subjects AM	Number of subjects ATM
1000Brains	712	276	276
CamCAN	435	146	146
CoRR	1152	168	168
DLBS	198	50	50
GOBS_CIVET	595	192	192
HCP	1113	396	396
IXI	462	96	96
OASIS3	237	76	76
PNC	296	120	120
eNKI	357	94	94

Table S5. Wilcoxon rank sum tests on the hold-out samples.

	GMV women classified as female vs. classified as male	GMV men classified as male vs. classified as female
	AM hold-out sample	
AM model	T = 12792, z = -1.3202, p = 0.1868, $\eta^2 = 0.0108$	T = 12519, z = 0.9763, p = 0.3289, $\eta^2 = 0.0059$
AM+cr model	T = 7352, z = 2.6727, p = 0.0075, $\eta^2 = 0.0444$	T = 9030, z = -2.0155, p = 0.0439, $\eta^2 = 0.0252$
ATM model	T = 10943, z = -0.7286, p = 0.4663, $\eta^2 = 0.0033$	T = 11371, z = -0.6838, p = 0.4941, $\eta^2 = 0.0029$
ATM+cr model	T = 11175, z = -0.0121, p = 0.9904, $\eta^2 < 0.001$	T = 10966, z = -1.0217, p = 0.3069, $\eta^2 = 0.0065$
	ATM hold-out sample	
AM model	T = 10722, z = -0.9762, p = 0.3290, $\eta^2 = 0.0059$	T = 10155, z = 1.6859, p = 0.0918, $\eta^2 = 0.0177$
AM+cr model	T = 8065, z = -0.4038, p = 0.6864, $\eta^2 = 0.0010$	T = 10438, z = -1.8801, p = 0.0601, $\eta^2 = 0.0220$
ATM model	T = 12437, z = 2.3654, p = 0.0180, $\eta^2 = 0.0348$	T = 12025, z = -0.2800, p = 0.7795, $\eta^2 < 0.001$
ATM+cr model	T = 12537, z = 2.5896, p = 0.0096, $\eta^2 = 0.0417$	T = 11884, z = -0.6422, p = 0.5207, $\eta^2 = 0.0026$

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

Table S6. Wilcoxon rank sum tests on the application samples.

a)	GMV CW classified as female vs. classified as male	GMV CM classified as male vs. classified as female
AM model	T = 202, z = -1.9065, p = 0.0566, $\eta^2 = 0.1454$	T = 276, z = 0.0522, p = 0.9584, $\eta^2 < 0.001$
AM+cr model	T = 226, z = -0.4539, p = 0.6499, $\eta^2 = 0.0082$	T = 240, z = -0.7359, p = 0.4618, $\eta^2 = 0.0226$
ATM model	T = 255, z = -1.2972, p = 0.1946, $\eta^2 = 0.0673$	no CM classified as female
ATM+cr model	T = 255, z = -1.2972, p = 0.1946, $\eta^2 = 0.0673$	T = 282, z = -0.7223, p = 0.4701, $\eta^2 = 0.0217$
	GMV TM classified as female vs. classified as male	GMV TW classified as male vs. classified as female
AM model	T = 515, z = 0.2818, p = 0.7781, $\eta^2 = 0.0024$	T = 540, z = -0.3676, p = 0.7132, $\eta^2 = 0.0041$
AM+cr model	T = 456, z = 1.9201, p = 0.0548, $\eta^2 = 0.1117$	T = 432, z = 0.2731, p = 0.7848, $\eta^2 = 0.0023$
ATM model	T = 505, z = 2.1238, p = 0.0337, $\eta^2 = 0.1367$	no TW classified as female
ATM+cr model	T = 527, z = 2.5356, p = 0.0112, $\eta^2 = 0.1948$	T = 529, z = 0.1132, p = 0.9099, $\eta^2 < 0.001$
b)	GMV CW classified as female vs. classified as male	GMV CM classified as male vs. classified as female
AM model	T = 230, z = 0, p = 1, $\eta^2 = 0$	T = 183, z = 1.6605, p = 0.0968, $\eta^2 = 0.1451$
AM+cr model	T = 196, z = 1.6213, p = 0.1050, $\eta^2 = 0.1195$	T = 162, z = -0.9963, p = 0.3191, $\eta^2 = 0.0522$
ATM model	T = 240, z = 1.0850, p = 0.2779, $\eta^2 = 0.0535$	T = 181, z = 0.0913, p = 0.9273, $\eta^2 < 0.001$
ATM+cr model	T = 240, z = 1.0850, p = 0.2779, $\eta^2 = 0.0535$	T = 139, z = -1.0500, p = 0.2937, $\eta^2 = 0.0580$
	GMV TM classified as female vs. classified as male	GMV TW classified as male vs. classified as female
AM model	no TM classified as male	T = 147, z = 1.7143, p = 0.0865, $\eta^2 = 0.1729$
AM+cr model	T = 286, z = 2.6404, p = 0.0083, $\eta^2 = 0.2404$	T = 125, z = 0.8492, p = 0.3958, $\eta^2 = 0.0424$
ATM model	T = 409, z = 1.3248, p = 0.1852, $\eta^2 = 0.0605$	no TW classified as female
ATM+cr model	T = 409, z = 1.3248, p = 0.1852, $\eta^2 = 0.0605$	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).

References

1. Gignac, G.E., & Szodorai, E. T., *Effect size guidelines for individual differences researchers*. Personality and individual differences, 2016. **102**: p. 74-78.
2. Caspers, S., et al., *Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS*. Front Aging Neurosci, 2014. **6**: p. 149.
3. Taylor, J.R., et al., *The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample*. Neuroimage, 2017. **144**(Pt B): p. 262-269.
4. Zuo, X.N., et al., *An open science resource for establishing reliability and reproducibility in functional connectomics*. Sci Data, 2014. **1**: p. 140049.
5. Zheng, F., et al., *The Volume of Hippocampal Subfields in Relation to Decline of Memory Recall Across the Adult Lifespan*. Front Aging Neurosci, 2018. **10**: p. 320.
6. Betzel, R.F., et al., *Changes in structural and functional connectivity among resting-state networks across the human lifespan*. Neuroimage, 2014. **102 Pt 2**: p. 345-57.
7. Nooner, K.B., et al., *The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry*. Frontiers in neuroscience, 2012. **6**: p. 152.
8. Olvera, R.L., et al., *Common genetic influences on depression, alcohol, and substance use disorders in Mexican-American families*. Am J Med Genet B Neuropsychiatr Genet, 2011. **156B**(5): p. 561-8.
9. Jahanshad, N., et al., *Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA-DTI working group*. Neuroimage, 2013. **81**: p. 455-469.

10. Van Essen, D.C., et al., *The Human Connectome Project: a data acquisition perspective*. Neuroimage, 2012. **62**(4): p. 2222-31.
11. LaMontagne, P.J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., ... & Marcus, D., *OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease*. MedRxiv, 2019.
12. Marcus, D.S., et al., *Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults*. J Cogn Neurosci, 2007. **19**(9): p. 1498-507.
13. Tavares, V., D. Prata, and H.A. Ferreira, *Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study*. J Neurosci Methods, 2019. **334**: p. 108565.
14. Kwak, S., et al., *Multivariate neuroanatomical correlates of behavioral and psychological symptoms in dementia and the moderating role of education*. Neuroimage Clin, 2020. **28**: p. 102452.
15. Satterthwaite, T.D., et al., *The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth*. Neuroimage, 2016. **124**(Pt B): p. 1115-1119.
16. Satterthwaite, T.D., et al., *Neuroimaging of the Philadelphia neurodevelopmental cohort*. Neuroimage, 2014. **86**: p. 544-53.