Deep Learning Identifies Morphological Determinants of Sex Differences in the Pre-Adolescent Brain

Ehsan Adeli[a,1], Qingyu Zhao[a,1], Natalie M. Zahr[a,b], Aimee Goldstone[b], Adolf Pfefferbaum[a,b], Edith V. Sullivan[a], Kilian M. Pohl[a,b,*]

[a] *Department of Psychiatry & Behavioral Sciences, Stanford University, Stanford, CA 94305*
[b] *Center for Biomedical Sciences, SRI International, Menlo Park, CA 94205*

Abstract

The application of data-driven deep learning to identify sex differences in developing brain structures of pre-adolescents has heretofore not been accomplished. Here, the approach identifies sex differences by analyzing the minimally processed MRIs of the first 8,144 participants (age 9 and 10 years) recruited by the Adolescent Brain Cognitive Development (ABCD) study. The identified pattern accounted for confounding factors (i.e., head size, age, puberty development, socioeconomic status) and comprised cerebellar (corpus medullare, lobules III, IV/V, and VI) and subcortical (pallidum, amygdala, hippocampus, parahippocampus, insula, putamen) structures. While these have been individually linked to expressing sex differences, a novel discovery was that their grouping accurately predicted the sex in individual pre-adolescents. Another novelty was relating differences specific to the cerebellum to pubertal development. Finally, we found that reducing the pattern to a single score not only accurately predicted sex but also correlated with cognitive behavior linked to working memory. The predictive power of this score and the constellation of identified brain structures provide evidence for sex differences in pre-adolescent neurodevelopment and may augment understanding of sex-specific vulnerability or resilience to psychiatric disorders and presage sex-linked learning disabilities.

*Corresponding author
Email address: kilian.pohl@stanford.edu (Kilian M. Pohl)
[1] These authors contributed equally to this work.

Keywords: Deep learning, sex differences, adolescents, study confounders, pubertal development, cerebellum.

1. Introduction

The concept of sex differences is based on biology and genetics. Since the 1930s (e.g., [1]), identifying sex differences in the Central Nervous System (CNS) has been explored in animal models [2, 3, 4, 5, 6] and histology of postmortem human brain samples [7, 8, 9]. More recently, in vivo neuroimaging [10, 11, 12, 3, 13, 14, 15] and computer-based learning tools [16, 17, 18, 19, 20] have been implemented in the search for a CNS basis of sexual differentiation. Beyond sex-linked risks for disease [21, 22, 23, 24, 25, 7], this search is motivated by adolescence being a period of particular vulnerability to the emergence of sex-linked neuropsychiatric disorders such as schizophrenia [7, 26] and autism [27, 28, 29, 25, 30], which have a higher prevalence in boys than girls, and depression, which girls by age 15 develop twice as likely as boys [18, 31].

In vivo structural magnetic resonance imaging (MRI) studies characterize brain development as following heterogeneous growth trajectories [32, 33] during which sex-specific behaviors emerge [34]. While physical signs of sex differences are present at birth [35], brain structural and functional differences between the sexes continue to develop over childhood through late adolescence [36, 37, 38, 39, 40]. For example, both cortical and subcortical gray matter volumes exhibit inverted U-shaped trajectories reflecting growth followed by synaptic pruning, with boys showing a slightly larger rate of change throughout childhood and adolescence than girls [41]. With respect to white matter, the volume increases with age in both sexes, but boys generally show a more rapid increase during adolescence [41]. These sex specific changes in brain structure during adolescence [42] are accompanied with asexual developments, such as structural volume [43, 44, 45, 46, 47, 48], cortical thickness [48], cortical surface area [44, 48], individual's behavior [49], and testosterone effects [42].

Many of the differences in brain development between the sexes are actually

linked to head size [50, 51]. As boys on average have larger brains than girls, identifying sex differences in the brain beyond head size is challenging and might explain the inconsistent findings in the literature. For example, whether sex differences are present within the corpus callosum has been a matter of debate [52, 23, 53, 54, 55]. Beyond properly accounting for head size [53, 56, 39, 51], discrepancies in findings may be due to small sample sizes [57, 11, 58], wide age distributions (sometimes across several decades so age-specific sex differences are obscured) [52, 59], or a priori assumptions that reduce the rich information encoded in MRIs to a few brain measurements (e.g., volumes of a limited number of brain regions of interest (ROIs)) [52, 20]. The study presented herein accounts for these issues by building on recent advancements in the field of deep learning [60, 61, 19] to identify patterns not driven by study confounders, which are extraneous variables (such as age) that may induce undesired class differences if not properly controlled.

Specifically, we present a deep learning framework (see Figure 1) predicting sex from the minimally processed T1-weighted (T1w) MRIs [62] of 8,144 pre-adolescents (ages 9 and 10 years) of the ABCD study (http://abcdstudy.org). The variance in the prediction scores is related to the cognition test scores of the National Institutes of Health (NIH) Toolbox® [63]. Finally, we qualitatively assess the average *saliency map* [64] across all MRIs, which encodes the contribution of each voxel of the MRI in predicting sex while removing the effects driven by the confounders, i.e., age and pubertal and socioeconomic status.

2. Materials and Methods

2.1. ABCD Participants and Study Design

The model was evaluated on data collected by the ABCD study (http://abcdstudy.org). Demographic information (Table 1), cognitive test scores from the NIH toolbox (Table 2), and T1-weighted (T1w) MR images [62] from 8,670 participants were distributed by the ABCD-Neurocognitive Prediction Chal-

lenge (ABCD-NP-Challenge 2019) [65] via the National Database for Autism Research (NDAR) portal (Release 2.0), of which 8,144 subjects contained the data needed for this analysis. Socioeconomic status (SES) was estimated by identifying the maximum level of education across parents/guardians as done elsewhere [66]. Pubertal status was determined by self-assessment with the Pubertal Development Scale (PDS) [67, 68], a validated measure of pubertal stage that shows modest concordance with a physical exam and that correlates with basal gonadal hormone levels. An average PDS was calculated for each participant by adding up scales on five self-reports obtained from parents' responses to a questionnaire, where each scale ranged from 1 to 4. Based on this computation, PDS categorized ABCD youth as either 1) pre-pubertal, 2) early-pubertal, 3) mid-pubertal, 4) late-pubertal 5) post-pubertal. Participants of multiple ethnicities were categorized according to their minority ethnicity (e.g., a report of Asian and Caucasian was classified as Asian) [39]. Body Mass Index (BMI) was calculated based on published methods [69]. Observed Sex for all the participants was defined as the sex at birth.

Recruitment for the ABCD study closely represented the general U.S. population of 9 and 10 year-old children with respect to key demographic variables including sex, ethnicity, household income, parental education, and parental marital status [70]. Parents provided informed consent and were fluent in either English or Spanish; children had to be fluent in English and provide assent for participation. Exclusionary criteria included poor English-language proficiency; the presence of severe sensory, intellectual, medical, or neurological issues that would affect the validity of data or ability to comply with the protocol; and contraindications to MRI (see [71] for complete description of details regarding recruitment and inclusion/exclusion criteria).

2.2. MRI Data Acquisition and Processing

Details on T1w-MRI acquisition are provided by `https://abcdstudy.org/images/Protocol_Imaging_Sequences.pdf`. Processing of T1w-MRI were subjected to the ABCD minimal-processing pipeline [62] followed by noise removal

[72] and field-inhomogeneity correction via N4ITK (Version 2.1.0) [73]. Brain masks were determined via majority voting [74] over the segmentations generated by applying the following tools to both bias and non-bias corrected T1w-images: FSL BET (Version 5.0.6) [75], AFNI 3dSkullStrip (Version AFNI_2011_12_21_1014) [76], FreeSurfer mri-gcut (Version 5.3.0) [77], and Robust Brain Extraction (ROBEX) (Version 1.2) [78]. The resulting brain mask was used to refine correction for image-inhomogeneity and skull stripping. MRIs were then affinely registered to the SRI24 template [79], down-sampled to 2mm isotropic voxel size, and re-scaled to 64x64x64 volumes. The affine registrations ensured that all MRIs of the ABCD study had similar head size as measured by supratentorium volume (svol) (see also Table 1 for the resulting insignificant difference in head size between boys and girls).

Figure 1 outlines the deep learning framework used to predict sex from minimally processed MRI data. The framework was composed of a Predictor/Extractor and a Classifier [60, 80]. The Predictor/Extractor identified a set of Predictor variables $\mathbf{P} = \{\mathbf{P}^1, \mathbf{P}^2, ..., \mathbf{P}^M\}$ from MR images based on a deep convolutional network [61]. The Classifier was a set of fully connected layers reducing $\mathbf{P}$ into a continuous Prediction Score $\mathbf{S}$, which was the probability π computed by the classifier of an MRI being associated with a girl (i.e., $\pi(\text{girl}) = \mathbf{S}$) or a boy (i.e., $\pi(\text{boy}) = 1 - \mathbf{S}$). Appendix B provides a more in-depth description of the deep learning architecture.

The prediction accuracy of the model was determined in two steps. Assuming that sex affects the brain bilateral [81, 82, 83, 84, 85] and to simplify the interpretation of the findings, the left hemisphere was first flipped to create a 2nd "right" hemisphere. Then, 5-fold cross-validation [86] was performed by splitting the data based on subjects. At each iteration of the cross-validation, the four folds of the data used for training were first augmented to ensure that the learning was based on a balanced and sufficient number of boys and girls [87], *i.e.*, 5000 for each group. Data augmentation consisted of applying random rigid transformations (within one voxel shifting and 1° rotation along the three axes) to the minimally processed (and flipped) MRIs. On this augmented data set,

the entire deep model, which included the predictor extractor and the classifier, was trained from scratch in an end-to-end manner [60]. Next, the prediction of the individual's sex was recorded on the fifth fold (which was not augmented) by computing the average prediction score ($\mathbf{S}$) across both hemispheres. The training and testing processes were repeated until the prediction score was reported for each subject. The average accuracy of the method on all folds was then computed by first binarizing $\mathbf{S}$ of each participant to 1 (girl) or 0 (boy) and then comparing the predictions to their observed sex via commonly used metrics: balanced classification accuracy [88] (a.k.a. accuracy), true positive rate, false positive rate, and the area under the receiving operating characteristic curve.

To put the prediction accuracy in perspective and compare with widely used machine learning methods, the cross-validation was repeated with respect to logistic regression [89, 90], support vector machines [91] and random forest [92] applied to the volumes of 116 brain ROIs defined according to the SRI24 atlas [79]. Measuring the volumes of ROIs consisted of non-rigidly registering the SRI24 atlas to each brain-size corrected MRI via ANTS (Version: 2.1.0) [93] and overlaying parcellations with the tissue segmentations from Atropos [94]. The experiment was repeated using the 906 regional scores generated by Freesurfer based on the Destrieux atlas [95], which were provided by the ABCD Study Release 2.0 (`http://abcdstudy.org`). These regional scores consisted of cortical thickness, sulcal depth, surface area, and volume of cortical ROIs and the average T1 intensities within the white and gray matter.

In addition to the comparison to other methods, a sex-agnostic test correlated the prediction score $\mathbf{S}$ of the individuals with the test scores of the age-corrected NIH toolbox (see Table 4 left; significant p-value < 0.05 according to Pearson's R). Identifying variance in the prediction score partially induced by an NIH toolbox score (Figure 2, Table 4 right) was done via the partial mediation model [96]. Partial mediation required that 1) observed sex significantly correlated with the NIH toolbox score; 2) the NIH toolbox score significantly correlated with the prediction score when accounting for observed sex as an

additional covariate; and 3) the correlation between observed sex and the prediction score was significantly reduced (p-value inferred from a permutation test of 10000 permutations) when accounting for the NIH toolbox score as an additional covariate.

Finally, we performed bootstrapping (5 runs) to determine the effects of PDS (the most significant confounder of this study according to Table 1) on the sex predictions of our approach. Each of the 5 runs was defined by 5-fold cross-validation consisting of a unique random split of the data into 5-folds. The correctly classified subjects in all 5 runs were assigned to one group, and the ones that were incorrectly classified in all 5 runs were assigned to a second group. For each sex separately, differences in the PDS between the two groups were defined by the p-value of the χ^2 test [97]. Across the two groups, the prediction accuracy (for both boys and girls) was determined for cohorts confined to the same PDS. We then reported on the cohorts with a sufficient number of samples for each sex, which were the cohorts for PDS 1, 2, and 3.

All methods were implemented using Python 3.7.0 and its libraries including SciPy 1.1.0, NumPy 1.15.1, Scikit-Learn 0.19.2, pygrowup 0.8.2 toolbox [98], Tensorflow 1.7.0 [99], and Keras 2.2.2 [100]. The codes of our deep learning implementation are publicly available at `https://github.com/QingyuZhao/Confounder-Aware-CNN-Visualization` and the tests at `https://github.com/eadeli/ABCD_SexDiff`.

2.3. Identifying Confounder-Free Patterns and ROIs Relevant to Sex

To derive a single pattern informative for identifying sex differences, we retrained our proposed approach on the entire dataset. For each participant, the discriminative power of each voxel to predict sex was recorded using a saliency map [64]. The initial salience map was computed by applying the minimally processed and flipped MRI to the trained prediction model and then performing back-propagation [101]. Note, saliency computation did not require data augmentation nor estimating prediction accuracy.

Next, the map was further corrected for the effects of potential confounders

on the decision process of the model. Confounders were demographic factors significantly different between sexes according to Table 1, i.e., age ($\mathbf{z}^{age}$), PDS ($\mathbf{z}^{pds}$), and SES ($\mathbf{z}^{ses}$). To determine if a confounder significantly influenced the decision process of $\mathbf{S}$, a general linear model (GLM) [102] was fit across all samples with respect to each predictor variable $\mathbf{P}^j$ of $\mathbf{P}$:

$$\mathbf{P}^j = \beta_0 + \beta_1 \mathbf{S} + \beta_2 \mathbf{z}^{pds} + \beta_3 \mathbf{z}^{age} + \beta_4 \mathbf{z}^{ses}. \quad (1)$$

If the predictor variable $\mathbf{P}^j$ of the GLM significantly correlated ($p \leq 0.05$) with one of the demographic variables, the predictor was considered confounded and omitted from computing the saliency maps. The lenient p-value threshold of 0.05 was not corrected for multiple comparison as we wanted our analysis to be sensitive towards identifying confounded predictors so that the resulting pattern accurately represented sex differences. The pattern encoding the relevance of each voxel in predicting sex was defined by the average across the confounder-free saliency maps of all participants. Conversely, a pattern encoding the effect of a specific confounder was created by computing the saliency maps based on confounded predictors.

To relate the identified voxels to previously defined brain ROIs (using SRI24 atlas [79]), we computed the average saliency value of each ROI from the confounder-free saliency map of each participant. For each ROI, follow-up t-tests evaluated whether the average saliency value within that region was significantly different between groups (p-value < 0.05 with Bonferroni multiple comparison correction [103]).

3. Results

The accuracy of the prediction score in correctly assigning MRIs to either sex was 89.6%, which was significantly better than chance (p <0.001 according to a Fisher exact test [104]). The prediction accuracy was stable across 5 runs of 5-fold cross-validation based on random splitting of folds (89.6% $\pm$ 0.13%) but was slightly lower (87.3%) on a subset of 2464 boys and 2464 girls matched

on head size (matched according to [105]). Furthermore, the True Positive Rate (TPR) of the deep learning model was 87.4% and True Negative Rate (TNR) was 91.5% (girls=1, boys=0). Compared with the correctly classified pre-adolescents, misclassified boys had significantly higher PDS while misclassified girls had significantly lower PDS (p-value $< 10^{-6}$ according to χ^2 test). The prediction confined to individuals with the same PDS was 88.9% for participants with PDS = 1, 89.5% for PDS = 2, and 90.1% for PDS = 3.

The prediction of our approach was significantly more accurate (Delong test [106], p-value<0.001) than the results reported by Logistic Regression, Support Vector Machine, and Random Forest applied to the 116 ROI volume measures or the 906 Destrieux parcellation measures (see Table 3). To gain a better understanding of this improvement, we recomputed the accuracy of our model across 5 runs of 5-fold cross-validation with respect to the number of predictors. The average accuracy remained relatively high (86.5%) even when extracting only 128 predictors from each MRI (see Fig. 3(a)). Furthermore, similarly high accuracy was achieved by the other approaches when trained on the predictors extracted by our deep model (Fig. 3(b)).

A visual confirmation of the significant prediction accuracy of our model were the two distinct distributions shown in Figure 4(a), which plotted the Prediction Score ($\mathbf{S}$) of each participant as a function of their observed sex. Furthermore, projecting the high dimensional Predictors ($\mathbf{P}$) learned from one training run into 2D via the t-distributed Stochastic Neighbor Embedding (tSNE) [107] also resulted in a cluster for boys and a separate one for girls (Figure 4(b)).

Figure 5(a) visualizes the initial saliency map with voxel values above 0.1 before correcting for confounders. The highlighted area significantly contributed to predicting sex, which partly consisted of the temporal lobes, subcortical regions, cerebellum, and corresponding white matter. Figure 5(b) shows the area of sexual differentiation according to the confounder-free saliency map (i.e., with age, PDS, and SES removed), which is more spatially concentrated than the initial saliency map (Figure 5(a)). According to the confounder-free saliency values, the 10 ROIs most relevant for predicting sex were insula, pallidum, para

hippocampus, and putamen (larger in boys than girls); hippocampus, corpus medullare, and cerebellum VI (larger in girls than boys) (Figure 6). Although deep learning identified insula, amygdala, and cerebellar lobules III and IV/V as significant predictors of sex, their volume differences by sex were not forthcoming. The cerebellum was also the region mostly confounded by PDS (Figure 5(c)), the most significant confounder in the model.

Table 4 lists the correlation and mediation effect of NIH toolbox scores with respect to the prediction score **S**. Significant correlations (p-value < 0.05) between **S** and NIH toolbox scores were confined to the List Sorting Working Memory Test, Pattern Comparison Processing Speed, Picture Sequence Memory Test, and Picture Vocabulary Test. Further, a partial mediation model examined whether the NIH toolbox scores could partially explain the variance in **S** in addition to the observed sex (Fig. 2). Only the List Sorting Working Memory Test score met the 3 significance conditions of the mediation model (p-value < 0.05): 1) observed sex significantly correlated with the NIH toolbox score; 2) the NIH toolbox score significantly correlated with **S** when accounting for observed sex as an additional covariate; and 3) the correlation between observed sex and **S** was significantly reduced when accounting for the NIH toolbox score as an additional covariate.

4. Discussion

The deep learning model presented herein not only successfully predicted the sex of 8144 pre-adolescents from (head-size normalized) T1w MRI but also was more accurate than several other commonly used machine learning approaches, e.g., logistic regression, support vector machine, and random forest. While these machine learning approaches relied on *a priori* defined regional measurements (as is commonly used for neuroscience studies [43, 108, 105, 109, 59]), the improved accuracy of the deep learning model was mostly due to its ability to simultaneously extract predictors directly from the MRIs and perform classification (see Figure 3). A novel discovery of that search for discriminative infor-

mation was that sex could be accurately predicted in individual pre-adolescents through a pattern composed of subcortical and cerebellar regions. Also unknown for pre-adolescence was that the cerebellum was most strongly affected by PDS, the most significant confounder of the study. Finally, reducing the pattern to a single score revealed that its variance was not only explained by sex but also by cognitive behavior linked to working memory.

Critical for interpreting the pattern was the notion that sex differences on brain structure are bilateral [82, 83, 84, 85]. We modeled that by 'flipping' the left hemisphere and then training the algorithm on two 'right' hemispheres for each subject. When omitting flipping, the prediction accuracy was 89.1% when just trained on the left hemisphere, 88.5% when only trained on the right hemisphere, and 90.1% when trained on both hemispheres (omitting flipping). These accuracy scores were insignificantly different ($p > 0.1$; DeLong's test) from those of the 'flipped' approach confirming the bilateral nature of sex differences.

Another critical aspect in analyzing the pattern was computing a saliency map that displayed brain areas exhibiting sex differences while accounting for confounders; something that had not been attempted by prior data-driven analyses [16, 17, 50, 19, 20]. Removing confounding effects after training a machine learning model is potentially a more conservative approach compared with removing effects through preprocessing (e.g., matching), i.e., before the training. Unlike removing confounding effects after training, preprocessing generally cannot completely remove those effects so that learning approaches can still leverage the remaining confounding effects to 'improve' predictions [88]. Of the three confounders considered, PDS was the most significant one, which was generally larger in girls than in boys within the pre-adolescent age range (Table 1). While misclassified boys had significantly higher PDS and misclassified girls had significantly lower PDS than correctly classified individuals of the same sex, the prediction accuracy of our deep learning model was not affected by PDS as the overall accuracy of 89.6% remained stable when confining the evaluation to individuals with the same PDS. The region most confounded by PDS was the cerebellum (Figure 5(c)) suggesting that pubertal status may be specifically

associated with cerebellum development at this young age. This hypothesis is difficult to test on the baseline data of ABCD as the majority ($\sim$ 73%) of individuals were categorized as pre- or early pubescent. However, as the ABCD cohort ages, the variability in PDS will be considerably greater, and as such, will allow us to explore in more detail the potential interaction between sex and puberty in terms of cerebellar development.

In addition to the relationship to PDS, structures of the cerebellum were critical to predicting sex in individual, which is inline with a number of adult studies [108, 10, 110, 111]. However, sex differences in cerebellar volume became generally negligible once studies corrected for intracranial volume (e.g., [112, 113, 114]). More specifically, the corpus medullare of the cerebellum in this study was significantly larger in girls than boys. By contrast, the longitudinal study by [111] did not detect significant sex differences in the corpus medullare but reported that total cerebellar volume was larger in boys than girls, and that this total volume peaked at age 15.6 years in boys and at age 11.8 years in girls. The discrepancy in age range of the participants between that study (spanning pre-adolescents to young adults) and our analysis (ages 9 and 10 years) might reflect variance in cerebellar developmental trajectories during critical developmental years. Indeed, a recent review of the literature on language and brain development concluded that sex differences were most often found in studies limited to tight age ranges [52]. Sex differences in regional brain volumes may be apparent in some but negligible in other developmental stages, likely due to different rates of brain maturation between girls and boys [115].

Of the predictive regions within the subcortex, the hippocampus was larger in girls than boys after correcting for head size (see Fig. 6). The hippocampus has often been associated with sex-specific differences in memory and learning in adolescence [116, 117]. This observation comports with the finding that girls participating in the ABCD study had significantly higher scores on the NIH Toolbox Picture Sequence Memory Test, which is a validated measure of episodic memory [118]. The finding that girls had relatively larger hippocampi

than boys is also supported by MRI studies of young adults [11, 119, 114] that linked sex differences in hippocampal volume to hormonal responsivity [120, 121] and memory performance [81, 122, 123]. Other studies noted relations between hippocampal volumes and clinical characteristics of psychiatric disorders [22, 119, 124], where sleep disturbances are more severe [125], depressive episodes are more frequent and longer, and higher frequency of migraines occurs in depressed female compared to depressed male patients [126].

Other regions relevant for predicting sex included putamen, pallidum, and amygdala. These regions are frequently noted with reference to sex differences in brain maturation. An early imaging study of children aged 4-18 years suggested that while the caudate is relatively larger in girls, the pallidum is larger in boys [127]. A more recent study based on data from the Pediatric Imaging, Neurocognition, and Genetics (PING) study with 1,234 participants (ages 3 to 21 years) [128] showed that volumes of putamen and pallidum had greater variance in boys than girls: these differences may contribute to the variability in cognition and general intelligence in developing boys [129, 130]. Likewise, the amygdala has been linked to sex differences in animal and human studies across the lifespan [131, 109]. A surface-based modeling approach showed that men had a larger mean radius of amygdala subregions than women [59]. Further, sex differences in amygdala volume may contribute to the expression of selective psychotic disorders occurring more commonly in men than women [131] and depressive disorders, which are more common in women [18, 31].

Like the amygadala, the insula was important for predicting sex but its volume was insignificantly different between the two cohorts. Functional studies have frequently shown the significant role of these two regions in working memory performance [132]. Interestingly in our study, sex prediction by the deep learning model was mediated by the List Sorting Working Memory test score, which was higher for boys than girls (see Table 4). These results suggest that the deep learning approach of directly analyzing intensity values at a voxel level is potentially more powerful in extracting morphological characteristics linked to cognitive differences between the sexes than traditional approaches that focus

on specific measurements.

In addition to the mediation analysis, the predictive score was significantly correlated to most of the cognitive scores by the NIH Toolbox. These early and pervasive sex differences in neurocognitive measures echoed those identified on the 10,000 youth of the Philadelphia Neurodevelopmental Cohort (PNC) [133], in which girls performed better than boys on tasks assessing verbal memory and social cognition, whereas boys excelled on spatial processing and motor speed [133, 134]. Similar results were reported with the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) data, whose cognitive test battery included those of the PNC study [66]. Further consistency in sex differences on performance is forthcoming between our results and those published by the PING study, which, like the ABCD study, used the NIH Toolbox Battery. The PING study found that girls performed better than boys on tests assessing cognitive flexibility, problem solving, and episodic memory, whereas boys performed better on a list sorting task, assessing working memory for sorting and sequencing information [135]. Taken together, the consistency of sex differences in the development of component processes of selective cognitive skills transcended cohort differences and specific testing materials, which provide evidence for generalization of these identified sex differences.

Limitation. Our analysis did not detect significant sex differences in the cortex possibly because the MRIs were affinely aligned to a template, thereby minimizing headsize differences. While a common practice in end-to-end training [136, 60], affine registration might poorly align the cortical gyri and sulci given their high inter-subject variability [137]. Non-rigid registration achieves better voxel-wise correspondence across MRIs enabling learning algorithms to focus on fine-grained regional cues [138, 139]. Now identifying cues differentiating between groups highly depends on the 'stiffness' of the deformation field [140, 141], which can substantially modify the shape and appearance of brain structures. As affine and non-rigid registration can both negatively affect analysis, their effect on our deep learning findings needs to be further investigated.

5. Conclusion

The voxel-level analysis on the large number (N=8,144) of pre-adolescents (age 9 and 10) confirmed and extended the common finding of smaller neuroimaging studies that cerebellum and subcortical structures (including hippocampus, amygdala, pallidum, and putamen) differed in size between boys and girls. Not known before, however, was that the constellation of those brain structures accurately predicted the sex of individual pre-adolescents. The predictive power of the pattern provides evidence for sex differences in pre-adolescent, pubertal development, which may show even greater differentiation as the cohort ages. Tracking these disparities is a normative process that could augment understanding of sex-specific vulnerability or resilience to psychiatric disorders and presage sex-linked learning disabilities.

6. Acknowledgments

Funding for this study was received from the U.S. National Institutes Health (NIH) grants AA026762, DA041123, AA021697, and AA010723.

None of the authors has conflicts of interest with the reported data or their interpretation.

References

[1] C. A. Pfeiffer, Sexual differences of the hypophyses and their determination by the gonads, American Journal of Anatomy 58 (1) (1936) 195–225.

[2] L. A. Galea, M. D. Spritzer, J. M. Barker, J. L. Pawluski, Gonadal hormone modulation of hippocampal neurogenesis in the adult, Hippocampus 16 (3) (2006) 225–32. doi:10.1002/hipo.20154.
URL https://www.ncbi.nlm.nih.gov/pubmed/16411182

[3] J. M. Goldstein, L. J. Seidman, N. J. Horton, N. Makris, D. N. Kennedy, J. Caviness, V. S., S. V. Faraone, M. T. Tsuang, Normal sexual dimorphism of the

adult human brain assessed by in vivo magnetic resonance imaging, Cereb Cortex 11 (6) (2001) 490–7. doi:10.1093/cercor/11.6.490.
URL https://www.ncbi.nlm.nih.gov/pubmed/11375910

[4] B. S. McEwen, Gonadal steroid influences on brain development and sexual differentiation, Int Rev Physiol 27 (1983) 99–145.
URL https://www.ncbi.nlm.nih.gov/pubmed/6303978

[5] J. C. Woodson, R. A. Gorski, Structural sex differences in the mammalian brain: Reconsidering the male/female dichotomy, Sexual differentiation of the brain (2000) 229–255.

[6] J. B. Becker, G. F. Koob, Sex differences in animal models: focus on addiction, Pharmacological reviews 68 (2) (2016) 242–263.

[7] K. Vogeley, T. Schneider-Axmann, U. Pfeiffer, R. Tepest, T. A. Bayer, B. Bogerts, W. G. Honer, P. Falkai, Disturbed gyrification of the prefrontal region in male schizophrenic patients: a morphometric postmortem study, American Journal of Psychiatry 157 (1) (2000) 34–39.

[8] S. F. Witelson, Hand and sex differences in the isthmus and genu of the human corpus callosum: a postmortem morphological study, Brain 112 (3) (1989) 799–835.

[9] S. Witelson, H. Beresh, D. Kigar, Intelligence and brain size in 100 postmortem brains: sex, lateralization and age factors, Brain 129 (2) (2005) 386–398.

[10] L. Fan, Y. Tang, B. Sun, G. Gong, Z. J. Chen, X. Lin, T. Yu, Z. Li, A. C. Evans, S. Liu, Sexual dimorphism and asymmetry in human cerebellum: an mri-based morphometric study, Brain Res 1353 (2010) 60–73. doi:10.1016/j.brainres.2010.07.031.
URL https://www.ncbi.nlm.nih.gov/pubmed/20647004

[11] P. A. Filipek, C. Richelme, D. N. Kennedy, J. Caviness, V. S., The young adult human brain: an mri-based morphometric analysis, Cereb Cortex 4 (4) (1994) 344–60. doi:10.1093/cercor/4.4.344.
URL https://www.ncbi.nlm.nih.gov/pubmed/7950308

[12] M. Flaum, V. Swayze, D. O'leary, W. Yuh, J. Ehrhardt, S. Arndt, N. Andreasen, Brain morphology in schizophrenia: effects of diagnosis, laterality and gender, Am J Psychiatry 152 (1995) 704–714.

[13] J. Hänggi, A. Buchmann, C. R. Mondadori, K. Henke, L. Jäncke, C. Hock, Sexual dimorphism in the parietal substrate associated with visuospatial cognition independent of general intelligence, Journal of cognitive neuroscience 22 (1) (2010) 139–155.

[14] J. Sacher, J. Neumann, H. Okon-Singer, S. Gotowiec, A. Villringer, Sexual dimorphism in the human brain: evidence from neuroimaging, Magnetic resonance imaging 31 (3) (2013) 366–375.

[15] L. Wang, H. Shen, F. Tang, Y. Zang, D. Hu, Combined structural and resting-state functional mri analysis of sexual dimorphism in the young adult human brain: an mvpa approach, Neuroimage 61 (4) (2012) 931–940.

[16] D.-L. Feis, K. H. Brodersen, D. Y. von Cramon, E. Luders, M. Tittgemeyer, Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion mri data, Neuroimage 70 (2013) 250–257.

[17] M. Nieuwenhuis, H. G. Schnack, N. E. van Haren, J. Lappin, C. Morgan, A. A. Reinders, D. Gutierrez-Tordesillas, R. Roiz-Santiañez, M. S. Schaufelberger, P. G. Rosa, Multi-center mri prediction models: Predicting sex and illness course in first episode psychosis patients, Neuroimage 145 (2017) 246–253.

[18] J. Breslau, S. E. Gilman, B. D. Stein, T. Ruder, T. Gmelin, E. Miller, Sex differences in recent first-onset depression in an epidemiological sample of adolescents, Translational psychiatry 7 (5) (2017) e1139.

[19] M. J. Van Putten, S. Olbrich, M. Arns, Predicting sex from brain rhythms with deep learning, Scientific reports 8 (1) (2018) 3069.

[20] J. Xin, X. Y. Zhang, Y. Tang, Y. Yang, Brain differences between men and women: Evidence from deep learning, Frontiers in neuroscience 13 (2019) 185.

[21] B. Brie, M. C. Ramirez, C. De Winne, F. L. Vicchi, L. Villarruel, E. Sorianello, P. Catalano, A. M. Ornstein, D. Becu-Villalobos, Brain control of sex-

ually dimorphic liver function and disease: The endocrine connection, Cellular and molecular neurobiology 39 (2) (2019) 169–180.

[22] L. Egloff, C. Lenz, E. Studerus, F. Harrisberger, R. Smieskova, A. Schmidt, C. Huber, A. Simon, U. E. Lang, A. Riecher-Rössler, Sexually dimorphic subcortical brain volumes in emerging psychosis, Schizophrenia research 199 (2018) 257–265.

[23] N. Jahanshad, P. M. Thompson, Multimodal neuroimaging of male and female brain structure in health and disease across the life span, Journal of neuroscience research 95 (1-2) (2017) 371–379.

[24] K. E. Lind, E. J. Gutierrez, D. J. Yamamoto, M. F. Regner, S. A. McKee, J. Tanabe, Sex disparities in substance abuse research: Evaluating 23 years of structural neuroimaging studies, Drug and alcohol dependence 173 (2017) 92–98.

[25] A. Retico, A. Giuliano, R. Tancredi, A. Cosenza, F. Apicella, A. Narzisi, L. Biagi, M. Tosetti, F. Muratori, S. Calderoni, The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study, Molecular autism 7 (1) (2016) 5.

[26] F. Y. Womer, Y. Tang, M. P. Harms, C. Bai, M. Chang, X. Jiang, S. Wei, F. Wang, D. M. Barch, Sexual dimorphism of the cerebellar vermis in schizophrenia, Schizophrenia research 176 (2-3) (2016) 164–170.

[27] G. Golarai, K. Grill-Spector, A. L. Reiss, Autism and the development of face processing, Clinical neuroscience research 6 (3-4) (2006) 145–160.

[28] N. Liu, S. Cliffer, A. H. Pradhan, A. Lightbody, S. S. Hall, A. L. Reiss, Optical-imaging-based neurofeedback to enhance therapeutic intervention in adolescents with autism: methodology and initial data, Neurophotonics 4 (1) (2016) 011003.

[29] K. Pierce, V. H. Gazestani, E. Bacon, C. C. Barnes, D. Cha, S. Nalabolu, L. Lopez, A. Moore, S. Pence-Stophaeros, E. Courchesne, Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months, JAMA pediatrics 173 (6) (2019) 578–587.

[30] G. K. Strickler, P. W. Kreiner, J. F. Halpin, E. Doyle, L. J. Paulozzi, Opioid prescribing behaviors – prescription behavior surveillance system, 11 states, 2010-2016., MMWR Surveillance Summaries 69 (1) (2020) 1 – 14.

[31] J. M. Cyranowski, E. Frank, E. Young, M. K. Shear, Adolescent onset of the gender difference in lifetime rates of major depression: a theoretical model, Archives of general psychiatry 57 (1) (2000) 21–27.

[32] J. N. Giedd, Structural magnetic resonance imaging of the adolescent brain, Ann N Y Acad Sci 1021 (2004) 77–85. doi:10.1196/annals.1308.009.
URL https://www.ncbi.nlm.nih.gov/pubmed/15251877

[33] R. Petrican, M. J. Taylor, C. L. Grady, Trajectories of brain system maturation from childhood to older adulthood: Implications for lifespan cognitive functioning, Neuroimage 163 (2017) 125–149. doi:10.1016/j.neuroimage.2017.09.025.
URL https://www.ncbi.nlm.nih.gov/pubmed/28917697

[34] E. S. Johnson, A. C. Meade, Developmental patterns of spatial ability: an early sex difference, Child Dev 58 (3) (1987) 725–40.
URL https://www.ncbi.nlm.nih.gov/pubmed/3608645

[35] J. H. Gilmore, W. Lin, M. W. Prastawa, C. B. Looney, Y. S. Vetsa, R. C. Knickmeyer, D. D. Evans, J. K. Smith, R. M. Hamer, J. A. Lieberman, G. Gerig, Regional gray matter growth, sexual dimorphism, and cerebral asymmetry in the neonatal brain, J Neurosci 27 (6) (2007) 1255–60.
URL https://www.ncbi.nlm.nih.gov/pubmed/17287499

[36] J. N. Giedd, A. Raznahan, A. Alexander-Bloch, E. Schmitt, N. Gogtay, J. L. Rapoport, Child psychiatry branch of the national institute of mental health longitudinal structural magnetic resonance imaging study of human brain development, Neuropsychopharmacology 40 (1) (2015) 43.

[37] C. Mankiw, M. T. M. Park, P. Reardon, A. M. Fish, L. S. Clasen, D. Greenstein, J. N. Giedd, J. D. Blumenthal, J. P. Lerch, M. M. Chakravarty, Allometric analysis detects brain size-independent effects of sex and sex chromosome complement on human cerebellar organization, Journal of Neuroscience 37 (21) (2017) 5221–5231.

[38] A. Pfefferbaum, D. Kwon, T. Brumback, W. K. Thompson, K. Cummins, S. F. Tapert, S. A. Brown, I. M. Colrain, F. C. Baker, D. Prouty, M. D. De Bellis, D. B. Clark, B. J. Nagel, W. Chu, S. H. Park, K. M. Pohl, E. V. Sullivan, Altered brain developmental trajectories in adolescents after initiating drinking, Am J Psychiatry 175 (4) (2018) 370–380. doi:10.1176/appi.ajp.2017.17040469.
URL https://www.ncbi.nlm.nih.gov/pubmed/29084454

[39] A. Pfefferbaum, T. Rohlfing, K. M. Pohl, B. Lane, W. Chu, D. Kwon, B. Nolan Nichols, S. A. Brown, S. F. Tapert, K. Cummins, W. K. Thompson, T. Brumback, M. J. Meloy, T. L. Jernigan, A. Dale, I. M. Colrain, F. C. Baker, D. Prouty, M. D. De Bellis, J. T. Voyvodic, D. B. Clark, B. Luna, T. Chung, B. J. Nagel, E. V. Sullivan, Adolescent development of cortical and white matter structure in the ncanda sample: Role of sex, ethnicity, puberty, and alcohol drinking, Cereb Cortex 26 (10) (2016) 4101–21. doi:10.1093/cercor/bhv205.
URL http://www.ncbi.nlm.nih.gov/pubmed/26408800

[40] C. K. Tamnes, M. M. Herting, A.-L. Goddings, R. Meuwese, S.-J. Blakemore, R. E. Dahl, B. Güroğlu, A. Raznahan, E. R. Sowell, E. A. Crone, Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness, Journal of Neuroscience 37 (12) (2017) 3402–3412.

[41] R. K. Lenroot, N. Gogtay, D. K. Greenstein, E. M. Wells, G. L. Wallace, L. S. Clasen, J. D. Blumenthal, J. Lerch, A. P. Zijdenbos, A. C. Evans, P. M. Thompson, J. N. Giedd, Sexual dimorphism of brain developmental trajectories during childhood and adolescence, Neuroimage 36 (4) (2007) 1065–73. doi:10.1016/j.neuroimage.2007.03.053.
URL https://www.ncbi.nlm.nih.gov/pubmed/17513132

[42] L. M. Wierenga, M. G. Bos, E. Schreuders, F. vd Kamp, J. S. Peper, C. K. Tamnes, E. A. Crone, Unraveling age, puberty and testosterone effects on subcortical brain development across adolescence, Psychoneuroendocrinology 91 (2018) 105–114.

[43] B. Aubert-Broche, V. S. Fonov, D. Garcia-Lorenzo, A. Mouiha, N. Guizard, P. Coupé, S. F. Eskildsen, D. L. Collins, A new method for structural volume

analysis of longitudinal brain mri data and its application in studying the growth trajectories of anatomical brain structures in childhood, Neuroimage 82 (2013) 393–402.

[44] S. Ducharme, M. D. Albaugh, T.-V. Nguyen, J. J. Hudziak, J. Mateos-Pérez, A. Labbe, A. C. Evans, S. Karama, B. D. C. Group, Trajectories of cortical surface area and cortical volume maturation in normal brain development, Data in brief 5 (2015) 929–938.

[45] M. M. Herting, C. Johnson, K. L. Mills, N. Vijayakumar, M. Dennison, C. Liu, A.-L. Goddings, R. E. Dahl, E. R. Sowell, S. Whittle, Development of subcortical volumes across adolescence in males and females: A multisample study of longitudinal changes, NeuroImage 172 (2018) 194–205.

[46] K. L. Mills, F. Lalonde, L. S. Clasen, J. N. Giedd, S.-J. Blakemore, Developmental changes in the structure of the social brain in late childhood and adolescence, Social cognitive and affective neuroscience 9 (1) (2012) 123–131.

[47] K. Narvacan, S. Treit, R. Camicioli, W. Martin, C. Beaulieu, Evolution of deep gray matter volume across the human lifespan, Human brain mapping 38 (8) (2017) 3771–3790.

[48] N. Vijayakumar, N. B. Allen, G. Youssef, M. Dennison, M. Yücel, J. G. Simmons, S. Whittle, Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume, Human brain mapping 37 (6) (2016) 2027–2038.

[49] C. Wierenga, A. Bischoff-Grethe, A. J. Melrose, E. Grenesko-Stevens, Z. Irvine, A. Wagner, A. Simmons, S. Matthews, W.-Y. W. Yau, C. Fennema-Notestine, Altered bold response during inhibitory and error processing in adolescents with anorexia nervosa, PloS one 9 (3) (2014) e92017.

[50] A. N. Ruigrok, G. Salimi-Khorshidi, M.-C. Lai, S. Baron-Cohen, M. V. Lombardo, R. J. Tait, J. Suckling, A meta-analysis of sex differences in human brain structure, Neuroscience & Biobehavioral Reviews 39 (2014) 34–50.

[51] C. Sanchis Segura, N. Aguirre, A. J. Cruz-Gómez, N. Solozano, C. Forn, Do gender-related stereotypes affect spatial performance? exploring when, how and

to whom using a chronometric two-choice mental rotation task, Frontiers in psychology 9 (2018) 1261.

[52] A. Etchell, A. Adhikari, L. S. Weinberg, A. L. Choo, E. O. Garnett, H. M. Chow, S.-E. Chang, A systematic literature review of sex differences in childhood language and brain development, Neuropsychologia 114 (2018) 19–31.

[53] E. Luders, A. W. Toga, P. M. Thompson, Why size matters: differences in brain volume account for apparent sex differences in callosal anatomy: the sexual dimorphism of the corpus callosum, Neuroimage 84 (2014) 820–824.

[54] K. S. Sawyer, N. Maleki, G. Papadimitriou, N. Makris, M. Oscar-Berman, G. J. Harris, Cerebral white matter sex dimorphism in alcoholism: a diffusion tensor imaging study, Neuropsychopharmacology 43 (9) (2018) 1876.

[55] E. V. Sullivan, M. J. Rosenbloom, J. E. Desmond, A. Pfefferbaum, Sex differences in corpus callosum size: relationship to age and intracranial size, Neurobiology of aging 22 (4) (2001) 603–611.

[56] G. Perlaki, G. Orsi, E. Plozer, A. Altbacker, G. Darnai, S. A. Nagy, R. Horvath, A. Toth, T. Doczi, N. Kovacs, Are there any gender differences in the hippocampus volume after head-size correction? a volumetric and voxel-based morphometric study, Neuroscience letters 570 (2014) 119–123.

[57] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, M. R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nature Reviews Neuroscience 14 (5) (2013) 365.

[58] R. K. Lenroot, N. Gogtay, D. K. Greenstein, E. M. Wells, G. L. Wallace, L. S. Clasen, J. D. Blumenthal, J. Lerch, A. P. Zijdenbos, A. C. Evans, Sexual dimorphism of brain developmental trajectories during childhood and adolescence, Neuroimage 36 (4) (2007) 1065–1073.

[59] H. J. Kim, N. Kim, S. Kim, S. Hong, K. Park, S. Lim, J. M. Park, B. Na, Y. Chae, J. Lee, S. Yeo, I. H. Choe, S. Y. Cho, G. Cho, Sex differences in amygdala subregions: evidence from subregional shape analysis, Neuroimage 60 (4) (2012) 2054–61. doi:10.1016/j.neuroimage.2012.02.025.
URL https://www.ncbi.nlm.nih.gov/pubmed/22374477

[60] S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, E. Adeli, End-to-end alzheimer's disease diagnosis and biomarker identification, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 337–345.

[61] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012.

[62] S. N. Hatton, M. D. Cornejo, C. Makowski, D. A. Fair, A. S. Dick, M. T. Sutherland, B. Casey, D. M. Barch, M. P. Harms, R. Watts, J. M. Bjork, H. P. Garavan, L. Hilmer, C. J. Pung, C. S. Sicat, J. Kuperman, H. Bartsch, F. Xue, M. M. Heitzeg, A. R. Laird, T. T. Trinh, R. Gonzalez, S. F. Tapert, M. C. Riedel, L. M. Squeglia, L. W. Hyde, M. D. Rosenberg, E. A. Earl, K. D. Howlett, F. C. Baker, M. Soules, J. Diaz, O. R. d. Leon, W. K. Thompson, M. C. Neale, M. Herting, E. R. Sowell, R. P. Alvarez, S. W. Hawes, M. Sanchez, J. Bodurka, F. J. Breslin, A. S. Morris, M. P. Paulus, W. K. Simmons, J. R. Polimeni, A. v. d. Kouwe, A. S. Nencka, K. M. Gray, C. Pierpaoli, J. A. Matochik, A. Noronha, W. M. Aklin, K. Conway, M. Glantz, E. Hoffman, R. Little, M. Lopez, V. Pariyadath, S. R. Weiss, D. L. Wolff-Hughes, R. DelCarmen-Wiggins, S. W. F. Ewing, O. Miranda-Dominguez, B. J. Nagel, A. J. Perrone, D. T. Sturgeon, A. Goldstone, A. Pfefferbaum, K. M. Pohl, D. Prouty, K. Uban, S. Y. Bookheimer, M. Dapretto, A. Galvan, K. Bagot, J. Giedd, M. A. Infante, J. Jacobus, K. Patrick, P. D. Shilling, R. Desikan, Y. Li, L. Sugrue, M. T. Banich, N. Friedman, J. K. Hewitt, C. Hopfer, J. Sakai, J. Tanabe, L. B. Cottler, S. J. Nixon, L. Chang, C. Cloak, T. Ernst, G. Reeves, D. N. Kennedy, S. Heeringa, S. Peltier, J. Schulenberg, et al., Image processing and analysis methods for the adolescent brain cognitive development study, Neuroimage, In Press.

[63] M. Luciana, J. Bjork, B. Nagel, D. Barch, R. Gonzalez, S. Nixon, M. Banich, Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery, Developmental cognitive neuroscience 32 (2018) 67–79.

[64] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: International Conference on Learning Representations (ICLR) Workshop, 2014.

[65] K. M. Pohl, W. Thompson, E. Adeli, M. G. Linguraru, Adolescent Brain Cognitive Development Neurocognitive Prediction Challenge, Vol. 11791 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, Germany, 2019.

[66] E. V. Sullivan, T. Brumback, S. F. Tapert, R. Fama, D. Prouty, S. A. Brown, K. Cummins, W. K. Thompson, I. M. Colrain, F. C. Baker, Cognitive, emotion control, and motor performance of adolescents in the ncanda study: Contributions from alcohol consumption, age, sex, ethnicity, and family history of addiction, Neuropsychology 30 (4) (2016) 449.

[67] M. A. Carskadon, C. Acebo, A self-administered rating scale for pubertal development, Journal of Adolescent Health 14 (3) (1993) 190–195.

[68] A. C. Petersen, L. Crockett, M. Richards, A. Boxer, A self-report measure of pubertal status: Reliability, validity, and initial norms, Journal of youth and adolescence 17 (2) (1988) 117–133.

[69] D. S. Freedman, N. F. Butte, E. M. Taveras, E. A. Lundeen, H. M. Blanck, A. B. Goodman, C. L. Ogden, Bmi z-scores are a poor indicator of adiposity among 2- to 19-year-olds with very high bmis, nhanes 1999-2000 to 2013-2014, Obesity (Silver Spring) 25 (4) (2017) 739–746. doi:10.1002/oby.21782.
URL https://www.ncbi.nlm.nih.gov/pubmed/28245098

[70] W. K. Thompson, D. M. Barch, J. M. Bjork, R. Gonzalez, B. J. Nagel, S. J. Nixon, M. Luciana, The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: Findings from the abcd study's baseline neurocognitive battery, Dev Cogn Neurosci 36 (2019) 100606. doi:10.1016/j.dcn.2018.12.004.
URL https://www.ncbi.nlm.nih.gov/pubmed/30595399

[71] H. Garavan, H. Bartsch, K. Conway, A. Decastro, R. Z. Goldstein, S. Heeringa, T. Jernigan, A. Potter, W. Thompson, D. Zahs, Recruiting the abcd sample: Design considerations and procedures, Dev Cogn Neurosci 32 (2018) 16–22. doi:10.1016/j.dcn.2018.04.004.
URL https://www.ncbi.nlm.nih.gov/pubmed/29703560

[72] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, C. Barillot, An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images,

IEEE Trans Med Imaging 27 (4) (2008) 425–41. doi:10.1109/TMI.2007.906087.
URL https://www.ncbi.nlm.nih.gov/pubmed/18390341

[73] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4itk: improved n3 bias correction, IEEE Trans Med Imaging 29 (6) (2010) 1310–20. doi:10.1109/TMI.2010.2046908.
URL https://www.ncbi.nlm.nih.gov/pubmed/20378467

[74] T. Rohlfing, D. B. Russakoff, J. Maurer, C. R., Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation, IEEE Trans Med Imaging 23 (8) (2004) 983–94. doi:10.1109/TMI.2004.830803.
URL https://www.ncbi.nlm.nih.gov/pubmed/15338732

[75] S. M. Smith, Fast robust automated brain extraction, Hum Brain Mapp 17 (3) (2002) 143–55. doi:10.1002/hbm.10062.
URL http://www.ncbi.nlm.nih.gov/pubmed/12391568

[76] R. W. Cox, Afni: software for analysis and visualization of functional magnetic resonance neuroimages, Comput Biomed Res 29 (3) (1996) 162–73.
URL https://www.ncbi.nlm.nih.gov/pubmed/8812068

[77] S. A. Sadananthan, W. Zheng, M. W. Chee, V. Zagorodnov, Skull stripping using graph cuts, Neuroimage 49 (1) (2010) 225–39. doi:10.1016/j.neuroimage.2009.08.050.
URL https://www.ncbi.nlm.nih.gov/pubmed/19732839

[78] J. E. Iglesias, C. Y. Liu, P. M. Thompson, Z. Tu, Robust brain extraction across datasets and comparison with publicly available methods, IEEE Trans Med Imaging 30 (9) (2011) 1617–34. doi:10.1109/TMI.2011.2138152.
URL http://www.ncbi.nlm.nih.gov/pubmed/21880566

[79] T. Rohlfing, N. M. Zahr, E. V. Sullivan, A. Pfefferbaum, The sri24 multichannel atlas of normal adult human brain structure, Hum Brain Mapp 31 (5) (2010) 798–819. doi:10.1002/hbm.20906.
URL https://www.ncbi.nlm.nih.gov/pubmed/20017133

[80] D. Nie, H. Zhang, E. Adeli, L. Liu, D. Shen, 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients, Med Image Comput Comput Assist Interv 9901 (2016) 212–220. doi:10.1007/978-3-319-46723-8_25.
URL https://www.ncbi.nlm.nih.gov/pubmed/28149967

[81] A. C. Hill, A. R. Laird, J. L. Robinson, Gender differences in working memory networks: a brainmap meta-analysis, Biol Psychol 102 (2014) 18–29. doi:10.1016/j.biopsycho.2014.06.008.
URL https://www.ncbi.nlm.nih.gov/pubmed/25042764

[82] A. Phinyomark, B. A. Hettinga, S. T. Osis, R. Ferber, Gender and age-related differences in bilateral lower extremity mechanics during treadmill running, PLoS One 9 (8).

[83] M. Hirnstein, K. Hugdahl, M. Hausmann, Cognitive sex differences and hemispheric asymmetry: A critical review of 40 years of research, Laterality: Asymmetries of Body, Brain and Cognition 24 (2) (2019) 204–252.

[84] J. T. Weinhandl, M. Joshi, K. M. O'Connor, Gender comparisons between unilateral and bilateral landings, Journal of applied biomechanics 26 (4) (2010) 444–453.

[85] F. Román, F. A. García-Sánchez, J. M. Martínez-Selva, J. Gómez-Amor, E. Carrillo, Sex differences and bilateral electrodermal activity, The Pavlovian journal of biological science 24 (4) (1989) 150–155.

[86] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, Vol. 14, Montreal, Canada, 1995, pp. 1137–1145.

[87] I. Oksuz, B. Ruijsink, E. Puyol-Anton, J. R. Clough, G. Cruz, A. Bustin, C. Prieto, R. Botnar, D. Rueckert, J. A. Schnabel, A. P. King, Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning, Med Image Anal 55 (2019) 136–147. doi:10.1016/j.media.2019.04.009.
URL https://www.ncbi.nlm.nih.gov/pubmed/31055126

[88] S. H. Park, Y. Zhang, D. Kwon, Q. Zhao, N. M. Zahr, A. Pfefferbaum, E. V. Sullivan, K. M. Pohl, Alcohol use effects on adolescent brain development revealed by simultaneously removing confounding factors, identifying morphometric patterns, and classifying individuals, Scientific reports 8 (1) (2018) 8297.

[89] E. Adeli, X. Li, D. Kwon, Y. Zhang, K. Pohl, Logistic regression confined by cardinality-constrained sample and feature selection, IEEE transactions on pattern analysis and machine intelligence.

[90] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic regression, Springer, 2002.

[91] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 27.

[92] A. Liaw, M. Wiener, Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[93] B. B. Avants, C. L. Epstein, M. Grossman, J. C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Med Image Anal 12 (1) (2008) 26–41. doi:10.1016/j.media.2007.06.004.
URL https://www.ncbi.nlm.nih.gov/pubmed/17659998

[94] B. B. Avants, N. J. Tustison, J. Wu, P. A. Cook, J. C. Gee, An open source multivariate framework for n-tissue segmentation with evaluation on public data, Neuroinformatics 9 (4) (2011) 381–400. doi:10.1007/s12021-011-9109-y.
URL https://www.ncbi.nlm.nih.gov/pubmed/21373993

[95] C. Destrieux, B. Fischl, A. Dale, E. Halgren, Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature, Neuroimage 53 (1) (2010) 1–15.

[96] R. M. Baron, D. A. Kenny, The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, Journal of personality and social psychology 51 (6) (1986) 1173.

[97] K. Pearson, X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50 (302) (1900) 157–175.

[98] pygrowup (2017). [link].
URL https://pypi.org/project/pygrowup/

[99] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.

[100] A. Gulli, S. Pal, Deep Learning with Keras, Packt Publishing Ltd, 2017.

[101] R. Kotikalapudi, contributors, keras-vis, https://github.com/raghakot/keras-vis (2017).

[102] H. Madsen, P. Thyregod, Introduction to general and generalized linear models. CRC Press, 2010., CRC Press, 2010.

[103] J. P. Shaffer, Multiple hypothesis testing, Ann. Rev. Psych. 46 (1995) 561–584.

[104] R. A. Fisher, The logic of inductive inference, Journal of the Royal Statistical Society 98 (1) (1935) 43.

[105] E. Adeli, D. Kwon, Q. Zhao, A. Pfefferbaum, N. M. Zahr, E. V. Sullivan, K. M. Pohl, Chained regularization for identifying brain patterns specific to hiv infection, NeuroImage 183 (2018) 425–437.

[106] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (3) (1988) 837–45.
URL http://www.ncbi.nlm.nih.gov/pubmed/3203132

[107] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

[108] S.-C. Chung, B.-Y. Lee, G.-R. Tack, S.-Y. Lee, J.-S. Eom, J.-H. Sohn, Effects of age, gender, and weight on the cerebellar volume of korean people, Brain research 1042 (2) (2005) 233–235.

[109] T. Green, K. C. Fierro, M. M. Raman, L. Foland-Ross, D. S. Hong, A. L. Reiss, Sex differences in amygdala shape: Insights from turner syndrome, Hum Brain Mapp 37 (4) (2016) 1593–601. doi:10.1002/hbm.23122.
URL https://www.ncbi.nlm.nih.gov/pubmed/26819071

[110] N. Raz, F. Gunning-Dixon, D. Head, A. Williamson, J. D. Acker, Age and sex differences in the cerebellum and the ventral pons: a prospective mr study of healthy adults, American Journal of Neuroradiology 22 (6) (2001) 1161–1167.

[111] H. Tiemeier, R. K. Lenroot, D. K. Greenstein, L. Tran, R. Pierson, J. N. Giedd, Cerebellum development during childhood and adolescence: a longitudinal morphometric mri study, Neuroimage 49 (1) (2010) 63–70. doi:10.1016/j.neuroimage.2009.08.016.
URL https://www.ncbi.nlm.nih.gov/pubmed/19683586

[112] P. Nopoulos, M. Flaum, D. O'Leary, N. C. Andreasen, Sexual dimorphism in the human brain: evaluation of tissue volume, tissue composition and surface anatomy using magnetic resonance imaging, Psychiatry Research: Neuroimaging 98 (1) (2000) 1–13.

[113] E. V. Sullivan, T. Brumback, S. F. Tapert, S. A. Brown, F. C. Baker, I. M. Colrain, D. Prouty, M. D. De Bellis, D. B. Clark, B. J. Nagel, Disturbed cerebellar growth trajectories in adolescents who initiate alcohol drinking, Biological Psychiatry.

[114] C. A. Szabo, J. L. Lancaster, J. Xiong, C. Cook, P. Fox, Mr imaging volumetry of subcortical structures and cerebellar hemispheres in normal persons, American Journal of Neuroradiology 24 (4) (2003) 644–647.

[115] B. Luna, K. E. Garver, T. A. Urban, N. A. Lazar, J. A. Sweeney, Maturation of cognitive processes from late childhood to adulthood, Child Development 75 (5) (2004) 1357–1372.

[116] J. P. Aggleton, S. M. O'Mara, S. D. Vann, N. F. Wright, M. Tsanov, J. T. Erichsen, Hippocampal-anterior thalamic pathways for memory: uncovering a network of direct and indirect actions, Eur J Neurosci 31 (12) (2010) 2292–307. doi:10.1111/j.1460-9568.2010.07251.x.
URL https://www.ncbi.nlm.nih.gov/pubmed/20550571

[117] P. K. Pilly, M. D. Howard, R. Bhattacharyya, Modeling contextual modulation of memory associations in the hippocampus, Front Hum Neurosci 12 (2018) 442. doi:10.3389/fnhum.2018.00442.
URL https://www.ncbi.nlm.nih.gov/pubmed/30473660

[118] S. S. Dikmen, P. J. Bauer, S. Weintraub, D. Mungas, J. Slotkin, J. L. Beaumont, R. Gershon, N. R. Temkin, R. K. Heaton, Measuring episodic memory across the lifespan: Nih toolbox picture sequence memory test, Journal of the International Neuropsychological Society 20 (6) (2014) 611–619.

[119] T. Frodl, E. M. Meisenzahl, T. Zetzsche, C. Born, C. Groll, M. Jager, G. Leinsinger, R. Bottlender, K. Hahn, H. J. Moller, Hippocampal changes in patients with a first episode of major depression, Am J Psychiatry 159 (7) (2002) 1112–8. doi:10.1176/appi.ajp.159.7.1112.
URL https://www.ncbi.nlm.nih.gov/pubmed/12091188

[120] J. N. Giedd, A. C. Vaituzis, S. D. Hamburger, N. Lange, J. C. Rajapakse, D. Kaysen, Y. C. Vauss, J. L. Rapoport, Quantitative mri of the temporal lobe, amygdala, and hippocampus in normal human development: ages 4-18 years, J Comp Neurol 366 (2) (1996) 223–30. doi:10.1002/(SICI)1096-9861(19960304)366:2<223::AID-CNE3>3.0.CO;2-7.
URL https://www.ncbi.nlm.nih.gov/pubmed/8698883

[121] M. H. Teicher, S. L. Andersen, A. Polcari, C. M. Anderson, C. P. Navalta, D. M. Kim, The neurobiological consequences of early stress and childhood maltreatment, Neurosci Biobehav Rev 27 (1-2) (2003) 33–44.
URL https://www.ncbi.nlm.nih.gov/pubmed/12732221

[122] M. R. Trenerry, C. R. Jack Jr, G. D. Cascino, F. W. Sharbrough, R. J. Ivnik, Gender differences in post-temporal lobectomy verbal memory and relationships between mri hippocampal volumes and preoperative verbal memory, Epilepsy research 20 (1) (1995) 69–76.

[123] K. D. Young, P. S. Bellgowan, J. Bodurka, W. C. Drevets, Functional neuroimaging of sex differences in autobiographical memory recall, Hum Brain Mapp 34 (12) (2013) 3320–32. doi:10.1002/hbm.22144.
URL https://www.ncbi.nlm.nih.gov/pubmed/22807028

[124] X. Yang, Z. Peng, X. Ma, Y. Meng, M. Li, J. Zhang, X. Song, Y. Liu, H. Fan, L. Zhao, W. Deng, T. Li, X. Ma, Sex differences in the clinical characteristics and brain gray matter volume alterations in unmedicated patients with major depressive disorder, Sci Rep 7 (1) (2017) 2515. doi:10.1038/s41598-017-02828-4.
URL https://www.ncbi.nlm.nih.gov/pubmed/28559571

[125] X. Yang, Z. Peng, X. Ma, Y. Meng, M. Li, J. Zhang, X. Song, Y. Liu, H. Fan, L. Zhao, Sex differences in the clinical characteristics and brain gray matter volume alterations in unmedicated patients with major depressive disorder, Scientific reports 7 (1) (2017) 2515.

[126] E. F. Saunders, R. Nazir, M. Kamali, K. A. Ryan, S. Evans, S. Langenecker, A. J. Gelenberg, M. G. McInnis, Gender differences, clinical correlates, and longitudinal outcome of bipolar disorder with comorbid migraine, J Clin Psychiatry 75 (5) (2014) 512–9. doi:10.4088/JCP.13m08623.
URL https://www.ncbi.nlm.nih.gov/pubmed/24816075

[127] J. N. Giedd, F. X. Castellanos, J. C. Rajapakse, A. C. Vaituzis, J. L. Rapoport, Sexual dimorphism of the developing human brain, Progress in Neuro-Psychopharmacology and Biological Psychiatry 21 (8) (1997) 1185–1201.

[128] L. M. Wierenga, J. A. Sexton, P. Laake, J. N. Giedd, C. K. Tamnes, N. Pediatric Imaging, S. Genetics, A key characteristic of sex differences in the developing brain: Greater variability in brain structure of boys than girls, Cereb Cortex 28 (8) (2018) 2741–2751. doi:10.1093/cercor/bhx154.
URL https://www.ncbi.nlm.nih.gov/pubmed/28981610

[129] R. Arden, R. Plomin, Sex differences in variance of intelligence across childhood, Personality and Individual Differences 41 (1) (2006) 39–48.

[130] A. Baye, C. Monseur, Gender differences in variability and extreme scores in an international context, Large-scale Assessments in Education 4 (1) (2016) 1.

[131] S. R. Blume, M. Freedberg, J. E. Vantrease, R. Chan, M. Padival, M. J. Record, M. R. DeJoseph, J. H. Urban, J. A. Rosenkranz, Sex- and estrus-dependent differences in rat basolateral amygdala, J Neurosci 37 (44) (2017) 10567–10586. doi:10.1523/JNEUROSCI.0758-17.2017.
URL https://www.ncbi.nlm.nih.gov/pubmed/28954870

[132] V. Menon, L. Uddin, Saliency, switching, attention and control: A network model of insula function, Brain structure & function 214 (2010) 655–67. doi:10.1007/s00429-010-0262-0.

[133] R. E. Gur, R. C. Gur, Sex differences in brain and behavior in adolescence: Findings from the philadelphia neurodevelopmental cohort, Neuroscience & Biobehavioral Reviews 70 (2016) 159–170.

[134] R. C. Gur, R. E. Gur, Complementarity of sex differences in brain and behavior: From laterality to multimodal neuroimaging, Journal of neuroscience research 95 (1-2) (2017) 189–199.

[135] N. Akshoomoff, E. Newman, W. K. Thompson, C. McCabe, C. S. Bloss, L. Chang, D. G. Amaral, B. Casey, T. M. Ernst, J. A. Frazier, The nih toolbox cognition battery: Results from a large normative developmental sample (ping), Neuropsychology 28 (1) (2014) 1.

[136] K. Bäckström, M. Nazari, I. Gu, A. Jakola, An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images, 2018, pp. 149–153. doi:10.1109/ISBI.2018.8363543.

[137] A. Llera, T. Wolfers, P. Mulders, C. Beckmann, Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior, eLife 8. doi:10.7554/eLife.44443.

[138] W. Lin, T. Tong, Q. Gao, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment, Frontiers in Neuroscience 12 (2018) 777. doi:10.3389/fnins.2018.00777.

[139] M. Liu, J. Zhang, D. Nie, P.-T. Yap, Anatomical landmark based deep feature representation for mr images in brain disease diagnosis, IEEE Journal of Biomedical and Health Informatics PP (2018) 1–1. doi:10.1109/JBHI.2018.2791863.

[140] M. C. Murphy, D. T. Jones, C. R. Jack Jr, K. J. Glaser, M. L. Senjem, A. Manduca, J. P. Felmlee, R. E. Carter, R. L. Ehman, J. Huston III, Regional brain stiffness changes across the alzheimer's disease spectrum, NeuroImage: Clinical 10 (2016) 283–290.

[141] A. Wittek, G. Joldes, M. Couton, S. K. Warfield, K. Miller, Patient-specific non-linear finite element modelling for predicting soft organ deformation in real-time; application to non-rigid neuroimage registration, Progress in biophysics and molecular biology 103 (2-3) (2010) 292–303.

[142] R. C. Gershon, M. V. Wagster, H. C. Hendrie, N. A. Fox, K. F. Cook, C. J. Nowinski, Nih toolbox for assessment of neurological and behavioral function, Neurology 80 (11 Supplement 3) (2013) S2–S6.

[143] R. J. Hodes, T. R. Insel, S. C. Landis, The nih toolbox: Setting a standard for biomedical research, Neurology 80 (11 Supplement 3) (2013) S1–S1.

[144] K. B. Casaletto, A. Umlauf, J. Beaumont, R. Gershon, J. Slotkin, N. Akshoomoff, R. K. Heaton, Demographically corrected normative standards for the english version of the nih toolbox cognition battery, Journal of the International Neuropsychological Society 21 (5) (2015) 378–391.

[145] R. C. Gershon, K. F. Cook, D. Mungas, J. J. Manly, J. Slotkin, J. L. Beaumont, S. Weintraub, Language measures of the nih toolbox cognition battery, Journal of the International Neuropsychological Society 20 (6) (2014) 642–651.

[146] R. C. Gershon, J. Slotkin, J. J. Manly, D. L. Blitz, J. L. Beaumont, D. Schnipke, K. Wallner-Allen, R. M. Golinkoff, J. B. Gleason, K. Hirsh-Pasek, Iv. nih toolbox cognition battery (cb): Measuring language (vocabulary comprehension and reading decoding), Monographs of the Society for Research in Child Development 78 (4) (2013) 49–69.

[147] D. Mungas, R. Heaton, D. Tulsky, P. D. Zelazo, J. Slotkin, D. Blitz, J.-S. Lai, R. Gershon, Factor structure, convergent validity, and discriminant validity of the nih toolbox cognitive health battery (nihtb-chb) in adults, Journal of the International Neuropsychological Society 20 (6) (2014) 579–587.

[148] N. E. Carlozzi, J. L. Beaumont, D. S. Tulsky, R. C. Gershon, The nih toolbox pattern comparison processing speed test: normative data, Archives of Clinical Neuropsychology 30 (5) (2015) 359–368.

[149] N. E. Carlozzi, D. S. Tulsky, N. D. Chiaravalloti, J. L. Beaumont, S. Weintraub, K. Conway, R. C. Gershon, Nih toolbox cognitive battery (nihtb-cb):

the nihtb pattern comparison processing speed test, Journal of the International Neuropsychological Society 20 (6) (2014) 630–641.

[150] N. E. Carlozzi, D. S. Tulsky, R. V. Kail, J. L. Beaumont, Vi. nih toolbox cognition battery (cb): measuring processing speed, Monographs of the Society for Research in Child Development 78 (4) (2013) 88–102.

[151] T. A. Salthouse, R. L. Babcock, R. J. Shaw, Effects of adult age on structural and operational capacities in working memory, Psychology and aging 6 (1) (1991) 118.

[152] J. M. Gold, C. Carpenter, C. Randolph, T. E. Goldberg, D. R. Weinberger, Auditory working memory and wisconsin card sorting test performance in schizophrenia, Archives of general psychiatry 54 (2) (1997) 159–165.

[153] D. S. Tulsky, N. Carlozzi, N. D. Chiaravalloti, J. L. Beaumont, P. A. Kisala, D. Mungas, K. Conway, R. Gershon, Nih toolbox cognition battery (nihtb-cb): List sorting test to measure working memory, Journal of the International Neuropsychological Society 20 (6) (2014) 599–610.

[154] D. S. Tulsky, N. E. Carlozzi, N. Chevalier, K. A. Espy, J. L. Beaumont, D. Mungas, V. nih toolbox cognition battery (cb): measuring working memory, Monographs of the Society for Research in Child Development 78 (4) (2013) 70–87.

[155] P. J. Bauer, S. S. Dikmen, R. K. Heaton, D. Mungas, J. Slotkin, J. L. Beaumont, Iii. nih toolbox cognition battery (cb): measuring episodic memory, Monographs of the Society for Research in Child Development 78 (4) (2013) 34–48.

[156] S. S. Dikmen, P. J. Bauer, S. Weintraub, D. Mungas, J. Slotkin, J. L. Beaumont, R. Gershon, N. R. Temkin, R. K. Heaton, Measuring episodic memory across the lifespan: Nih toolbox picture sequence memory test, Journal of the International Neuropsychological Society 20 (6) (2014) 611–619.

[157] B. A. Eriksen, C. W. Eriksen, Effects of noise letters upon the identification of a target letter in a nonsearch task, Perception & psychophysics 16 (1) (1974) 143–149.

[158] J. Fan, B. D. McCandliss, T. Sommer, A. Raz, M. I. Posner, Testing the efficiency and independence of attentional networks, Journal of cognitive neuroscience 14 (3) (2002) 340–347.

[159] M. R. Rueda, J. Fan, B. D. McCandliss, J. D. Halparin, D. B. Gruber, L. P. Lercari, M. I. Posner, Development of attentional networks in childhood, Neuropsychologia 42 (8) (2004) 1029–1040.

[160] P. D. Zelazo, The dimensional change card sort (dccs): A method of assessing executive function in children, Nature protocols 1 (1) (2006) 297.

Appendix A. Descriptions of the NIH Toolbox® Cognitive Tests

The NIH Toolbox® cognition measures were developed as part of the NIH Blueprint for Neuroscience Research (http://www.nihtoolbox.org). The tests assess episodic memory, executive function, attention, working memory, processing speed, and language abilities, enabling generation of composite scores [142, 143]. Use of a common tool for cognitive assessment valid for ages spanning the ABCD cohort's current and future range allows for longitudinal tracking of the developmental trajectories of this cohort in addition to harmonization and comparison of cognitive performance with numerous other studies. The tasks were selected based on a consensus building process and developed and validated using assessment methods that included item response theory (IRT) and computerized adaptive testing (CAT) where appropriate and feasible. Each Toolbox® task produces a number of scores, some of which are adjusted for age, sex, and ethnicity. All tasks provide raw scores, uncorrected standard scores, and age-corrected standard scores based on a normative sample of 2917 children and adolescents [144]. This study used age-corrected measures to compare the two cohorts of boys and girls, as there was a significant difference between our two cohorts. These tests are comprehensively described elsewhere [63] and briefly below.

1. **Language/Vocabulary Comprehension:** The Toolbox Picture Vocabulary Task® (TPVT) is a variant of the Peabody Picture Vocabulary Test

(PPTV) [145, 146, 147].

2. **Language/Reading Decoding:** The Toolbox Oral Reading Recognition Task® (TORRT) asks individuals to pronounce single letters or words presented in the middle of the iPad screen [145, 146] and measures exposure to language materials and cognitive skills involved in reading.
3. **Processing Speed:** The Toolbox Pattern Comparison Processing Speed Test® (TPCPST) [148, 149, 150] was modeled on the Pattern Comparison Task developed by Salthouse [151] and is a measure of rapid visual processing.
4. **Working Memory:** The Toolbox List Sorting Working Memory Test® (TLSWMT) is a variant of the letter-number sequencing test [152] that uses pictures rather than words or letters [153, 154].
5. **Episodic Memory:** The Toolbox Picture Sequence Memory Test® (TPSMT) was modeled after memory tests asking children to imitate a sequence of actions with props developed by Bauer and colleagues. [155, 156].
6. **Executive Function/Attention/Inhibition:** The Toolbox Flanker Task® (TFT), a variant of the Eriksen Flanker task [157], was adapted from the Attention Network Task [158, 159] and assessed response inhibition.
7. **Executive Function/Cognitive Flexibility:** The Toolbox Dimensional Change Card Sort Task® (TDCCS) was based on the work of Zelazo and colleagues [160] and measures problem solving and cognitive flexibility.

Appendix B. Deep Learning Model Architecture & Hyperparameters

Input to the deep learning model was the 3D MRI of one hemisphere of size $64 \times 64 \times 32$. The predictor extraction network contained 4 stacks of $3 \times 3 \times 3$ convolutional layers, ReLu activation, batch normalization, and $2 \times 2 \times 2$ max-pooling layers. The size of feature channel for the 4 convolution layers was $(16, 32, 64, 128)$. Then the resulting 4096 features were fed into a set of fully connected layers (Multi-Layer Perceptron) classifier composed of three Fully Connected (FC) layers of dimension $(64, 32, 1)$. `tanh` activation was used for

the first two FC layers, and `sigmoid` activation was used for the last FC layer resulting in the final prediction score $\mathbf{S} \in [0, 1]$. An L2 regularization of weight 0.1 was applied to the FC layers (see Fig. 7).

Appendix C. Receiver Operating Characteristic Curve

As included in the main paper, our deep learning framework led to an accuracy of nearly 90% for predicting the sex of individuals based their structural MRI data. The receiver operating characteristic (ROC) curve of this classification model is depicted in Appendix Figure 8, which shows an area under the curve (AUC) of 0.96.

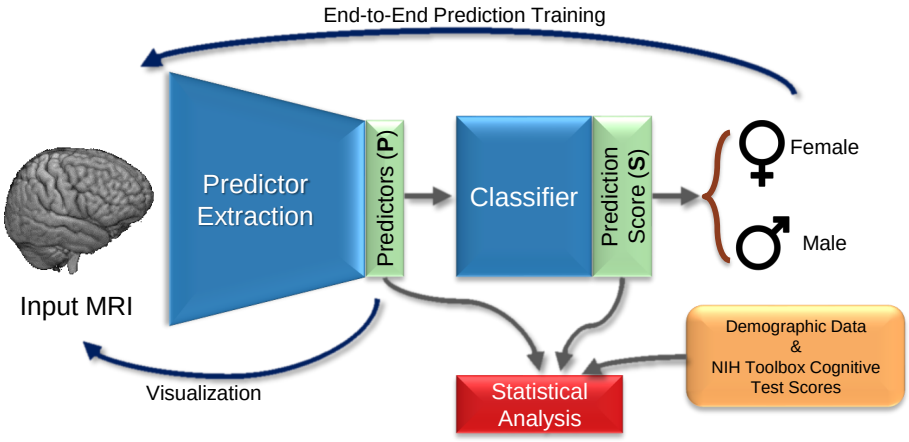


Figure 1: Overview of the proposed analysis. The convolutional neural network (CNN) automatically extracts predictors (**P**) from the minimally processed MRI. Based on **P**, the classifier computes a prediction score (**S**) that assigns the MRI to either sex. This deep learning analysis operates directly on voxel-level data omitting any hypothesis or assumption related to brain regions or tissue measurements (like regional volumes). Statistical analysis relates obtained results to NIH Toolbox cognitive test scores, creates confounder-free visualization of the patterns predicting sex (a.k.a. saliency map), and examines volume scores of those regions that contribute significantly to the prediction according to the saliency map.

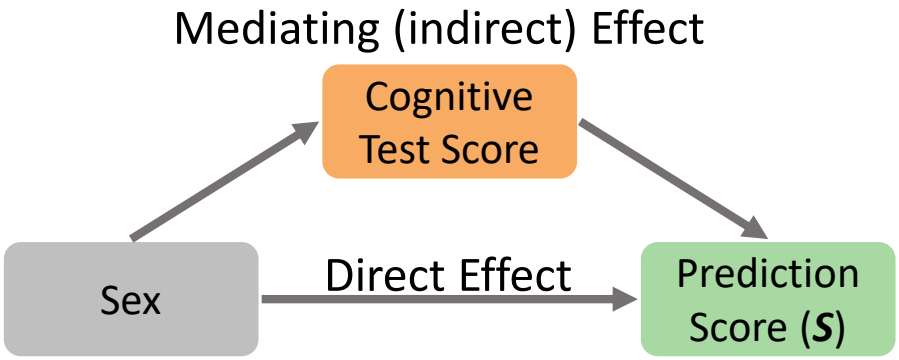
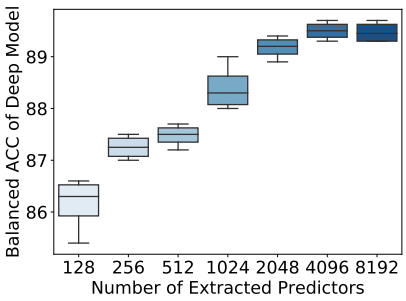
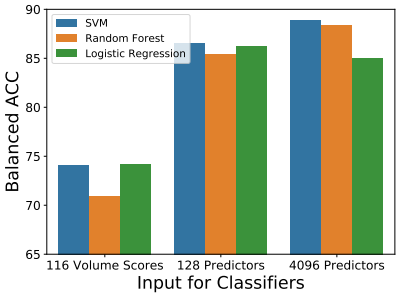


Figure 2: Mediation analysis to observe how much of the variance in the prediction score was explained by the observed sex and how much was influenced by the NIH toolbox score.

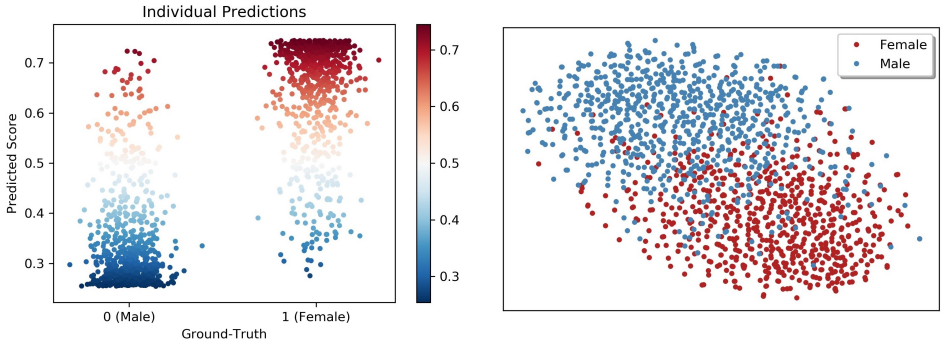


(a) Accuracy of our deep learning classification as a function of the number of extracted predictors, $\mathbf{P}$.



(b) Accuracy of different models using 116 regional volumetric measures (left) and predictors extracted by the deep learning (middle & right).

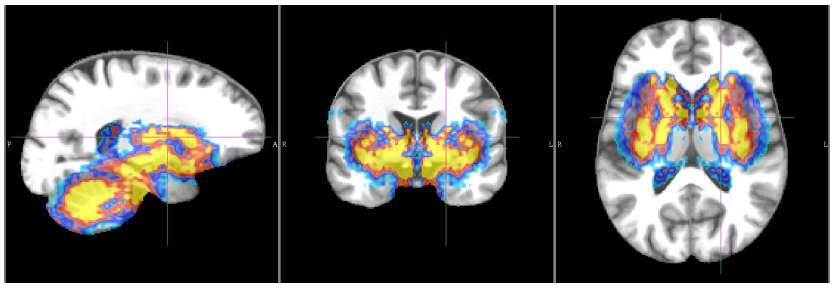
Figure 3: Results of the deep learning model predicting sex with different numbers of predictors (a), and different classifiers (b).



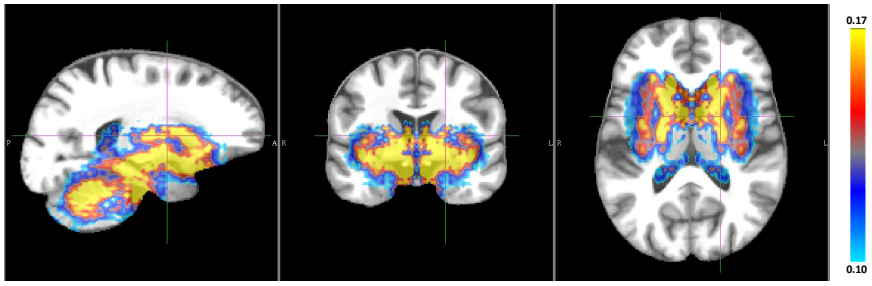
(a) Prediction Score (**S**)

(b) 2D Projection of Extracted Predictors (**P**)

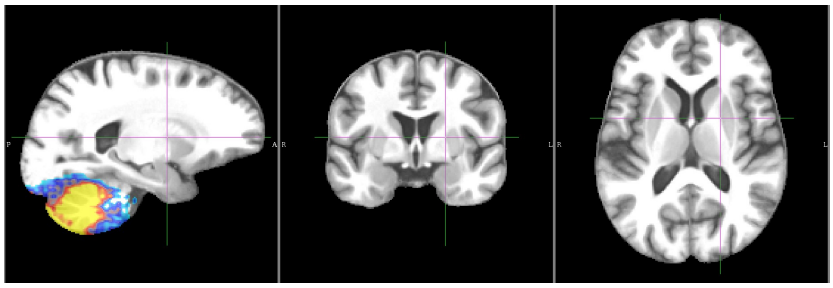
Figure 4: Visualization of **Predictors** and the **Prediction Score** as determined by the deep learning model. (a) Prediction Score (**S**) of each participant as a function of their observed sex. These two figures show that our deep learning model can effectively reduce the MRIs to a vector of predictors (**P**) and then to a scalar value (**S**) that distinguishes girls from boys. (b) t-Distributed Stochastic Neighbor Embedding (tSNE) [107] projection of extracted Predictors (**P**) in 2D space. Each point indicates one adolescent; color represents sex. The axes show the relative location of each individual with respect to their neighbors in 2D. tSNE preserves the neighborhood of the high dimensional space.



(a) Initial saliency map

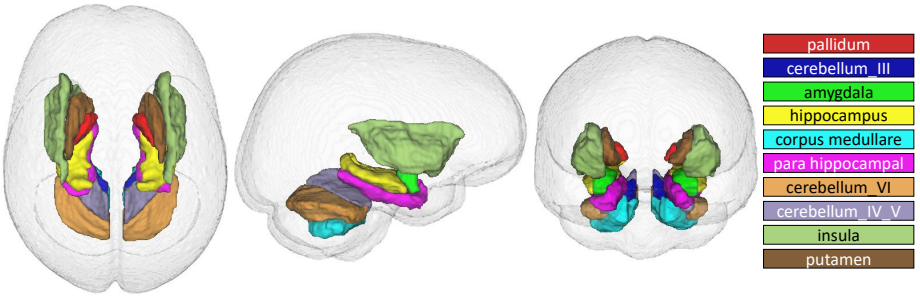


(b) Saliency map corrected for age, PDS, and SES



(c) Saliency maps confounded by PDS

Figure 5: Saliency maps defining predictive brain areas for distinguishing boys from girls in the ABCD study; (a) original and (b) corrected for confounding factors. In the developing brain of 9 and 10-year-olds, the factors distinguishing boys from girls mainly lie in the subcortical and cerebellar regions. (c) Regional brain pattern of sex differences confounded by PDS. Note, computing saliency maps requires scaling of the maps so that the resulting importance values are only meaningful within one saliency map but cannot be directly compared across maps.



Regions of Interest (ROIs)	**Mean Saliency**	**Volume Mean ± SD**		**Cohen's d**	**Group Differences**	
		Female	**Male**		**p-value**	**Direction**
Pallidum	0.1605	80.00 ± 29.25	81.53 ± 29.12	0.053	0.02	F < M
Cerebellum III	0.1500	803.53 ± 115.78	805.16 ± 120.03	0.014	NS	F = M
Amygdala	0.1460	1169.53 ± 94.08	1173.33 ± 94.87	0.040	NS	F = M
Hippocampus	0.1398	3408.35 ± 244.99	3322.98 ± 249.41	0.350	$< 10^{-6}$	F > M
Corpus Medullare	0.1387	7720.72 ± 660.95	7493.03 ± 663.39	0.343	$< 10^{-6}$	F > M
Para Hippocampal	0.1384	4700.50 ± 400.66	4726.10 ± 416.05	0.064	0.005	F < M
Cerebellum VI	0.1384	9902.93 ± 1207.29	9478.25 ± 1151.66	0.360	$< 10^{-6}$	F > M
Cerebellum IV/V	0.1340	5440.99 ± 585.16	5441.95 ± 582.88	0.002	NS	F = M
Insula	0.1308	7322.80 ± 559.90	7305.74 ± 539.40	0.030	NS	F = M
Putamen	0.1298	2237.28 ± 502.88	2344.62 ± 502.94	0.213	$< 10^{-6}$	F < M

Figure 6: Top 10 regions relevant for distinguishing sex as determined by the deep learning framework. Some of these regions are smaller in girls (cerebellar lobules III and IV/V, amygdala; and insula, pallidum, para hippocampus, and putamen), while hippocampus, corpus medullare, and cerebellar lobule VI are smaller in boys. P-values of group differences of ROI volumes were calculated using two sample t-test. NS denotes *not significant*.

Figure 7: Architecture of our deep learning model.

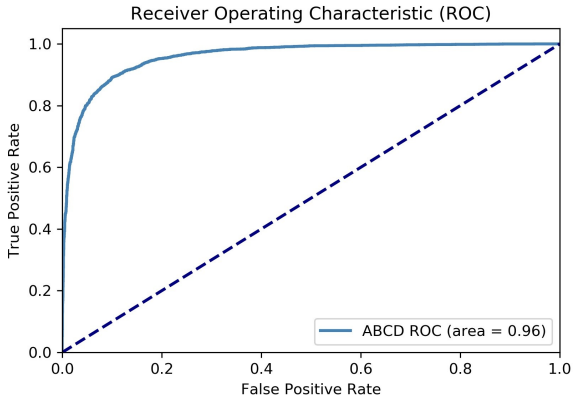


Figure 8: Receiver Operating Characteristics (ROC) curve of the classifier differentiating boys and girls based on MR images. The blue curve shows the results of the model based on ABCD data.

Table 1: Demographic information (mean ± standard deviation).

Measure	**Female (F)**	**Male (M)**	p-**value**[a]	**Group Difference**
Total Subjects	3,895	4,249	-	-
Age (years)	9.92±0.62	9.95±0.62	0.04	F < M
Head size[b] (svol cm^3)	1341.5±15.3	1342.0±16.0	NS	F=M
Socioeconomic Status (SES)	18.0±66.8	18.5±67.8	NS	F = M
Pubertal Development Scale (PDS)	2.0±1.0	1.3±0.6	$\leq 10^{-6}$	F > M
Ethnicity (%)[c]:				
Asian/African American/Caucasian/Other	269/631/2650/346	282/629/2975/363	NS	F = M
Body Mass Index (BMI) z-scores[d]	0.23±5.6	0.15±14.1	NS	F = M

[a]measured by χ^2-test or t-test: NS '=' not significantly different by p=0.05; '<' or '>' significantly different at $p \leq 0.05$.

[b]head size was measured after being affinely registered to the SRI24 template.

[c]individuals who self-identified as Hispanic were included in the Caucasian group: 493 girls and 574 boys.

[d]z-scores of the BMI (instead of percentile) are calculated by the *pygrowup* toolbox [98] for each individual to enable group comparison using t-test.

Table 2: Scores of cognitive tests (mean ± standard deviation).

Test	**Cognitive Process**	**Female (F) N=3895**	**Male (M) N = 4249**	**Cohen's d**	**p-value***	**Group Difference***
NIH Toolbox: Flanker®	cognitive control; attention	96.29±13.37	97.09±14.39	0.058	NS	F = M
NIH Toolbox: List Sorting Working Memory Test®	working memory; categorization; information processing	101.68±14.08	102.64±14.68	0.067	0.038	F < M
NIH Toolbox Dimensional Change Card Sort®	flexible thinking; concept formation; set shifting	98.89±15.07	97.44±15.54	0.095	0.3429e-2	F > M
NIH Toolbox Oral Reading Recognition Test®	reading ability; language; academic achievement	104.45±19.53	103.65±18.52	0.042	NS	F = M
NIH Toolbox: Pattern Comparison Processing Speed®	processing speed; information processing	96.70±20.92	93.72±22.18	0.140	0.1875e-4	F > M
NIH Toolbox: Picture Sequence Memory Test®	visuospatial sequencing & memory	103.47±16.47	100.62±15.81	0.180	$< 10^{-6}$	F > M
NIH Toolbox: Picture Vocabulary Test®	language; verbal intellect	108.53±16.98	109.35±17.05	0.056	NS	F = M

*measured by χ^2-test or t-test: NS '=' not significant; '<' or '>' significant at $p \leq 0.05$.

Table 3: Accuracy (Acc), true positive rate (TPR), true negative rate (TNR), area under the ROC curve (AUC) of different methods for predicting sex from MRIs.

Method	**Acc**	**TPR**	**TNR**	**AUC**
Ours (end-to-end deep learning)	**89.6%**	**87.4%**	**91.5%**	**0.96**
116 SRI24 volume Scores				
Logistic Regression	74.2%	74.3%	74.0%	0.80
Support Vector Machine	74.2%	73.0%	75.5%	0.81
Random Forest	70.9%	66.7%	74.5%	0.75
906 Destrieux Parcellation Measures				
Logistic Regression	80.0%	80.8%	79.2%	0.88
Support Vector Machine	79.1%	78.1%	79.9%	0.84
Random Forest	74.2%	72.2%	76.0%	0.79

Table 4: p-values of the correlation and mediation analysis with respect to the NIH Toolbox Scores. Correlation analysis was examined by Pearson's R. Mediation analysis examined the indirect effect of NIH Toolbox scores on sex prediction; Significant mediation effect (p <0.05 for all 3 conditions of the partial mediation model) is marked by bold typeface. NS denotes *not significant*.

Test	**Correlation**	**Mediation**		
	Prediction Score	Observed Sex (Condition 1)	Prediction Score (Condition 2)	Correlation Reduction (Condition 3)
NIH Toolbox Flanker®	NS	NS	NS	NS
NIH Toolbox List Sorting Working Memory Test®	0.001	**0.03817**	**0.0037**	**0.0005**
NIH Toolbox Dimensional Change Card Sort®	NS	0.00342	0.011	NS
NIH Toolbox Oral Reading Recognition Test®	NS	NS	0.036	NS
NIH Toolbox Pattern Comparison Processing Speed®	0.0025	0.00002	NS	NS
NIH Toolbox Picture Sequence Memory Test®	0.00001	$< 10^{-6}$	NS	< NS
NIH Toolbox Picture Vocabulary Test®	0.0309	NS	NS	0.018