HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 January 04.

Published in final edited form as:

Neuroimage. 2020 December ; 223: 117293. doi:10.1016/j.neuroimage.2020.117293.

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, USA

[b]Center for Biomedical Sciences, SRI International, Menlo Park, CA 94025, USA

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 8144 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.

Keywords

Deep learning; Sex differences; Adolescents; Study confounders; Pubertal development; Cerebellum

*Corresponding author. kilian.pohl@stanford.edu (K.M. Pohl).

Credit author statement

E.A. and Q.Z. developed the idea and implemented the methods. E.A. Q.Z., N.M.Z., and K.M.P. wrote the paper and analyzed the technical results. A.G. helped with data gathering and preparation as well as analyzing the demographic variables for the study. N.M.Z., A.P., and E.V.S. analyzed the clinical results and wrote the discussion on the clinical findings of the work.

[1]These authors contributed equally to this work.

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

1. Introduction

The concept of sex differences is based on biology and genetics. Since the 1930s (e.g., Pfeiffer, 1936), identifying sex differences in the Central Nervous System (CNS) has been explored in animal models (Becker and Koob, 2016; Galea et al., 2006; Goldstein et al., 2001; McEwen, 1983; Woodson and Gorski, 2000) and histology of postmortem human brain samples (Vogeley et al., 2000; Witelson et al., 2005; Witelson, 1989). More recently, in vivo neuroimaging (Fan et al., 2010; Filipek et al., 1994; Flaum et al., 1995; Goldstein et al., 2001; Hänggi et al., 2010; Sacher et al., 2013; Wang et al., 2012) and computer-based learning tools (Breslau et al., 2017; Feis et al., 2013; Nieuwenhuis et al., 2017; Van Putten et al., 2018; Xin et al., 2019) have been implemented in the search for a CNS basis of sexual differentiation. Beyond sex-linked risks for disease (Brie et al., 2019; Egloff et al., 2018; Jahanshad and Thompson, 2017; Lind et al., 2017; Retico et al., 2016; Vogeley et al., 2000), this search is motivated by adolescence being a period of particular vulnerability to the emergence of sex-linked neuropsychiatric disorders such as schizophrenia (Vogeley et al., 2000; Womer et al., 2016) and autism (Golarai et al., 2006; Liu et al., 2016; Pierce et al., 2019; Retico et al., 2016; Strickler et al., 2020), which have a higher prevalence in boys than girls, and depression, which girls by age 15 develop twice as likely as boys (Breslau et al., 2017; Cyranowski et al., 2000).

In vivo structural magnetic resonance imaging (MRI) studies characterize brain development as following heterogeneous growth trajectories (Giedd, 2004; Petrican et al., 2017) during which sex-specific behaviors emerge (Johnson and Meade, 1987). While physical signs of sex differences are present at birth (Gilmore et al., 2007), brain structural and functional differences between the sexes continue to develop over childhood through late adolescence (Giedd et al., 2015; Mankiw et al., 2017; Pfefferbaum et al., 2016, 2018; Tamnes et al., 2017). 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 (Lenroot et al., 2007b). With respect to white matter, the volume increases with age in both sexes, but boys generally show a more rapid increase during adolescence (Lenroot et al., 2007b). These sex specific changes in brain structure during adolescence (Wierenga et al., 2018a) are accompanied with asexual developments, such as structural volume (Aubert-Broche et al., 2013; Ducharme et al., 2015; Herting et al., 2018; Mills et al., 2012; Narvacan et al., 2017; Vijayakumar et al., 2016), cortical thickness (Vijayakumar et al., 2016), cortical surface area (Ducharme et al., 2015; Vijayakumar et al., 2016), individual's behavior (Wierenga et al., 2014), and testosterone effects (Wierenga et al., 2018a).

Many of the differences in brain development between the sexes are actually linked to head size (Ruigrok et al., 2014; Sanchis Segura et al., 2018). 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 (Etchell et al., 2018; Jahanshad and Thompson, 2017; Luders et al., 2014; Sawyer et al., 2018; Sullivan et al., 2001). Beyond properly accounting for head size (Luders et al., 2014; Perlaki et al., 2014; Pfefferbaum et al., 2016; Sanchis Segura et al., 2018), discrepancies in findings may

be due to small sample sizes (Button et al., 2013; Filipek et al., 1994; Lenroot et al., 2007a), wide age distributions (sometimes across several decades so age-specific sex differences are obscured) (Etchell et al., 2018; Kim et al., 2012), 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)) (Etchell et al., 2018; Xin et al., 2019). The study presented herein accounts for these issues by building on recent advancements in the field of deep learning (Esmaeilzadeh et al., 2018; Krizhevsky et al., 2012; Van Putten et al., 2018) 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 Fig. 1) predicting sex from the minimally processed T1-weighted (T1w) MRIs (Hagler et al., 2019) of 8144 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® (Luciana et al., 2018). Finally, we qualitatively assess the average *saliency map* (Simonyan et al., 2014) 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, details are explained in Appendix A), and T1-weighted (T1w) MR images (Hagler et al., 2019) from 8670 participants were distributed by the ABCD-Neurocognitive Prediction Challenge (ABCD-NP-Challenge 2019) (Pohl et al., 2019) via the National Database for Autism Research (NDAR) portal (Release 2.0), of which 8144 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 (Sullivan et al., 2016). Pubertal status was determined by self-assessment with the Pubertal Development Scale (PDS) (Carskadon and Acebo, 1993; Petersen et al., 1988), 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) (Pfefferbaum et al., 2016). Body Mass Index (BMI) was calculated based on published methods (Freedman et al., 2017). 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 (Thompson et al., 2019).

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 Garavan et al., 2018) 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 (Hagler et al., 2019) followed by noise removal (Coupe et al., 2008) and field-inhomogeneity correction via N4ITK (Version 2.1.0) (Tustison et al., 2010). Brain masks were determined via majority voting (Rohlfing et al., 2004) over the segmentations generated by applying the following tools to both bias and non-bias corrected T1w-images: FSL BET (Version 5.0.6) (Smith, 2002), AFNI 3dSkullStrip (Version AFNI_2011_12_21_1014) (Cox, 1996), FreeSurfer mri-gcut (Version 5.3.0) (Sadananthan et al., 2010), and Robust Brain Extraction (ROBEX) (Version 1.2) (Iglesias et al., 2011). The resulting brain mask was used to refine correction for image-inhomogeneity and skull stripping. MRIs were then affinely registered to the SRI24 template (Rohlfing et al., 2010), down-sampled to 2 mm isotropic voxel size, and re-scaled to 64 × 64 × 64 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).

Fig. 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 (Esmaeilzadeh et al., 2018; Nie et al., 2016). The Predictor/Extractor identified a set of Predictor variables $\mathbf{P} = \{\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^M\}$ from MR images based on a deep convolutional network (Krizhevsky et al., 2012). The Classifier was a set of fully connected layers reducing **P** into a continuous Prediction Score **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 (Hill et al., 2014; Hirnstein et al., 2019; Phinyomark et al., 2014; Román et al., 1989; Weinhandl et al., 2010) 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 (Kohavi, 1995) 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 (Oksuz et al., 2019), *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 (Esmaeilzadeh et al., 2018). Next, the prediction of the

individual's sex was recorded on the fifth fold (which was not augmented) by computing the average prediction score (**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 **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 (Park et al., 2018) (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 (Adeli et al., 2020; Kleinbaum and Klein, 2002), support vector machines (Chang and Lin, 2011) and random forest (Liaw and Wiener, 2002) applied to the volumes of 116 brain ROIs defined according to the SRI24 atlas (Rohlfing et al., 2010). 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) (Avants et al., 2008) and overlaying parcellations with the tissue segmentations from Atropos (Avants et al., 2011). The experiment was repeated using the 906 regional scores generated by Freesurfer based on the Destrieux atlas (Destrieux et al., 2010), 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 **S** of the individuals with the test scores of the age-corrected NIH toolbox (significant p-value < 0.05 according to Pearson's R). Identifying variance in the prediction score partially induced by an NIH toolbox score (Fig. 2) was done via the partial mediation model (Baron and Kenny, 1986). 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 10,000 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 (Pearson, 1900). 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 (pygrowup, 2017), Tensorflow

1.7.0 (Abadi et al., 2016), and Keras 2.2.2 (Gulli and Pal, 2017). 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 re-trained 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 (Simonyan et al., 2014). The initial salience map was computed by applying the minimally processed and flipped MRI to the trained prediction model and then performing back-propagation (Kotikalapudi and contributors, 2017). 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) (Madsen and Thyregod, 2010) 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, Rohlfing et al., 2010), 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 Shaffer, 1995).

3. Results

The accuracy of the prediction score in correctly assigning MRIs to either sex was 89.6% (Receiver operating characteristic curve in Appendix C), which was significantly better than chance ($p < 0.001$ according to a Fisher exact test, Fisher, 1935). The prediction accuracy was stable across 5 runs of 5-fold cross-validation based on random splitting of folds (89.6%

± 0.13%) but was slightly lower (87.3%) on a subset of 2464 boys and 2464 girls matched on head size (matched according to Adeli et al., 2018). 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, DeLong et al., 1988, 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 Fig. 4 (a), which plotted the Prediction Score (**S**) of each participant as a function of their observed sex. Furthermore, projecting the high dimensional Predictors (**P**) learned from one training run into 2D via the t-distributed Stochastic Neighbor Embedding (tSNE) (Maaten and Hinton, 2008) also resulted in a cluster for boys and a separate one for girls (Fig. 4 (b)).

Fig. 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. Fig. 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 (Fig. 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) (Fig. 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 (Fig. 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, Adeli et al., 2018; Aubert-Broche et al., 2013; Chung et al., 2005; Green et al., 2016; Kim et al., 2012), 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 Fig. 3). A novel discovery of that search for discriminative information 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 (Hirnstein et al., 2019; Phinyomark et al., 2014; Román et al., 1989; Weinhandl et al., 2010). 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 (Feis et al., 2013; Nieuwenhuis et al., 2017; Ruigrok et al., 2014; Van Putten et al., 2018; Xin et al., 2019). 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 (Park et al., 2018). 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 (Fig. 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 (~ 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 (Chung et al., 2005; Fan et al., 2010; Raz et al., 2001; Tiemeier et al., 2010). However, sex differences in cerebellar volume became generally negligible once studies corrected for intracranial volume (e.g., Nopoulos et al., 2000; Sullivan et al., 2020; Szabo et al., 2003). More specifically, the corpus medullare of the cerebellum in this study was significantly larger in girls than boys. By contrast, the longitudinal study by Tiemeier et al., 2010 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 (Etchell et al., 2018). 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 (Luna et al., 2004).

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 (Aggleton et al., 2010; Pilly et al., 2018). 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 (Dikmen et al., 2014a). The finding that girls had relatively larger hippocampi than boys is also supported by MRI studies of young adults (Filipek et al., 1994; Frodl et al., 2002; Szabo et al., 2003) that linked sex differences in hippocampal volume to hormonal responsivity (Giedd et al., 1996; Teicher et al., 2003) and memory performance (Hill et al., 2014; Trenerry et al., 1995; Young et al., 2013). Other studies noted relations between hippocampal volumes and clinical characteristics of psychiatric disorders (Egloff et al., 2018; Frodl et al., 2002; Yang et al., 2017), where sleep disturbances are more severe (Yang et al., 2017), depressive episodes are more frequent and longer, and higher frequency of migraines occurs in depressed female compared to depressed male patients (Saunders et al., 2014).

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 (Giedd et al., 1997). A more recent study based on data from the Pediatric Imaging, Neurocognition, and Genetics (PING) study with 1234 participants (ages 3 to 21 years) (Wierenga et al., 2018b) 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 (Arden and Plomin, 2006; Baye and Monseur, 2016). Likewise, the amygdala has been linked to sex differences in animal and human studies across the lifespan (Blume et al., 2017; Green et al., 2016). A surface-based modeling approach showed that men had a larger mean radius of amygdala subregions than women (Kim et al., 2012). Further, sex differences in amygdala volume may contribute to the expression of selective psychotic disorders occurring more commonly in men than women (Blume et al., 2017) and depressive disorders, which are more common in women (Breslau et al., 2017; Cyranowski et al., 2000).

Like the amygdala, 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 (Menon and Uddin, 2010). 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) (Gur and Gur, 2016), in which girls performed better than boys on tasks assessing verbal memory and social cognition, whereas boys excelled on spatial processing and motor speed (Gur and Gur, 2017; 2016). 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 (Sullivan et al., 2016). 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 (Akshoomoff et al., 2014). 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 (Bäckström et al., 2018; Esmaeilzadeh et al., 2018), affine registration might poorly align the cortical gyri and sulci given their high inter-subject variability (Llera et al., 2019). Non-rigid registration achieves better voxel-wise correspondence across MRIs enabling learning algorithms to focus on fine-grained regional

cues (Lin et al., 2018; Liu et al., 2018). Now identifying cues differentiating between groups highly depends on the 'stiffness' of the deformation field (Murphy et al., 2016; Wittek et al., 2010), which can substantially modify the shape and appearance of brain structures. One possible data driven approach for setting the stiffness with respect to the cortex is to first parcellate the structure (via a surface based segmentation tool, Dale et al., 1999; Onofrey et al., 2018; Zhao et al., 2019) and then perform an ROI-based registration for the whole brain (such as Lopez-Garcia et al., 2006; Yi et al., 2006). As any of these registration can 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 ($N = 8144$) 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.

Acknowledgments

Funding for this study was received from the U.S. National Institutes Health (NIH) grants AA026762, DA041123, AA021697, and AA010723 This study also benefited from the Stanford Institute for Human-centered Artificial Intelligence (HAI) AWS Cloud Credit.

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 (Gershon et al., 2013b; Hodes et al., 2013). 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 (Casaletto et al., 2015). 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 (Luciana et al., 2018) and briefly below.

1. *Language/Vocabulary Comprehension:* The Toolbox Picture Vocabulary Task® (TPVT) is a variant of the Peabody Picture Vocabulary Test (PPTV) (Gershon et al., 2014; 2013a; Mungas et al., 2014).
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 (Gershon et al., 2014; 2013a) and measures exposure to language materials and cognitive skills involved in reading.
3. *Processing Speed:* The Toolbox Pattern Comparison Processing Speed Test® (TPCPST) (Carlozzi et al., 2015; 2014; 2013) was modeled on the Pattern Comparison Task developed by Salthouse (Salthouse et al., 1991) 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 (Gold et al., 1997) that uses pictures rather than words or letters (Tulsky et al., 2013, 2014).
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 et al. (2013) and Dikmen et al. (2014b)
6. *Executive Function/Attention/Inhibition:* The Toolbox Flanker Task® (TFT), a variant of the Eriksen Flanker task (Eriksen and Eriksen, 1974), was adapted from the Attention Network Task (Fan et al., 2002; Rueda et al., 2004) 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 (Zelazo, 2006) and measures problem solving and cognitive flexibility.

Appendix B.: Deep learning model architecture and 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).

Fig. 7.
Architecture of our deep learning model.

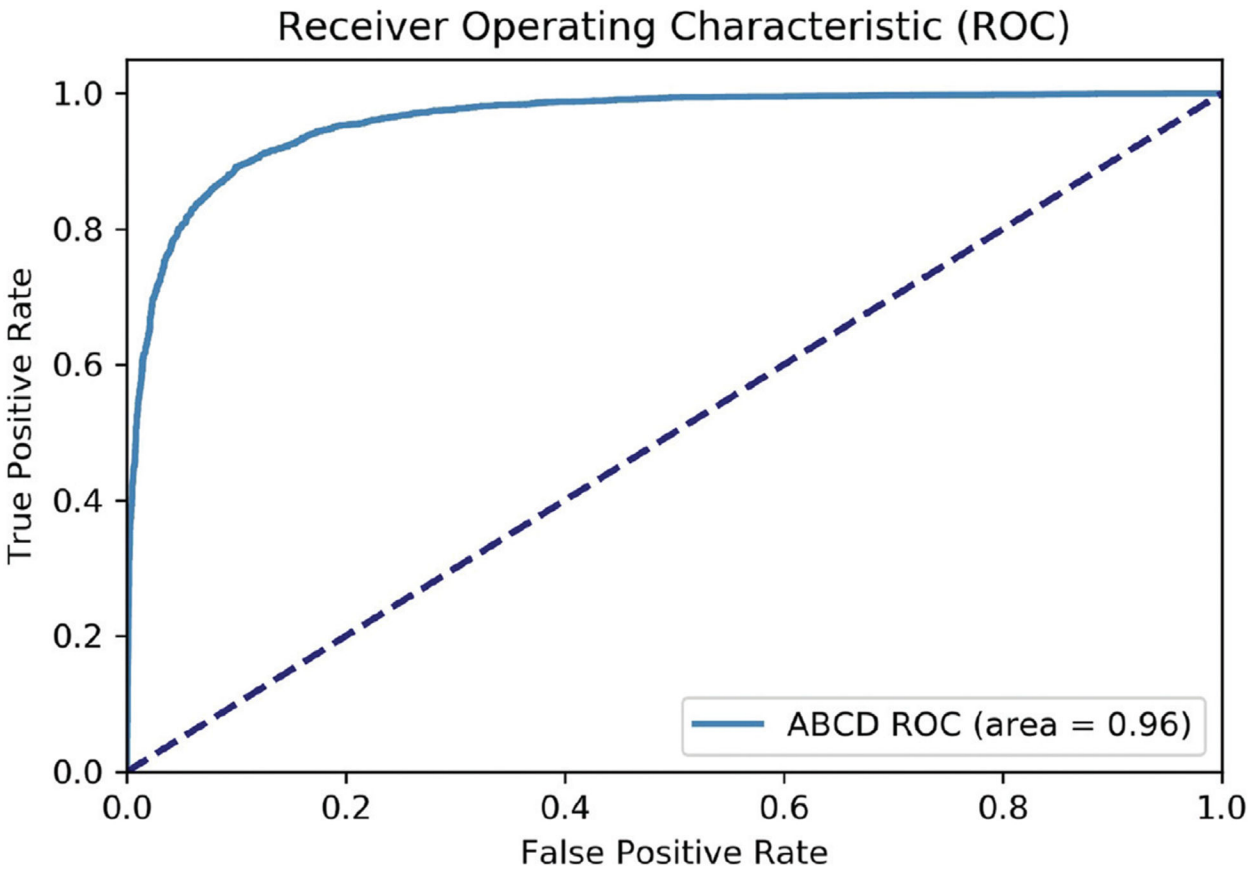


Fig. C.1.
Receiver Operating Characteristics (ROC) curve of the classifier differentiating boys and girls based on MR images. The blue curve shows the results of the model based on ABCD data.

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 Fig. C.1, which shows an area under the curve (AUC) of 0.96.

References

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

Adeli E, Kwon D, Zhao Q, Pfefferbaum A, Zahr NM, Sullivan EV, Pohl KM, 2018 Chained regularization for identifying brain patterns specific to HIV infection. NeuroImage 183, 425–437 [PubMed: 30138676]

Adeli E, Li X, Kwon D, Zhang Y, Pohl K, 2020 Logistic regression confined by cardinality-constrained sample and feature selection. IEEE Trans. Pattern Anal. Mach. Intell 42 (7), 1713–1728 [PubMed: 30835210]

Aggleton JP, O'Mara SM, Vann SD, Wright NF, Tsanov M, Erichsen JT, 2010 Hippocampal-anterior thalamic pathways for memory: uncovering a network of direct and indirect actions. Eur. J. Neurosci 31 (12), 2292–2307. doi: 10.1111/j.1460-9568.2010.07251.x [PubMed: 20550571]

Akshoomoff N, Newman E, Thompson WK, McCabe C, Bloss CS, Chang L, Amaral DG, Casey B, Ernst TM, Frazier JA, 2014 The NIH Toolbox Cognition Battery: Results from a large normative developmental sample (PING). Neuropsychology 28 (1), 1 [PubMed: 24219608]

Arden R, Plomin R, 2006 Sex differences in variance of intelligence across childhood. Person. Individ. Diff 41 (1), 39–48

Aubert-Broche B, Fonov VS, Garcia-Lorenzo D, Mouiha A, Guizard N, Coupé P, Eskildsen SF, Collins DL, 2013 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, 393–402 [PubMed: 23719155]

Avants BB, Epstein CL, Grossman M, Gee JC, 2008 Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal 12 (1), 26–41. doi: 10.1016/j.media.2007.06.004 [PubMed: 17659998]

Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC, 2011 An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400. doi: 10.1007/s12021-011-9109-y [PubMed: 21373993]

Bäckström K, Nazari M, Gu I, Jakola A, 2018 An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) IEEE, pp. 149–153

Baron RM, Kenny DA, 1986 The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Person. Soc. Psychol 51 (6), 1173–1182

Bauer PJ, Dikmen SS, Heaton RK, Mungas D, Slotkin J, Beaumont JL, 2013 III. NIH Toolbox Cognition Battery (CB): measuring episodic memory. Monogr. Soc. Res. Child Dev 78 (4), 34–48 [PubMed: 23952201]

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

Becker JB, Koob GF, 2016 Sex differences in animal models: focus on addiction. Pharmacol. Rev 68 (2), 242–263 [PubMed: 26772794]

Blume SR, Freedberg M, Vantrease JE, Chan R, Padival M, Record MJ, De-Joseph MR, Urban JH, Rosenkranz JA, 2017 Sex- and estrus-dependent differences in rat basolateral amygdala. J. Neurosci 37 (44), 10567–10586. doi: 10.1523/JNEUROSCI.0758-17.2017 [PubMed: 28954870]

Breslau J, Gilman SE, Stein BD, Ruder T, Gmelin T, Miller E, 2017 Sex differences in recent first-onset depression in an epidemiological sample of adolescents. Transl. Psychiatry 7 (5), e1139 [PubMed: 28556831]

Brie B, Ramirez MC, De Winne C, Vicchi FL, Villarruel L, Sorianello E, Catalano P, Ornstein AM, Becu-Villalobos D, 2019 Brain control of sexually dimorphic liver function and disease: The endocrine connection. Cell. Mol. Neurobiol 39 (2), 169–180 [PubMed: 30656469]

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafó MR, 2013 Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci 14 (5), 365–376 [PubMed: 23571845]

Carlozzi NE, Beaumont JL, Tulsky DS, Gershon RC, 2015 The NIH toolbox pattern comparison processing speed test: normative data. Arch. Clin. Neuropsychol 30 (5), 359–368. [PubMed: 26025230]

Carlozzi NE, Tulsky DS, Chiaravalloti ND, Beaumont JL, Weintraub S, Conway K, Gershon RC, 2014 NIH toolbox cognitive battery (NIHTB-CB): the NIHTB pattern comparison processing speed test. J. Int. Neuropsychol. Soc 20 (6), 630–641 [PubMed: 24960594]

Carlozzi NE, Tulsky DS, Kail RV, Beaumont JL, 2013 VI. NIH Toolbox Cognition Battery (CB): measuring processing speed. Monogr. Soc. Res. Child Dev 78 (4), 88–102 [PubMed: 23952204]

Carskadon MA, Acebo C, 1993 A self-administered rating scale for pubertal development. J. Adolesc. Health 14 (3), 190–195 [PubMed: 8323929]

Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, Heaton RK, 2015 Demographically corrected normative standards for the English version of the NIH Toolbox Cognition Battery. J. Int. Neuropsychol. Soc 21 (5), 378–391 [PubMed: 26030001]

Chang C-C, Lin C-J, 2011 Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol 2 (3), 1–27

Chung S-C, Lee B-Y, Tack G-R, Lee S-Y, Eom J-S, Sohn J-H, 2005 Effects of age, gender, and weight on the cerebellar volume of Korean people. Brain Res 1042 (2), 233–235 [PubMed: 15854595]

Coupe P, Yger P, Prima S, Hellier P, Kervrann C, Barillot C, 2008 An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441. doi: 10.1109/TMI.2007.906087 [PubMed: 18390341]

Cox RW, 1996 Afni: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res 29 (3), 162–173 [PubMed: 8812068]

Cyranowski JM, Frank E, Young E, Shear MK, 2000 Adolescent onset of the gender difference in lifetime rates of major depression: a theoretical model. Arch. Gen. Psychiatry 57 (1), 21–27 [PubMed: 10632229]

Dale AM, Fischl B, Sereno MI, 1999 Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage 9 (2), 179–194 [PubMed: 9931268]

DeLong ER, DeLong DM, Clarke-Pearson DL, 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845 [PubMed: 3203132]

Destrieux C, Fischl B, Dale A, Halgren E, 2010 Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage 53 (1), 1–15 [PubMed: 20547229]

Dikmen SS, Bauer PJ, Weintraub S, Mungas D, Slotkin J, Beaumont JL, Gershon R, Temkin NR, Heaton RK, 2014a Measuring episodic memory across the lifespan: NIH toolbox picture sequence memory test. J. Int. Neuropsychol. Soc 20 (6), 611–619 [PubMed: 24960230]

Dikmen SS, Bauer PJ, Weintraub S, Mungas D, Slotkin J, Beaumont JL, Gershon R, Temkin NR, Heaton RK, 2014b Measuring episodic memory across the lifespan: NIH toolbox picture sequence memory test. J. Int. Neuropsychol. Soc 20 (6), 611–619 [PubMed: 24960230]

Ducharme S, Albaugh MD, Nguyen T-V, Hudziak JJ, Mateos-Pérez J, Labbe A, Evans AC, Karama S, Group BDC, 2015 Trajectories of cortical surface area and cortical volume maturation in normal brain development. Data Brief 5, 929–938 [PubMed: 26702424]

Egloff L, Lenz C, Studerus E, Harrisberger F, Smieskova R, Schmidt A, Huber C, Simon A, Lang UE, Riecher-Rössler A, 2018 Sexually dimorphic subcortical brain volumes in emerging psychosis. Schizophr. Res 199, 257–265 [PubMed: 29605160]

Eriksen BA, Eriksen CW, 1974 Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept. Psychophys 16 (1), 143–149

Esmaeilzadeh S, Belivanis DI, Pohl KM, Adeli E, 2018 End-to-end Alzheimers disease diagnosis and biomarker identification In: Proceedings of the International Workshop on Machine Learning in Medical Imaging In: Lecture Notes in Computer Science, 11046. Springer, pp. 337–345

Etchell A, Adhikari A, Weinberg LS, Choo AL, Garnett EO, Chow HM, Chang S-E, 2018 A systematic literature review of sex differences in childhood language and brain development. Neuropsychologia 114, 19–31 [PubMed: 29654881]

Fan J, McCandliss BD, Sommer T, Raz A, Posner MI, 2002 Testing the efficiency and independence of attentional networks. J. Cognit. Neurosci 14 (3), 340–347 [PubMed: 11970796]

Fan L, Tang Y, Sun B, Gong G, Chen ZJ, Lin X, Yu T, Li Z, Evans AC, Liu S, 2010 Sexual dimorphism and asymmetry in human cerebellum: an MRI-based morphometric study. Brain Res 1353, 60–73. doi: 10.1016/j.brainres.2010.07.031 [PubMed: 20647004]

Feis D-L, Brodersen KH, von Cramon DY, Luders E, Tittgemeyer M, 2013 Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. NeuroImage 70, 250–257 [PubMed: 23298750]

Filipek PA, Richelme C, Kennedy DN, Caviness VSJ, 1994 The young adult human brain: an MRI-based morphometric analysis. Cereb. Cortex 4 (4), 344–360. doi: 10.1093/cercor/4.4.344 [PubMed: 7950308]

Fisher RA, 1935 The logic of inductive inference. J. R. Stat. Soc 98 (1), 39–82

Flaum M, Swayze V, O'leary D, Yuh W, Ehrhardt J, Arndt S, Andreasen N, 1995 Brain morphology in schizophrenia: effects of diagnosis, laterality and gender. Am. J. Psychiatry 152, 704–714 [PubMed: 7726310]

Freedman DS, Butte NF, Taveras EM, Lundeen EA, Blanck HM, Goodman AB, Ogden CL, 2017 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 25 (4), 739–746. doi: 10.1002/oby.21782 [PubMed: 28245098]

Frodl T, Meisenzahl EM, Zetzsche T, Born C, Groll C, Jager M, Leinsinger G, Bottlender R, Hahn K, Moller HJ, 2002 Hippocampal changes in patients with a first episode of major depression. Am. J. Psychiatry 159 (7), 1112–1118. doi: 10.1176/appi.ajp.159.7.1112 [PubMed: 12091188]

Galea LA, Spritzer MD, Barker JM, Pawluski JL, 2006 Gonadal hormone modulation of hippocampal neurogenesis in the adult. Hippocampus 16 (3), 225–232. doi: 10.1002/hipo.20154 [PubMed: 16411182]

Garavan H, Bartsch H, Conway K, Decastro A, Goldstein RZ, Heeringa S, Jernigan T, Potter A, Thompson W, Zahs D, 2018 Recruiting the ABCD sample: Design considerations and procedures. Dev. Cogn. Neurosci 32, 16–22. doi: 10.1016/j.dcn.2018.04.004 [PubMed: 29703560]

Gershon RC, Cook KF, Mungas D, Manly JJ, Slotkin J, Beaumont JL, Weintraub S, 2014 Language measures of the NIH toolbox cognition battery. J. Int. Neuropsychol. Soc 20 (6), 642–651. [PubMed: 24960128]

Gershon RC, Slotkin J, Manly JJ, Blitz DL, Beaumont JL, Schnipke D, Wallner-Allen K, Golinkoff RM, Gleason JB, Hirsh-Pasek K, 2013a IV. NIH Toolbox Cognition Battery (CB): Measuring language (vocabulary comprehension and reading decoding). Monogr. Soc. Res. Child Dev 78 (4), 49–69 [PubMed: 23952202]

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ, 2013b NIH toolbox for assessment of neurological and behavioral function. Neurology 80 (11 Supplement 3), S2–S6 [PubMed: 23479538]

Giedd JN, 2004 Structural magnetic resonance imaging of the adolescent brain. Ann. N. Y. Acad. Sci 1021, 77–85. doi: 10.1196/annals.1308.009 [PubMed: 15251877]

Giedd JN, Castellanos FX, Rajapakse JC, Vaituzis AC, Rapoport JL, 1997 Sexual dimorphism of the developing human brain. Progr. Neuro-Psychopharmacol. Biol. Psychiatry 21 (8), 1185–1201

Giedd JN, Raznahan A, Alexander-Bloch A, Schmitt E, Gogtay N, Rapoport JL, 2015 Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development. Neuropsychopharmacology 40 (1), 43–49 [PubMed: 25195638]

Giedd JN, Vaituzis AC, Hamburger SD, Lange N, Rajapakse JC, Kaysen D, Vauss YC, Rapoport JL, 1996 Quantitative MRI of the temporal lobe, amygdala, and hippocampus in normal human development: ages 4–18 years. J. Comp. Neurol 366 (2), 223–230. doi: 10.1002/(SICI)1096-9861(19960304)366:2<223::AID-CNE3>3.0.CO;2-7 [PubMed: 8698883]

Gilmore JH, Lin W, Prastawa MW, Looney CB, Vetsa YS, Knickmeyer RC, Evans DD, Smith JK, Hamer RM, Lieberman JA, Gerig G, 2007 Regional gray matter growth, sexual dimorphism, and cerebral asymmetry in the neonatal brain. J. Neurosci 27 (6), 1255–1260 [PubMed: 17287499]

Golarai G, Grill-Spector K, Reiss AL, 2006 Autism and the development of face processing. Clin. Neurosci. Res 6 (3–4), 145–160 [PubMed: 18176635]

Gold JM, Carpenter C, Randolph C, Goldberg TE, Weinberger DR, 1997 Auditory working memory and wisconsin card sorting test performance in schizophrenia. Arch. Gen. Psychiatry 54 (2), 159–165 [PubMed: 9040284]

Goldstein JM, Seidman LJ, Horton NJ, Makris N, Kennedy DN, Caviness VSJ, Faraone SV, Tsuang MT, 2001 Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. Cereb. Cortex 11 (6), 490–497. doi: 10.1093/cercor/11.6.490 [PubMed: 11375910]

Green T, Fierro KC, Raman MM, Foland-Ross L, Hong DS, Reiss AL, 2016 Sex differences in amygdala shape: Insights from turner syndrome. Hum. Brain Mapp 37 (4), 1593–1601. doi: 10.1002/hbm.23122 [PubMed: 26819071]

Gulli A, Pal S, 2017 Deep Learning with Keras Packt Publishing Ltd

Gur RC, Gur RE, 2017 Complementarity of sex differences in brain and behavior: From laterality to multimodal neuroimaging. J. Neurosci. Res 95 (1–2), 189–199 [PubMed: 27870413]

Gur RE, Gur RC, 2016 Sex differences in brain and behavior in adolescence: Findings from the philadelphia neurodevelopmental cohort. Neurosci. Biobehav. Rev 70, 159–170 [PubMed: 27498084]

Hagler DJ, Hatton SN, Cornejo MD, Makowski C, Fair DA, Dick AS, Sutherland MT, Casey B, Barch DM, Harms MP, Watts R, Bjork JM, Garavan HP, Hilmer L, Pung CJ, Sicat CS, Kuperman J, Bartsch H, Xue F, Heitzeg MM, Laird AR, Trinh TT, Gonzalez R, Tapert SF, Riedel MC, Squeglia LM, Hyde LW, Rosenberg MD, Earl EA, Howlett KD, Baker FC, Soules M, Diaz J, Leon O.R.d., Thompson WK, Neale MC, Herting M, Sowell ER, Alvarez RP, Hawes SW, Sanchez M, Bodurka J, Breslin FJ, Morris AS, Paulus MP, Simmons WK, Polimeni JR, v.d. Kouwe A, Nencka AS, Gray KM, Pierpaoli C, Matochik JA, Noronha A, Aklin WM, Conway K, Glantz M, Hoffman E, Little R, Lopez M, Pariyadath V, Weiss SR, Wolff-Hughes DL, DelCarmen-Wiggins R, Ewing SWF, Miranda-Dominguez O, Nagel BJ, Perrone AJ, Sturgeon DT, Goldstone A, Pfefferbaum A, Pohl KM, Prouty D, Uban K, Bookheimer SY, Dapretto M, Galvan A, Bagot K, Giedd J, Infante MA, Jacobus J, Patrick K, Shilling PD, Desikan R, Li Y, Sugrue L, Banich MT, Friedman N, Hewitt JK, Hopfer C, Sakai J, Tanabe J, Cottler LB, Nixon SJ, Chang L, Cloak C, Ernst T, Reeves G, Kennedy DN, Heeringa S, Peltier S, Schulenberg J, et al., 2019 Image processing and analysis methods for the adolescent brain cognitive development study. NeuroImage 202, 116091 [PubMed: 31415884]

Herting MM, Johnson C, Mills KL, Vijayakumar N, Dennison M, Liu C, Goddings A-L, Dahl RE, Sowell ER, Whittle S, 2018 Development of subcortical volumes across adolescence in males and females: A multisample study of longitudinal changes. NeuroImage 172, 194–205 [PubMed: 29353072]

Hill AC, Laird AR, Robinson JL, 2014 Gender differences in working memory networks: a brainmap meta-analysis. Biol. Psychol 102, 18–29. doi: 10.1016/j.biopsycho.2014.06.008 [PubMed: 25042764]

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

Hodes RJ, Insel TR, Landis SC, 2013 The NIH toolbox: Setting a standard for biomedical research. Neurology 80 (11 Supplement 3) S1–S1

Hänggi J, Buchmann A, Mondadori CR, Henke K, Jäncke L, Hock C, 2010 Sexual dimorphism in the parietal substrate associated with visuospatial cognition independent of general intelligence. J. Cognit. Neurosci 22 (1), 139–155 [PubMed: 19199407]

Iglesias JE, Liu CY, Thompson PM, Tu Z, 2011 Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634. doi: 10.1109/TMI.2011.2138152 [PubMed: 21880566]

Jahanshad N, Thompson PM, 2017 Multimodal neuroimaging of male and female brain structure in health and disease across the life span. J. Neurosci. Res 95 (1–2), 371–379 [PubMed: 27870421]

Johnson ES, Meade AC, 1987 Developmental patterns of spatial ability: an early sex difference. Child Dev 58 (3), 725–740 [PubMed: 3608645]

Kim HJ, Kim N, Kim S, Hong S, Park K, Lim S, Park JM, Na B, Chae Y, Lee J, Yeo S, Choe IH, Cho SY, Cho G, 2012 Sex differences in amygdala subregions: evidence from subregional shape analysis. NeuroImage 60 (4), 2054–2061. doi: 10.1016/j.neuroimage.2012.02.025 [PubMed: 22374477]

Kleinbaum DG, Klein M, 2002 Logistic Regression Springer

Kohavi R, 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1137–1143

Kotikalapudi R and contributors 2017 keras-vis https://github.com/raghakot/keras-vis

Krizhevsky A, Sutskever I, Hinton GE, 2012 Imagenet classification with deep convolutional neural networks In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (Eds.), Proceedings of the Advances in Neural Information Processing Systems 25 Curran Associates, Inc., pp. 1097–1105

Lenroot RK, Gogtay N, Greenstein DK, Wells EM, Wallace GL, Clasen LS, Blumenthal JD, Lerch J, Zijdenbos AP, Evans AC, 2007 Sexual dimorphism of brain developmental trajectories during childhood and adolescence. NeuroImage 36 (4), 1065–1073 [PubMed: 17513132]

Lenroot RK, Gogtay N, Greenstein DK, Wells EM, Wallace GL, Clasen LS, Blumenthal JD, Lerch J, Zijdenbos AP, Evans AC, Thompson PM, Giedd JN, 2007 Sexual dimorphism of brain developmental trajectories during childhood and adolescence. NeuroImage 36 (4), 1065–1073. doi: 10.1016/j.neuroimage.2007.03.053 [PubMed: 17513132]

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

Lin W, Tong T, Gao Q, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X, 2018 Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front. Neurosci 12, 777. doi: 10.3389/fnins.2018.00777 [PubMed: 30455622]

Lind KE, Gutierrez EJ, Yamamoto DJ, Regner MF, McKee SA, Tanabe J, 2017 Sex disparities in substance abuse research: Evaluating 23 years of structural neuroimaging studies. Drug Alcohol Depend 173, 92–98 [PubMed: 28212516]

Liu M, Zhang J, Nie D, Yap P, Shen D, 2018 Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J. Biomed. Health Inform 22 (5), 1476–1485. doi: 10.1109/JBHI.2018.2791863 [PubMed: 29994175]

Liu N, Cliffer S, Pradhan AH, Lightbody A, Hall SS, Reiss AL, 2016 Optical-imaging-based neurofeedback to enhance therapeutic intervention in adolescents with autism: methodology and initial data. Neurophotonics 4 (1), 011003 [PubMed: 27570790]

Llera A, Wolfers T, Mulders P, Beckmann C, 2019 Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. eLife 8. doi: 10.7554/eLife.44443

Lopez-Garcia P, Aizenstein HJ, Snitz BE, Walter RP, Carter CS, 2006 Automated ROI-based brain parcellation analysis of frontal and temporal brain volumes in schizophrenia. Psychiatry Res.: Neuroimaging 147 (2–3), 153–161

Luciana M, Bjork J, Nagel B, Barch D, Gonzalez R, Nixon S, Banich M, 2018 Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. Dev. Cognit. Neurosci 32, 67–79 [PubMed: 29525452]

Luders E, Toga AW, Thompson PM, 2014 Why size matters: differences in brain volume account for apparent sex differences in callosal anatomy: the sexual dimorphism of the corpus callosum. NeuroImage 84, 820–824 [PubMed: 24064068]

Luna B, Garver KE, Urban TA, Lazar NA, Sweeney JA, 2004 Maturation of cognitive processes from late childhood to adulthood. Child Dev 75 (5), 1357–1372 [PubMed: 15369519]

v.d. Maaten L, Hinton G, 2008 Visualizing data using t-SNE. J. Mach. Learn. Res 9 (Nov), 2579–2605

Madsen H, Thyregod P, 2010 Introduction to General and Generalized Linear Models CRC Press

Mankiw C, Park MTM, Reardon P, Fish AM, Clasen LS, Greenstein D, Giedd JN, Blumenthal JD, Lerch JP, Chakravarty MM, 2017 Allometric analysis detects brain size-independent effects of sex and sex chromosome complement on human cerebellar organization. J. Neurosci 37 (21), 5221–5231 [PubMed: 28314818]

McEwen BS, 1983 Gonadal steroid influences on brain development and sexual differentiation. Int. Rev. Physiol 27, 99–145 [PubMed: 6303978]

Menon V, Uddin L, 2010 Saliency, switching, attention and control: A network model of insula function. Brain Struct. Funct 214, 655–667. doi: 10.1007/s00429-010-0262-0 [PubMed: 20512370]

Mills KL, Lalonde F, Clasen LS, Giedd JN, Blakemore S-J, 2012 Developmental changes in the structure of the social brain in late childhood and adolescence. Soc. Cognit. Affect. Neurosci 9 (1), 123–131 [PubMed: 23051898]

Mungas D, Heaton R, Tulsky D, Zelazo PD, Slotkin J, Blitz D, Lai J-S, Gershon R, 2014 Factor structure, convergent validity, and discriminant validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in adults. J. Int. Neuropsychol. Soc 20 (6), 579–587 [PubMed: 24960474]

Murphy MC, Jones DT, Jack CR Jr, Glaser KJ, Senjem ML, Manduca A, Felmlee JP, Carter RE, Ehman RL, Huston J III, 2016 Regional brain stiffness changes across the Alzheimer's disease spectrum. NeuroImage: Clin 10, 283–290 [PubMed: 26900568]

Narvacan K, Treit S, Camicioli R, Martin W, Beaulieu C, 2017 Evolution of deep gray matter volume across the human lifespan. Hum. Brain Mapp 38 (8), 3771–3790 [PubMed: 28548250]

Nie D, Zhang H, Adeli E, Liu L, Shen D, 2016 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. Med. Image Comput. Comput. Assist. Interv., Lecture Notes in Computer Science 9901, 212–220. doi: 10.1007/978-3-319-46723-8_25

Nieuwenhuis M, Schnack HG, van Haren NE, Lappin J, Morgan C, Reinders AA, Gutierrez-Tordesillas D, Roiz-Santiañez R, Schaufelberger MS, Rosa PG, 2017 Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. NeuroImage 145, 246–253 [PubMed: 27421184]

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

Oksuz I, Ruijsink B, Puyol-Anton E, Clough JR, Cruz G, Bustin A, Prieto C, Botnar R, Rueckert D, Schnabel JA, King AP, 2019 Automatic cnn-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. Med. Image Anal 55, 136–147. doi: 10.1016/j.media.2019.04.009 [PubMed: 31055126]

Onofrey JA, Staib LH, Papademetris X, 2018 Segmenting the brain surface from ct images with artifacts using locally oriented appearance and dictionary learning. IEEE Trans. Med. Imaging 38 (2), 596–607 [PubMed: 30176584]

Park SH, Zhang Y, Kwon D, Zhao Q, Zahr NM, Pfefferbaum A, Sullivan EV, Pohl KM, 2018 Alcohol use effects on adolescent brain development revealed by simultaneously removing confounding factors, identifying morphometric patterns, and classifying individuals. Sci. Rep 8 (1), 8297 [PubMed: 29844507]

Pearson K, 1900 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. Lond. Edinb. Dublin Philos. Mag. J. Sci 50 (302), 157–175

Perlaki G, Orsi G, Plozer E, Altbacker A, Darnai G, Nagy SA, Horvath R, Toth A, Doczi T, Kovacs N, 2014 Are there any gender differences in the hippocampus volume after head-size correction? A volumetric and voxel-based morphometric study. Neurosci. Lett 570, 119–123 [PubMed: 24746928]

Petersen AC, Crockett L, Richards M, Boxer A, 1988 A self-report measure of pubertal status: Reliability, validity, and initial norms. J. Youth Adolesc 17 (2), 117–133 [PubMed: 24277579]

Petrican R, Taylor MJ, Grady CL, 2017 Trajectories of brain system maturation from childhood to older adulthood: Implications for lifespan cognitive functioning. NeuroImage 163, 125–149. doi: 10.1016/j.neuroimage.2017.09.025 [PubMed: 28917697]

Pfefferbaum A, Kwon D, Brumback T, Thompson WK, Cummins K, Tapert SF, Brown SA, Colrain IM, Baker FC, Prouty D, De Bellis MD, Clark DB, Nagel BJ, Chu W, Park SH, Pohl KM, Sullivan EV, 2018 Altered brain developmental trajectories in adolescents after initiating drinking. Am. J. Psychiatry 175 (4), 370–380. doi: 10.1176/appi.ajp.2017.17040469 [PubMed: 29084454]

Pfefferbaum A, Rohlfing T, Pohl KM, Lane B, Chu W, Kwon D, Nolan Nichols B, Brown SA, Tapert SF, Cummins K, Thompson WK, Brumback T, Meloy MJ, Jernigan TL, Dale A, Colrain IM, Baker FC, Prouty D, De Bellis MD, Voyvodic JT, Clark DB, Luna B, Chung T, Nagel BJ, Sullivan EV, 2016 Adolescent development of cortical and white matter structure in the NCANDA sample: Role of sex, ethnicity, puberty, and alcohol drinking. Cereb. Cortex 26 (10), 4101–4121. doi: 10.1093/cercor/bhv205 [PubMed: 26408800]

Pfeiffer CA, 1936 Sexual differences of the hypophyses and their determination by the gonads. Am. J. Anatomy 58 (1), 195–225

Phinyomark A, Hettinga BA, Osis ST, Ferber R, 2014 Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. PLoS One 9 (8), e105246 [PubMed: 25137240]

Pierce K, Gazestani VH, Bacon E, Barnes CC, Cha D, Nalabolu S, Lopez L, Moore A, Pence-Stophaeros S, Courchesne E, 2019 Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. JAMA Pediatr 173 (6), 578–587 [PubMed: 31034004]

Pilly PK, Howard MD, Bhattacharyya R, 2018 Modeling contextual modulation of memory associations in the hippocampus. Front. Hum. Neurosci 12, 442. doi: 10.3389/fnhum.2018.00442 [PubMed: 30473660]

Pohl KM, Thompson W, Adeli E, Linguraru MG, 2019 Adolescent Brain Cognitive Development Neurocognitive Prediction Challenge Lecture Notes in Computer Science, 11791 Springer-Verlag, Berlin, Germany

pygrowup, 2017, https://pypi.org/project/pygrowup/, Retrieved August 26, 2020.

Raz N, Gunning-Dixon F, Head D, Williamson A, Acker JD, 2001 Age and sex differences in the cerebellum and the ventral pons: A prospective MR study of healthy adults. Am. J. Neuroradiol 22 (6), 1161–1167 [PubMed: 11415913]

Retico A, Giuliano A, Tancredi R, Cosenza A, Apicella F, Narzisi A, Biagi L, Tosetti M, Muratori F, Calderoni S, 2016 The effect of gender on the neuroanatomy of children with autism spectrum disorders: A support vector machine case-control study. Mol. Autism 7 (1), 5 [PubMed: 26788282]

Rohlfing T, Russakoff DB, Maurer CRJ, 2004 Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging 23 (8), 983–994. doi: 10.1109/TMI.2004.830803 [PubMed: 15338732]

Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A, 2010 The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp 31 (5), 798–819. doi: 10.1002/hbm.20906 [PubMed: 20017133]

Román F, Garcia-Sánchez FA, Martinez-Selva JM, Gómez-Amor J, Carrillo E, 1989 Sex differences and bilateral electrodermal activity. Pavlov. J. Biol. Sci 24 (4), 150–155 [PubMed: 2616227]

Rueda MR, Fan J, McCandliss BD, Halparin JD, Gruber DB, Lercari LP, Posner MI, 2004 Development of attentional networks in childhood. Neuropsychologia 42 (8), 1029–1040 [PubMed: 15093142]

Ruigrok AN, Salimi-Khorshidi G, Lai M-C, Baron-Cohen S, Lombardo MV, Tait RJ, Suckling J, 2014 A meta-analysis of sex differences in human brain structure. Neurosci. Biobehav. Rev 39, 34–50 [PubMed: 24374381]

Sacher J, Neumann J, Okon-Singer H, Gotowiec S, Villringer A, 2013 Sexual dimorphism in the human brain: Evidence from neuroimaging. Magn. Reson. Imaging 31 (3), 366–375 [PubMed: 22921939]

Sadananthan SA, Zheng W, Chee MW, Zagorodnov V, 2010 Skull stripping using graph cuts. NeuroImage 49 (1), 225–239. doi: 10.1016/j.neuroimage.2009.08.050 [PubMed: 19732839]

Salthouse TA, Babcock RL, Shaw RJ, 1991 Effects of adult age on structural and operational capacities in working memory. Psychol. Aging 6 (1), 118 [PubMed: 2029360]

Sanchis Segura C, Aguirre N, Cruz-Gómez AJ, Solozano N, Forn C, 2018 Do gender-related stereotypes affect spatial performance? Exploring when, how and to whom using a chronometric two-choice mental rotation task. Front. Psychol 9, 1261 [PubMed: 30087637]

Saunders EF, Nazir R, Kamali M, Ryan KA, Evans S, Langenecker S, Gelenberg AJ, McInnis MG, 2014 Gender differences, clinical correlates, and longitudinal outcome of bipolar disorder with comorbid migraine. J. Clin. Psychiatry 75 (5), 512–519. doi: 10.4088/JCP.13m08623 [PubMed: 24816075]

Sawyer KS, Maleki N, Papadimitriou G, Makris N, Oscar-Berman M, Harris GJ, 2018 Cerebral white matter sex dimorphism in alcoholism: A diffusion tensor imaging study. Neuropsychopharmacology 43 (9), 1876 [PubMed: 29795404]

Shaffer JP, 1995 Multiple hypothesis testing. Ann. Rev. Psych 46, 561–584

Simonyan K, Vedaldi A, Zisserman A, 2014 Deep inside convolutional networks: Visualising image classification models and saliency maps Proceedings of the International Conference on Learning Representations Workshop ICLR, pp. 1–8

Smith SM, 2002 Fast robust automated brain extraction. Hum. Brain Mapp 17 (3), 143–155. doi: 10.1002/hbm.10062 [PubMed: 12391568]

Strickler GK, Kreiner PW, Halpin JF, Doyle E, Paulozzi LJ, 2020 Opioid prescribing behaviors - prescription behavior surveillance system, 11 states, 2010–2016.. MMWR Surveill. Summ 69 (1), 1–14

Sullivan EV, Brumback T, Tapert SF, Brown SA, Baker FC, Colrain IM, Prouty D, De Bellis MD, Clark DB, Nagel BJ, 2020 Disturbed cerebellar growth trajectories in adolescents who initiate alcohol drinking. Biol. Psychiatry 87 (7), 632–644 [PubMed: 31653477]

Sullivan EV, Brumback T, Tapert SF, Fama R, Prouty D, Brown SA, Cummins K, Thompson WK, Colrain IM, Baker FC, 2016 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), 449–473 [PubMed: 26752122]

Sullivan EV, Rosenbloom MJ, Desmond JE, Pfefferbaum A, 2001 Sex differences in corpus callosum size: Relationship to age and intracranial size. Neurobiol. Aging 22 (4), 603–611 [PubMed: 11445261]

Szabo CA, Lancaster JL, Xiong J, Cook C, Fox P, 2003 MR imaging volumetry of subcortical structures and cerebellar hemispheres in normal persons. Am. J. Neuroradiol 24 (4), 644–647 [PubMed: 12695196]

Tamnes CK, Herting MM, Goddings A-L, Meuwese R, Blakemore S-J, Dahl RE, Guroglu B, Raznahan A, Sowell ER, Crone EA, 2017 Development of the cerebral cortex across adolescence: A multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness. J. Neurosci 37 (12), 3402–3412 [PubMed: 28242797]

Teicher MH, Andersen SL, Polcari A, Anderson CM, Navalta CP, Kim DM, 2003 The neurobiological consequences of early stress and childhood maltreatment. Neurosci. Biobehav. Rev 27 (1–2), 33–44 [PubMed: 12732221]

Thompson WK, Barch DM, Bjork JM, Gonzalez R, Nagel BJ, Nixon SJ, Luciana M, 2019 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, 100606. doi: 10.1016/j.dcn.2018.12.004 [PubMed: 30595399]

Tiemeier H, Lenroot RK, Greenstein DK, Tran L, Pierson R, Giedd JN, 2010 Cerebellum development during childhood and adolescence: A longitudinal morphometric MRI study. NeuroImage 49 (1), 63–70. doi: 10.1016/j.neuroimage.2009.08.016 [PubMed: 19683586]

Trenerry MR, Jack CR Jr, Cascino GD, Sharbrough FW, Ivnik RJ, 1995 Gender differences in post-temporal lobectomy verbal memory and relationships between MRI hippocampal volumes and preoperative verbal memory. Epilepsy Res 20 (1), 69–76 [PubMed: 7713061]

Tulsky DS, Carlozzi NE, Chevalier N, Espy KA, Beaumont JL, Mungas D, 2013 V. NIH toolbox cognition battery (cb): Measuring working memory. Monogr. Soc. Res. Child Dev 78 (4), 70–87 [PubMed: 23952203]

Tulsky DS, Carlozzi N, Chiaravalloti ND, Beaumont JL, Kisala PA, Mungas D, Conway K, Gershon R, 2014 NIH Toolbox Cognition Battery (NIHTB-CB): List sorting test to measure working memory. J. Int. Neuropsychol. Soc 20 (6), 599–610 [PubMed: 24959983]

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC, 2010 N4ITK: Improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. doi: 10.1109/TMI.2010.2046908 [PubMed: 20378467]

Van Putten MJ, Olbrich S, Arns M, 2018 Predicting sex from brain rhythms with deep learning. Sci. Rep 8 (1), 3069 [PubMed: 29449649]

Vijayakumar N, Allen NB, Youssef G, Dennison M, Yücel M, Simmons JG, Whittle S, 2016 Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume. Hum. Brain Mapp 37 (6), 2027–2038 [PubMed: 26946457]

Vogeley K, Schneider-Axmann T, Pfeiffer U, Tepest R, Bayer TA, Bogerts B, Honer WG, Falkai P, 2000 Disturbed gyrification of the prefrontal region in male schizophrenic patients: a morphometric postmortem study. Am. J. Psychiatry 157 (1), 34–39 [PubMed: 10618010]

Wang L, Shen H, Tang F, Zang Y, Hu D, 2012 Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain: An MVPA approach. NeuroImage 61 (4), 931–940 [PubMed: 22498657]

Weinhandl JT, Joshi M, O'Connor KM, 2010 Gender comparisons between unilateral and bilateral landings. J. Appl. Biomech 26 (4), 444–453 [PubMed: 21245504]

Wierenga C, Bischoff-Grethe A, Melrose AJ, Grenesko-Stevens E, Irvine Z, Wagner A, Simmons A, Matthews S, Yau W-YW, Fennema-Notestine C, 2014 Altered bold response during inhibitory and error processing in adolescents with anorexia nervosa. PloS One 9 (3), e92017 [PubMed: 24651705]

Wierenga LM, Bos MG, Schreuders E, vd Kamp F, Peper JS, Tamnes CK, Crone EA, 2018 Unraveling age, puberty and testosterone effects on subcortical brain development across adolescence. Psychoneuroendocrinology 91, 105–114 [PubMed: 29547741]

Wierenga LM, Sexton JA, Laake P, Giedd JN, Tamnes CK, Pediatric Imaging N, Genetics S, 2018 A key characteristic of sex differences in the developing brain: Greater variability in brain structure of boys than girls. Cereb Cortex 28 (8), 2741–2751. doi: 10.1093/cercor/bhx154 [PubMed: 28981610]

Witelson S, Beresh H, Kigar D, 2005 Intelligence and brain size in 100 postmortem brains: sex, lateralization and age factors. Brain 129 (2), 386–398 [PubMed: 16339797]

Witelson SF, 1989 Hand and sex differences in the isthmus and genu of the human corpus callosum: a postmortem morphological study. Brain 112 (3), 799–835 [PubMed: 2731030]

Wittek A, Joldes G, Couton M, Warfield SK, Miller K, 2010 Patient-specific non-linear finite element modelling for predicting soft organ deformation in real-time; application to non-rigid neuroimage registration. Progr. Biophys. Mol. Biol 103 (2–3), 292–303

Womer FY, Tang Y, Harms MP, Bai C, Chang M, Jiang X, Wei S, Wang F, Barch DM, 2016 Sexual dimorphism of the cerebellar vermis in schizophrenia. Schizophr. Res 176 (2–3), 164–170 [PubMed: 27401530]

Woodson JC, Gorski RA, 2000 Structural sex differences in the mammalian brain: Reconsidering the male/female dichotomy. Sex. Differ. Brain 229–255

Xin J, Zhang XY, Tang Y, Yang Y, 2019 Brain differences between men and women: Evidence from deep learning. Front. Neurosci 13, 185 [PubMed: 30906246]

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

Yi W-J, Heo M-S, Lee S-S, Choi S-C, Huh K-H, 2006 ROI-based image registration for digital subtraction radiography. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontol 101 (4), 523–529

Young KD, Bellgowan PS, Bodurka J, Drevets WC, 2013 Functional neuroimaging of sex differences in autobiographical memory recall. Hum. Brain Mapp 34 (12), 3320–3332. doi: 10.1002/hbm.22144 [PubMed: 22807028]

Zelazo PD, 2006 The dimensional change card sort (dccs): A method of assessing executive function in children. Nat. Protoc 1 (1), 297–301 [PubMed: 17406248]

Zhao F, Xia S, Wu Z, Duan D, Wang L, Lin W, Gilmore JH, Shen D, Li G, 2019 Spherical U-Net on cortical surfaces: Methods and applications In: Proceedings of the International Conference on Information Processing in Medical Imaging In: Lecture Notes in Computer Science, 11492. Springer, pp. 855–866

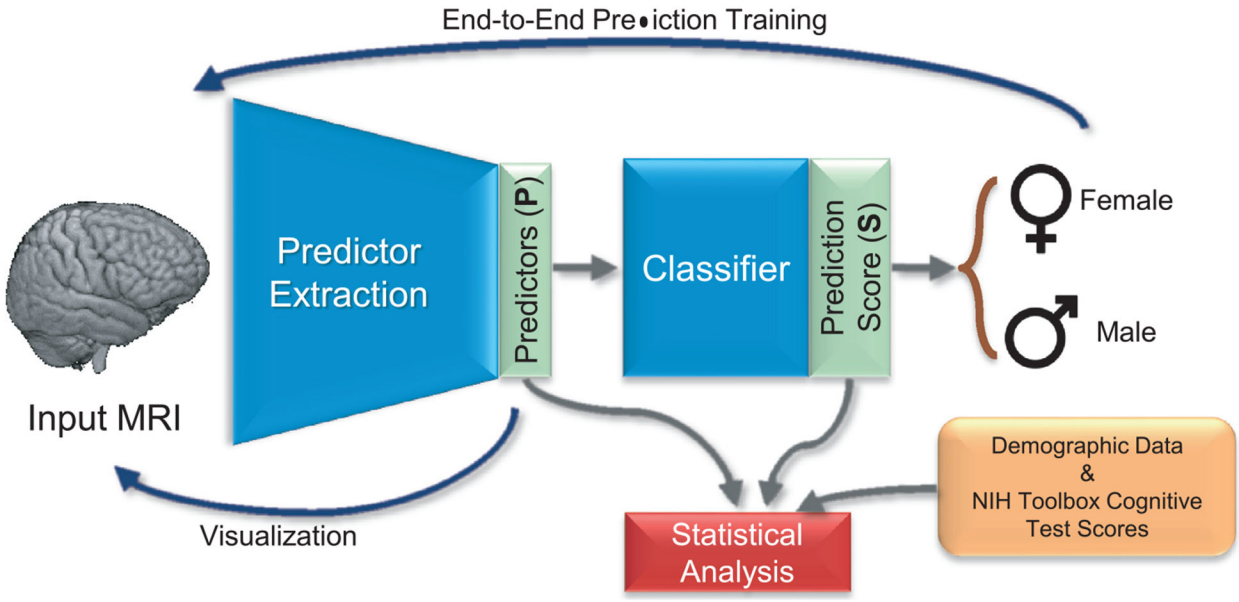


Fig. 1.

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.

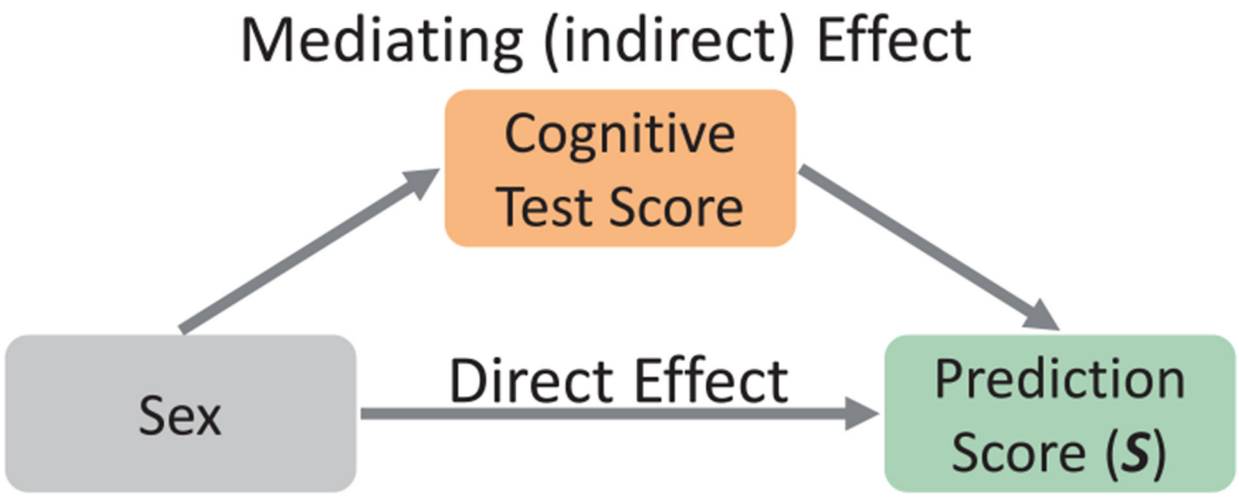


Fig. 2.

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.

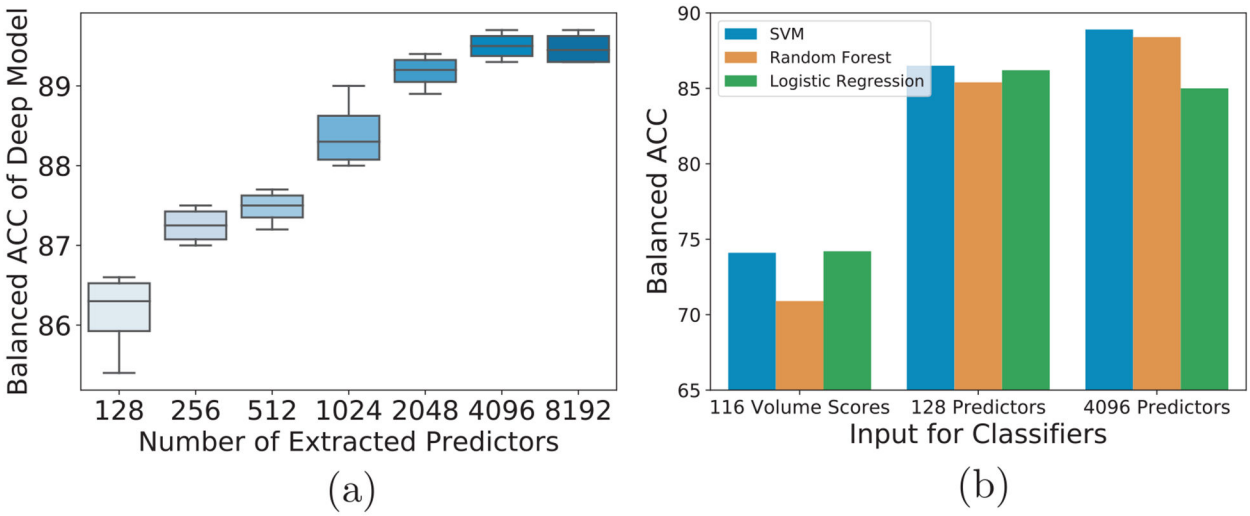


Fig. 3.

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

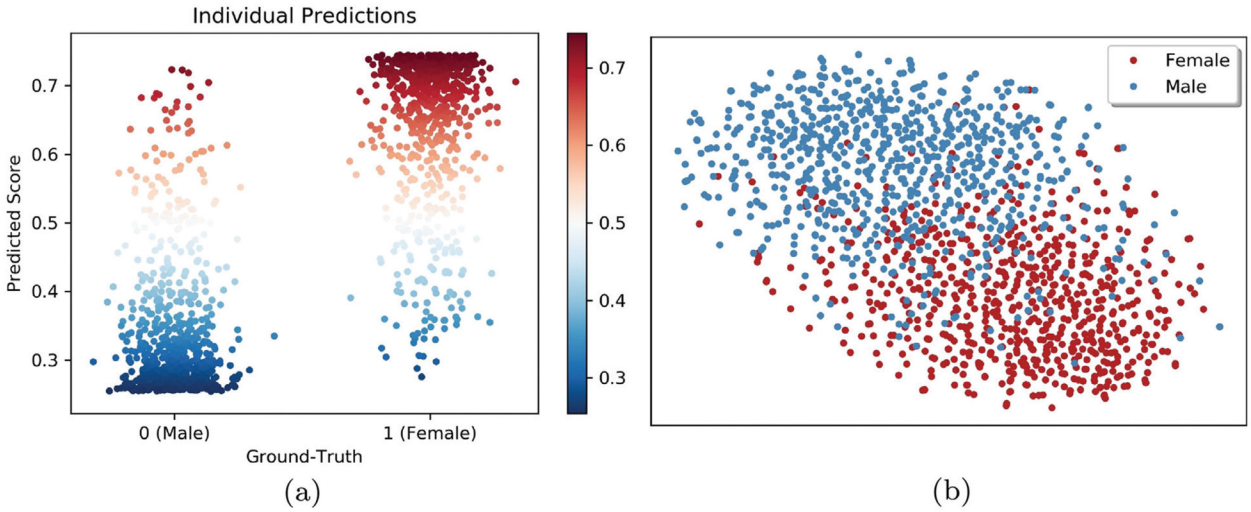
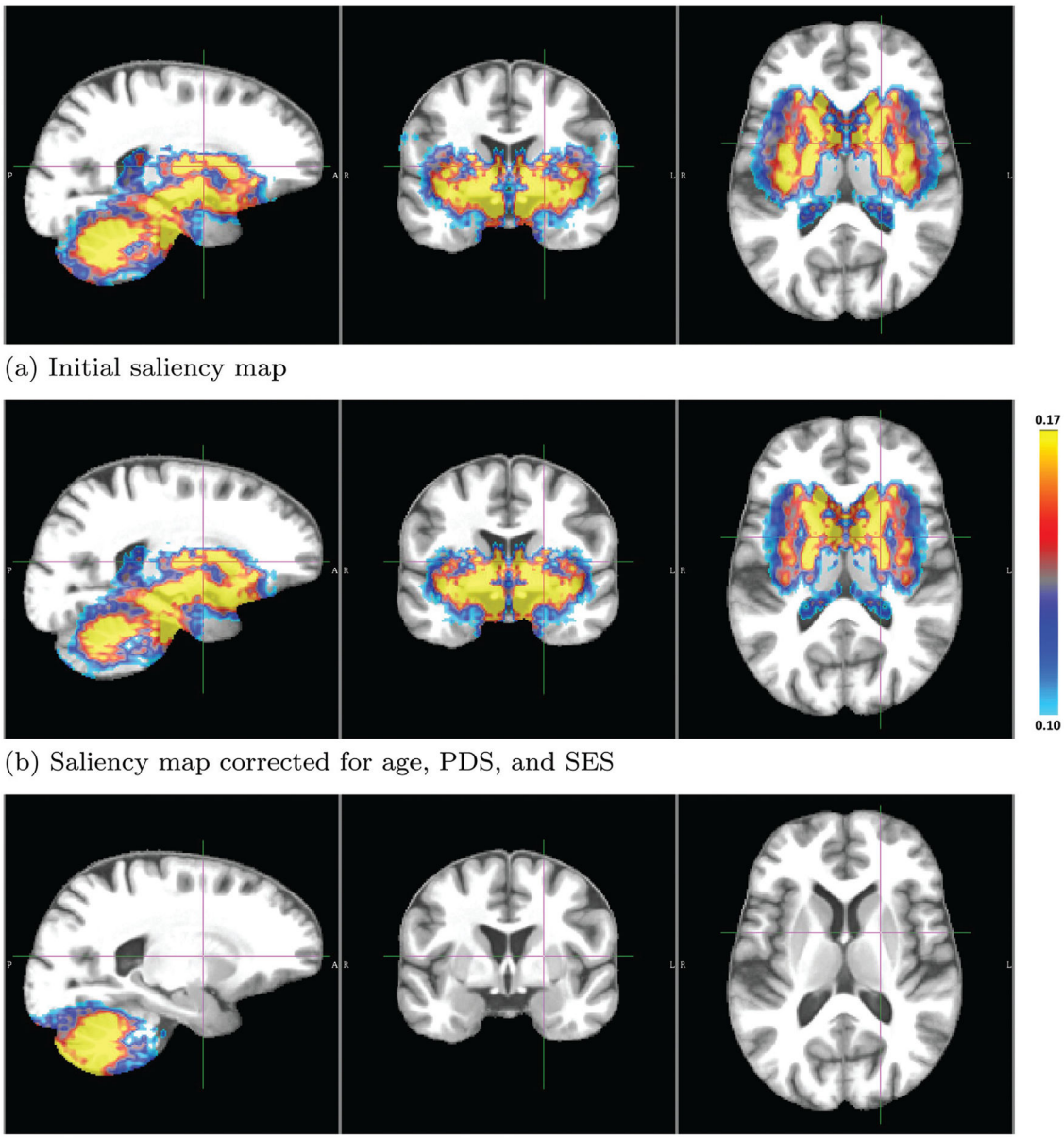


Fig. 4.

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) (Maaten and Hinton, 2008) 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 with neighborhoods reflecting those of the high dimensional space (according to Maaten and Hinton, 2008).



(a) Initial saliency map

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

(c) Saliency maps confounded by PDS

Fig. 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.

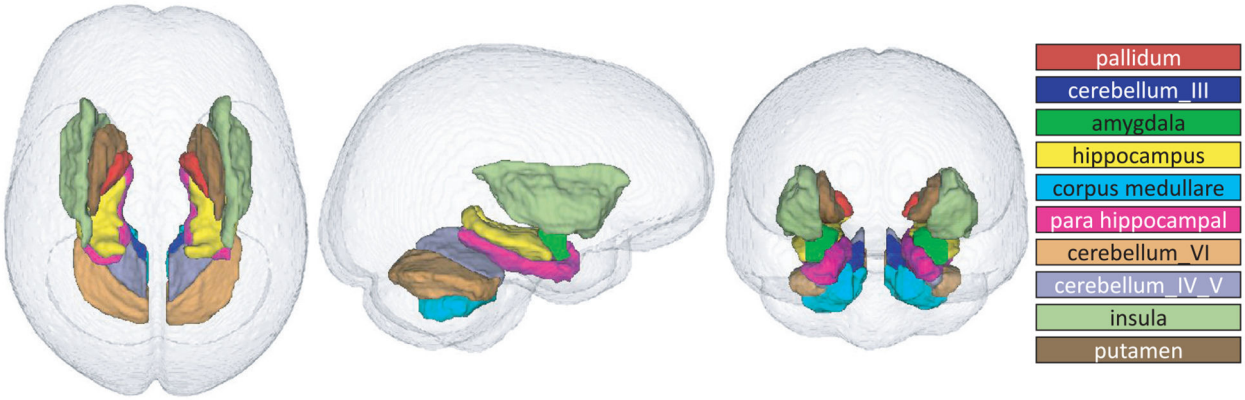


Fig. 6.

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*

Table 1

Demographic information (mean ± standard deviation).

Measure	Female (F)	Male (M)	*p*-value[a]	Group difference
Total subjects	**3895**	**4249**	-	-
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 (BM1) *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 (pygrowup, 2017) 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 and 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