

1 **Title:** Development of subcortical volumes across adolescence in males and females: A  
2 multisample study of longitudinal changes

3  
4 **Authors:** Megan M. Herting<sup>1</sup>, Cory Johnson<sup>1</sup>, Kathryn L. Mills<sup>2</sup>, Nandita Vijayakumar<sup>2</sup>, Meg  
5 Dennison<sup>3</sup>, Chang Liu<sup>1</sup>, Anne-Lise Goddings<sup>4</sup>, Ronald E. Dahl<sup>5</sup>, Elizabeth R. Sowell<sup>6</sup>, Sarah  
6 Whittle<sup>7</sup>, Nicholas B. Allen<sup>2</sup>, Christian K. Tamnes<sup>8</sup>

7  
8 <sup>1</sup> Department of Preventive Medicine, Keck School of Medicine, University of Southern  
9 California, Los Angeles, CA, USA

10 <sup>2</sup> Department of Psychology, University of Oregon, Eugene, OR, USA

11 <sup>3</sup> Phoenix Australia: Centre for Posttraumatic Mental Health, Department of Psychiatry, The  
12 University of Melbourne, Melbourne, AU

13 <sup>4</sup> Institute of Child Health, University College London, London, UK

14 <sup>5</sup> Institute of Human Development, University of California Berkeley, Berkeley, CA, USA

15 <sup>6</sup> Department of Pediatrics, Keck School of Medicine, University of Southern California, and  
16 Children's Hospital Los Angeles, Los Angeles, CA, USA,

17 <sup>7</sup> Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne  
18 and Melbourne Health, Melbourne, AU

19 <sup>8</sup> Department of Psychology, University of Oslo, Oslo, Norway

20  
21 Corresponding author: Megan M. Herting, Department of Preventive Medicine, University of  
22 Southern California, 2001 N Soto, Los Angeles, CA, 90032, USA; Email: [herting@usc.edu](mailto:herting@usc.edu)

23  
24

1 **Abstract**

2 The developmental patterns of subcortical brain volumes in males and females observed in  
3 previous studies have been inconsistent. To help resolve these discrepancies, we examined  
4 developmental trajectories using three independent longitudinal samples of participants in the  
5 age-span of 8-22 years (total 216 participants and 467 scans). These datasets, including  
6 *Pittsburgh* (PIT; University of Pittsburgh, USA), *NeuroCognitive Development* (NCD; University  
7 of Oslo, Norway), and *Orygen Adolescent Development Study* (OADS; The University of  
8 Melbourne, Australia), span three countries and were analyzed together and in parallel using  
9 mixed-effects modeling with both generalized additive models and general linear models. For all  
10 regions and across all samples, males were found to have significantly larger volumes as  
11 compared to females, and significant sex differences were seen in age trajectories over time.  
12 However, direct comparison of sample trajectories and sex differences identified within samples  
13 were not consistent. The trajectories for the amygdala, putamen, and nucleus accumbens were  
14 most consistent between the three samples. Our results suggest that even after using similar  
15 preprocessing and analytic techniques, additional factors, such as image acquisition or sample  
16 composition may contribute to some of the discrepancies in sex specific patterns in subcortical  
17 brain changes across adolescence, and highlight region-specific variations in congruency of  
18 developmental trajectories.

19  
20  
21

22 **Keywords:** neurodevelopment; subcortical; adolescence; sex differences; replication;  
23 longitudinal; Magnetic Resonance Imaging; amygdala; hippocampus

24  
25  
26  
27

1

## 2 **1. Introduction**

3           Developmental patterns of brain morphology, and sex differences in this structural  
4 variation, exist due to both global and local maturational changes (Sowell, Thompson et al.  
5 2004, Tamnes, Walhovd et al. 2013, Erus, Battapady et al. 2015, Giedd, Raznahan et al. 2015,  
6 Narvacan, Treit et al. 2017). Determining when and how sex differences emerge in the  
7 developing brain is essential to understanding differential risk for disease, especially  
8 psychopathology (Kessler, McGonagle et al. 1993, Kessler, Berglund et al. 2005), as well as  
9 life-long sex differences in various cognitive and behavioral traits (Choudhury, Blakemore et al.  
10 2006, Rose and Rudolph 2006, Roalf, Gur et al. 2014, Gur and Gur 2016). For example, late  
11 childhood and adolescence is a time period when many forms of psychopathology begin to  
12 emerge and do so in a sex-specific fashion, with disproportionate increases in rates of anxiety  
13 and depression seen in girls and a higher prevalence of externalizing behaviors and substance  
14 use disorders in boys (Kessler, Berglund et al. 2005, Kuhn 2015). Given that structural and  
15 functional abnormalities in subcortical regions have been associated with these various mental  
16 health problems, it is thought that plausible sex differences in the development of subcortical  
17 structures may be pertinent to explaining sex differences in onset, prevalence, and progression  
18 of mental health disorders (Paus, Keshavan et al. 2008, Gogtay and Thompson 2010, Shaw,  
19 Gogtay et al. 2010). As such, a number of sex differences have been reported in structural  
20 magnetic resonance imaging (MRI) growth trajectories of subcortical structures. However,  
21 developmental patterns observed in these structures have been inconsistent across studies,  
22 and there has yet to be a consensus as to how these patterns differ between sexes (Sowell,  
23 Trauner et al. 2002, Lenroot, Gogtay et al. 2007, Ostby, Tamnes et al. 2009, Dennison, Whittle  
24 et al. 2013, Wierenga, Langen et al. 2014, Narvacan, Treit et al. 2017).

25           To date, studies have reported discrepant findings including growth versus reduction of  
26 the thalamus and basal ganglia beginning in late childhood, as well as stability versus  
27 continuing growth of the amygdala and hippocampus across adolescence (Giedd, Vaituzis et al.  
28 1996, Sowell, Trauner et al. 2002, Ostby, Tamnes et al. 2009, Koolschijn and Crone 2013,

1 Wierenga, Langen et al. 2014). Similarly, reported sex differences in these trajectories remain  
2 variable. From a study design perspective, it is believed that longitudinal study designs that are  
3 able to better account for both within- and between- individual differences over time may help to  
4 improve our understanding of cross-sectional findings that focus on mean group differences  
5 between the sexes (Crone and Elzinga 2015). As such, longitudinal MRI studies using raw  
6 volumes (uncorrected for whole brain size or other allometric scaling) consistently show larger  
7 volumes in males as compared to females (i.e. main effects) (Dennison, Whittle et al. 2013,  
8 Raznahan, Shaw et al. 2014, Wierenga, Langen et al. 2014, Narvacan, Treit et al. 2017).  
9 However, findings are less clear in terms of sex differences in the trajectories (i.e. slopes) of  
10 development seen across childhood and adolescence. Based on using raw volume estimates  
11 (i.e. trajectories reported without including allometric scaling), some studies report sex  
12 differences in neurodevelopmental trajectories of subcortical regions (Dennison, Whittle et al.  
13 2013, Goddings, Mills et al. 2014, Raznahan, Shaw et al. 2014), whereas other studies find no  
14 difference between the sexes (Wierenga, Langen et al. 2014, Narvacan, Treit et al. 2017).

15         These discrepant observations in studies of subcortical volume development and sex  
16 differences in these patterns may be due to a number of factors, including cohort effects  
17 inherent to the sample, variation in study design, image acquisition and preprocessing, and/or  
18 statistical modeling approaches. In terms of image processing, dissimilarities have been  
19 reported in the absolute volume estimates as well as in the reliability of subcortical brain  
20 structures across different freely available automated segmentation software (Morey, Selgrade  
21 et al. 2010, Makowski, Beland et al. 2017). In addition, software packages vary in their  
22 methodology for processing longitudinal scans. For example, FreeSurfer's longitudinal pipeline  
23 includes creating an unbiased within-subject template space to help reduce random variation  
24 and improve the sensitivity of detecting changes over time (Reuter, Schmansky et al. 2012).  
25 Recently, a longitudinal cortical thickness pipeline has also been developed as part of the ANTs  
26 software (Tustison, Holbrook et al. Unpublished). To our knowledge, other commonly used  
27 software packages for structural analysis (i.e. CIVET (Zijdenbos, Forghani et al. 2002), MAGeT

1 (Chakravarty, Steadman et al. 2013), and FSL (Zhang, Brady et al. 2001)) do not account for  
2 within-subject variance in a similar fashion during the preprocessing stream. Beyond software,  
3 differences in quality control (QC) procedures utilized across studies may also impact the  
4 results (Ducharme, Albaugh et al. 2016).

5 From a statistical perspective, the inclusion of covariates and/or statistical model vary  
6 widely by study and may impact results (Vijayakumar, Mills et al. Accepted). For example,  
7 during statistical testing the inclusion of a 'global' or 'allometric' covariate to account for *between*  
8 *subject* differences in body size or weight (Sanfilippo, Benedict et al. 2004) may directly influence  
9 sex differences that are identified (Lenroot, Gogtay et al. 2007, Dennison, Whittle et al. 2013).  
10 Moreover, despite sex differences in allometric variables (i.e. whole brain or intracranial  
11 volume), recent findings suggest that the variability of anatomical volumes are not equal  
12 between the sexes (males show larger variance expressed at both upper and lower extremities  
13 of the distributions) (Wierenga, Sexton et al. 2017), allometric covariates follow non-linear  
14 developmental patterns from childhood to adulthood (Mills, Goddings et al. 2016, Reardon,  
15 Clasen et al. 2016), and regions including the thalamus, striatum, and pallidum show  
16 hypoallometric scaling with whole brain size (i.e. volumes become proportionately smaller with  
17 increasing head size) (Reardon, Clasen et al. 2016). Moreover, the inclusion of an allometric  
18 term may be redundant when examining longitudinal change using hierarchical modeling, as  
19 each subject receives its own intercept and slope (Crone and Elzinga 2015). Thus, the *between-*  
20 *subject* variance due to individual differences in head size is captured at the individual level over  
21 time; allowing for better characterization of changes in regional volume estimates over time.

22 Study results may vary based on the type of statistical analytic techniques employed.  
23 Although longitudinal studies have typically used linear mixed effect modeling (LME) to describe  
24 age-related changes, the model terms are diverse (Vijayakumar, Mills et al. Accepted). For  
25 example, studies have differed in their modeling approach, including use of polynomial terms  
26 (e.g. quadratic or cubic), model selection strategy (e.g. top-down or likelihood indices), testing  
27 males and females separately and/or including sex as an interaction term, as well as regarding

1 the inclusion of other confounding factors (Ruigrok, Salimi-Khorshidi et al. 2014). Moreover,  
2 while LME including polynomial terms remains a popular approach, polynomials are rather  
3 restrictive, whereas other modeling techniques, such as general additive modeling (GAMM),  
4 may allow for a more flexible fit of a curve to the data. Specifically, GAMM replaces the linear  
5 slope parameters with 'smooth' functions to find the optimal functional form between the  
6 predictor and response (Jones and Almond 1992). Given the existing discrepancies in the  
7 existing literature and the vast array of methodology (including software, quality checking  
8 procedures, and model terms) utilized between studies, there remains an important gap in our  
9 knowledge regarding the reproducibility of possible sex differences in subcortical  
10 neurodevelopmental trajectories across childhood and adolescence.

11 The goal of the current study was to utilize identical image processing and analysis  
12 methods in three independent longitudinal neuroimaging samples to describe the development  
13 of uncorrected subcortical volumes for males and females from late childhood into young  
14 adulthood. This study is part of an international collaboration project intended to improve the  
15 reliability and efficiency of neurodevelopmental research by simultaneously analyzing multiple  
16 existing neuroimaging datasets (Mills, Goddings et al. 2016, Tamnes, Herting et al. 2017). By  
17 keeping longitudinal preprocessing methods, quality control procedures, and statistical methods  
18 constant across samples, we can assess and interpret the potential impact of sample and  
19 acquisition differences on brain development patterns in males and females. Moreover, given  
20 inherent study design differences between the longitudinal samples (e.g. age ranges and scan  
21 follow-up), we explored age and age by sex relationships in each sample using both the more  
22 flexible general additive modeling (GAMM) approach as well as the more common general  
23 mixed-effects modeling (LME). Because LME is the most commonly used approach in  
24 longitudinal MRI studies (Vijayakumar, Mills et al. Accepted), LME estimates in the current study  
25 were included in order to help directly compare our results with those reported in previous  
26 studies. Thus, we aimed to examine the consistency and reproducibility of neurodevelopmental

1 change for subcortical gray matter regions, including the thalamus, caudate, putamen, pallidum,  
2 hippocampus, amygdala, and nucleus accumbens in males and females.

## 3 **2. Materials and Methods**

### 4 **2.1 Participants**

5 This study analyzed data from typically developing youth from three separate cohorts collected  
6 utilizing longitudinal designs at three separate sites in independent research projects: *Pittsburgh*  
7 (PIT; University of Pittsburgh, USA), *NeuroCognitive Development* (NCD; University of Oslo,  
8 Norway), and *Orygen Adolescent Development Study* (OADS; The University of Melbourne,  
9 Australia). Each project was approved by their respective local review board and informed  
10 consent/assent was obtained from parents and children prior to data collection. In order to best  
11 account for within-subject variance, only participants with  $\geq 2$  scans from each cohort were  
12 included in analyses. Details regarding participant recruitment in each project have been  
13 previously described (Yap, Allen et al. 2011, Tamnes, Walhovd et al. 2013, Herting, Gautam et  
14 al. 2014). By study design, all projects enrolled typically developing children and adolescents at  
15 baseline, although OADS over-sampled children at both high and low temperamental risk of  
16 developing psychopathology. Only data passing QC procedures from typically developing youth  
17 were included in the current study. Demographic information and sample distributions for each  
18 sample are presented in Table 1 and Figure 1. For the PIT dataset, 126 participants were  
19 recruited and scanned at baseline, with 20 not completing their follow-up visit, and 33 excluded  
20 due to poor image quality of the MRI (see additional details of QC procedures in section 2.2).  
21 For NCD, 111 participants were recruited and scanned at baseline; 26 were unable to complete  
22 their follow-up visit, and 9 were excluded due to poor image quality. For OADS, 177 participants  
23 completed a baseline visit, of which 45 did not complete any additional follow-up visits, 61 were  
24 excluded due to psychiatric history or medical illness and 4 were excluded due to poor image  
25 quality. The final samples thus included 73 participants from PIT, 76 from NCD, and 67 from  
26 OADS. In total, the present study included 216 participants (110 females) and 467 scans  
27 covering the age range of 8 to 22 years.

## 1 **2.2 Image Acquisition and Analysis**

2 T1-weighted anatomical scans were obtained at the three sites using different MRI scanners  
3 and sequences (see Supplementary Material). At each site, a radiologist reviewed all scans for  
4 incidental findings of gross abnormalities. Image processing, including whole brain  
5 segmentation with automated labeling of different neuroanatomical structures, was performed  
6 using the longitudinal pipeline of FreeSurfer 5.3 (<http://surfer.nmr.mgh.harvard.edu>; (Fischl,  
7 Salat et al. 2002, Reuter, Schmansky et al. 2012). The longitudinal pipeline includes creating an  
8 unbiased within-subject template space and image using inverse consistent registration. Skull  
9 stripping, Talariach transform and atlas registration, and parcellations are initialized in the  
10 common within-subject template, which increases reliability and statistical power. Similar  
11 standard QC procedures were carried out between sites. QC details were as follows: 1) all raw  
12 images were visually inspected for motion prior to processing, 2) post-processed images were  
13 visually inspected by trained operators for accuracy of subcortical segmentation by the  
14 longitudinal pipeline for each scan per participant, 3) images with inaccurate segmentation were  
15 excluded (number of participants excluded during QC is outlined above in section 2.1). No  
16 manual edits were made to subcortical regions of interests. Regions of interest for the present  
17 study included the thalamus, caudate, putamen, pallidum, amygdala, hippocampus, and  
18 nucleus accumbens for each hemisphere.

## 19 **2.3 Statistical Analyses**

20 Given previous findings highlighting hemispheric differences (Dennison, Whittle et al. 2013,  
21 Herting, Gautam et al. 2014), we first examined if patterns of change differed by hemisphere by  
22 plotting LOESS (locally weighted scatterplot smoothing) curves to each dataset. Overall,  
23 trajectories were similar between hemispheres (see Supplementary Material SFigures 1-7), and  
24 therefore left and right hemisphere volume estimates were averaged for all subsequent  
25 analyses. Given that one of the primary aims of the study was to examine the distinct  
26 developmental trajectories in males and females, and preliminary exploratory LOESS plots



1 confirmed the shape of developmental trajectories varied between males and females (see  
2 Supplementary Material SFigures 1-7).

3 To more fully understand sex differences in subcortical volume changes from late  
4 childhood and throughout adolescence, analyses were performed to examine age, sex, and age  
5 by sex relationships both together across sites, as well as in each sample separately, using  
6 mixed-effects modeling with both generalized additive models (GAMM) (mgcv package version  
7 1.8-17) and general linear mixed effects modeling (LME) (R version 3.4.0; nlme package  
8 version 3.1-131). Follow-up analyses were also conducted to examine age trajectories in each  
9 sex separately, both based on the 3 samples as well as in each sample separately.

### 10 2.3.1 GAMM

11 Unlike parametric general linear modeling, GAMM does not require *a priori* knowledge of the  
12 functional form for the data and is an extension of LME; rather GAMM replaces one or more of  
13 the linear predictor terms with a 'smooth' function term. The non-linear smooth function  
14 describes the best relationship between the covariate(s) and the outcome variable of interest.  
15 For GAMM analyses on subcortical structure volume, the main predictor was age. GAMM can  
16 be represented by the following formula:

$$G(y) = X^* \alpha + \sum_{j=1}^p f_j(x_j) + Zb + \varepsilon$$

17 where  $G(y)$  is a monotonic differentiable link function,  $\alpha$  is the vector of regression coefficients  
18 for the fixed parameters;  $X^*$  is the fixed-effects matrix;  $f_j$  is the smooth function of the covariate  
19  $x_j$ ;  $Z$  is the random-effects model matrix;  $b$  is the vector of random-effects coefficients; and  $\varepsilon$  is  
20 the residual error vector.

21 Using this approach, GAMM models were implemented to examine age, sex, as well as  
22 an age\*sex interaction to test a sex difference in the intercept (main effect of sex) as well as a  
23 sex difference in the trajectory or slope of age (age\*sex interaction), respectively; while also

1 controlling for sample at the level of intercept (main effect of sample) and slope (age\*sample  
2 term). Importantly, sex was coded as a factor (male=0, female=1), allowing for each term to  
3 reflect the following: sex term reflected the difference in intercept in females as compared to  
4 males; age term reflected the slope of age for males; age\*sex term reflected the difference in  
5 slope of females as compared to males. To better understand significant differences in age  
6 trajectories between the sexes, GAMM estimates for age (controlling for sample) were also  
7 implemented in each sample separately. Lastly, sample was converted to an ordered factor and  
8 GAMM models were implemented to directly test significant differences in the slopes of age  
9 between each sample for each region of interest. Thus, GAMM age, sex, and age\*sex models  
10 were updated to use the previous covariates of sample and sample\*age as contrasting factors.  
11 Sample was coded as a factor and two models were implemented to test sample differences:  
12 one model included sample as a factor with OADS=1, NCD=2, and PIT=3 (in order to compare  
13 OADS vs. NCD and OADS vs. PIT), and the second model included sample as a factor with  
14 NCD=1, OADS=2, and PIT=3 (in order to compare NCD vs. PIT).

### 15 2.3.2 LME

16 LME estimates the fixed effect of measured variables on subcortical volume while including  
17 within-person variation as nested random effects in the regression model. This is done to  
18 account for individual subject effects and correlation of the data inherent to longitudinal analysis.  
19 LME can be represented by the following formula for linear changes with age both between and  
20 within participants:

$$Volume_{ij} = Intercept_{0i} + \alpha_{1i}(age)_{ij} + \varepsilon$$

21 where  $Volume_{ij}$  represents the volume in an ROI at the  $j^{\text{th}}$  timepoint for the  $i^{\text{th}}$  participant, the  
22  $intercept_{0i}$  represents the grand mean at the centered age (age 15),  $\alpha_{1i}$  is the grand mean slope  
23 of age (linear); and  $\varepsilon$  is the residual error and reflects within-person variance. All models also  
24 included a random intercept for each participant. The linear model was then built upon to also

1 include quadratic and cubic fixed terms to assess linear versus more complex patterns of  
2 change. The linear, quadratic, and cubic models were as follows:

3 1. *Linear model: Volume = Intercept +  $\alpha(\text{age}) + \epsilon$*

4 2. *Quadratic model: Volume = Intercept +  $\alpha(\text{age}) + \beta(\text{age}^2) + \epsilon$*

5 3. *Cubic model: Volume = Intercept +  $\alpha(\text{age}) + \beta(\text{age}^2) + \gamma(\text{age}^3) + \epsilon$*

6 where  $\alpha$ ,  $\beta$ , and  $\gamma$  represent the effects of each fixed term. Likelihood ratio tests and Akaike  
7 Information Criterion (AIC) were used to compare the models and to determine which had the  
8 best fit. All models were tested against a null model that included only the intercept term, but not  
9 the fixed effect of age. The model with the lowest AIC that was also significantly different from  
10 the less complex model as determined by the likelihood ratio test was chosen as the best fit  
11 model (e.g. linear had to have a lower AIC and be significantly different from null; quadratic had  
12 to have a lower AIC and be significantly different from both the null and linear model).

13 Using LME, models of age were implemented on each sample separately to examine  
14 sex differences by including a term for the main effect of sex as well as an age\*sex interaction  
15 to each model to test a sex difference in the intercept (main effect of sex) as well as a sex  
16 difference in the trajectory or slope of age (age\*sex interaction), respectively. In the cases  
17 where polynomial LME best fits were different between males and females, sex difference were  
18 only tested by using the highest polynomial fit. That is, if a linear best fit was detected for  
19 females but a quadratic best fit for males, a quadratic fit was tested between sexes.

## 20 **3. Results**

### 21 **3.1 Description of Developmental Age Trajectories using GAMM**

22 GAMM estimates of developmental trajectories for volume for each region of interest in females  
23 and males based on the three independent samples are presented in Figure 2. GAMM models  
24 included age, sex, as well as an age\*sex interaction to test a sex difference in the intercept  
25 (main effect of sex) as well as a sex difference in the trajectory or slope of age (age\*sex

1 interaction), while covarying for sample (sample and age\*sample). A significant sex difference  
2 was detected for the smoothed slope of age for all seven regions of interest (Table 2). To better  
3 understand these differences, we examined age trajectories in each sex separately, while again  
4 covarying for sample. These results are presented in Table 3, and we below describe these  
5 developmental trajectories for each subcortical structure in females and males.

### 6 3.1.1 Thalamus

7 Overall, females showed smaller thalamus volumes as compared to males across the entire age  
8 range of 8 to 22 years. Moreover, both males and females showed a nonlinear change with age,  
9 with decreases seen during mid-adolescence and into adulthood; however, when tested  
10 separately, the slope for age was only significant in males, and not in females.

### 11 3.1.2 Pallidum

12 Age trajectories for each sex displayed greater divergence in pallidum volumes from ages 8 to  
13 22 years, with males showing larger volumes as compared to females beginning in early  
14 adolescence thru young adulthood. However, when each sex was examined separately, age  
15 trajectories for the pallidum did not reach statistical significance in either females or males  
16 alone.

### 17 3.1.3 Caudate

18 Males and females displayed similar volumes during late childhood and early adolescence,  
19 whereas females had smaller volumes compared to males by young adulthood. When each sex  
20 was examined separately, the sex difference detected in young adulthood was a result of  
21 females showing a decrease in caudate volumes across adolescence, with no significant  
22 change seen in volumes with age in males.

### 23 3.1.4 Putamen

24 Similar to the caudate, sex differences in the putamen volumes also emerged with age, with  
25 greater sex differences seen in later adolescence and young adulthood. When examined

1 separately, putamen volume showed a nonlinear decrease with age in females, whereas  
2 changes in volumes did not reach significance in males.

### 3 3.1.5 Nucleus Accumbens

4 Nucleus accumbens volumes were similar in males and females from late childhood to mid-  
5 adolescence, with sex differences emerging during late adolescence and young adulthood.  
6 Examining the sexes separately revealed a significant decrease in nucleus accumbens volumes  
7 for females, with no changes in volume in males from ages 8 to 22 years.

### 8 3.1.6 Hippocampus

9 Females showed smaller hippocampal volumes by 10 years of age compared to males.  
10 Moreover, females and males both showed significant nonlinear patterns of hippocampal growth  
11 with age, with an emergent divergence between the sexes across adolescence and into young  
12 adulthood. These nonlinear changes with age reached significant in males and females when  
13 examined separately.

### 14 3.1.7 Amygdala

15 Females showed smaller amygdala volumes at age 8 years compared to males, with greater  
16 separation in volumes seen between the sexes with age. Again, these nonlinear changes with  
17 age reached significant in males and females when examined separately.

## 18 **3.2. Testing Between Sample Differences**

19 Spaghetti plots of GAMM estimates of developmental trajectories for volume for each region of  
20 interest in females and males for each of the three independent samples are presented in  
21 Figures 3-5. To directly test sample differences, previous GAMM models were updated to  
22 change the covariates of sample and sample\*age as contrasting factors (sample: OADS=1,  
23 NCD=2, PIT=3). This allows for directly comparing the main effect of sample as well as if the  
24 samples have significantly different trajectories over age. GAMM estimates for each of these  
25 smooth terms are presented in Table 4 and described below.

1 3.2.1 Thalamus

2 The main effect of sample was not significant. However, the OADS and NCD samples showed  
3 significant sample differences in growth trajectories for the thalamus, whereas trajectories were  
4 not significantly different between PIT and OADS or NCD for this region.

5 3.2.2 Pallidum

6 A main effect of sample was seen with OADS having significantly smaller volumes compared to  
7 NCD and PIT at baseline. Sample differences were also seen in age trajectories, with significant  
8 differences noted between each of the three samples (OADS vs. PIT, OADS vs. NCD, and PIT  
9 vs. NCD).

10 3.2.3 Caudate

11 A main effect of sample was seen with PIT having significantly smaller volumes compared to  
12 NCD and OADS. Significant sample differences were seen in age trajectories between OADS  
13 and NCD as well as OADS and PIT samples, whereas trajectories were not significantly  
14 different between PIT and NCD.

15 3.2.4 Putamen

16 A main effect of sample was seen with OADS having significantly larger volumes compared to  
17 NCD and PIT. However, the samples did not have significantly different age trajectories.

18 3.2.5 Nucleus Accumbens

19 A main effect of sample was detected reflecting the largest volumes seen in NCD, followed by  
20 PIT, and OADS (all  $p$ 's  $<0.05$ ). However, the samples did not have significantly different age  
21 trajectories.

22 3.2.6 Hippocampus

23 A main effect of sample included OADS having larger hippocampal volumes as compared to  
24 NCD and PIT. In addition, OADS and PIT samples showed significant sample differences in

1 growth trajectories for the hippocampus, whereas trajectories were not significantly different  
2 between OADS and NCD or PIT and NCD.

### 3 3.2.7 Amygdala

4 A main effect of sample was seen with PIT having significantly larger volumes compared to  
5 NCD, but no significant difference detected between PIT vs. OADS or NCD vs. OADS. The  
6 samples did not significantly differ for age trajectories for the amygdala.

### 7 **3.3 Testing of Developmental Models using LME**

8 Linear, quadratic, and cubic LME were used to determine best fit models for females and males  
9 of each sample. The highest-order polynomial model for each brain region is summarized in  
10 Table 5 (for AIC comparisons, see Supplementary Tables 1-7). LME best fits were also different  
11 between samples in both sexes for most ROIs, except for the caudate and nucleus accumbens.  
12 For the caudate, both males and females in each sample showed similar trajectories with age  
13 (PIT: linear, NCD: quadratic, and OADS: linear). For the nucleus accumbens, no significant  
14 change was found in all samples, except for OADS females, which showed a linear decrease  
15 with age. Overall, LME best fit models per sample were largely in agreement with the GAMM  
16 trajectories; the exception to this are highlighted in Supplementary Table 8 and include the  
17 pallidum for NCD females (LME=cubic, GAMM=n.s.), the hippocampus for PIT females  
18 (LME=cubic, GAMM=n.s.), the amygdala for PIT females (LME=quadratic, GAMM=n.s.) and the  
19 hippocampus for NCD males (LME=n.s., GAMM=nonlinear). Full model details for LME results  
20 when testing sex, age, and age\*sex using linear and polynomial LME best fit models are  
21 presented in Supplementary Tables 10-16. Using LME, a few models that were identified as  
22 significant using GAMM did not reach significance using best fit models including the pallidum  
23 and putamen in the OADS sample and the putamen in the NCD sample (as shown in  
24 Supplementary Table 17).

### 25 **4. Discussion**

1           This is the first study to examine longitudinal subcortical neurodevelopmental trajectories  
2 in males and females using a multisample approach spanning ages 8 to 22 years. The current  
3 study is an extension from an on-going international collaboration project aiming to improve the  
4 understanding, reliability, and efficiency of neurodevelopmental research by simultaneously  
5 analyzing multiple existing longitudinal neuroimaging datasets (Mills, Goddings et al. 2016,  
6 Tamnes, Herting et al. 2017). By utilizing the identical longitudinal preprocessing pipeline, QC  
7 procedures, and statistical methods across samples, we aimed to shed light on the potential  
8 impact of sample and acquisition differences on our ability to detect sex differences in  
9 subcortical developmental patterns. While sex differences in patterns of subcortical  
10 development across adolescence were identified using all three datasets, divergent results were  
11 also seen for both within-sex and between-sex differences when comparing estimates from  
12 each of the independent datasets. Below we describe the findings using all samples as well as  
13 the differences detected between samples, as well as highlight the additional factors that may  
14 continue to contribute to mixed findings in our understanding of sex differences in subcortical  
15 neurodevelopment.

#### 16 **4.1 Patterns of Age Related Changes in Males and Females**

17 GAMM estimates highlight an overall sex difference in patterns of development of subcortical  
18 volumes based on data from three independent samples (Figure 2). Significant non-linear  
19 changes were seen with age in the thalamus, curvilinear growth of the pallidum and amygdala,  
20 and decreases in the caudate, putamen, and nucleus accumbens. The average trajectories from  
21 our longitudinal datasets are largely in agreement with previous research that sensory, motor,  
22 and cognitive related subcortical regions, such as the caudate and the thalamus, undergo  
23 reduction into young adulthood (Lenroot, Gogtay et al. 2007, Raznahan, Shaw et al. 2014), as  
24 well as increases in amygdala volumes (Goddings, Mills et al. 2014).

25           Estimated sex differences across all participants confirmed previous findings of overall  
26 smaller volumes in females compared to males in all subcortical regions examined in the



1 present study. In addition, significant sex differences were detected for changes with age for all  
2 regions of interest (Table 2). Of these results, replication across all three independent samples  
3 was relatively poor, as assessed by statistical results comparing age trajectory GAMM  
4 estimates. In fact, when directly testing between sample differences in age trajectories, only the  
5 putamen, nucleus accumbens, and amygdala showed no significant differences in trajectories of  
6 age development between the samples (Table 4). These findings may suggest greater  
7 generalizability of the sex differences in curvilinear amygdala growth across childhood and  
8 adolescence, with males showing significantly steeper increases compared to females. In  
9 addition, across samples, females displayed decreases in nucleus accumbens and putamen  
10 volumes, whereas males showed no change (nucleus accumbens) or less change (putamen)  
11 with age from 8 to 22 years.

## 12 **4.2 LME versus GAMM Modeling**

13 LME is perhaps the most commonly used statistical approaches to determining both between  
14 and within-person changes in longitudinal neuroimaging studies (Vijayakumar, Mills et al.  
15 Accepted). At the outset of the study, it was assumed that using LME might therefore allow for a  
16 more direct comparison of our results and previous studies. However, previous studies have  
17 also shown that the shape of growth trajectories can vary when examining each sex separately  
18 (Goddings, Mills et al. 2014). When using LME, this creates a challenge because in the case  
19 where the shape of the trajectory may differ between groups, putting both males and females in  
20 the same model may incorrectly assume similar shapes in growth in both sexes. For these  
21 reasons, GAMM may allow for a more flexible fit, given that it does not assume the curve to the  
22 data at the time of fitting the model. For these reasons, we chose to examine each sex  
23 separately as well as together using both LME and GAMM. Overall, strong similarities were  
24 seen in the ability for GAMM and LME modeling strategies to detect significant age-related  
25 changes in each sex separately across the three independent samples. However, when testing  
26 significant differences in changes in volumes with age between the sexes, GAMM identified  
27 changes in the pallidum for NCD and OADS and the putamen for NCD as significantly different

1 between males and females, whereas LME models had  $p$ 's > 0.05. Thus, GAMM models may be  
2 able to help reframe and bring additional clarity in understanding group differences in patterns of  
3 neurodevelopment, especially when there are presumed sex differences in the shape of  
4 trajectories in males versus females.

### 5 **4.3 Sample Consistencies and Differences**

6 Despite our best efforts to minimize between sample effects by utilizing similar  
7 preprocessing and analytic techniques, both within-sex and between-sex trajectories of  
8 neurodevelopment were nevertheless significantly different between samples. This may suggest  
9 that factors such as population differences, sampling strategy, scanning protocols, as well as  
10 age-range and statistical power may influence best model fits in longitudinal studies of  
11 subcortical development. Because studies often differ in their age-ranges, scan intervals, and  
12 sample size, the conclusion that a particular region shows a linear, quadratic, or even cubic  
13 developmental pattern will not necessarily generalize to another study. For example, the current  
14 LME results show both linear and quadratic best fits for caudate volumes in males and females,  
15 and in contrast the putamen showed linear and cubic fits in females, but linear and quadratic for  
16 males. Similarly, previous single-sample longitudinal studies have also reported linear  
17 decreases in caudate and putamen volumes in both sexes (7-24 years;  $n=223$  scans from 147  
18 individuals; (Wierenga, Langen et al. 2014)); but also no change for caudate and quadratic for  
19 putamen volumes from 5 to 27 years ( $n=175$  scans from 84 individuals; (Narvacan, Treit et al.  
20 2017)) or quadratic for females and cubic for males from 3 to 26 years ( $n=829$  scans from 387  
21 individuals (Lenroot, Gogtay et al. 2007)). This poses a challenge for the field, especially given  
22 that we are rarely interested in exact ages, but rather periods of development across the life  
23 span. Furthermore, while GAMM models may provide greater precision in the description of  
24 volumetric change by moving away from more traditional polynomial assumptions in  
25 neurodevelopmental growth trajectories, significant differences were also found in the current  
26 study when using these models to examine developmental changes in subcortical volumes with  
27 age across the included samples. Given how sensitive these analyses can be to sample

1 differences, it may be more useful as a field to focus on patterns of change (i.e. periods of  
2 relative stability/change and direction of change, as opposed to using model terms) when trying  
3 to understand overall developmental patterns, as well as providing access to statistical code in  
4 order to allow for directly testing prediction accuracy of previously published models on new  
5 datasets.

6 Besides inherent study population and sample differences, power is likely an issue when  
7 examining each sample separately (N's ranging from 67-76 per study versus N=216 together).  
8 Interestingly, despite being able to better account for within-subject variability, these longitudinal  
9 findings are in agreement with recent reports that sample composition can alter age  
10 associations in large cross-sectional study designs (LeWinn, Sheridan et al. 2017).  
11 Furthermore, when examining sex differences in each sample separately (Supplementary Table  
12 9), OADS was also found to show significant sex differences for each region of interest,  
13 whereas the other two samples were more variable. The ability for OADS to detect similar sex  
14 differences as seen with the larger combined sample may in fact be due to better within-subject  
15 estimates due to three waves of data collection as compared to the two waves design used for  
16 NCD and PIT.

17 Although using a commonly employed longitudinal preprocessing pipeline stream, the  
18 degree to which automated segmentation programs may contribute to the seen sample  
19 differences remains a concern. Given that manual tracing requires availability of multiple highly  
20 trained raters without intra- and inter-rater drift over time, manual tracing becomes exceeding  
21 time intensive for even medium scaled longitudinal studies that span multiple years, such as  
22 those included here (PIT=146, NCD=152, and OADS=169 scans). For these same reasons,  
23 poor segmentations are often excluded from the analyses rather than performing manual edits  
24 to the FreeSurfer subcortical volume segmentation (e.g. aseg) (as done in the current study).  
25 Thus, large scaled studies often implement automated software and in the current study we  
26 implemented the FreeSurfer longitudinal pipeline given that it was specifically created to better  
27 capture within-subject changes over time in the subcortical regions examined (as shown by

1 intraclass correlation coefficients ranging from .90 for the left amygdala to .99 for the right  
2 caudate and putamen; note that nucleus accumbens volumes were not included in this report)  
3 (Reuter, Schmansky et al. 2012). While to our knowledge, no study has been published  
4 comparing the longitudinal pipeline estimates with manual tracing, cross-sectional studies have  
5 found that automated software tend to overestimate subcortical volumes as compare to manual  
6 tracings (Schoemaker, Buss et al. 2016, Makowski, Beland et al. 2017). Moreover, for  
7 longitudinal studies if an over estimation in volumes consistently occurs at both the between-  
8 and within-subject levels, relative differences are likely to still be meaningful (Chepkoech,  
9 Walhovd et al. 2016). However, variation in volume estimates between scanners types and  
10 acquisition parameters is likely more problematic, especially if this interacts with age. In this  
11 regard, cross-sectional studies have reported the pallidum to have poor reliability out of  
12 subcortical volume measurements using FreeSurfer, with volume estimates for this region  
13 impacted by MP-RAGE acquisition parameters, such as isotropic versus anisotropic voxel size  
14 (Wonderlick, Ziegler et al. 2009). Furthermore, low reliability of pallidum volumes, specifically,  
15 have been attributed to the T1-weighted contrast profile of the pallidum which is less distinct  
16 from its surrounding white matter as compared to other subcortical regions such as the  
17 thalamus or caudate (Fischl, Salat et al. 2002, Wonderlick, Ziegler et al. 2009). With dedication  
18 to automated software continuing to improve (e.g. FreeSurfer 6.0 was released mid-way through  
19 the current project), future studies will benefit from reductions in such potential software  
20 confounds.

#### 21 **4.4 Limitations**

22 Recent studies document the impact of motion on structural measures (Reuter, Tisdall et al.  
23 2015, Alexander-Bloch, Clasen et al. 2016, Ducharme, Albaugh et al. 2016), and this likely  
24 represents an especially important confound for developmental studies. We therefore conducted  
25 detailed QC of all raw and processed images and excluded participants with excessive motion.  
26 Nonetheless, future studies could benefit from employing standardized and well-documented  
27 QC procedures (Backhausen, Herting et al. 2016), and/or methods for tracking in-scanner

1 motion, automated QC assessment, and motion correction procedures (further discussed in  
2 (Vijayakumar, Mills et al. Accepted)). Previous studies suggest that within-subject changes in  
3 puberty (both physical and hormonal) are important factors for amygdala growth in male and  
4 female adolescents (Goddings, Mills et al. 2014, Herting, Gautam et al. 2014); with very similar  
5 curvilinear amygdala growth patterns seen as reported here when raw volumes were estimated  
6 based on Tanner stage in males and females separately (Goddings, Mills et al. 2014). Puberty  
7 has also been found to relate to nucleus accumbens volumes (Goddings, Mills et al. 2014),  
8 although the trajectories do not mirror the patterns of volumetric change identified in the current  
9 study. Unfortunately, pubertal metrics were not consistent across the three cohorts; making it  
10 impossible for us to investigate how puberty may contribute to differences in amygdala and  
11 nucleus accumbens trajectories in males and females in the current study. Thus, future  
12 research is warranted to examine the contributions of hormones to sex differences in the  
13 neurodevelopmental trajectories of the amygdala and nucleus accumbens. In addition,  
14 FreeSurfer 6.0 was released mid-way through the current project, after the preprocessing for the  
15 current study was complete. Replication studies are always warranted, and should consider also  
16 examining the potential impact of the longitudinal processing stream of FreeSurfer 5.3 versus  
17 FreeSurfer 6.0, especially given the efforts of this new version on estimating the putamen.

## 18 **5. Conclusions**

19 Across all participants from the three independent samples, sex differences in age trajectories  
20 of volumetric development for the thalamus, pallidum, caudate, putamen, nucleus accumbens,  
21 hippocampus, and amygdala were apparent. Using a multisite approach with consistent  
22 longitudinal preprocessing (software and quality checking) and statistical analyses,  
23 generalizable patterns were found for age changes across adolescence in the amygdala,  
24 putamen, and nucleus accumbens. However, conspicuous sample differences were seen for the  
25 thalamus, pallidum, caudate, and hippocampus; perhaps a cautionary limitation when  
26 attempting to generalize subcortical findings from these regions of interests in longitudinal  
27 samples with different age ranges. Efforts aimed at improving our ability to replicate trajectories

1 in typical development, such as the current study, are ultimately necessary in order to be able to  
2 further focus our inquiry on the factors influencing sex differences and individual differences in  
3 subcortical growth; including genetic, and/or environmental effects that may contribute to the  
4 observed differences at the group and individual-level. Furthermore, improving our ability to  
5 assess the 'age residual' of within-subject changes in deep gray matter structures is crucial in  
6 our ability to understand risk and resilience for psychopathology during development.

7

8

1    **Acknowledgments:** The authors thank all participating individuals and families in these  
2    longitudinal studies.  
3  
4    **Funding:** This study was supported by the National Institutes of Health [R01 DA018910 (RED),  
5    R01 HD053893 and R01 MH087563 (ERS), F32 HD0708084 (MMH), K01 MH1087610 (MMH),  
6    R01 MH107418 (NV)]; the Research Council of Norway and the University of Oslo [FRIMEDBIO  
7    230345 to CKT]; the Colonial Foundation, the National Health and Medical Research Council  
8    [NHMRC; Australia; Program Grant 350241; Career Development Fellowship 1007716 (SW)],  
9    the Australian Research Council [ARC; Discovery Grant DP0878136].

## 1   **References**

- 2
- 3   Alexander-Bloch, A., L. Clasen, M. Stockman, L. Ronan, F. Lalonde, J. Giedd and A. Raznahan  
4   (2016). "Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo  
5   MRI." Hum Brain Mapp **37**(7): 2385-2397.
- 6   Backhausen, L. L., M. M. Herting, J. Buse, V. Roessner, M. N. Smolka and N. C. Vetter (2016).  
7   "Quality Control of Structural MRI Images Applied Using FreeSurfer-A Hands-On Workflow to  
8   Rate Motion Artifacts." Front Neurosci **10**: 558.
- 9   Chakravarty, M. M., P. Steadman, M. C. van Eede, R. D. Calcott, V. Gu, P. Shaw, A. Raznahan,  
10   D. L. Collins and J. P. Lerch (2013). "Performing label-fusion-based segmentation using multiple  
11   automatically generated templates." Hum Brain Mapp **34**(10): 2635-2654.
- 12   Chepkoech, J. L., K. B. Walhovd, H. Grydeland, A. M. Fjell and I. Alzheimer's Disease  
13   Neuroimaging (2016). "Effects of change in FreeSurfer version on classification accuracy of  
14   patients with Alzheimer's disease and mild cognitive impairment." Hum Brain Mapp **37**(5): 1831-  
15   1841.
- 16   Choudhury, S., S. J. Blakemore and T. Charman (2006). "Social cognitive development during  
17   adolescence." Soc Cogn Affect Neurosci **1**(3): 165-174.
- 18   Crone, E. A. and B. M. Elzinga (2015). "Changing brains: how longitudinal functional magnetic  
19   resonance imaging studies can inform us about cognitive and social-affective growth  
20   trajectories." Wiley Interdiscip Rev Cogn Sci **6**(1): 53-63.
- 21   Dennison, M., S. Whittle, M. Yucel, N. Vijayakumar, A. Kline, J. Simmons and N. B. Allen  
22   (2013). "Mapping subcortical brain maturation during adolescence: evidence of hemisphere-  
23   and sex-specific longitudinal changes." Dev Sci **16**(5): 772-791.
- 24   Ducharme, S., M. D. Albaugh, T. V. Nguyen, J. J. Hudziak, J. M. Mateos-Perez, A. Labbe, A. C.  
25   Evans, S. Karama and G. Brain Development Cooperative (2016). "Trajectories of cortical  
26   thickness maturation in normal brain development--The importance of quality control  
27   procedures." Neuroimage **125**: 267-279.
- 28   Erus, G., H. Battapady, T. D. Satterthwaite, H. Hakonarson, R. E. Gur, C. Davatzikos and R. C.  
29   Gur (2015). "Imaging patterns of brain development and their relationship to cognition." Cereb  
30   Cortex **25**(6): 1676-1684.
- 31   Fischl, B., D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R.  
32   Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen and A. M. Dale (2002).  
33   "Whole brain segmentation: automated labeling of neuroanatomical structures in the human  
34   brain." Neuron **33**(3): 341-355.



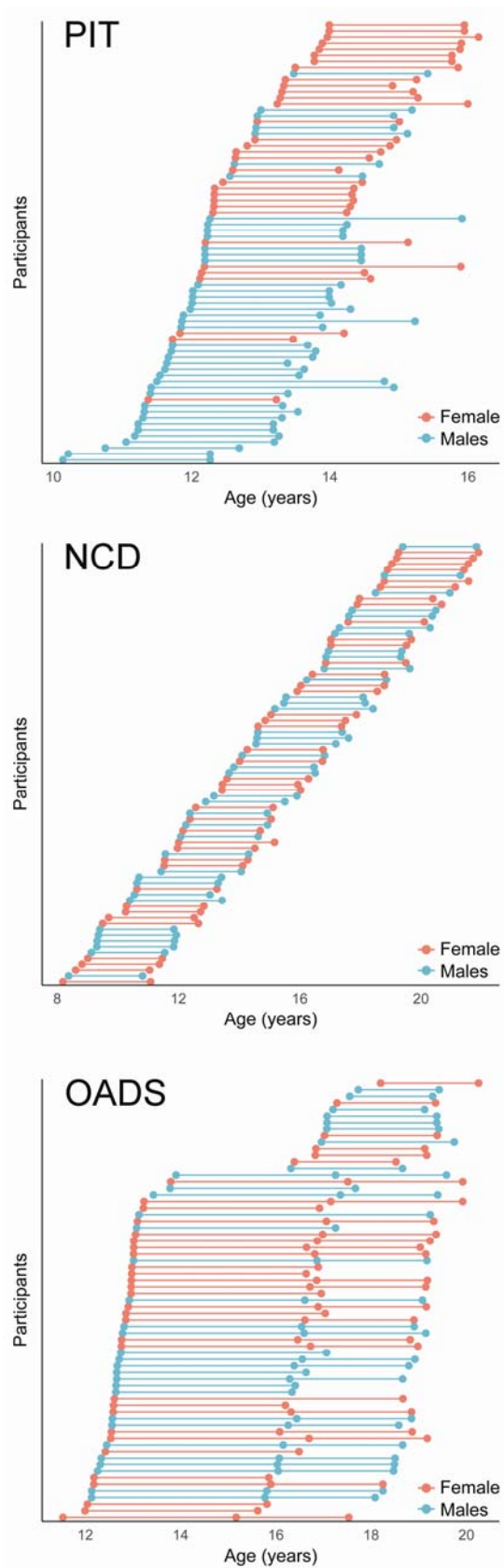
- 1 Giedd, J. N., A. Raznahan, A. Alexander-Bloch, E. Schmitt, N. Gogtay and J. L. Rapoport  
2 (2015). "Child psychiatry branch of the National Institute of Mental Health longitudinal structural  
3 magnetic resonance imaging study of human brain development." Neuropsychopharmacology  
4 **40**(1): 43-49.
- 5 Giedd, J. N., A. C. Vaituzis, S. D. Hamburger, N. Lange, J. C. Rajapakse, D. Kaysen, Y. C.  
6 Vauss and J. L. Rapoport (1996). "Quantitative MRI of the temporal lobe, amygdala, and  
7 hippocampus in normal human development: ages 4-18 years." J Comp Neurol **366**(2): 223-  
8 230.
- 9 Goddings, A. L., K. L. Mills, L. S. Clasen, J. N. Giedd, R. M. Viner and S. J. Blakemore (2014).  
10 "The influence of puberty on subcortical brain development." Neuroimage **88**: 242-251.
- 11 Gogtay, N. and P. M. Thompson (2010). "Mapping gray matter development: implications for  
12 typical development and vulnerability to psychopathology." Brain Cogn **72**(1): 6-15.
- 13 Gur, R. E. and R. C. Gur (2016). "Sex differences in brain and behavior in adolescence:  
14 Findings from the Philadelphia Neurodevelopmental Cohort." Neurosci Biobehav Rev **70**: 159-  
15 170.
- 16 Herting, M. M., P. Gautam, J. M. Spielberg, E. Kan, R. E. Dahl and E. R. Sowell (2014). "The  
17 role of testosterone and estradiol in brain volume changes across adolescence: a longitudinal  
18 structural MRI study." Hum Brain Mapp **35**(11): 5633-5645.
- 19 Jones, K. and S. Almond (1992). "Moving out of the Linear Rut: The Possibilities of Generalized  
20 Additive Models." Transactions of the Institute of British Geographers **17**(4): 434-447.
- 21 Kessler, R. C., P. Berglund, O. Demler, R. Jin, K. R. Merikangas and E. E. Walters (2005).  
22 "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National  
23 Comorbidity Survey Replication." Arch Gen Psychiatry **62**(6): 593-602.
- 24 Kessler, R. C., K. A. McGonagle, M. Swartz, D. G. Blazer and C. B. Nelson (1993). "Sex and  
25 depression in the National Comorbidity Survey. I: Lifetime prevalence, chronicity and  
26 recurrence." J Affect Disord **29**(2-3): 85-96.
- 27 Koolschijn, P. C. and E. A. Crone (2013). "Sex differences and structural brain maturation from  
28 childhood to early adulthood." Dev Cogn Neurosci **5**: 106-118.
- 29 Kuhn, C. (2015). "Emergence of sex differences in the development of substance use and  
30 abuse during adolescence." Pharmacol Ther **153**: 55-78.
- 31 Lenroot, R. K., N. Gogtay, D. K. Greenstein, E. M. Wells, G. L. Wallace, L. S. Clasen, J. D.  
32 Blumenthal, J. Lerch, A. P. Zijdenbos, A. C. Evans, P. M. Thompson and J. N. Giedd (2007).

- 1 "Sexual dimorphism of brain developmental trajectories during childhood and adolescence."  
2 Neuroimage **36**(4): 1065-1073.
- 3 LeWinn, K. Z., M. A. Sheridan, K. M. Keyes, A. Hamilton and K. A. McLaughlin (2017). "Sample  
4 composition alters associations between age and brain structure." Nat Commun **8**(1): 874.
- 5 Makowski, C., S. Beland, P. Kostopoulos, N. Bhagwat, G. A. Devenyi, A. K. Malla, R. Joober, M.  
6 Lepage and M. M. Chakravarty (2017). "Evaluating accuracy of striatal, pallidal, and thalamic  
7 segmentation methods: Comparing automated approaches to manual delineation." Neuroimage.
- 8 Mills, K. L., A. L. Goddings, M. M. Herting, R. Meuwese, S. J. Blakemore, E. A. Crone, R. E.  
9 Dahl, B. Guroglu, A. Raznahan, E. R. Sowell and C. K. Tamnes (2016). "Structural brain  
10 development between childhood and adulthood: Convergence across four longitudinal  
11 samples." Neuroimage **141**: 273-281.
- 12 Morey, R. A., E. S. Selgrade, H. R. Wagner, 2nd, S. A. Huettel, L. Wang and G. McCarthy  
13 (2010). "Scan-rescan reliability of subcortical brain volumes derived from automated  
14 segmentation." Hum Brain Mapp **31**(11): 1751-1762.
- 15 Narvacan, K., S. Treit, R. Camicioli, W. Martin and C. Beaulieu (2017). "Evolution of deep gray  
16 matter volume across the human lifespan." Hum Brain Mapp **38**(8): 3771-3790.
- 17 Ostby, Y., C. K. Tamnes, A. M. Fjell, L. T. Westlye, P. Due-Tonnessen and K. B. Walhovd  
18 (2009). "Heterogeneity in subcortical brain development: A structural magnetic resonance  
19 imaging study of brain maturation from 8 to 30 years." J Neurosci **29**(38): 11772-11782.
- 20 Paus, T., M. Keshavan and J. N. Giedd (2008). "Why do many psychiatric disorders emerge  
21 during adolescence?" Nat Rev Neurosci **9**(12): 947-957.
- 22 Raznahan, A., P. W. Shaw, J. P. Lerch, L. S. Clasen, D. Greenstein, R. Berman, J. Pipitone, M.  
23 M. Chakravarty and J. N. Giedd (2014). "Longitudinal four-dimensional mapping of subcortical  
24 anatomy in human development." Proc Natl Acad Sci U S A **111**(4): 1592-1597.
- 25 Reardon, P. K., L. Clasen, J. N. Giedd, J. Blumenthal, J. P. Lerch, M. M. Chakravarty and A.  
26 Raznahan (2016). "An Allometric Analysis of Sex and Sex Chromosome Dosage Effects on  
27 Subcortical Anatomy in Humans." J Neurosci **36**(8): 2438-2448.
- 28 Reuter, M., N. J. Schmansky, H. D. Rosas and B. Fischl (2012). "Within-subject template  
29 estimation for unbiased longitudinal image analysis." Neuroimage **61**(4): 1402-1418.
- 30 Reuter, M., M. D. Tisdall, A. Qureshi, R. L. Buckner, A. J. van der Kouwe and B. Fischl (2015).  
31 "Head motion during MRI acquisition reduces gray matter volume and thickness estimates."  
32 Neuroimage **107**: 107-115.

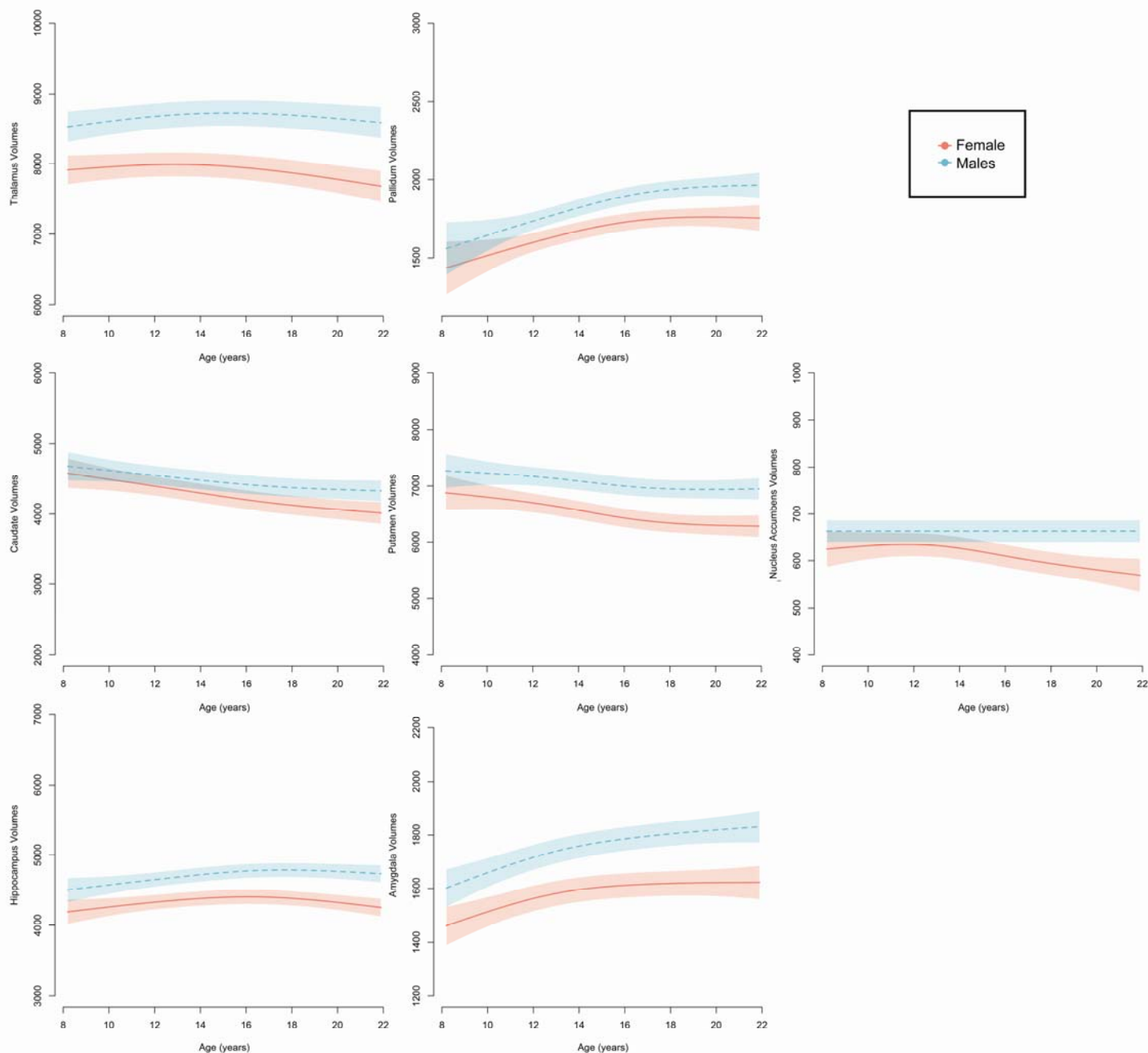
- 1 Roalf, D. R., R. E. Gur, K. Ruparel, M. E. Calkins, T. D. Satterthwaite, W. B. Bilker, H.  
2 Hakonarson, L. J. Harris and R. C. Gur (2014). "Within-individual variability in neurocognitive  
3 performance: age- and sex-related differences in children and youths from ages 8 to 21."  
4 Neuropsychology **28**(4): 506-518.
- 5 Rose, A. J. and K. D. Rudolph (2006). "A review of sex differences in peer relationship  
6 processes: potential trade-offs for the emotional and behavioral development of girls and boys."  
7 Psychol Bull **132**(1): 98-131.
- 8 Ruigrok, A. N., G. Salimi-Khorshidi, M. C. Lai, S. Baron-Cohen, M. V. Lombardo, R. J. Tait and  
9 J. Suckling (2014). "A meta-analysis of sex differences in human brain structure." Neurosci  
10 Biobehav Rev **39**: 34-50.
- 11 Sanfilipo, M. P., R. H. Benedict, R. Zivadinov and R. Bakshi (2004). "Correction for intracranial  
12 volume in analysis of whole brain atrophy in multiple sclerosis: the proportion vs. residual  
13 method." Neuroimage **22**(4): 1732-1743.
- 14 Schoemaker, D., C. Buss, K. Head, C. A. Sandman, E. P. Davis, M. M. Chakravarty, S.  
15 Gauthier and J. C. Pruessner (2016). "Hippocampus and amygdala volumes from magnetic  
16 resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual  
17 segmentation." Neuroimage **129**: 1-14.
- 18 Shaw, P., N. Gogtay and J. Rapoport (2010). "Childhood psychiatric disorders as anomalies in  
19 neurodevelopmental trajectories." Hum Brain Mapp **31**(6): 917-925.
- 20 Sowell, E. R., P. M. Thompson and A. W. Toga (2004). "Mapping changes in the human cortex  
21 throughout the span of life." Neuroscientist **10**(4): 372-392.
- 22 Sowell, E. R., D. A. Trauner, A. Gamst and T. L. Jernigan (2002). "Development of cortical and  
23 subcortical brain structures in childhood and adolescence: a structural MRI study."  
24 Developmental Medicine and Child Neurology **44**(1): 4-16.
- 25 Tamnes, C. K., M. M. Herting, A. L. Goddings, R. Meuwese, S. J. Blakemore, R. E. Dahl, B.  
26 Guroglu, A. Raznahan, E. R. Sowell, E. A. Crone and K. L. Mills (2017). "Development of the  
27 Cerebral Cortex across Adolescence: A Multisample Study of Inter-Related Longitudinal  
28 Changes in Cortical Volume, Surface Area, and Thickness." J Neurosci **37**(12): 3402-3412.
- 29 Tamnes, C. K., K. B. Walhovd, A. M. Dale, Y. Ostby, H. Grydeland, G. Richardson, L. T.  
30 Westlye, J. C. Roddey, D. J. Hagler, Jr., P. Due-Tonnessen, D. Holland, A. M. Fjell and I.  
31 Alzheimer's Disease Neuroimaging (2013). "Brain development and aging: overlapping and  
32 unique patterns of change." Neuroimage **68**: 63-74.
- 33 Tustison, N., A. Holbrook, B. Avants, J. Roberts, P. Cook, Z. Reagh, J. Stone, D. Gillen and Y.  
34 MA (Unpublished). The ANTs Longitudinal Cortical Thickness Pipeline.

- 1 Vijayakumar, N., K. L. Mills, A. Alexander-Bloch, C. K. Tamnes and S. Whittle (Accepted).  
2 "Structural brain development: a review of methodological approaches and best practices."
- 3 Wierenga, L., M. Langen, S. Ambrosino, S. van Dijk, B. Oranje and S. Durston (2014). "Typical  
4 development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24."  
5 Neuroimage **96**: 67-72.
- 6 Wierenga, L. M., J. A. Sexton, P. Laake, J. N. Giedd, C. K. Tamnes and N. the Pediatric  
7 Imaging, and Genetics Study (2017). "A Key Characteristic of Sex Differences in the Developing  
8 Brain: Greater Variability in Brain Structure of Boys than Girls." Cerebral Cortex **1**(11).
- 9 Wonderlick, J. S., D. A. Ziegler, P. Hosseini-Varnamkhasti, J. J. Locascio, A. Bakkour, A. van  
10 der Kouwe, C. Triantafyllou, S. Corkin and B. C. Dickerson (2009). "Reliability of MRI-derived  
11 cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry,  
12 and parallel imaging." Neuroimage **44**(4): 1324-1333.
- 13 Yap, M. B., N. B. Allen, M. O'Shea, P. di Parsia, J. G. Simmons and L. Sheeber (2011). "Early  
14 adolescents' temperament, emotion regulation during mother-child interactions, and depressive  
15 symptomatology." Dev Psychopathol **23**(1): 267-282.
- 16 Zhang, Y., M. Brady and S. Smith (2001). "Segmentation of brain MR images through a hidden  
17 Markov random field model and the expectation-maximization algorithm." IEEE Trans Med  
18 Imaging **20**(1): 45-57.
- 19 Zijdenbos, A. P., R. Forghani and A. C. Evans (2002). "Automatic "pipeline" analysis of 3-D MRI  
20 data for clinical trials: application to multiple sclerosis." IEEE Trans Med Imaging **21**(10): 1280-  
21 1291.
- 22

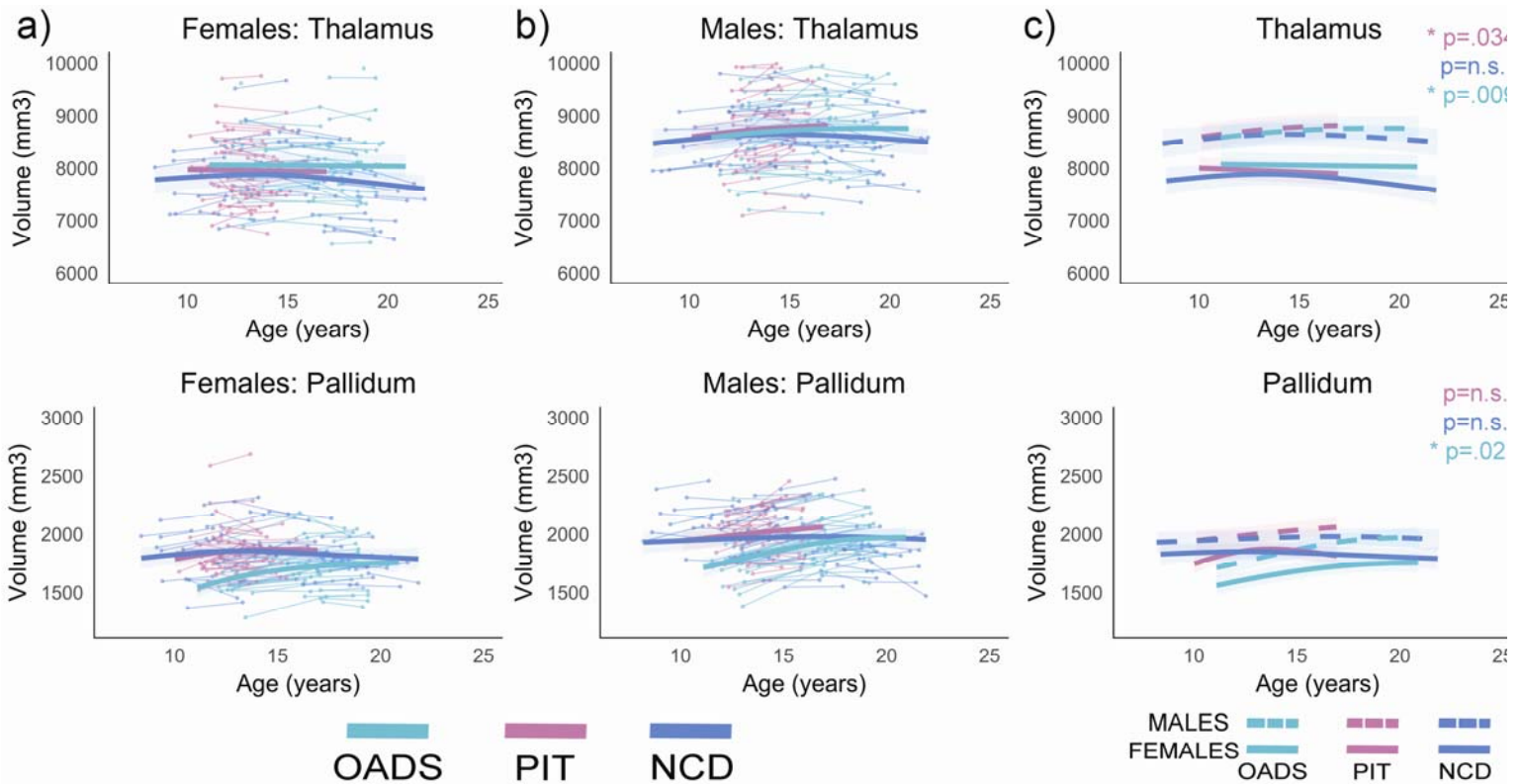
**Figure 1.** Age and sex distributions for each sample.



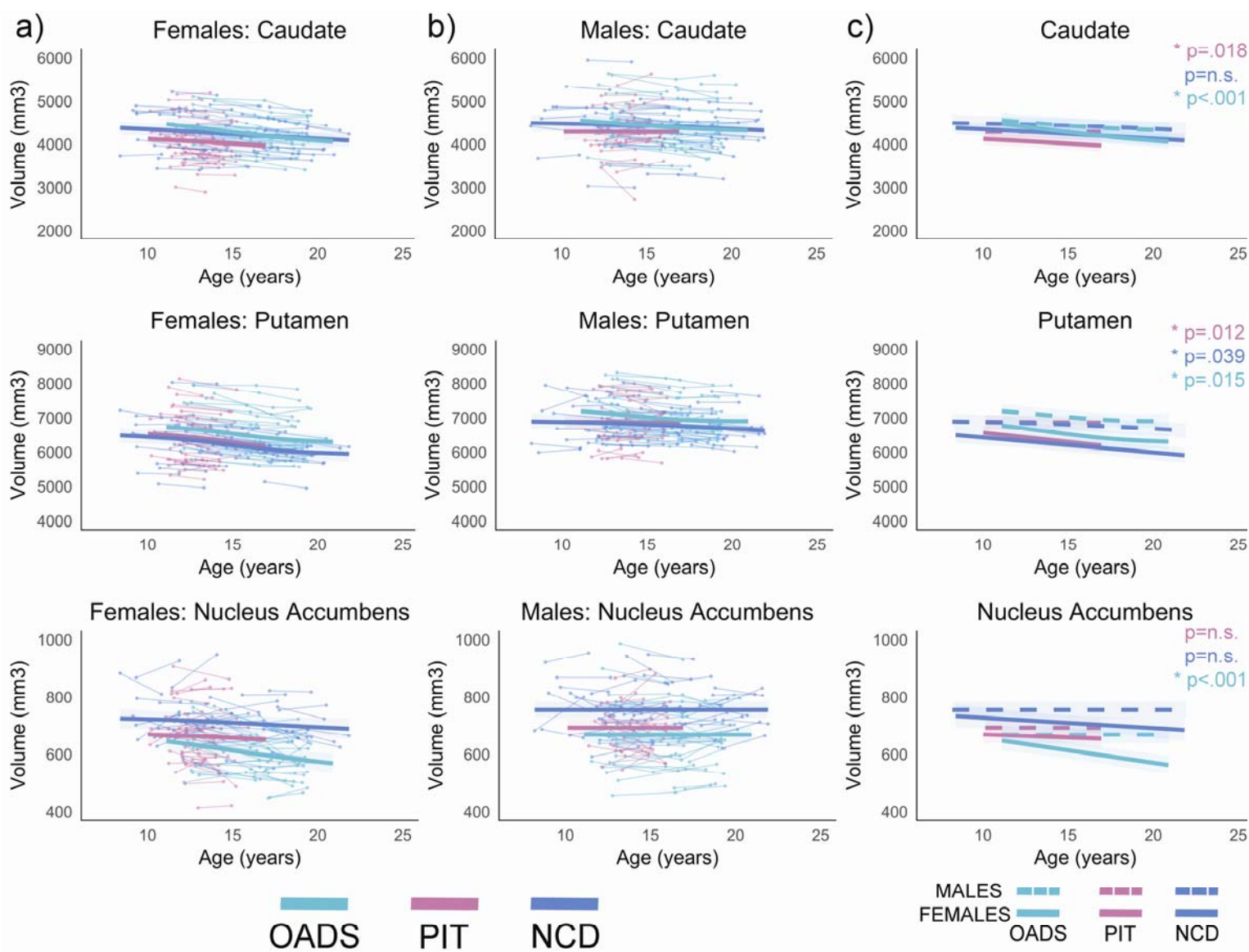
**Figure 2.** Sex differences in the developmental age trajectories for subcortical volumes based on three independent samples.



**Figure 3.** Developmental age trajectories for the thalamus and pallidum. a) Females and b) males are plotted separately. Individual datapoints are shown, connected for each participant, in the appropriate sample color. The bolded colored lines represent the GAMM fitting for each sample with 95% confidence intervals. c) Representation of GAMM fits (with 95% confidence intervals) for each sex per sample plotted together, p-values represent sex differences per sample.



**Figure 4.** Developmental age trajectories for the caudate, putamen, and nucleus accumbens. a) Females and b) males are plotted separately. Individual datapoints are shown, connected for each participant, in the appropriate sample color. The bolded colored lines represent the GAMM fitting for each sample with 95% confidence intervals. c) Representation of GAMM fits (with 95% confidence intervals) for each sex per sample plotted together; p-values represent sex differences per sample.





**Figure 5.** Developmental age trajectories for the hippocampus and amygdala. a) Females and b) males are plotted separately. Individual datapoints are shown, connected for each participant, in the appropriate sample color. The bolded colored lines represent the GAMM fitting for each sample with 95% confidence intervals. c) Representation of GAMM fits (with 95% confidence intervals) for each sex per sample plotted together; p-values represent sex differences per sample.

