Predicting Developmental Norms from Baseline Cortical Thickness in Longitudinal Studies

Philipp Seidel[1,2], Tobias Kaufmann[1,2,3,*], Thomas Wolfers[1,2,*]

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

[2]German Center for Mental Health (DZPG), partner site Tübingen, Germany

[3]Centre for Precision Psychiatry, University of Oslo, Norway

[*]Shared senior authorship

Abstract

Normative models have gained popularity in computational psychiatry for studying individual-level differences relative to population norms in biological data such as brain imaging, where measures like cortical thickness are typically predicted from variables such as age and sex. Nearly all published models to date are based on cross-sectional data, limiting their ability to predict longitudinal change. Here, we used longitudinal brain data from the Adolescent Brain Cognitive Development (ABCD) study, comprising cortical thickness measures from 180 regions per hemisphere in youths at baseline (N=6179; 47% females), 2-year (N=6179; 47% females), and 4-year (N=805; 45% females) follow-up. A training set was established from baseline and 2-year follow-up data (N=5374; 47% females), while data from individuals with all three time points available served as an independent test set (N=805; 45% females). We developed sex-specific Baseline-Integrated Norms (B-Norms) that predict brain region thickness at follow-up based on baseline thickness, baseline age, and follow-up age, and compared them to sex-specific standard Cross-Sectional Norms (C-Norms) based on age alone. Out-of-sample testing in 2-year and 4-year follow-up data showed that B-Norms consistently provided better fits than C-Norms for nearly all cortical regions. Explained variance was higher in B-Norms than in C-Norms. We found no significant differences between time points ($p = 0.45$). Repeated measures ANOVA revealed differences in higher-order moments (e.g., skewness and kurtosis) for both models; for example, skewness varied by model, sex, time point, and their interactions. While improved fit alone does not necessarily indicate a superior normative model - since normative models aim to capture population variance rather than simply optimize fit - we demonstrated that four regions were associated with pubertal changes in B-Norms but not in C-Norms, suggesting enhanced sensitivity of B-Norms to developmental processes. Together, our findings highlight the potential of B-Norms for capturing normative variation in longitudinal structural brain change.

Intro

The predominant analytical approach in neuroimaging for identifying disease-related brain alterations has been case-control comparisons. Over the past decade, this group-level approach has increasingly been complemented by normative modelling - a framework that enables the quantification of individual deviations from expected neurobiological patterns relative to a reference population (A. F. Marquand et al., 2016, 2019). In brief, normative models estimate centiles of variation across neurobiological or behavioral measures, allowing inferences about an atypical development at the individual level without the need for predefined diagnostic groups.

A key strength of normative modelling lies in its flexibility to model diverse mappings across different phenotypic domains, ranging from neuroimaging measures (e.g., cortical thickness/volume) to behavioral or demographic traits (Rutherford, Fraza, et al., 2022). Therefore, this modeling approach is particularly well-suited for lifespan research, where inter-individual variability often reflects subtle and complex developmental or degenerative processes (Bethlehem et al., 2022; Franke & Gaser, 2019; Rutherford, Fraza, et al., 2022; Rutherford et al., 2023). For example, deviations from normative neurodevelopmental trajectories have been implicated in the pathogenesis of psychiatric and neurodevelopmental conditions (Insel, 2014). More recent applications have used normative models to investigate cognitive decline and deterioration or increases of structural brain measures in aging and neurodegenerative disorders (Bethlehem et al., 2022; Di Biase et al., 2023; Frangou et al., 2022; Jack et al., 2010; Karas et al., 2004; Kjelkenes et al., 2022, 2023). These models further provided valuable insights into heterogeneity in psychiatric conditions such as attention-deficit/hyperactivity disorder (Kia & Marquand, 2019; Wolfers et al., 2020), schizophrenia (Kia & Marquand, 2019; Pinaya et al., 2019; Wolfers et al., 2018, 2021), autism spectrum disorder (Bethlehem et al., 2020; Pinaya et al., 2019; Wolfers et al., 2019; Zabihi et al., 2019, 2020), and Alzheimer's/dementia (Alden et al., 2022; Pinaya et al., 2021; Verdi et al., 2021, 2023, 2024).

Despite these advances, most normative modelling studies have relied on cross-sectional data. To some extent, this limits their capacity to accurately predict within-subject longitudinal change. A recent study showed evidence that cross-sectional models may underestimate dynamic brain changes over time (Di Biase et al., 2023; Korbmacher et al., 2024; Vidal-Pineiro et al., 2025). Although recent studies have begun to incorporate and investigate longitudinal data, some of which, however, still apply cross-sectional models to individual data points, rather than leveraging longitudinal information (Gaiser et al., 2024; Janssen et al., 2024; Rehák Bučková et al., 2025; Verdi et al., 2024). Among other disciplines developmental neuroscience has emphasized the need for longitudinal designs and investigations to capture changes during brain maturation - a crucial developmental period for which mounting evidence of vulnerability to psychiatric illness has been put forth (Bayer et al., 2022;

Dehestani et al., 2023; Rogers & De Brito, 2016; Solmi et al., 2022; Whittle et al., 2020). Recent findings suggest that modelling deviations in the timing of pubertal onset - such as early puberty onset - can enhance the prediction of later mental health outcomes (Dehestani et al., 2023, 2024). Such evidence highlights the translational potential of (longitudinal) normative models for early detection and thus possible interventions.

To this end, we estimated Baseline-integrated Norms (B-Norms) that use baseline cortical thickness data and age to predict cortical thickness at later timepoints and compared them to Cross-sectional Norms (C-Norms). Specifically, we used data from the Adolescent Brain Cognitive Development (ABCD) Study (Casey et al., 2018), comprising baseline data from 5374 participants (2515 female; baseline: mean±SD age = 118.78±7.44 months), and follow-up data at 2-year follow-up as well as baseline, 2-year, and 4-year follow-up data for 805 participants (366 female; baseline: mean±SD age = 119.28±7.34; 2-year: mean±SD age = 143.22±7.49; 4-year: mean±SD age=168.45±7.77). Cortical thickness measurements were derived from 360 regions of interest (ROIs) based on the Glasser atlas (Glasser, 2016). For both B- and C-Norm models, we trained them for each brain region separately, hypothesizing that B-Norm models utilizing longitudinal data would show enhanced sensitivity to developmental changes within the respective developmental time period. To validate this, we examined the relationship between model-derived deviation scores and pubertal development as determined by the Pubertal Development Scale (PDS) scores. Validation was performed in the held-out test set. PDS scores acquired at 2-year and 4-year follow-up were statistically associated with the respective model-derived deviation scores of C- and B-Norm models. Our findings aim to contribute to the ongoing refinement of normative modelling approaches, especially in the context of individual-level longitudinal developmental predictions.

Results

Integrating Baseline Thickness Measures as Predictors for Norms in Logitudinal Designs

While C-Norms predict cortical thickness based on age, we here examined if additionally incorporating information on cortical thickness at baseline alongside age at follow-up as features yields more accurate predictions (B-Norms), and if the resulting deviation scores are associated with external variables that are sensitive to change, such as pubertal development. We trained B-Norms and C-Norms models in the same training data set, separately for males and females to account for potential sex specific trajectories. To validate the trained models, we compared explained variance for our models when applied to the 2-year and 4-year follow-up held-out test data. This allowed us to assess the accuracy of fits (i.e., the center of the distribution). Additionally, we assessed differences in higher order moments (i.e., skewness and kurtosis) of the resulting z-scores to investigate the shape and calibration of the centiles of deviation. The latter is important as it ensures that the models accurately fit to the overall distribution and thereby allow for reliable inferences (Marquand et al., 2024). We found marked differences in the mean fit between models. As expected, B-Norms yielded much higher variance explained for both 2-year and 4-year follow up compared to C-Norms (**Figure 1A**). The overall pattern was similar across sexes, with higher variance explained for B-Norms (mean±SD across timepoints, males: 0.635±0.125; females: 0.633±0.123) compared to C-Norms (males: 0.023±0.038; females: 0.026±0.04). On top of these large B-vs-C-Norm differences, we observed a small model-by-timepoint interaction, indicating that the accuracy of mean fits decreased from 2-year to 4-year follow up, with largest effects in B-Norms and in males (model-by-timepoint interaction, males: $F_{1,359}$=25.86, $p<5.91e^{-7}$, η^2=0.0672; females: $F_{1,359}$=26.04, $p<5.43e^{-7}$, η^2=0.0676; post-hoc t-tests males: *C-Norm* mean-diff=0.0038, t=3.268, $p<1.18e^{-3}$, *B-Norm* mean-diff=0.0154, t=7.016, $p<1.14e^{-11}$; post-hoc tests in females: *C-Norm* mean-diff=0.0006, t=0.613, p<0.54; *B-Norm*: mean-diff=0.0129, t=5.618, $p<3.88e^{-8}$). For the shape of the distribution and the centiles of the predicted z-scores, we investigated skewness and excess kurtosis (i.e., kurtosis values below or greater than 0), where values close to zero would indicate a standard normal distribution. On average, we found that skewness was closer to zero for B-Norms than C-Norms, however, variance was larger in B-Norms. We observed larger excess kurtosis for B-Norms as compared to C-Norms. The subpanels in **Figure 1A** depict the distributions and the corresponding statistics for explained variance, skewness, and kurtosis, among other statistics, from our repeated measures ANOVA are detailed in the Supplement (see Figures S6-19 and Tables S1-8). **Figure 1B** illustrates the spatial distribution of the performance metrics, exemplarily for females at 2-year follow-up. In the depicted female B-Norms, we found highest explained variance in large parts of the occipital (left/right PIT, VMV1, and left MT) and temporal regions (right PHT, PH, TP0J2 and left

TP0J1, TE1m and STSda), and lowest explained variance scores in frontal (left/right pOFC, right OFC, 13l, and 25 and left 6d) and insular (left AAIC and PoI2, and left/right FOP3) areas. Similar results were found for the female 4-year follow-up data. Results for male B-Norms, were mostly consistent with females (see supplementary Figures 21-22). Interestingly male C-Norms showed different highest explained variance scores in parietal and frontal areas for the 2-year and 4-year follow-up data. Furthermore, we saw differences between lowest explained variances in male C-Norms for the 2-year and 4-year follow-up data. Specifically, we found different areas of the frontal lobe that showed explained variances below 0 (i.e., the model performs worse than chance) for the 2- and 4-year follow-up data as well as some temporal regions in the 2-year and parietal regions in the 4-year follow-up data. These results were inconsistent with the predictions made by female C-Norms on 2-year follow-up data with largest explained variance in frontal and temporal regions. Supplementary Figures S21 and 22 provide surface maps for the female 4-year and the male 2- and 4-year data. Supplementary file *test_metrics.csv* provides all metrics. Largest and smallest skewness and excess kurtosis also differed between male and female models for both timepoints. These results suggest that there may be sex differences in estimating ROI-wise C- or B-Norms.

ROI-wise comparisons of mean fits for the C-Norms and B-Norms reveal differential effects, that is, some ROI models explained more variance in the 2-year than in the 4-year follow-up or vice versa. For example, in the female longitudinal B-Norm model for a section of the posterior cingulate cortex (i.e., 31pv) of the right hemisphere, we found higher explained variance in the 2-year than in the 4-year follow-up data, whereas for the right PEF (premotor eye field) ROI the model explained more variance in the 4-year follow-up data. The full range of these results is available in the supplements (see supplementary Figures S8, 9, 14, 15, 17, and 18).

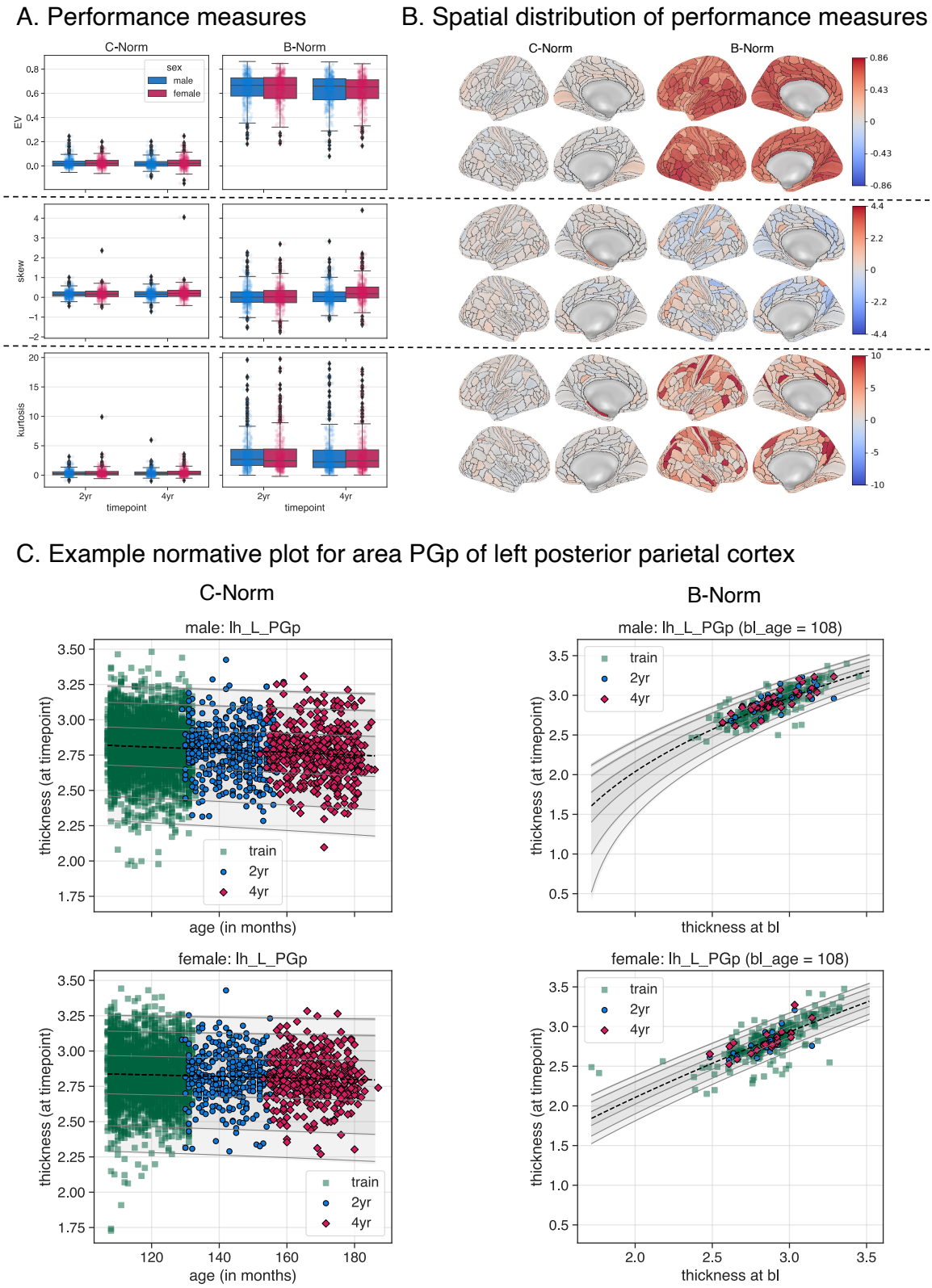


Fig 1. Cross Sectional Norms (C-Norm)- vs Baseline Integrated Norms (B-Norm). A) depicts three performance measures of the different normative models: explained variance (top row), skew (middle row), and kurtosis (bottom row) for both models (columns). B) The first two columns correspond to the lateral and medial views of the C-Norms, the last two to the B-Norms. Warmer colors indicate higher explained variance, positive skewness or kurtosis; colder colors indicate negative skewness or kurtosis. Panel C) We also provide normative plots of an example ROI (here the PGp, as defined by the Glasser atlas). The columns represent normative trajectories for the standard C-Norms (left) and the B-Norms (right). Rows correspond to sex.

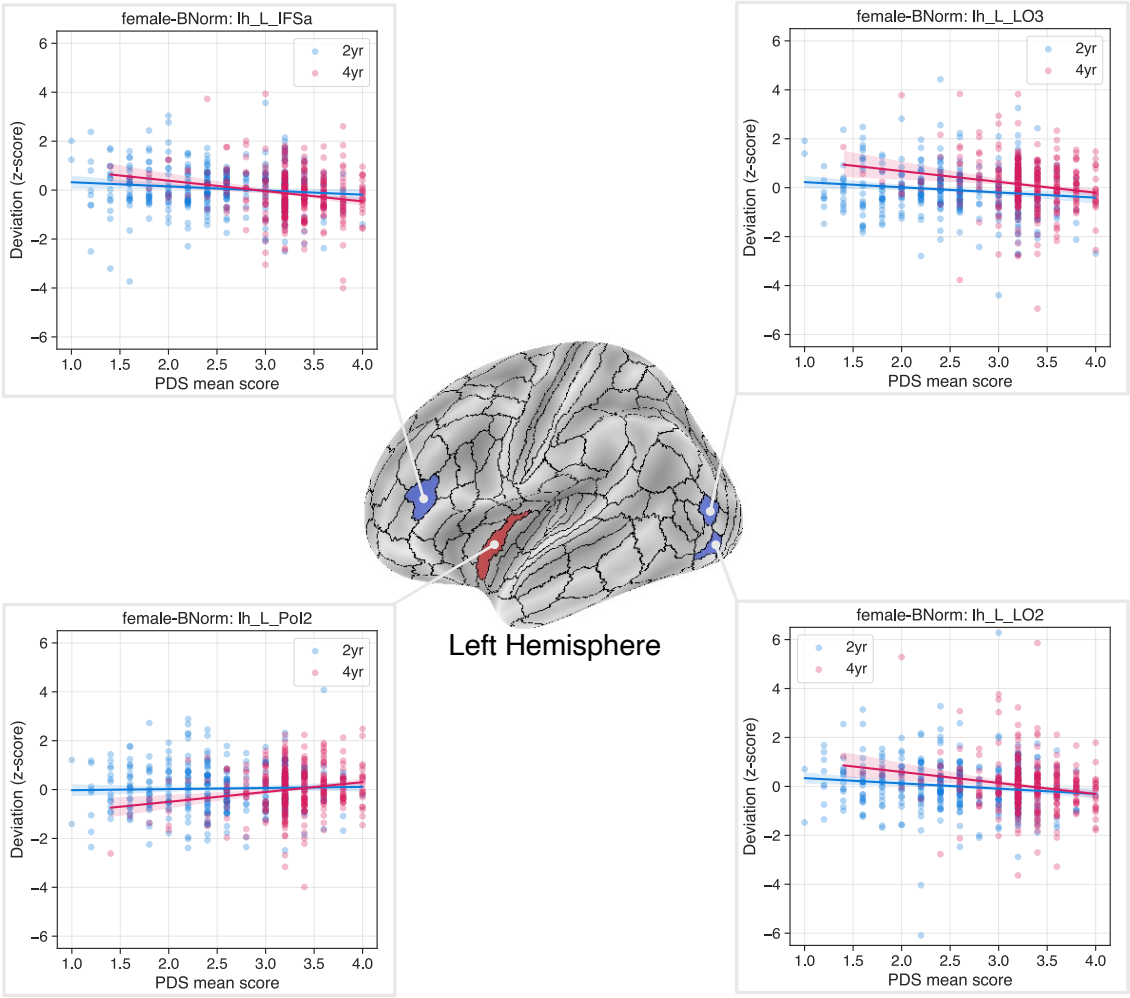
Within each graph, green squares indicate the training data. The black dashed lines indicate the median. The blue circles and red diamonds correspond to the 2-year and 4-year follow-up data, respectively. Grey lines indicate the centiles (1%, 5%, 25%, 75%, 95%, and 99%) and gray patches around these lines indicate the respective uncertainty. Note: given the nature of the B-Norms, we needed to plot one normative plot per baseline age. Only one baseline age (108 months) is depicted here. Others can be found in the supplementary material (See Figure S20).

Association with puberty scores

We investigated whether deviation scores of the 2-year or 4-year follow-up test data were associated with puberty stage as rated by the youths' caregivers. Specifically, for each of the 360 ROIs, for each model and for each sex, we tested for linear associations between pubertal scores (PDS) and the z-scores from B-Norm and C-Norm model, with Benjamini-Hochberg (BH) false discovery correction applied across tests.

For C-Norms, only one association survived correction for multiple comparisons; specifically in 2-year follow-up data for females in the left hemisphere's area 31pv, a part of the left posterior cingulate cortex ($t=-3.985$, $p<0.0244$, BH corrected). For female B-Norms, four areas of the left hemisphere survived correction; specifically for two subsections of the lateral occipital cortex (LO2: $t=-3.597$, $p<0.0362$, BH corrected and LO3: $t=-3.540$, $p<0.0362$, BH corrected), an area in the lateral frontal lobe (IFSa: $t=-3.619$, $p<0.0362$, BH corrected), and an area of the insula (PoI2: $t=3.632$, $p<0.0362$, BH corrected). These results suggest, that with progression through puberty, areas LO2, LO3, and the IFSa exhibit decrease in cortical thickness, whereas parts of the insula appeared to show positive deviation. A corresponding surface map for the significant B-Norm areas is depicted in **Figure 2A** (for the 31pv of C-Norms see supplementary figure S23)**.** Furthermore, the distributions of t-scores across ROIs in Figure 2B suggest that there is an average negative shift for the females in the 2-year data for both the C-Norms and B-Norms. This negative shift for the females is lost for the C-Norms in the 4-year follow-up data but remains for the B-Norms, indicating that such models may better capture a relationship between puberty progression and changes in cortical thickness.

A. Association between deviation- and PDS scores (female longitudal model)



B. Association of z-scores with puberty scores as rated by caregivers

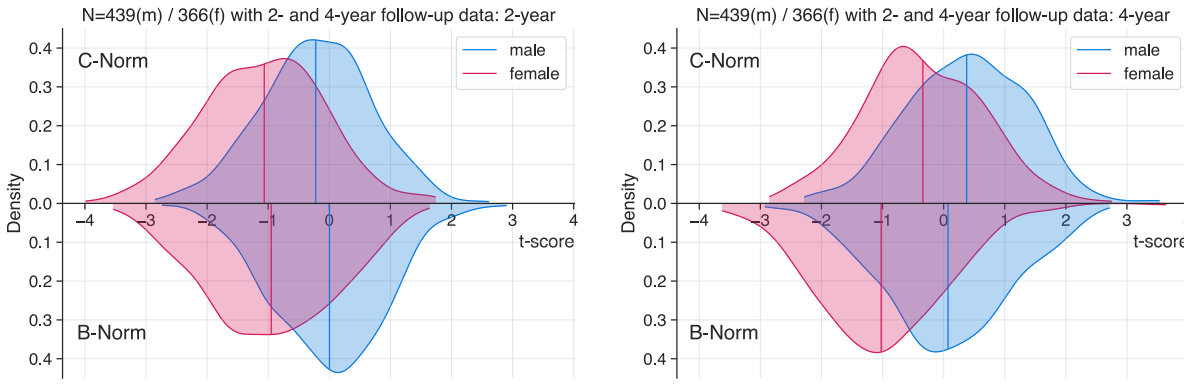


Fig 2. Association of deviation scores with puberty.
Panel A shows a spatial surface map with the statistically significant areas for the female B-Norm models and the respective scatter plots and regression lines. The three areas marked in blue correspond to negative t-scores, whereas the red marked area corresponds to the positive t-score. The blue (2-year) and red (4-year) colored dots within the scatter plots correspond to a single individual. The colored lines reflect the respective regression line of the data; the patches around the lines indicate the 95th confidence interval. The scatter plot for the female C-Norm models can be found in the supplementary material (Fig S23). Panel B shows the overall distributions of the t-scores for this analysis. Colors indicate sex (red = female, blue = male), curves above the x-axis correspond to the associations for the C-Norms, whereas those below correspond to those of the B-Norms.

Validation of normative models using puberty subgroups

We further evaluated whether positive or negative deviation counts, defined as the sum of regions with deviation scores greater than or less than z = ±1.96, were associated with specific pubertal stages. To this end, we grouped participants according to their pubertal stage (pre, early, mid, late and post pubertal, **Figure 3**, panel A) and analyzed deviations at the level of lobes (**Figure 3**, panel B). For females (see **Figure 3**, panel C top), 11 Kruskal-Wallis omnibus tests yielded significant group effects. In cross-sectional normative models (C-Norms), significant group differences were found among negative deviators at the 2-year follow-up in the left occipital ($H(4)=9.96$, $p=0.041$) and parietal ($H(4)=14.38$, $p=0.0062$) lobes; Dunn's post-hoc tests suggested a mid- vs late-pubertal group difference in the parietal lobe ($p=0.0053$) but not in any of the pubertal stages for the occipital lobe (min $p=0.139$ for mid- vs late-pubertal). For positive deviators, we found significant effects in the left cingulate for the 2-year follow-up data ($H(4)=15.33$, $p=0.0041$), with early- vs late- ($p=0.036$) and mid- vs late-pubertal ($p=0.0101$) contrasts, and in the right cingulate for the 4-year follow-up ($H(2)=7.10$, $p=0.0288$), though pairwise comparisons were not significant (min $p=0.0705$ for late- vs post-pubertal).

B-Norms showed significant effects for negative deviators in the right occipital ($H(4)=15.95$, $p=0.0031$; mid- vs late-pubertal $p=0.0031$), right parietal ($H(4)=10.53$, $p=0.0324$; min pairwise $p=0.069$ for early- vs mid-pubertal), and the left cingulate ($H(4)=10.276$, $p<0.0360$, min $p=0.221$ for mid- vs late-pubertal) after two years. At 4-years follow-up, significant overall pubertal stage differences were found for negative deviators in the left ($H(2)=9.07$, $p=0.0108$) and right ($H(2)=8.30$, $p=0.016$) temporal lobes, and the right frontal lobe ($H(2)=8.61$, $p=0.0135$), all driven by mid- vs late-pubertal contrasts ($p=0.095$, $p=0.012$, and $p=0.01$, respectively). Corresponding p-values are visualized in **Figure 3C**; significant cells are hatched in red. For Dunn's post-hoc tests Bonferroni correction was used to adjust p-values for multiple comparisons. These results suggest, that for females, the largest differences between pubertal stages in relation to extreme deviations from the norm can be found between mid- and late pubertal participants.

For males (see **Figure 3C** bottom), we only found three significant Kruskal-Wallis omnibus tests. That is, for the C-Norms, we found significant group differences only the left parietal lobe of negative deviators in the 2-year follow-up data ($H(3)=11.176$, $p<0.0109$, specifically for early- vs mid-pubertal $p=0.0134$). For B-Norms, we found significant effects for positive deviators in the left temporal lobe ($H(3)=8.213$, $p<0.042$, specifically for pre- vs early pubertal $p=0.0481$) in the 2-year follow-up data and in the left occipital lobe ($H(4)=14.364$, $p<0.0062$, specifically for mid- vs late-pubertal $p=0.038$) for 4-year follow-up data.

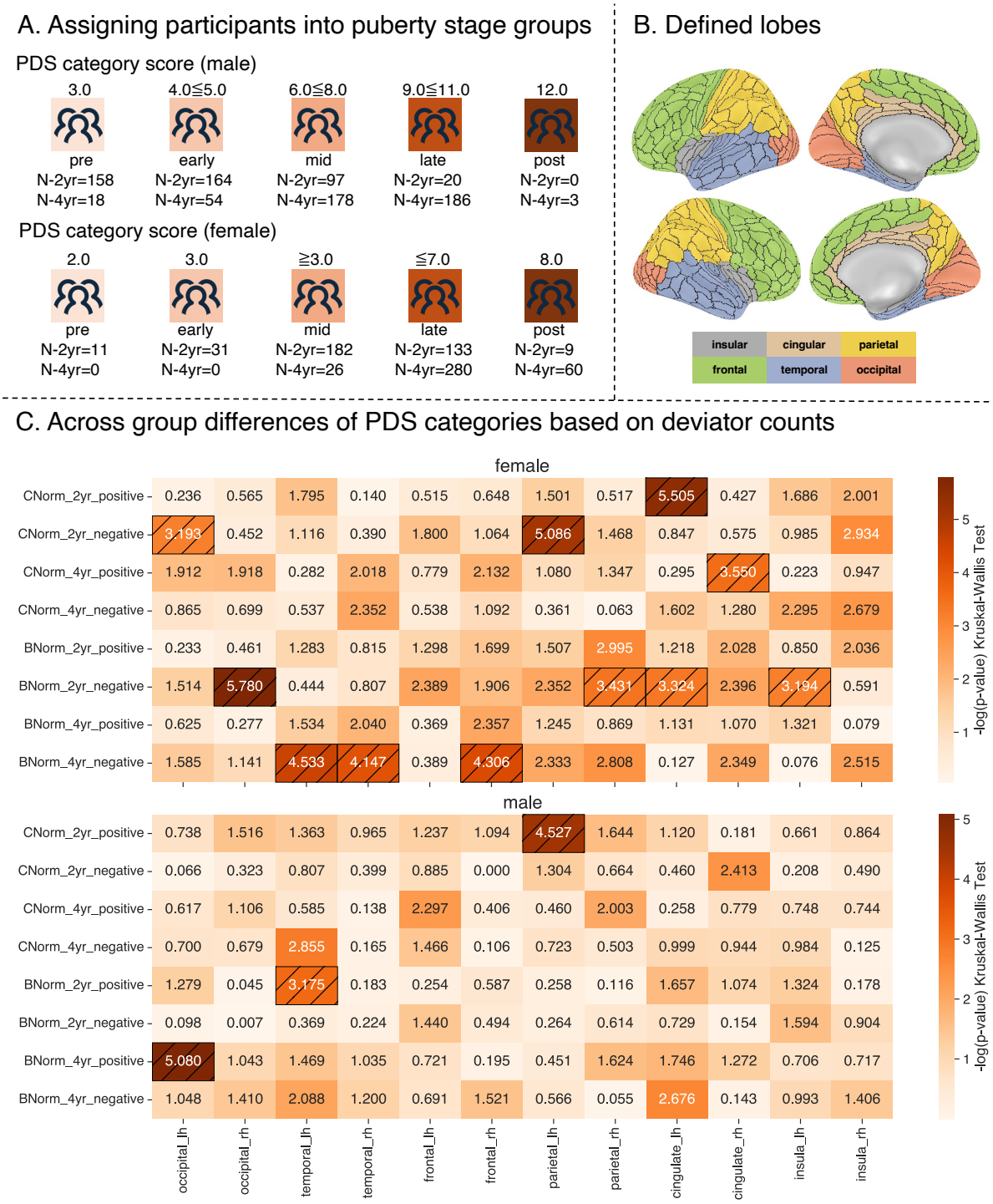


Fig 3. Definition and differences in pubertal stages for lobe-wise positive and negative deviation counts. Panel A) shows the definiton of pubertal stages/categories based on puberty scores (see Herting et al., and Kraft et al.). Panel B) illustrates which regions of the Glasser atlas were combined into "lobe" regions (definition based on (Kaufmann et al., 2019) Panel C) shows -log-transformed p-values according to Kruskal-Wallis omnibus test for differences in positive/negative deviation counts between pubertal categories (pre-, early-, mid-, late-, and post-pubertal) for both timepoints using the C-Norms or B-Norms for female (top) and male (bottom) data.

Validation of normative models using percentile shifts

As a final validation step, similar to longitudinal changes used in pediatrics, we validated our normative models by categorizing participants into three groups based on percentile shifts between the 2-year and 4-year follow-up data: negative (zDiff < -1), stable (-1 < zDiff < 1), and positive (zDiff > 1), with zDiff representing the change in ROI-specific z-scores over time. For each ROI, participants were assigned to one of these groups, and a Kruskal-Wallis test was conducted to assess group differences in delta PDS scores (4-year minus 2-year). No group differences survived multiple comparison correction using the Benjamini-Hochberg procedure (male: min p-uncorrected = 0.001142, p-corrected = 0.2723 for right PFop in the B-Norms; female: min p-uncorrected = 0.001286, p-corrected = 0.4526 for left PHA2 in the C-Norms). Nevertheless, regions showing uncorrected p-values < 0.05 are displayed in **Figure 4** for the female C-Norms (left) and B-Norms (right), along with corresponding boxplots of delta PDS scores for the most notable ROIs (TGv and PHA2 in the C-Norms; p24pr and PHA2 in the B-Norms, all left hemisphere). Of note, due to insufficient sample sizes within deviation groups, Kruskal-Wallis tests could not be performed in the following regions of the female C-Norms: MT, VMV1, and L31a (left hemisphere), and TE1a, PHT, PH, TPOJ1, VMV1, VMV2, and TE1m (right hemisphere). Figure S24 of the supplements shows a similar graph for the male models.

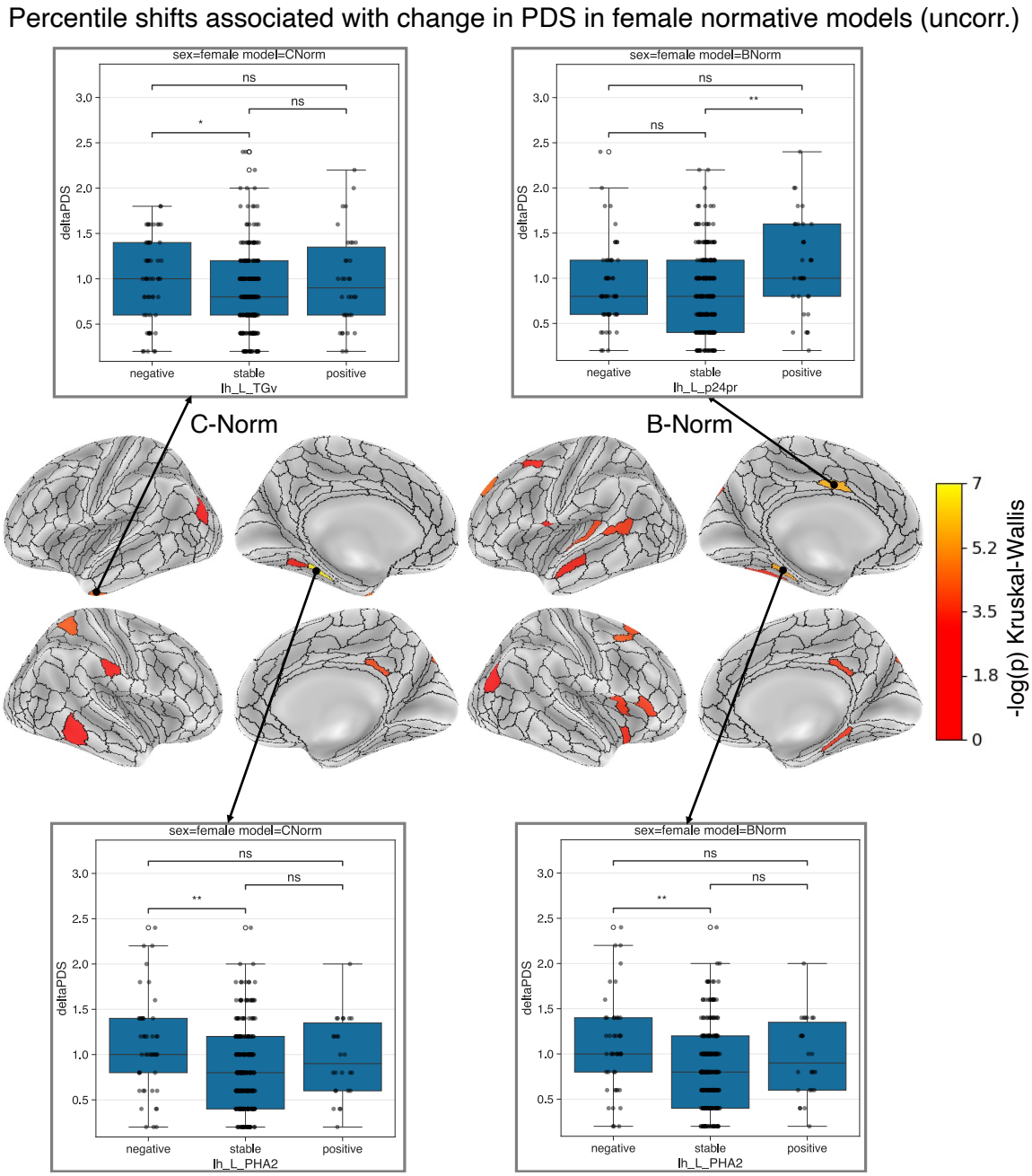


Fig 4. Percentile shifts as indicators of individual level brain development and pubertal progress
Uncorrected significant -log-transformed p-values of a Kruskal-Wallis test plotted on the surface brain for female C-Norms (left) and B-Norms (right). Arrows connecting ROIs with boxplots indicate the two largest effects for an area in the left temporal pole (TGv) and left peri-hippocampal area PHA2 for the C-Norms or for parts of the left cingulate cortex (area p24pr) and area PHA2 for the B-Norms. The boxplots show delta PDS scores across the three groups: negative (zDiff < -1), stable (-1 < zDiff < 1), and positive (zDiff > 1). Black dots represent individual participants. Asterisks above connecting boxes indicate the significance of pair-wise Dunn's tests (n.s.: p>0.05, *: p<0.05, **: p<0.01).

Discussion

Our study shows that Baseline-Integrated Norms (B-Norms), which incorporate baseline cortical thickness, baseline age, and age at follow-up, outperform Cross-Sectional Norms (C-Norms) in

capturing meaningful associations between brain structure and pubertal development in a longitudinal context such as the ABCD study (Casey et al., 2018). B-Norms revealed region-specific deviation patterns associated with pubertal stage progression in females, particularly at later follow-ups, including four years after baseline. In contrast, these associations were absent or markedly weaker when using C-Norms, highlighting the added value of integrating individual baseline measures into longitudinal normative models. The enhanced sensitivity of B-Norms likely stems from their ability to capture developmental changes occurring between baseline and follow-up, which are obscured when relying solely on age-based cross-sectional norms. Notably, the link between the number of negatively deviating regions within lobes and pubertal stage progression underscores the potential of B-Norms to detect subtle, regionally specific brain changes during adolescence. Conversely, growth chart-style analyses did not reveal significant associations with pubertal status at this stage, suggesting that models incorporating baseline measures are better suited for detecting nuanced developmental trajectories.

Performance differences between the two model types indicated that B-Norms substantially outperformed C-Norms, explaining approximately 60% more variance in cortical thickness at both 2-year and 4-year follow-ups. Of note, C-Norms in our analysis performed worse than what has been reported in prior work using similar models trained across the full lifespan (Rutherford, Fraza, et al., 2022). It is likely that the narrow age range of the ABCD study sample accounts for the lower performance overall. However, this age range attributed performance drop cannot explain the superiority of B-Norms over C-Norms in our work, given that the training data used for C-Norms and B-Norms is identical. Investigating the spatial distribution of performances across the brain surface allowed us to further investigate performance differences between C-Norms and B-Norms. While B-Norms achieved largest explained variances in the occipital (Glasser: left/right PIT, VMV1, and left MT) and temporal (Glasser: right PHT, PH, TP0J2 and left TP0J1, TE1m and STSda) regions with similar patterns across sexes, explained variance was lowest in parietal, insular and to some degree in frontal regions. Prior work linked parts of the occipital cortex to pubertal changes, including negative associations with testosterone in females (Bramen et al., 2012). Although a positive association was reported in males, the relatively early pubertal stage of ABCD male participants may explain the absence of this effect here. Interestingly, Wierenga et al. (2022) reported an interaction of cortical thickness with age in the left insula for males (Wierenga et al., 2022). Here, our B-Norms fit insular regions less well in females, possibly reflecting more variability in females. Furthermore, we found inconsistent results in frontal and parietal regions for male and female C-Norms, areas previously tied to pubertal development (Beck et al., 2023). This could be a potential reason for the lower predictive performances of C-Norms compared to B-Norms.

As was highlighted in a recent commentary, it is important to investigate higher-order moments of model derived deviation score distributions in addition to the mean fit to properly assess model performance (A. Marquand et al., 2024). Therefore, we also investigated differences in skewness and excess kurtosis for our C- and B-Norms. This analysis suggested that deviation scores obtained from C-Norm models were more normally distributed than for B-Norm models (excess kurtosis and skewness are closer to 0). The kurtosis results suggest that regional B-Norms are generally more sensitive to changes between visits, which can be seen by the narrower centiles of the normative plots as opposed to those of the C-Norms (see **Fig 1, panel C**).

While better performance in some circumstances is desirable it is important to validate models using meaningful data. To this end, we examined whether deviation scores obtained from our models show a relationship with progression through puberty. The three regions that we found to be significantly associated all showed a negative association between pubertal process and deviation scores and were all in B-Norms of females. These negative associations with PDS may be consistent with previous findings in relation with estradiol levels (Brouwer et al., 2015). Area left PoI2 in the insula was associated with positive deviations (i.e., an increase in cortical thickness) with progression through puberty (higher PDS). Overall, this analysis suggested that deviation scores as obtained by our B-Norms were slightly more negatively associated with pubertal development in females at the 4-year timepoint than the cross-sectional models. This can be seen by the leftward shift of the respective distribution as seen on the right side of **Fig. 2B**. These results suggest that while a better mean fit of the respective population technically means less variance in the residuals which often leads to fewer associations with phenotypic variables and may therefore not be useful in clinical settings. This issue has previously been reported for brain age prediction where models with very high fit can yield weaker associations with clinical variables (Bashyam et al., 2020 and see Hahn et al., 2021). Despite this effect our B-Norms still pick up on unexpected or unusual (e.g., early or late puberty onset in relation to the baseline cortical thickness and age) changes between the baseline and 4-year follow-up. Additionally, such changes could be quite small and yet lead to large deviations. This may be due to the high excess kurtosis, which indicates narrow distributions where even small changes produce large deviation scores. That these small changes align with pubertal progress suggests that incorporating baseline measures (B-Norms) increases sensitivity. Longer intervals between baseline and follow-up may amplify this effect by detecting more atypical changes, which should be tested in future, possibly clinical designs.

Similar to previous studies that applied normative modelling to clinical populations, we investigated whether aggregated deviations from normative brain development, summed across regions of interest (similar to total outlier count, see Verdi et al., 2024; Wolfers et al., 2018), were sensitive to distinct

pubertal stages. We observed sex-specific differences in how deviation counts varied across pubertal stages. Female participants showed greater effects: 4 lobes exhibited significant differences across stages in the C-Norms, and 7 in the B-Norms, compared to only 1 and 2 lobes respectively in males. Pairwise comparisons in the female 2-year follow-up cohort revealed that deviation scores differed significantly between mid- and late-pubertal stages in the right occipital and left cingulate lobes, and between early- and mid-pubertal stages in the right parietal lobe. Interestingly, in the 4-year follow-up data, negative deviations, which reflect cortical thinning relative to normative expectations, were observed in the bilateral temporal and right frontal lobes, primarily distinguishing mid- from late-pubertal females. These lobes are known to undergo significant cortical thinning during adolescence (Shaw et al., 2008; Tamnes et al., 2017). Thus, our findings are in line with previously established findings that this process is stage specific, with late pubertal stages marked by increased deviation from normative developmental trajectories given a certain baseline and age. These findings align with previous literature emphasizing more dynamic neurodevelopmental trajectories in females during puberty, possibly due to earlier onset and more rapid progression of hormonal changes. Reduced cortical thickness during puberty has been previously linked to synaptic pruning, myelination, and hormonal influences (Herting et al., 2015; Vijayakumar et al., 2016). In particular, thinning in temporal and frontal areas has been associated with cognitive maturation and may reflect the refinement of higher-order processes such as emotion regulation and social cognition. Such cognitive functions are especially sensitive to pubertal timing (Crone & Dahl, 2012; Mills et al., 2014).

In a final validation step, using our normative models related to growth charts used in pediatrics, we categorized participants according to percentile shifts into negative, stable and positive deviators by computing the difference between the deviation scores (i.e., z-scores) between the 2- and 4-year follow-up data. This analysis did not reveal any significant effects after correcting for multiple comparisons rendering these results difficult to interpret. We believe that one reason may be our chosen zDiff threshold – which is one of many possible thresholds. Without correcting for multiple comparisons, we found nominal differences in associated regions between females (see Fig 4) and males (see Fig S24). Interestingly, if these results were to hold, ROI-wise male C-Norm and B-Norm models could show stronger and more differences than female models in this percentile shift analysis. This direction should be investigated further with future ABCD releases.

Limitations

A limitation of this study, much like for any other current development of longitudinal methods, is restricted by the availability of large-scale datasets covering the age-range of interest. The degree to which our results translate beyond the population characteristics of the ABCD study cohort (Casey et

al., 2018) remains to be investigated with future releases of new longitudinal datasets. Fortunately, such models can easily be extended, adapted, or retrained with new data releases.

Sample characteristics also apply when interpreting our results on sex differences. The majority of the male participants of the ABCD dataset did not yet fully progress through puberty, which likely explains the lack of puberty related associations with deviation scores. We argue that this has no implications for the results of the female models, as we have trained sex specific models and because it has been shown previously that brain development during puberty is different for males and females (Wierenga et al., 2022).

Another limitation of the current study is the sole focus on puberty related changes. While interesting in their own rights, validation of C- & B-Norms using clinical variables is necessary to estimate their utility in, for example, detecting and predicting developmental or age-related disease trajectories. Thus, future studies should incorporate clinical values and extended our proposed models. Furthermore, interpretation of the tested B-Norm models may be difficult due to their complexity. Given the required baseline features, cortical thickness and age as well as the age at a future timepoint, the model trajectories and centiles are estimated based on variable baseline cortical thickness depending on the baseline and future age. In theory this means that for each baseline and follow-up age per region of interest a growth chart must be generated (see Supplementary Figure S20 for two additional examples).

Conclusion

In this study, we present a developmental application of normative modeling and demonstrate that Baseline-Integrated Norms (B-Norms) utilizing longitudinal data yield deviation scores that are meaningfully associated with pubertal development. These associations were more pronounced in females than in males, likely reflecting the greater puberty-related variance in females in the respective age period. While still at an early stage, this baseline-integrated modeling framework substantially improved model fit compared to cross-sectional approaches. These findings underscore the potential of this modeling approach to capture individual variability in different developmental phases.

Methods

Dataset

We made use of longitudinally acquired data of the Adolescent Brain Cognitive Development (ABCD, Casey et al., 2018) study. In the ABCD dataset children are recruited at ages 9 to 10 with the aim of

characterizing brain developmental trajectories. To this date more than 11.000 children were recruited across 21 different sites in the United States of America. Study procedures have been approved by either the local site Institutional Review Board (IRB) or by local IRB reliance agreements with the central IRB at the University of California. All participants and their parents or legal caregivers provided written informed consent. Data for the current study was obtained from ABCD release 5.1 utilizing phenotypic and imaging data from the baseline, 2-year, and 4-year follow-up visits.

Data selection and preprocessing

Demographics. The ABCD project provides a multitude of tabulated data. Here, we made use of the following files: *abcd_p_demo* to extract sex and ethnicity; *abcd_y_lt* to extract the interview age at baseline and follow-up visits; the *ph_y_anthro* file to compute body mass index (BMI) which we use to exclude participants with unusually large BMI; and *mri_y_adm_info* for information about scan sites which we use as covariates during model training and prediction. We additionally computed mean puberty score (PDS) as rated by the youths' caregivers from data of the *ph_p_pds* file to associate deviation scores with youth's pubertal progress.

Puberty scores. We calculated a summary statistic representing progress in pubertal development using items of the Pubertal Development Scale (PDS; Herting et al., 2021; Kraft et al., 2023). This rating scale was designed to reflect Tanner stages without the need for physical examination (Cheng et al., 2021; Petersen et al., 1988). In this questionnaire, a child's pubertal development is assessed using a four-point Likert scale ranging from 'has not begun' to 'completed'. These items were specific to certain physical characteristics (including skin changes, breast development, deepening in voice, etc.). Please note that some items were administered based on sex. For example, the onset of menarche was exclusively asked for females and is a binary (i.e., either 1 or 4) rating. The ratings are either provided by the children or their caregivers. In this study, we focus on ratings provided by the caregivers for two reasons: a) the self-reported ratings appeared less reliable and b) more data is available for caregiver ratings (Cheng et al., 2021).

Cortical thickness data. We preprocessed the raw structural data of the ABCD on an in-house cluster computer (Ubuntu 22.04) using the recon-all functionality of Freesurfer v7.4.1 (Fischl, 2012). Cortical thickness values were calculated and extracted for 180 regions per hemisphere as defined by the Glasser atlas (Glasser et al., 2016). In addition, we stored Euler numbers which we used to exclude badly reconstructed data during preprocessing.

Preprocessing. Before preprocessing, data comprised of 11868 participants with a baseline measure, 10908 participants with 2-year follow-up data, and 4688 participants with 4-year follow-up data. We excluded all those participants which had only a single measurement. Additionally, we excluded all participants with BMIs of less than 10 or larger than 50. We included this step because BMI has previously been associated with changes in cortical thickness in adolescents (Laurent et al., 2020) as well as in adults (Veit et al., 2014; Westwater et al., 2019). Furthermore, we excluded all those participants who had no valid or no pubertal development score (PDS) rating and cortical thickness measures available. We then separated those participants with only baseline- and 2-year follow-up data into a subset and those with baseline-, 2-year, and 4-year data into another. For both subsets we used the Euler numbers and computed their mean and standard deviation for each of the ABCDs scan-sites. We then excluded all participants whose Euler numbers were larger than 6 SDs from the mean for a given scan-site. This was to ensure that no extremely badly reconstructed data was used (for overview plots see Supplementary Figures S1 and S2). Lastly, we performed a similar exclusion step for cortical thickness values. We again calculated the mean cortical thickness and standard deviation across subjects for each ROI and excluded all participants who exceeded the 6 SDs threshold. This was done to prevent strange overfitting phenomena due to extreme outliers.

Train test and splits. After preprocessing we designated those participants with only baseline (female: N=2515, age=118.30±7.39 [mean±SD]; male: N=2859, age=119.06±7.49) and 2-year follow-up measurements (female: age=142.92±7.81; male: age=143.74±7.81) as the training dataset. Participants with baseline (female: N=366, age=118.87±7.47; male: N=439, age=119.62±7.41), 2-year follow-up (female: age=142.73±7.31; male: age=143.64±7.62), and 4-year follow-up (female: age=167.82±7.48; male: age=168.98±7.99) data were used for testing. These splits did not differ significantly on core demographic variables such as age or BMI at baseline but showed some statistical differences as determined by a Welch's t-test in PDS scores. We argue that this difference, while unfortunate, is barely avoidable as stratifying for these variables would have resulted in a much smaller test set. For more details see supplementary figures S3-5 and the corresponding text.

Normative modelling with Bayesian Linear Regression

We employed normative modeling using Python 3.12.9 and the PCNtoolkit (Rutherford, Kia, et al., 2022) package (version 0.33). Bayesian Linear Regression (BLR) with likelihood warping (Fraza et al., 2021) was used to predict cortical thickness from a covariance matrix including "age_{bl} and site" for the standard Cross-Sectional normative models (C-Norms) and "cortical $thickness_{bl}$, age_{bl}, $age_{follow\text{-}up}$, and site" for the Baseline-Integrated normative models (B-Norms). Sinarcsinsh was employed for warping

(Jones & Pewsey, 2009). For each of the 180 brain regions for both hemispheres as defined by the Glasser atlas (Glasser et al., 2016), cortical thickness is predicted as:

$$CT_{bl} = age_{bl} + site_covariates$$

for the C-Norms, and

$$CT_{2yr} = CT_{bl} + age_{bl} + age_{2yr} + \ site_covariates$$

for the B-Norms.

Both models were optimized using the powell algorithm, and results are based on models trained on the reference sample and evaluated on the independent test set. We assessed the model fit for each brain region using several metrics, including Pearson's correlation between observed and predicted measures, root-mean-squared error (RMSE), standardized mean-squared error (SMSE), explained variance (EV), and mean squared log-loss (MSLL). Additionally, we evaluated skewness and kurtosis to estimate higher-order moments beyond the mean in the test set (Dinga et al., 2021).

Comparing performance measures

We used the ANOVA function from pingouin python package to perform 2-factor repeated-measures analysis of variance (rm-ANOVA) to statistically assess differences in performance metrics of the C- and B-Norms separately for both males and females. The dependent variables were the performances measures as defined earlier. Our within-subject factors were MODEL (levels: C-Norm and B-Norm) and TIMEPOINT (levels: 2-year and 4-year follow-up), the identifier variable were the regions of interest as defined by the Glasser atlas. We show statistics for explained variance, skewness, and kurtosis in the main text. Mean squared log loss (MSLL), (standardized) mean squared error ([S]MSE), root-mean squared error (RMSE), and Bayesian Information Criteria (BIC) are presented in the supplementary material.

Validations of normative models against puberty scores

We performed an association analysis to reveal relationships between puberty scores as rated by the youths' caregivers and the deviation scores (i.e., z-scores) obtained from the normative models. We conducted this analysis using the statsmodels Python-package, specifically the glm function. For each sex-specific model and each ROI, we computed the association $PDS_t \sim zROI_t$, where PDS are vectorized mean PDS scores per participants and $zROI$ is a vector containing subjects' deviation scores as obtained by the C- or B-Norms at follow-up time point *t* (i.e., 2-year or 4-year). This resulted in 360 t-scores and p-values (which we corrected for multiple comparisons using the Benjamini-Hochberg (Benjamini & Hochberg, 1995) false discovery method), one for each ROI, for each model, and sex. Statistics were then projected onto a surface brain and thresholded according to the critical BH values.

Validation of normative models using puberty subgroups

To examine differences in positive and negative deviations across ROIs, participants were grouped into five pubertal stages (pre-, early-, mid-, late-, and post-pubertal), following Herting et al. and Kraft et al. (Herting et al., 2021; Kraft et al., 2023). Glasser ROIs were aggregated into six lobe-level regions: occipital, frontal, temporal, parietal, insular, and cingulate – based on previously proposed definitions (Kaufmann et al., 2019). For each participant and lobe, we counted the number of regions exceeding z-scores of +/- 1.96, producing lobe-specific deviation vectors. These count vectors were stratified by pubertal stage, and Kruskal-Wallis tests were conducted across stages for each lobe, deviation type (positive/negative), model (cross-sectional/baseline-integrated), and timepoint (2- and 4-year follow-ups), totaling 192 tests.

Association of stable, positive or negative percentile shifts with Puberty progression

We additionally validated the normative models by categorizing participants into three groups based on percentile shifts between the 2-year and 4-year follow-up data. Specifically, we defined three groups: negative (zDiff < -1), stable (-1 < zDiff < 1), and positive (zDiff > 1), with zDiff representing the change in ROI-specific z-scores over time. For each ROI, participants were assigned to one of these groups; the number of participants per group therefore varies. For each ROI a Kruskal-Wallis test was conducted to assess group differences in delta PDS scores (4-year minus 2-year).

Acknowledgements

The study was funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung [BMBF]) and the ministry of Baden-Würtemberg within the initial phase of the German Center for Mental Health (DZPG) (grant: 01EE2306). T.K. and T.W. are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764. TW acknowledges funding from German Research Foundation (DFG) Emmy Noether: 513851350. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). The authors used data from the Adolescent Brain Cognitive Development^{SM} Study (ABCD, abcdstudy.org). ABCD data, held in the NIMH Data Archive (NDA), is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016,

U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. The ABCD consortium investigators designed and implemented the respective studies and/or provided data but did not participate in the analysis or writing of this report. This manuscript reflects the views of the authors and does not necessarily reflect the opinions or views of any other agency, organization, employer or company.

Author Contributions

P.S. was responsible for data curation, code, analyses, and first draft of the manuscript. T.K. was responsible for preprocessing. T.K. and T.W. were responsible for conceptualization. All authors were responsible for writing and revising the manuscript.

Data availability

Data from the Adolescent Brain Cognitive Development^{SM} Study (ABCD, abcdstudy.org) is available upon application on the NIH website.

Code availability

Code is available on P.S. github-repository: https://github.com/PhilippS893/b-norm_modelling.

References

Alden, E. C., Lundt, E. S., Twohy, E. L., Christianson, T. J., Kremers, W. K., Machulda, M. M., Jack, C. R., Knopman, D. S., Mielke, M. M., Petersen, R. C., & Stricker, N. H. (2022). Mayo normative studies: A conditional normative model for longitudinal change on the Auditory Verbal Learning Test and preliminary validation in preclinical Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *14*(1), e12325. https://doi.org/10.1002/dad2.12325

Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I. M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., … Davatzikos, C. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, *143*(7), 2312–2324. https://doi.org/10.1093/brain/awaa160

Bayer, J. M. M., Dinga, R., Kia, S. M., Kottaram, A. R., Wolfers, T., Lv, J., Zalesky, A., Schmaal, L., & Marquand, A. (2022). Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. *NeuroImage*, *264*, 119699. https://doi.org/10.1016/j.neuroimage.2022.119699

Beck, D., Ferschmann, L., MacSweeney, N., Norbom, L. B., Wiker, T., Aksnes, E., Karl, V., Dégeilh, F., Holm, M., Mills, K. L., Andreassen, O. A., Agartz, I., Westlye, L. T., Von Soest, T., & Tamnes, C. K. (2023). Puberty differentially predicts brain maturation in male and female youth: A longitudinal ABCD Study. *Developmental Cognitive Neuroscience*, *61*, 101261. https://doi.org/10.1016/j.dcn.2023.101261

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bethlehem, R. A. I., Seidlitz, J., Romero-Garcia, R., Trakoshis, S., Dumas, G., & Lombardo, M. V. (2020). A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Communications Biology*, *3*(1), 486. https://doi.org/10.1038/s42003-020-01212-9

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y

Bramen, J. E., Hranilovich, J. A., Dahl, R. E., Chen, J., Rosso, C., Forbes, E. E., Dinov, I. D., Worthman, C. M., & Sowell, E. R. (2012). Sex Matters during Adolescence: Testosterone-Related Cortical Thickness Maturation Differs between Boys and Girls. *PLoS ONE*, *7*(3), e33850. https://doi.org/10.1371/journal.pone.0033850

Brouwer, R. M., Koenis, M. M. G., Schnack, H. G., Van Baal, G. C., Van Soelen, I. L. C., Boomsma, D. I., & Hulshoff Pol, H. E. (2015). Longitudinal Development of Hormone Levels and Grey Matter Density in 9 and 12-Year-Old Twins. *Behavior Genetics*, *45*(3), 313–323. https://doi.org/10.1007/s10519-015-9708-8

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001

Cheng, T. W., Magis-Weinberg, L., Williamson, V. G., Ladouceur, C. D., Whittle, S. L., Herting, M. M., Uban, K. A., Byrne, M. L., Barendse, M. E. A., Shirtcliff, E. A., & Pfeifer, J. H. (2021). A Researcher's Guide to the Measurement and Modeling of Puberty in the ABCD Study® at Baseline. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.608575

Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nature Reviews Neuroscience*, *13*(9), 636–650. https://doi.org/10.1038/nrn3313

Dehestani, N., Vijayakumar, N., Ball, G., Mansour L, S., Whittle, S., & Silk, T. J. (2024). "Puberty age gap": New method of assessing pubertal timing and its association with mental health problems. *Molecular Psychiatry*, *29*(2), 221–228. https://doi.org/10.1038/s41380-023-02316-4

Dehestani, N., Whittle, S., Vijayakumar, N., & Silk, T. J. (2023). Developmental brain changes during puberty and associations with mental health problems. *Developmental Cognitive Neuroscience*, *60*, 101227. https://doi.org/10.1016/j.dcn.2023.101227

Di Biase, M. A., Tian, Y. E., Bethlehem, R. A. I., Seidlitz, J., Alexander-Bloch, Aaron. F., Yeo, B. T. T., & Zalesky, A. (2023). Mapping human brain charts cross-sectionally and longitudinally. *Proceedings of the National Academy of Sciences*, *120*(20), e2216798120. https://doi.org/10.1073/pnas.2216798120

Dinga, R., Fraza, C. J., Bayer, J. M. M., Kia, S. M., Beckmann, C. F., & Marquand, A. F. (2021). *Normative modeling of neuroimaging data using generalized additive models of location scale and shape*. https://doi.org/10.1101/2021.06.14.448106

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Frangou, S., Modabbernia, A., Williams, S. C. R., Papachristou, E., Doucet, G. E., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnæs, D., Alpert, K. I., Andersson, M., Andreasen, N. C., Andreassen, O. A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., … Dima, D. (2022). Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3–90 years. *Human Brain Mapping*, *43*(1), 431–451. https://doi.org/10.1002/hbm.25364

Franke, K., & Gaser, C. (2019). Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00789

Fraza, C. J., Dinga, R., Beckmann, C. F., & Marquand, A. F. (2021). Warped Bayesian linear regression for normative modelling of big data. *NeuroImage*, *245*, 118715. https://doi.org/10.1016/j.neuroimage.2021.118715

Gaiser, C., Berthet, P., Kia, S. M., Frens, M. A., Beckmann, C. F., Muetzel, R. L., & Marquand, A. F. (2024). Estimating cortical thickness trajectories in children across different scanners using transfer learning from normative models. *Human Brain Mapping*, *45*(2), e26565. https://doi.org/10.1002/hbm.26565

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Hahn, T., Fisch, L., Ernsting, J., Winter, N. R., Leenings, R., Sarink, K., Emden, D., Kircher, T., Berger, K., & Dannlowski, U. (2021). From 'loose fitting' to high-performance, uncertainty-aware brain-age modelling. *Brain*, *144*(3), e31–e31. https://doi.org/10.1093/brain/awaa454

Herting, M. M., Gautam, P., Spielberg, J. M., Dahl, R. E., & Sowell, E. R. (2015). A Longitudinal Study: Changes in Cortical Thickness and Surface Area during Pubertal Maturation. *PLOS ONE*, *10*(3), e0119774. https://doi.org/10.1371/journal.pone.0119774

Herting, M. M., Uban, K. A., Gonzalez, M. R., Baker, F. C., Kan, E. C., Thompson, W. K., Granger, D. A., Albaugh, M. D., Anokhin, A. P., Bagot, K. S., Banich, M. T., Barch, D. M., Baskin-Sommers, A., Breslin, F. J., Casey, B. J., Chaarani, B., Chang, L., Clark, D. B., Cloak, C. C., … Sowell, E. R. (2021). Correspondence Between Perceived Pubertal Development and Hormone Levels in 9-10 Year-Olds From the Adolescent Brain Cognitive Development Study. *Frontiers in Endocrinology*, *11*, 549928. https://doi.org/10.3389/fendo.2020.549928

Insel, T. R. (2014). Mental Disorders in Childhood: Shifting the Focus From Behavioral Symptoms to Neurodevelopmental Trajectories. *JAMA*, *311*(17), 1727. https://doi.org/10.1001/jama.2014.1193

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., & Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, *9*(1), 119–128. https://doi.org/10.1016/S1474-4422(09)70299-6

Janssen, J., Gallego, A. G., Díaz-Caneja, C. M., Lois, N. G., Janssen, N., González-Peñas, J., Gordaliza, P. M., Buimer, E. E. L., Van Haren, N. E. M., Arango, C., Kahn, R. S., Hulshoff Pol, H. E., & Schnack, H. G. (2024). *Heterogeneity of morphometric similarity networks in health and schizophrenia*. Neuroscience. https://doi.org/10.1101/2024.03.26.586768

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, *96*(4), 761–780. https://doi.org/10.1093/biomet/asp053

Karas, G. B., Scheltens, P., Rombouts, S. A. R. B., Visser, P. J., Van Schijndel, R. A., Fox, N. C., & Barkhof, F. (2004). Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, *23*(2), 708–716. https://doi.org/10.1016/j.neuroimage.2004.07.006

Kaufmann, T., Karolinska Schizophrenia Project (KaSP), Van Der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M. K., Bøen, E., Borgwardt, S., Brandt, C. L., Buitelaar, J., Celius, E. G., Cervenka, S., … Westlye, L. T. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, *22*(10), 1617–1623. https://doi.org/10.1038/s41593-019-0471-7

Kia, S. M., & Marquand, A. F. (2019). *Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data* (No. arXiv:1812.04998). arXiv. https://doi.org/10.48550/arXiv.1812.04998

Kjelkenes, R., Wolfers, T., Alnæs, D., Norbom, L. B., Voldsbekk, I., Holm, M., Dahl, A., Berthet, P., Tamnes, C. K., Marquand, A. F., & Westlye, L. T. (2022). Deviations from normative brain white and gray matter structure are associated with psychopathology in youth. *Developmental Cognitive Neuroscience*, *58*, 101173. https://doi.org/10.1016/j.dcn.2022.101173

Kjelkenes, R., Wolfers, T., Alnæs, D., Van Der Meer, D., Pedersen, M. L., Dahl, A., Voldsbekk, I., Moberget, T., Tamnes, C. K., Andreassen, O. A., Marquand, A. F., & Westlye, L. T. (2023). Mapping Normative Trajectories of Cognitive Function and Its Relation to Psychopathology Symptoms and Genetic Risk in Youth. *Biological Psychiatry Global Open Science*, *3*(2), 255–263. https://doi.org/10.1016/j.bpsgos.2022.01.007

Korbmacher, M., Vidal-Pineiro, D., Wang, M.-Y., Van Der Meer, D., Wolfers, T., Nakua, H., Eikefjord, E., Andreassen, O. A., Westlye, L. T., & Maximov, I. I. (2024). *Cross-sectional brain age assessments are limited in predicting future brain change*. Neuroscience. https://doi.org/10.1101/2024.09.11.612523

Kraft, D., Alnæs, D., & Kaufmann, T. (2023). Domain adapted brain network fusion captures variance related to pubertal brain development and mental health. *Nature Communications*, *14*(1), 6698. https://doi.org/10.1038/s41467-023-41839-w

Laurent, J. S., Watts, R., Adise, S., Allgaier, N., Chaarani, B., Garavan, H., Potter, A., & Mackey, S. (2020). Associations Among Body Mass Index, Cortical Thickness, and Executive Function in Children. *JAMA Pediatrics*, *174*(2), 170. https://doi.org/10.1001/jamapediatrics.2019.4708

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, *24*(10), 1415–1424. https://doi.org/10.1038/s41380-019-0441-1

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Marquand, A., Rutherford, S., & Dinga, R. (2024). Fairly evaluating the performance of normative models. *The Lancet Digital Health*, *6*(11), e775. https://doi.org/10.1016/S2589-7500(24)00200-0

Mills, K. L., Goddings, A.-L., Clasen, L. S., Giedd, J. N., & Blakemore, S.-J. (2014). The Developmental Mismatch in Structural Brain Maturation during Adolescence. *Developmental Neuroscience*, *36*(3–4), 147–160. https://doi.org/10.1159/000362328

Petersen, A. C., Crockett, L., Richards, M., & Boxer, A. (1988). A self-report measure of pubertal status: Reliability, validity, and initial norms. *Journal of Youth and Adolescence*, *17*(2), 117–133. https://doi.org/10.1007/BF01537962

Pinaya, W. H. L., Mechelli, A., & Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human Brain Mapping*, *40*(3), 944–954. https://doi.org/10.1002/hbm.24423

Pinaya, W. H. L., Scarpazza, C., Garcia-Dias, R., Vieira, S., Baecker, L., F Da Costa, P., Redolfi, A., Frisoni, G. B., Pievani, M., Calhoun, V. D., Sato, J. R., & Mechelli, A. (2021). Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study. *Scientific Reports*, *11*(1), 15746. https://doi.org/10.1038/s41598-021-95098-0

Rehák Bučková, B., Fraza, C., Rehák, R., Kolenič, M., Beckmann, C., Španiel, F., Marquand, A., & Hlinka, J. (2025). *Using normative models pre-trained on cross-sectional data to evaluate intra-individual longitudinal changes in neuroimaging data*. https://doi.org/10.7554/eLife.95823.3

Rogers, J. C., & De Brito, S. A. (2016). Cortical and Subcortical Gray Matter Volume in Youths With Conduct Problems: A Meta-analysis. *JAMA Psychiatry*, *73*(1), 64. https://doi.org/10.1001/jamapsychiatry.2015.2423

Rutherford, S., Barkema, P., Tso, I. F., Sripada, C., Beckmann, C. F., Ruhe, H. G., & Marquand, A. F. (2023). Evidence for embracing normative modeling. *eLife*, *12*, e85082. https://doi.org/10.7554/eLife.85082

Rutherford, S., Fraza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L. K., Bayer, J. M., Dazzan, P., McGuire, P., Mocking, R. T., Schene, A., Sripada, C., Tso, I. F., Duval, E. R., … Marquand, A. F. (2022). Charting brain growth and aging at high spatial precision. *eLife*, *11*, e72904. https://doi.org/10.7554/eLife.72904

Rutherford, S., Kia, S. M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H. G., Beckmann, C. F., & Marquand, A. F. (2022). The normative modeling framework for computational psychiatry. *Nature Protocols*, *17*(7), 1711–1734. https://doi.org/10.1038/s41596-022-00696-5

Shaw, P., Kabani, N. J., Lerch, J. P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J. L., Giedd, J. N., & Wise, S. P. (2008). Neurodevelopmental Trajectories of the Human Cerebral Cortex. *The Journal of Neuroscience*, *28*(14), 3586–3594. https://doi.org/10.1523/JNEUROSCI.5309-07.2008

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar De Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: Large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*(1), 281–295. https://doi.org/10.1038/s41380-021-01161-7

Tamnes, C. K., Herting, M. M., Goddings, A.-L., Meuwese, R., Blakemore, S.-J., Dahl, R. E., Güroğlu, B., Raznahan, A., Sowell, E. R., Crone, E. A., & Mills, K. L. (2017). Development of the Cerebral Cortex across Adolescence: A Multisample Study of Inter-Related Longitudinal Changes in Cortical Volume, Surface Area, and Thickness. *The Journal of Neuroscience*, *37*(12), 3402–3412. https://doi.org/10.1523/JNEUROSCI.3302-16.2017

Veit, R., Kullmann, S., Heni, M., Machann, J., Häring, H.-U., Fritsche, A., & Preissl, H. (2014). Reduced cortical thickness associated with visceral fat and BMI. *NeuroImage: Clinical*, *6*, 307–311. https://doi.org/10.1016/j.nicl.2014.09.013

Verdi, S., Kia, S. M., Yong, K. X. X., Tosun, D., Schott, J. M., Marquand, A. F., & Cole, J. H. (2023). Revealing Individual Neuroanatomical Heterogeneity in Alzheimer Disease Using Neuroanatomical Normative Modeling. *Neurology*, *100*(24). https://doi.org/10.1212/WNL.0000000000207298

Verdi, S., Marquand, A. F., Schott, J. M., & Cole, J. H. (2021). Beyond the average patient: How neuroimaging models can address heterogeneity in dementia. *Brain*, *144*(10), 2946–2953. https://doi.org/10.1093/brain/awab165

Verdi, S., Rutherford, S., Fraza, C., Tosun, D., Altmann, A., Raket, L. L., Schott, J. M., Marquand, A. F., Cole, J. H., & for the Alzheimer's Disease Neuroimaging Initiative. (2024). Personalizing progressive changes to brain structure in Alzheimer's disease using normative modeling. *Alzheimer's & Dementia*, *20*(10), 6998–7012. https://doi.org/10.1002/alz.14174

Vidal-Pineiro, D., Sorensen, O., Stromstad, M., Amlien, I. K., Baare, W. F. C., Bartres-Faz, D., Brandmaier, A. M., Cattaneo, G., Duzel, S., Ghisletta, P., Henson, R. N., Kuhn, S., Lindenberger, U., Mowinckel, A. M., Nyberg, L., Pascual-Leone, A., Roe, J. M., Solana-Sanchez, J., Sole-Padulles, C., … Fjell, A. M. (2025). *Vulnerability to memory decline in aging. A mega-analysis of structural brain change.* Neuroscience. https://doi.org/10.1101/2025.03.27.642988

Vijayakumar, N., Allen, N. B., Youssef, G., Dennison, M., Yücel, M., Simmons, J. G., & Whittle, S. (2016). Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume. *Human Brain Mapping*, *37*(6), 2027–2038. https://doi.org/10.1002/hbm.23154

Westwater, M. L., Vilar-López, R., Ziauddeen, H., Verdejo-García, A., & Fletcher, P. C. (2019). Combined effects of age and BMI are related to altered cortical thickness in adolescence and adulthood. *Developmental Cognitive Neuroscience*, *40*, 100728. https://doi.org/10.1016/j.dcn.2019.100728

Whittle, S., Vijayakumar, N., Simmons, J. G., & Allen, N. B. (2020). Internalizing and Externalizing Symptoms Are Associated With Different Trajectories of Cortical Development During Late Childhood. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59*(1), 177–185. https://doi.org/10.1016/j.jaac.2019.04.006

Wierenga, L. M., Doucet, G. E., Dima, D., Agartz, I., Aghajani, M., Akudjedu, T. N., Albajes-Eizagirre, A., Alnæs, D., Alpert, K. I., Andreassen, O. A., Anticevic, A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D. I., … Tamnes, C. K. (2022). Greater male than female variability in regional brain structure across the lifespan. *Human Brain Mapping*, *43*(1), 470–499. https://doi.org/10.1002/hbm.25204

Wolfers, T., Beckmann, C. F., Hoogman, M., Buitelaar, J. K., Franke, B., & Marquand, A. F. (2020). Individual differences *v.* the average patient: Mapping the heterogeneity in ADHD using normative models. *Psychological Medicine*, *50*(2), 314–323. https://doi.org/10.1017/S0033291719000084

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J. K., Ueland, T., Melle, I., Franke, B., Andreassen, O. A., Beckmann, C. F., Westlye, L. T., & Marquand, A. F. (2018). Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry*, *75*(11), 1146. https://doi.org/10.1001/jamapsychiatry.2018.2467

Wolfers, T., Floris, D. L., Dinga, R., Van Rooij, D., Isakoglou, C., Kia, S. M., Zabihi, M., Llera, A., Chowdanayaka, R., Kumar, V. J., Peng, H., Laidi, C., Batalle, D., Dimitrova, R., Charman, T., Loth, E., Lai, M.-C., Jones, E., Baumeister, S., … Beckmann, C. F. (2019). From pattern classification to stratification: Towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neuroscience & Biobehavioral Reviews*, *104*, 240–254. https://doi.org/10.1016/j.neubiorev.2019.07.010

Wolfers, T., Rokicki, J., Alnæs, D., Berthet, P., Agartz, I., Kia, S. M., Kaufmann, T., Zabihi, M., Moberget, T., Melle, I., Beckmann, C. F., Andreassen, O. A., Marquand, A. F., & Westlye, L. T. (2021). Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder. *Human Brain Mapping*, *42*(8), 2546–2555. https://doi.org/10.1002/hbm.25386

Zabihi, M., Floris, D. L., Kia, S. M., Wolfers, T., Tillmann, J., Arenas, A. L., Moessnang, C., Banaschewski, T., Holt, R., Baron-Cohen, S., Loth, E., Charman, T., Bourgeron, T., Murphy, D., Ecker, C., Buitelaar, J. K., Beckmann, C. F., Marquand, A., & The EU-AIMS LEAP Group. (2020). Fractionating autism based on neuroanatomical normative modeling. *Translational Psychiatry*, *10*(1), 384. https://doi.org/10.1038/s41398-020-01057-0

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., Holt, R., Baron-Cohen, S., Durston, S., Bölte, S., Murphy, D., Ecker, C., Buitelaar, J. K., Beckmann, C. F., & Marquand, A. F. (2019). Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(6), 567–578. https://doi.org/10.1016/j.bpsc.2018.11.013