Estimating cortical thickness trajectories in children across different scanners using transfer learning from normative models

Gaiser C.*[a,b], Berthet P.*[c,d], Kia S.M.[e,f,g], , Frens M. A.[a], Beckmann C. F.[e,h,i]., Muetzel R. L.[j,k], Marquand A.[e,h]

Affiliations

[a] Department of Neuroscience, Erasmus MC, University Medical Centre Rotterdam, Rotterdam, The Netherlands.

[b] The Generation R Study Group, Erasmus MC - Sophia Children's Hospital, University Medical Centre Rotterdam, Rotterdam, The Netherlands.

[c] Department of Psychology, University of Oslo, Oslo, Norway.

[d] Norwegian Center for Mental Disorders Research (NORMENT), University of Oslo, and Oslo University Hospital, Oslo, Norway.

[e] Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, the Netherlands.

[f] Department of Psychiatry, Utrecht University Medical Center, Utrecht, the Netherlands.

[g] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, the Netherlands.

[h] Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands.

[i] Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK.

[j] Department of Child and Adolescent Psychiatry, Erasmus MC - Sophia Children's Hospital, University Medical Centre Rotterdam, Rotterdam, The Netherlands.

[k] Department of Radiology and Nuclear Medicine, Erasmus MC - Sophia Children's Hospital, University Medical Centre Rotterdam, Rotterdam, The Netherlands.

* Contributed equally to this work

Keywords

Normative modelling
MRI
Neurodevelopment
Brain age
Cortical thickness

Abstract

This work illustrates the use of normative models in a longitudinal neuroimaging study of children aged 6-17 years and demonstrates how such models can be used to make meaningful comparisons in longitudinal studies, even when individuals are scanned with different scanners across successive study waves. More specifically, we first estimated a large-scale reference normative model using hierarchical Bayesian regression from N=40,435 individuals across the lifespan and from dozens of sites. We then transfer these models to a longitudinal developmental cohort (N=5,985) with three measurement waves acquired on two different scanners that were unseen during estimation of the reference models. We show that the use of normative models provides individual deviation scores that are independent of scanner effects and efficiently accommodate inter-site variations. Moreover, we provide empirical evidence to guide the optimization of sample size for the transfer of prior knowledge about the distribution of regional cortical thicknesses. We show that a transfer set containing as few as 25 samples per site can lead to good performance metrics on the test set. Finally, we demonstrate the clinical utility of this approach by showing that deviation scores obtained from the transferred normative models are able to detect and chart morphological heterogeneity in individuals born pre-term.

1. Introduction

Identifying structural or functional biomarkers of psychiatric and neurological illnesses across the lifespan has received increasing attention in recent years. Many of these disorders present symptoms that begin during childhood and adolescence (Bayer et al., 2021; Rogers & de Brito, 2016; Solmi et al., 2022; Whittle et al., 2020). There is however large inter-individual heterogeneity in symptoms and underlying biology (DeLisi, 2008; Fuhrmann et al., 2022; Mills et al., 2021; Tamnes et al., 2017), making it challenging to pinpoint the precise underlying neurobiological substrates. Longitudinal datasets provide particularly valuable insights on the temporal evolution of brain development and offer considerable potential to understand the emergence of psychopathology and to parse this heterogeneity across individuals.

To detect and understand this heterogeneity and atypicality, there is a need to better characterize typical neurodevelopment (Insel, 2014; Volpe, 2009). In recent years, the availability of large datasets has greatly assisted efforts to understand inter-individual variability in brain development (Bethlehem et al., 2022; Rutherford, Fraza, et al., 2022). For example, large scale studies using cortical volume, cortical thickness (CT) and surface area have identified a general decrease in these metrics with age, after adolescence (Bethlehem et al., 2022; Frangou et al., 2022; Rutherford, Fraza, et al., 2022; Tamnes et al., 2017; Thambisetty et al., 2010). CT has been shown to more accurately reflect underlying pathophysiologic mechanisms than gray matter volume analysis (Clarkson et al., 2011; Hutton et al., 2009; Pereira et al., 2012; Zhao et al., 2022). However, these large data resources have expanded in scale via large, long-running longitudinal cohort studies. While the benefits of these large and unique cohorts are obvious, such studies impose particularly difficulties. For example, data must often be aggregated across multiple study centers, necessitating dealing with site effects and across developmental time scale, subjects are often scanned with different scanner hardware and/or software at successive timepoints. As a result, there is often little or no overlap in terms of age of participants and site effects in successive acquisition waves. Such non-trivial differences across sites, scanners, and timepoints have been difficult to account for statistically in analyses. Therefore, in addition to longitudinal data, novel methodological tools that map inter-individual differences are needed to generate new insights.

Normative modeling approaches have recently emerged as a tool for better understanding longitudinal developments with neuroimaging data (Marquand et al., 2019; Marquand, Rezek, et al., 2016). These approaches produce statistical inference at the individual level, without relying on strong assumptions about clustering of individuals or population structure (Antoniades et al., 2021; Cole, 2012; Marquand et al., 2019). Instead, symptoms in individual patients can be related to extreme deviation from the normative range (Fraza et al., 2021; Marquand, Wolfers, et al., 2016; Zabihi et al., 2019). This has shown the potential to detect morphological differences in patient populations which were not evident using standard techniques (Remiszewski et al., 2022). Additionally, a *Hierarchical Bayesian Regression* (HBR) approach to normative modeling has been shown to efficiently accommodate inter-site variation and to provide good computational scaling, which is useful when using large studies, longitudinal studies, or combining smaller studies together, that are acquired across multiple sites (Bayer et al., 2021; Kia et al., 2022; Rutherford, Fraza, et al., 2022). It also supports federated (*i.e.* decentralized) multi-site normative modeling to transfer previously trained models onto unseen sites, while benefiting from the training on the large reference datasets (Kia et al., 2022; Rutherford, Kia, et al., 2022). This is especially interesting given that in longitudinal studies running over several years, changes of scanner hardware, software and/or scan protocols are the norm rather than the exception, which generates a need to correct for the resulting scanner effects.

In this work, we provide a case study in using the transfer of prior knowledge about CT distributions from normative models derived from a large reference (e.g. lifespan) cohort to better estimate parameters on a smaller target (e.g. clinical) cohort. For this, we use longitudinal CT data from the Generation R study (Jaddoe et al., 2006; Kooijman et al., 2016; White et al., 2018), which contains data from children aged 6-17 years scanned in two different scanners, unseen by the reference models. The narrow age-range makes this study a good candidate for transfer learning in that it is necessary to transfer information learned from a large lifespan cohort to obtain precise estimates of the slope or trajectory of developmental effects across this narrow age range. This method provides important benefits: one, it allows meaningful comparison of individuals scanned on different scanners, while taking advantage of previous knowledge, built from large publicly available datasets to set informed hyperpriors: expected mean and variance of the distribution of samples for each ROIs. This, in turn, provide three benefits to the study, one providing more accurate predictions from the models thanks to the use of the mentioned informed priors; second this enables to reduce the ratio of training samples necessary to learn developmental trajectories for to the unseen sites, thereby enabling more participants to be allocated to the test set, and thus improving statistical power (Pan & Yang, 2010). Third, we will show that it provides a means to draw meaningful inferences within individuals across timepoints, even when follow-up scans are derived from a different scanner. This work also aims to offer some guidance on the methodology, *e.g.* providing empirical estimates of the number of samples required for the transfer of knowledge from previous learnings and choices in transfer configurations, *e.g.* factors included as batch effects. Finally, we provide a demonstration of the clinical utility of this approach by using it to understand inter-individual differences in brain morphology resulting from pre-term birth.

2. Methods

2.1 Normative modeling

We estimated normative models using *Hierarchical Bayesian Regression* (HBR) to predict cortical thickness from age, sex and scanner site, for each *region of interest* (ROI) using the PCNtoolkit python package, version 0.22 (Rutherford, Kia, et al., 2022).

2.1.1 Reference models

We assembled a large reference cohort (Kia et al., 2022) containing n=40,435 (95%) healthy individuals to train the normative models before validating this model on n=2,548 controls and patients (5%, stratified by sites) from a collection of mostly publicly available MRI datasets across 77 sites and 42,983 participants: ABIDE-1, ADHD-200, CAMCAM, PNC, CNP, HCP-Aging, HCP-Dev, HCP-EP, OASIS, OPN, IXI, NKI-RS, UKBB, ABCD and CMI-HBN. The reference model is available on the PCNportal (https://pcnportal.dccn.nl/). Cortical thickness measures were obtained from FreeSurfer processing (versions 5.3 or 6.0), as referred in the publications associated with the datasets (Dale et al., 1999; Fischl et al., 1999, 2002; Fischl & Dale, 2000). Parcellation of the brain was made with the Destrieux atlas (Destrieux et al., 2010). One normative model was estimated per ROI. Linear HBR models were estimated using fixed effects of age and batch (*i.e.* random) effects for site and sex. In practice, this allows each site and sex to have different slopes, intercepts and variances. We included only data from the first visit when multiple visits were available (*i.e.* UKBB and ABCD). Any single missing individual ROI data (less than 0.1% of the samples per ROI) was imputed as the site and sex specific ROI mean.

Estimated reference models performed well according to accuracy metrics (explained variance: mean=0.44, SD=0.13, *standardized mean squared error* (SMSE): mean=0.55, SD=0.13, and *mean standardized log loss* (MSLL): mean=-0.37, SD=0.14). Outputs include hyperparameters defining the mean and variance of the site-specific mean effects and variance, estimated during the training over the collection of datasets. This can be used as informed priors when adapting the normative models to unseen target sites. These hyperparameters are adapted to the unseen site using a holdout subset of the target dataset*, i.e.* the adaptation set. This allows to reduce the number of samples used for adaptation while retaining a low variance of the estimations.

2.1.2 Target cohort

As target cohort, 8,523 T_1-weighted MRI scans from the population-based longitudinal Generation R study (Jaddoe et al., 2006; Kooijman et al., 2016) were used. In short, the Generation R study is a prospective cohort study from fetal life until adulthood that is designed to find early markers for typical and atypical development, growth, and health. Almost 10,000 pregnant women living in Rotterdam, the Netherlands, were enrolled in the study between 2002 and 2006. Data from the children and caregivers was collected at several time points and written informed consent and/or assent was obtained from all participants. The imaging protocol and quality assessment is extensively described by White, Muetzel and colleagues (White et al., 2018). MRI scans were acquired in 3 waves using 2 different scanners, making the cohort an ideal validation set to investigate the transfer of hyperparameters from a reference dataset to an unseen target set. In longitudinal studies running over several years, changes of scanner hardware, software and/or scan protocols are often inevitable, which generates a need to correct for the resulting scanner effects. In the first wave, 1,033 participants (484 female, age range: [6-10]) were imaged with a 3T MR750 Discovery MRI scanner, while in the second (n=3,920, 1,977 female, age range: [9-12]) and third wave (n=3,570, 1,866 female, age range: [13-17]) a 3T MR750w Discovery scanner

(General Electric, Milwaukee, WI, USA) was used. After exclusion of scans with incidental findings (n=58), braces (n=1067), and low-quality visual inspection ratings of FreeSurfer reconstructions (n=2067), a total of 6,285 scans were included in the target dataset. Figure 1 shows a histogram of age and scanner distributions in the target dataset.

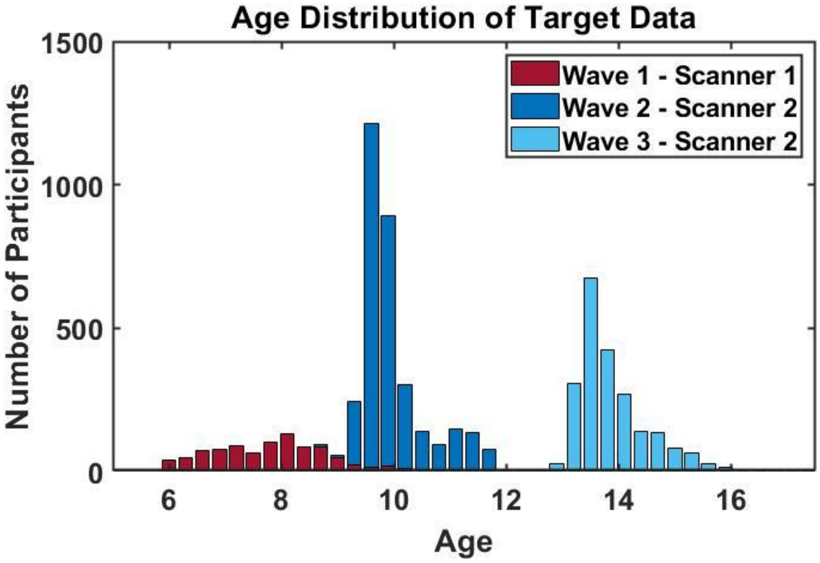

Figure 1. Histogram of the scanning waves and age distributions in the Generation R target dataset.

2.2. Transfer of hyperparameters from reference models to target cohort

By making use of the Generation R study cohort, we set out to show the advantage of transferring the hyperparameters to an unseen site by 1) determining the optimal number of samples needed for adaptation to the target cohort, 2) validating the recalibration of data to the target cohort and successful removal of site-effects by comparing raw and scanner corrected values, and 3) illustrating the utility of site-corrected deviations scores to uncover changes in morphology between groups and individuals. In the following, these three aims are described in more detail.

2.2.1. Optimal sample size for parameter adaptation

In order to determine the optimal number of samples in the adaptation set we leveraged the large amount of data available in the Generation R dataset. As described above, to prevent bias, heldout data should be used for adapting the parameters of the normative model to the target cohort (see Kia et al., 2020, 2022 for details). The number of scans in the adaptation set was varied ranging from 5 to 300 scans and model metrics (explained variance, SMSE, MSLL) of the subsequent models was calculated for each sample size. The resulting information is particularly useful for small imaging cohorts, since cohorts with smaller sample sizes can employ the current approach to boost power by making use of the hyperpriors inferred from large data. Yet, this is only viable if the samples needed to recalibrate the models can be kept to an optimal minimum.

2.2.2 Validation of adaptation

Additionally, two aspects of the Generation R study design make the cohort an ideal target set to validate the successful recalibration of the normative models to an unseen site. First, scans of participants that have repeated measurements over all 3 scanning time points are present. Uncorrected CT values of a participant with scans across all 3 measurement waves (and therefore across both scanners) show heterogeneity over time points that is partly due to changes over time and partly due to confounding site-effects. After successful recalibration of the normative model, we expect resulting z-scores, which are in principle free of site-effects, of the same participant to be in a similar range while raw values will differ. Second, there is an overlapping age range (8,6 - 10,7 years of age), in which scans from both scanners were obtained (Fig. 1). Z-scores of participants from wave 1 (scanner 1), that fall in the overlapping age range of wave 1 and wave 2 should be distributed similarly after recalibration as z-scores in the same age range of wave 2 (scanner 2) while raw, uncorrected values differ due to scanner effects. Therefore, scans of participants with measurements at all 3 scanning time points (n=1,317) and scans from the first imaging wave that fall in the overlapping age range (n=211) were withheld from the adaptation set that was used to recalibrate the reference normative model to the new unseen site. As outlined above, these scans hold valuable information that will be used to determine the successful calibration of the models by comparing raw CT before adaptation and corrected estimates after adaptation.

2.2.3 Clinical application of normative estimates

Lastly, we used the resulting site-effect free estimates to illustrate their potential to uncover morphological deviations in clinical cohorts by contrasting estimates in CT per ROI between participants in the Generation R cohort born pre-term (gestational age < 37 weeks, n=339) and children born at term (n=5,646). Pre-term birth interrupts a vulnerable period for brain development, as processes such as synaptogenesis, axonal growth, and neuronal migration, take place during the third semester (Volpe, 2009). Therefore, deviation scores from the normative models can for instance be used to explore the variability in CT within children born pre-term, but also to find ROIs that differ between children born pre-term and at term. Notably, these deviation scores are free of site-effects and therefore especially suited for longitudinal MRI designs, as it is the case with the Generation R study.

3. Results

3.1. Transfer results

3.1.1 Optimal number of samples for parameter adaptation

We first determined the optimal number of subjects needed in the adaptation set. Figure 2 shows evaluation metrics for each ROI as the sample size of the adaptation set increases. Performance of the model reaches a plateau around 100 subjects. We thus adapted the initial reference models to the unseen sites of the Generation R study on n=300 (4.8%) (n=100 for scanner 1 in wave 1; n=200 for scanner 2 in wave 2 and 3) and tested the models on the remaining participants (n=5,985; n=813 for scanner 1 in wave 1, n=5,172 for scanner 2 in wave 2 and 3). Subjects from wave 1 and 3 were sampled randomly, whereas subjects from wave 2 were sampled pseudo-randomly to ensure a uniform cover of the full range of the narrow and highly peaked age distribution in this wave (Fig. 1). While model performance reached a performance ceiling at approximately 100 scans per scanner/wave in the adaptation sample, only slight concessions in model performance are present as adaptation sample size decreases to only 25 scans.

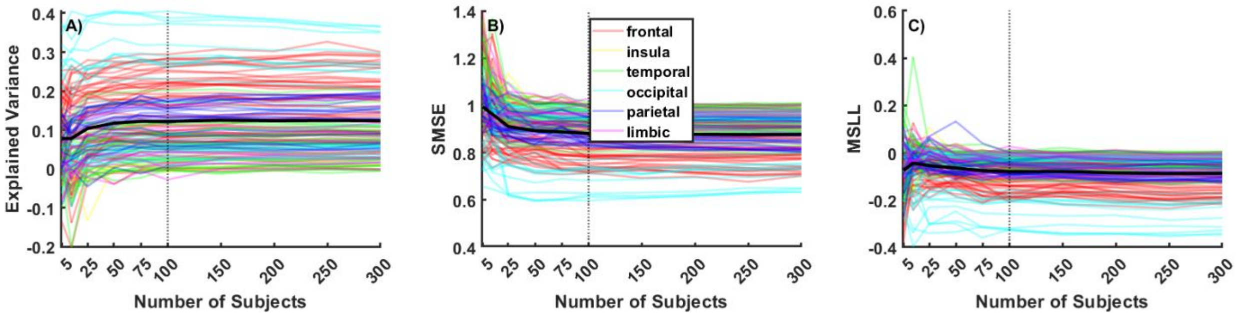

Figure 2.: Comparison of model performance as the number of subjects in adaptation set increases. Colored lines show evaluation metrics per ROI, color coded according to cerebral area. The black line illustrates the mean across all ROIs. Model performance reaches a plateau at approximately 100 scans per wave in the adaptation sample (vertical dotted line).

3.1.2 Adaptation settings

We furthermore tested different adaptation settings. Despite the fact that scans in wave 2 and 3 of the target cohort were acquired on the same scanner, we compared adaptation settings treating waves 2 and 3 as the same but also as different sites. When treated as same sites, we found a slight bias for higher deviation scores (z-scores) when running the adaptation to a wave 2+3 test set with wave 3 subjects compared to wave 2 subjects only, in particular for frontal ROIs. The effect of the different adaptation settings on all ROIs is shown in Supplementary Figure 1 and is explicitly illustrated in an example ROI in Figure 3. Panel A shows that the model is more successful in reparametrizing the raw data to centiles when each time point of measurement is handled as a separate site-effect. Possible sources for such effects might stem from changes in scanner software, changes in image quality with age (i.e. motion artifacts), or sample variability. In our target cohort, scanner software was upgraded after the first 370 scans of wave 2 but was otherwise identical in wave 2 and wave 3. However, age-related improvements in images quality are frequently reported in the literature and quality assurance, measured as topological defects in the surface reconstruction for FreeSurfer processed MRI data (https://github.com/Deep-MI/qatools-python), does show improvements in image quality with age across the three waves ($\text{Mean}_{\text{wave 1}}$=229.06, $\text{SD}_{\text{wave 1}}$=98.56; $\text{Mean}_{\text{wave 2}}$=213.89, $\text{SD}_{\text{wave 2}}$=67.15; $\text{Mean}_{\text{wave 3}}$=166.97, $\text{SD}_{\text{wave 3}}$=67.15).

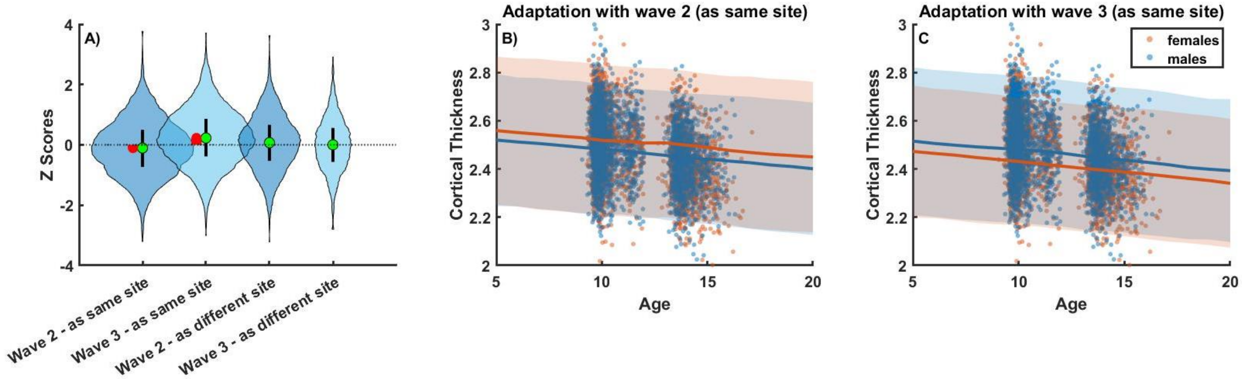

Figure 3. Effects of different recalibration configurations on the target cohort illustrated in an example ROI (inferior frontal sulcus). Panel A) shows z-score distributions when measurement waves 2 and 3 of the Generation R target cohort are treated as the same (same scanner) or different sites in a frontal example ROI, the inferior frontal sulcus. Median and interquartile range are represented by green dots and black bars, respectively. For each measurement wave, we would expect the median z-score to be around 0. However, this is not the case if measurement wave 2 and 3 are treated as the same site. The difference from 0 is indicated by red bars. By examining the CT trajectories in panel B) and C), we see that this might be due to a misestimation of mean and variance in females when both waves are treated as the same site.

3.1.3. Adaptation Validation

After choosing for a adaptation setting treating the three measurement waves as different batch effects, we validated the success of the adaptation of the reference model to the target cohort by examining the differences between raw CT values and corrected deviations (z-scores) after transfer of the subjects which were withheld from the adaptation sets (Fig. 4). Scans of participants with repeated measurements at all imaging waves (a random sample of ten participants is depicted by green lines) show a decline over time in raw CT. As expected, thinning of the cortex can be observed with age, however, the raw CT values are confounded by noise stemming from site-effects of the different measurement waves. In the resulting z-scores of the withheld subjects, these site-effects are removed as demonstrated by stable deviations from the normative model within a participant (Fig. 4B). The same holds true for the withheld subjects from measurement wave 1 that fall in the overlapping age range (8,6 - 10,7 years of age) of wave 1 (scanner 1) and 2 (scanner 2). While raw CT values in the overlapping age range vary vastly between the two measurement waves ($t(2874)=13.4$, $p<0.001$), with a tendency of higher values in measurement wave 1 compared to wave 2 (Fig. 4C), this difference is slightly reduced when correcting for sex (Fig. 4D) ($t(2874)=11.4$, $p<0.001$) and practically absent in the sex- and additionally site-effect corrected z-scores (Fig. 4E) ($t(2874)=1.0$, $p=0.324$). Therefore, we can meaningfully compare individuals on the basis of z-scores, bearing in mind that the z-scores are defined with respect to a lifespan based normative model.

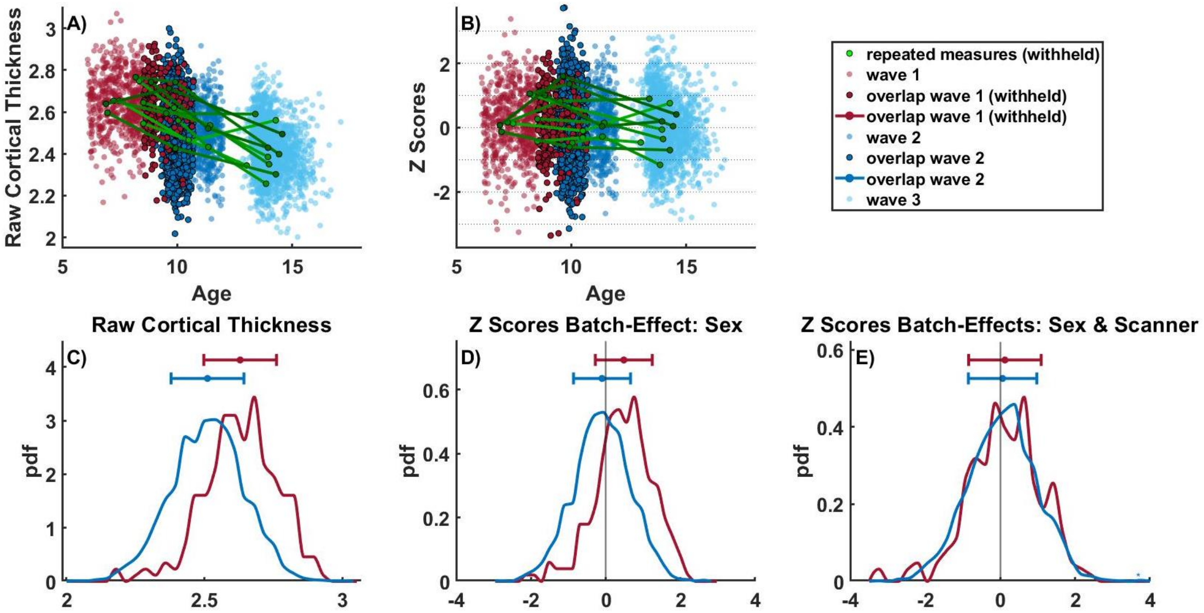

Figure 4: Validation of transferring the reference normative model to the target cohort using two groups of subjects that were withheld from the adaptation set: 1) subjects with repeated measurements at all 3 imaging waves (random sample of ten participants depicted by green lines, panels A and B; 2) subjects from imaging waves 1 and 2 that fall in the overlapping age range of both scanners [8,6 - 10,7] (depicted by darker shaded red and blue dots and lines, panels A-E). Panel A and C show raw cortical thickness values. Panels B, D, and E show sex-effect (panel D) or sex- and site-effect corrected z-scores of the same participants (panels B&E). For consistency, the same ROI (inferior frontal sulcus) as in the previous figures is illustrated.

3.2. Relating site-effect corrected z-scores to gestational age

To illustrate the usefulness of the resulting models, we compared extreme deviations, acquired at the level of individuals, between children born pre-term and children born at term in the target cohort. Percentages of individuals with an extreme z-score (larger/smaller than 2) per ROI are shown in Figure 5. In the children born at term, we find approximately 2.5% of children with extreme negative and extreme positive z-scores respectively across ROIs. Exceptions are primarily smaller ROIs (sulcus intermedius primus (left & right), posterior ramus of the lateral sulcus (left), anterior transverse collateral sulcus (right), orbital sulcus (right)) where areas with thicker cortices than expected can be observed. Importantly, extreme deviations are much more prevalent with children born pre-term with the most pronounced extreme positive deviations (thicker cortex than expected) found in the left pericallosal sulcus and lateral aspect of the superior temporal gyrus, as well as in the anterior part of cingulate gyrus and sulcus (ACC) of both hemispheres. The most striking extreme negative deviations (thinner cortex than expected) can be seen on the left hemisphere in the superior and inferior temporal sulcus, lingual sulcus, superior part of the precentral sulcus, supramarginal gyrus, and on the right hemisphere in the superior and inferior part of the precentral sulcus, superior frontal sulcus, angular gyrus, precentral gyrus and the precuneus.

These regions are consistent with previous findings on CT differences in adolescents born pre-term. Pronounced cortical thinning has been found persistently in areas surrounding the central sulcus and temporal lobes (Martinussen et al., 2005; Nagy et al., 2011; Zubiaurre-Elorza et al., 2012) as well as thicker cortices in frontal regions surrounding the anterior cingulate cortex (Bjuland et al., 2013). The current approach has been shown to capture structural deviations better than case-control studies as they are more sensitive to individual heterogeneity (Remiszewski et al., 2022). It also offers improved insights in longitudinal cohorts, as these deviation scores are not cofounded by site-effects.

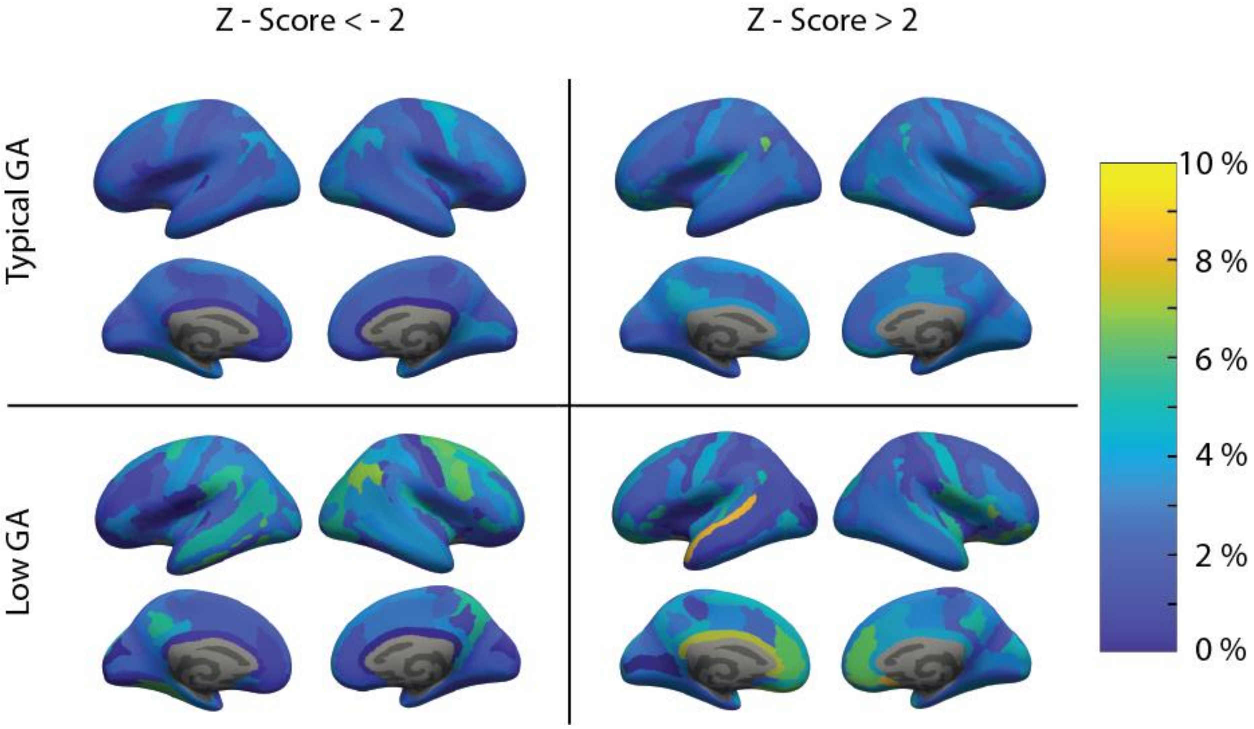

Figure 5.: Differences in site-effect corrected z-scores between children born pre-term (low gestational age (GA)) and children born term (typical gestational age (GA)). On the left side, extreme negative deviations (cortex thinner than expected) are illustrated. On the right, extreme positive deviations (cortex thicker than expected) are shown.

4. Discussion

In this study, we used information from normative models that were initially trained on a large number of samples, scanned over 77 sites, as prior knowledge for the parameters of the CT distributions when adapting these models to the two scanners of the longitudinal Generation R study.

We report three main findings: first, transfer learning is successful and allows for meaningful comparisons between individuals from different scanners, and sexes, as previously reported (Kia et al., 2022). Second, we quantified the number of samples in the transfer set needed to obtain good performance metrics on the test set and show that relatively few samples are sufficient for good performance (approximately n=25). This provides the added benefit of improving the statistical power of statistical analyses on the resulting larger test set. While we used 100 samples per measurement wave in the adaptation site, slightly smaller adaptation samples decreased the evaluation metrics only marginally. Third, we show that the deviations from these normative models are meaningful in that they are altered in a highly individualized manner in individuals born pre-term.

Our results support the finding that normative models capture the general trend of decreasing cortical thickness with age, as reported in previous studies (Bethlehem et al., 2022; Frangou et al., 2022; Rutherford, Fraza, et al., 2022; Tamnes et al., 2017; Thambisetty et al., 2010). Interestingly, we found that the model performed better when each measurement wave of the transfer cohort was treated as a separate site-effect, even though two of three waves were acquired on the same scanner. This could be due to sample variability or a misestimation of parameters in the female cohort, possibly linked to the fact that scan quality tends to improve with age. For future studies, it may be useful to treat distinct measurement intervals as separate batch-effects, resulting in a factorial design of sex x scanners x waves, even if the scanner setup has not changed, to produce more precise models. Our recommendations might differ for longer timescales, non-linear or non-Gaussian lifespan trajectories, which usually requires more data (de Boer et al., 2022). However, the methods we introduce can be used to determine the optimal number of subjects for such cases.

The successful validation of the use of transfer learning with normative models opens the door for further investigations exploring the relationship between deviation scores and various phenotypes. Individual-level deviations, as obtained through normative models, have been shown to provide stronger effects than typical case-control studies using uncorrected raw measurements (Rutherford, Fraza, et al., 2022) and are therefore particularly suitable for exploring and investigating individual differences within and across datasets. The used federated learning framework makes it possible to use the models presented in this work as informed priors (models are available online via PCNportal [https://pcnportal.dccn.nl/] to investigate CT in smaller and/or clinical cohorts.

In this study, we provide an example how normative models can be used to investigate clinical phenotypes, by investigating the relation between extreme deviations scores and the gestational age at birth, that is between children born at-term and pre-term. While children born at-term show an expected distribution of approximately 2.5% with z-scores higher than 2 or lower than -2 respectively, children born pre-term are more likely to have extreme deviations in specific ROIs, which are consistent with previous literature showing pronounced differences in particular in frontal and temporal cortices. While we show a comparison between groups, the current approach does not require clustering of individuals into groups but instead can be used to make inferences about heterogeneity within clinical groups as well as about deviations on an individual level.

4.1. Limitations and future directions

Although we demonstrate that evaluation metrics level off after 100 scans in the adaptation set, as few as 25 scans can still lead to effective transfer of knowledge. However, this limitation prevents small cohorts from utilizing the current approach due to the fact that the median sample size of neuroimaging studies typically includes 25 participants (Marek et al., 2022).
Furthermore, our work estimates normative models on a single ROI, thereby neglecting any spatial interdependencies between brain regions. Moreover, other image-derived phenotypes, such as the cerebellum, could also be considered.

5. Conclusion

Using longitudinal cortical thickness data from the Generation R study on children aged 6 to 17 years old, we present an application of transfer learning of large-scale normative models which produce good performance metrics with even a limited size of adaptation set. The resulting deviation scores per age and ROIs, allow for meaningful comparison inter sites and inter sex.

Using these obtained deviation scores, we were able to show specifically localized differences in cortical thickness between children born pre-term and children born at-term.

References

Antoniades, M., Haas, S. S., Modabbernia, A., Bykowsky, O., Frangou, S., Borgwardt, S., & Schmidt, A. (2021). Personalized Estimates of Brain Structural Variability in Individuals with Early Psychosis. *Schizophrenia Bulletin*, *47*(4), 1029–1038. https://doi.org/10.1093/schbul/sbab005

Bayer, J. M. M., Dinga, R., Kia, S. M., Kottaram, A. R., Wolfers, T., Lv, J., Zalesky, A., Schmaal, L., & Marquand, A. (2021). Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. *BioRxiv*, 2021.02.09.430363. https://www.biorxiv.org/content/10.1101/2021.02.09.430363v2%0Ahttps://www.biorxiv.org/content/10.1101/2021.02.09.430363v2.abstract

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

Bjuland, K. J., Løhaugen, G. C. C., Martinussen, M., & Skranes, J. (2013). Cortical thickness and cognition in very-low-birth-weight late teenagers. *Early Human Development*, *89*(6), 371–380. https://doi.org/10.1016/j.earlhumdev.2012.12.003

Clarkson, M. J., Cardoso, M. J., Ridgway, G. R., Modat, M., Leung, K. K., Rohrer, J. D., Fox, N. C., & Ourselin, S. (2011). A comparison of voxel and surface based cortical thickness estimation methods. *NeuroImage*, *57*(3), 856–865. https://doi.org/10.1016/j.neuroimage.2011.05.053

Cole, T. J. (2012). The development of growth references and growth charts. *Annals of Human Biology*, *39*(5), 382–394. https://doi.org/10.3109/03014460.2012.694475

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. II. Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage*, *9*(2), 179–194. http://linkinghub.elsevier.com/retrieve/pii/S1053811998903950%0Ahttp://surfer.nmr.mgh.harvard.edu/ftp/articles/fischl99b-recon2.pdf de Boer, A. A. A., Kia, S. M., Rutherford, S., Zabihi, M., Fraza, C., Barkema, P., Westlye, L. T., Andreassen, O. A., Hinne, M., Beckmann, C. F., & Marquand, A. (2022). Non-Gaussian Normative Modelling With Hierarchical Bayesian Regression. *BioRxiv*, 2022.10.05.510988. https://doi.org/10.1101/2022.10.05.510988

DeLisi, L. E. (2008). The concept of progressive brain change in schizophrenia: Implications for understanding schizophrenia. *Schizophrenia Bulletin*, *34*(2), 312–321. https://doi.org/10.1093/schbul/sbm164

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, *97*(20), 11050–11055. https://doi.org/10.1073/pnas.200033797

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 195–207. http://www.ncbi.nlm.nih.gov/pubmed/9931269

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

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

Fuhrmann, D., Madsen, K. S., Johansen, L. B., Baaré, W. F. C., & Kievit, R. A. (2022). The midpoint of cortical thinning between late childhood and early adulthood differs across individuals and regions : Evidence from longitudinal modelling in a 12-wave sample. *BioRxiv*, *261*(March), 1–23. https://doi.org/10.1016/j.neuroimage.2022.119507

Hutton, C., Draganski, B., Ashburner, J., & Weiskopf, N. (2009). A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage*, *48*(2), 371–380. https://doi.org/10.1016/j.neuroimage.2009.06.043

Insel, T. R. (2014). Mental Disorders in Childhood. *JAMA*, *311*(17), 1727. https://doi.org/10.1001/jama.2014.1193

Jaddoe, V. W. V., Mackenbach, J. P., Moll, H. A., Steegers, E. A. P., Tiemeier, H., Verhulst, F. C., Witteman, J. C. M., & Hofman, A. (2006). The Generation R Study: Design and cohort profile. *European Journal of Epidemiology*, *21*(6), 475–484. https://doi.org/10.1007/s10654-006-9022-0

Kia, S. M., Huijsdens, H., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2020). *Hierarchical Bayesian Regression for Multi-Site Normative Modeling of Neuroimaging Data*. 1–12. http://arxiv.org/abs/2005.12055

Kia, S. M., Huijsdens, H., Rutherford, S., de Boer, A., Dinga, R., Wolfers, T., Berthet, P., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2022). Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression. *PLOS ONE*, *17*(12), e0278776. https://doi.org/10.1371/journal.pone.0278776

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van IJzendoorn, M. H., de Jongste, J. C., Klaver, C. C. W., van der Lugt, A., Mackenbach, J. P., Moll, H. A., Peeters, R. P., Raat, H., Rings, E. H. H. M., Rivadeneira, F., van der Schroeff, M. P., Steegers, E. A. P., Tiemeier, H., Uitterlinden, A. G., … Jaddoe, V. W. V. (2016). The Generation R Study: design and cohort update 2017. *European Journal of Epidemiology*, *31*(12), 1243–1264. https://doi.org/10.1007/s10654-016-0224-9

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *August 2022*. https://doi.org/10.1038/s41586-022-04492-9

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. F., Wolfers, T., Mennes, M., Buitelaar, J., & Beckmann, C. F. (2016). Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 433–447. https://doi.org/10.1016/j.bpsc.2016.04.002

Martinussen, M., Fischl, B., Larsson, H. B., Skranes, J., Kulseng, S., Vangberg, T. R., Vik, T., Brubakk, A.-M., Haraldseth, O., & Dale, A. M. (2005). Cerebral cortex thickness in 15-year-old adolescents with low birth weight measured by an automated MRI-based method. *Brain*, *128*(11), 2588–2596. https://doi.org/10.1093/brain/awh610

Mills, K. L., Siegmund, K. D., Tamnes, C. K., Ferschmann, L., Wierenga, L. M., Bos, M. G. N., Luna, B., Li, C., & Herting, M. M. (2021). Inter-individual variability in structural brain development from late childhood to young adulthood. *NeuroImage*, *242*(August), 118450. https://doi.org/10.1016/j.neuroimage.2021.118450

Nagy, Z., Lagercrantz, H., & Hutton, C. (2011). Effects of Preterm Birth on Cortical Thickness Measured in Adolescence. *Cerebral Cortex*, *21*(2), 300–306. https://doi.org/10.1093/cercor/bhq095

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pereira, J. B., Ibarretxe-Bilbao, N., Marti, M. J., Compta, Y., Junqué, C., Bargallo, N., & Tolosa, E. (2012). Assessment of cortical degeneration in patients with Parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. *Human Brain Mapping*, *33*(11), 2521–2534. https://doi.org/10.1002/hbm.21378

Remiszewski, N., Bryant, J. E., Rutherford, S. E., Marquand, A. F., Nelson, E., Askar, I., Lahti, A. C., & Kraguljac, N. V. (2022). Contrasting Case-Control and Normative Reference Approaches to Capture Clinically Relevant Structural Brain Abnormalities in Patients With First-Episode Psychosis Who Are Antipsychotic Naive. *JAMA Psychiatry*, *79*(11), 1133. https://doi.org/10.1001/jamapsychiatry.2022.3010

Rogers, J. C., & de Brito, S. A. (2016). Cortical and subcortical gray matter volume in youths with conduct problems ameta-Analysis. *JAMA Psychiatry*, *73*(1), 64–72. https://doi.org/10.1001/jamapsychiatry.2015.2423

Rutherford, S., Fraza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L. K. M., Bayer, J. M. 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*, 1–15. 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*(July). https://doi.org/10.1038/s41596-022-00696-5

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. *Journal of Neuroscience*, *37*(12), 3402–3412. https://doi.org/10.1523/JNEUROSCI.3302-16.2017

Thambisetty, M., Wan, J., Carass, A., An, Y., Prince, J. L., & Resnick, S. M. (2010). Longitudinal changes in cortical thickness associated with normal aging. *NeuroImage*, *52*(4), 1215–1223. https://doi.org/10.1016/j.neuroimage.2010.04.258

Volpe, J. J. (2009). Brain injury in premature infants: a complex amalgam of destructive and developmental disturbances. In *The Lancet Neurology* (Vol. 8, Issue 1, pp. 110–124). https://doi.org/10.1016/S1474-4422(08)70294-1

White, T., Muetzel, R. L., el Marroun, H., Blanken, L. M. E., Jansen, P., Bolhuis, K., Kocevska, D., Mous, S. E., Mulder, R., Jaddoe, V. W. V., van der Lugt, A., Verhulst, F. C., & Tiemeier, H. (2018). Paediatric population neuroimaging and the Generation R Study: the second wave. *European Journal of Epidemiology*, *33*(1), 99–125. https://doi.org/10.1007/s10654-017-0319-y

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 and Adolescent Psychiatry*, *59*(1), 177–185. https://doi.org/10.1016/j.jaac.2019.04.006

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

Zhao, Y., Zhang, Q., Shah, C., Li, Q., Sweeney, J. A., Li, F., & Gong, Q. (2022). Cortical Thickness Abnormalities at Different Stages of the Illness Course in Schizophrenia. *JAMA Psychiatry*, *79*(6), 560. https://doi.org/10.1001/jamapsychiatry.2022.0799

Zubiaurre-Elorza, L., Soria-Pastor, S., Junque, C., Sala-Llonch, R., Segarra, D., Bargallo, N., & Macaya, A. (2012). Cortical Thickness and Behavior Abnormalities in Children Born Preterm. *PLoS ONE*, *7*(7), e42148. https://doi.org/10.1371/journal.pone.0042148