Normative Modelling for Analysis of Brain Development Trajectory

Carlotta Sophia Hölzle[1]

Abstract

Creating reference models for population variations and identifying individual deviations are vital for understanding neurological disorders. This paper reviews normative modelling approaches for analyzing these developmental changes, focusing on algorithms and techniques for estimating anatomical parameters and detecting deviations from typical patterns. In the second part of the study, three different normative modeling approaches are compared and assessed based on their quantitative and qualitative performance. The findings highlight the strengths and limitations of each approach, offering valuable insights into their practical applications in both clinical and research settings.

1. Introduction

Numerous neurological disorders originate from atypical brain development, making monitoring and predicting changes in neuronal structures essential (Insel, 2014). While traditional group-level statistics can help identify broad trends and significant differences between populations in neurological disorder analysis, they fall short due to overlapping distributions between cases and controls, leading to unclear diagnoses. Pure MRI data can help diagnose and analyze neurological diseases, like schizophrenia, bipolar disorder, autism, and ADHD (Frisoni et al., 2010; Insel, 2014; Rutherford et al., 2022), by comparing them with reference markers from healthy populations. However, variability among evaluators, demographic factors, and image processing methods limit the broad application of reference models, emphasizing the need for a more standardized, individualized approach (Bethlehem et al., 2022).

To address these limitations, models predicting brain age have been used in clinical applications, reducing complex MRI and demographic data to a single number — the difference between a person's age and their predicted brain age (Ziegler et al., 2014; Brewer, 2009). However, a drawback of these models is their reliance on a group-based approach, which may overlook individual abnormalities. Clustering models that predict typical and atypical neuronal groups also assume clear divisions in clinical populations and fail to account for individual variability (Marquand et al., 2016b).

Contrarily, normative models provide a personalized approach to identifying individual patterns in brain development and aging rather than relying on group averages or clinical divisions. This approach has been applied in psychiatric research, offering insights into how mental disorders deviate from expected developmental trajectories (Marquand et al., 2016b; Erus et al., 2015; Bethlehem et al., 2018a), for conditions such as schizophrenia (Wolfers et al., 2018), attention-deficit/hyperactivity disorder (ADHD) (Wolfers et al., 2020), and autism (Bethlehem et al., 2018a; Zabihi et al., 2019). This research first examines different approaches to building normative models for detecting atypical neuronal developments. It then analyzes various methods for using the outputs from normative models for downstream analysis. Lastly, it compares a Bayesian linear regression approach to a GAM and GAMLSS model.

2. Normative Modelling

Normative models are statistical techniques used in neuroscience and developmental research to establish baselines or ’norms’ for developmental changes. In neurology, these models predict anatomical or functional brain values across different age ranges, covariates are typically age, sex, and MRI scanner site. Normative models generate estimates for healthy individuals over time, enabling the creation of reference growth curves. These curves can highlight abnormal development by comparing individual data to established healthy development graphs, including percentiles. Additionally, the predicted values can be used in further downstream analysis, integrating them with model or data uncertainties. By integrating variability into their framework, normative models provide insights into deviations from typical developmental trajectories, offering a more dynamic approach compared to traditional case-control studies that rely on static, single-point comparisons using basic statistics. However, significant challenges remain, such as the need for well-defined reference models capable of quantifying variability across the lifespan and ensuring the

[1] Technical University of Munich, Germany. Correspondence to: Carlotta Sophia Hölzle <carlotta.hoelzle@tum.de>.

comparability of results across different studies.

3. Estimation of Normative Models

While normative models are fundamentally mathematical constructs, various approaches exist for their estimation. This section explores the different mathematical methods used to estimate normative models in brain development research. The approaches are categorized into linear, non-linear, and Bayesian/hierarchical methods, with specific examples from the literature provided for each category.

3.1. Linear Models

Linear models assume a direct relationship between the dependent variable y and one or more independent variables x.

Gur et al. (2014) utilized regression analysis to predict individuals' neurocognitive age $\hat{y}$. The linear regression model is mathematically expressed as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where $x_1, ..., x_n$ are different cognitive performance measures, $\beta_0, ..., \beta_n$ are coefficients estimated by the model, and ϵ is the error term. The authors used the predictions to explore how each person's cognitive abilities compare to expected age-related norms, providing insights into both typical and atypical neurocognitive development. However, the pure linearity of their model, while interpretable and effective in making predictions, is not capable of detailed analysis of the overall variability within the dataset nor does it allow for anatomic individual-level statistical predictions.

3.1.1. Quadratic Regression

Polynomial regression analysis models the relationship between an independent variable and a dependent variable with an nth-degree polynomial, capturing non-linear relationships. Quadratic fit lines are a type of polynomial regression that adds a quadratic term to linear regression to model parabolic relationships.

Kessler et al. (2016) used quadratic regression to model the expression scores of brain connectivity components as functions of age. The model is represented mathematically as follows:

$$\hat{y_i} = \beta_0 + \beta_1 * age_i + \beta_2 * age_i^2 + \epsilon_i$$

Where $\hat{y_i}$ is the expression score of a component for participant i, age_i the age, $\beta_0, \beta_1, \beta_2$ are the coefficients to be estimated, and ϵ is the error term.

The covariates sex and motion were considered and removed, as the authors found none of these to exhibit a statistically significant age difference by sex interaction. With this model the authors build growth charts for brain development, which can identify deviations from expected growth patterns and their association with cognitive outcomes like attention performance and ADHD diagnosis. A limitation of this study is the machine learning-based data reduction method, which relies on parameter-tuning, and the assumptions of a non-Gaussian component distribution. The small sample size and abbreviated diagnostic interviews for ADHD patients result in reduced generalizability to clinical populations diagnosed through more standard methods.

3.1.2. Partial Least Squares Regression (PLSR)

PLSR extends linear regression by modelling relationships between observed variables through projections into a new space of latent variables, which optimally explain the covariance between dependent and independent variables.

Huizinga et al. (2018) employ PLSR to correlate brain deformations, derived from image registration, with subjects' ages. This method identifies age-related patterns in brain morphology and generates morphology scores, serving as basis for constructing percentile curves. These curves showed group differences between the morphology score distributions of the CN and AD subgroups. However, the diagnostic value of the morphology score alone is limited due to high individual variability and unaccounted registration errors.

3.1.3. Generalized Regression Models

Generalized regression models are an extensions of linear regression models. They allow for different types of response variables by using a link function, which linearly connects the expected value of the response to the predictors.

Gur et al. (2014) explored improving their traditional regression analysis by incorporating a General Adaptive Model (GAM) to address non-linearities through smooth functions of predictors. Their predictive model for neurocognitive age $\hat{y}$ is represented as:

$$\hat{y} = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_n(x_n) + \epsilon$$

Where $x_1, ..., x_n$ are cognitive performance measures, $f_0, ..., f_n$ are smooth functions applied to the predictors (including linear, squared, or more complex non-linear terms), β_0 is the intercept, and ϵ is the error term. The authors decided to not incorporate the GAM approach, as it did not significantly improve prediction accuracy.

3.1.4. Support Vector Regression (SVR)

SVR fits a regression line within a specified error margin, handling both linear and non-linear relationships through various kernels. Despite its flexibility, SVR is fundamentally a linear regression model as it finds a linear relationship in the transformed feature space defined by the kernel.

Erus et al. (2015) used SVR to predict brain development indices (BDI) from multimodal MRI data using LIBSVM (Chang & Lin, 2011). Separate models for males and females examined sex-specific brain development. The model is mathematically modelled as:

$$\hat{y} = w^T \Phi(x) + b$$

Where y is the predicted BDI, w the weight vector, x the input features, Φ a linear kernel, and b the bias term.

The study showed that BDI could create brain growth charts for detecting deviations in brain development, potentially identifying early neuropsychiatric biomarkers. Limitations included a focus on cognitive and structural brain changes, necessitating further validation in clinical settings and other datasets. SVR's computational demands grow with dataset size and kernel complexity, requiring careful parameter selection.

3.2. Non-Linear Models

Non-linear models capture complex relationships that cannot be represented by linear models by allowing for interactions and dependencies between variables that are not constrained to a straight-line relationship.

3.2.1. Local Polynomial Regression Fitting (LOESS)

LOESS fits low-degree polynomials to subsets of data, allowing for local modelling of complex relationships within datasets. Bethlehem et al. (2018b) utilized LOESS to analyze cortical thickness patterns across different scanner sites and demographics, employing weighted data points based on proximity to create multiple local models for predicting cortical thickness. LOESS provided age-specific mean and standard deviation of cortical thickness scores for various brain regions. The authors also used a linear mixed effects model, incorporating age as a significant factor influencing brain structure. Scanner site showed the highest variance in cortical thickness, followed by age and full-scale IQ, with other factors contributing minimally, underscoring the divers impact of technical and biological factors. However, LOESS's sensitivity to local data variations poses limitations particularly in sparse or uneven datasets where it may overfit or fail to capture global trends effectively.

3.2.2. Generalized Additive Models for Location, Scale, and Shape (GAMLSS)

GAMLSS models flexibly characterize the mean, variance, and shape of response variables, providing a scale parameter for each data point, effectively adjusting it to local variability. Bethlehem et al. (2022) applied GAMLSS to create normative brain charts, expanding on the work by Stasinopoulos & Rigby (2008). Their models used MRI data, demographic factors, and scanner software version to map non-linear growth trajectories for brain volumes like grey matter volume (GMV). These trajectories captured age, sex, and software version effects using fractional polynomials and random effects to account for individual variability.

The model for GMV is:

$$\begin{aligned}
GMV &\sim \text{GeneralizedGamma}(\mu, \sigma, \nu) \\
\log(\mu) &= \alpha_\mu + \alpha_{\mu,\text{sex}}(\text{sex}) + \alpha_{\mu,\text{ver}}(\text{ver}) \\
&\quad + \beta_{\mu,1}(\text{age})^{-2} + \beta_{\mu,2}(\text{age})^{-2} + \beta_{\mu,3}(\text{age})^{-2}\log(\text{age})^2 \\
&\quad + \gamma_{\mu,\text{study}} \\
\log(\sigma) &= \alpha_\sigma + \alpha_{\sigma,\text{sex}}(\text{sex}) + \beta_{\sigma,1}(\text{age})^{-2} \\
&\quad + \beta_{\sigma,2}(\text{age})^3 + \gamma_{\sigma,\text{study}} \\
\nu &= \alpha_\nu
\end{aligned}$$

Predictions and centile scores from the GAMLSS models were used to construct a multidimensional view of brain development. The authors identified key developmental milestones like growth trajectory peaks, and enabled detailed comparisons across demographic and clinical groups. However, focusing on global brain features and cross-sectional data limits clinical applicability and subtle analysis of neurodegenerative processes.

3.3. Bayesian & Hierachical Models

These models incorporate prior distributions, hierarchical structures, and inherent uncertainty estimation, for more complex data relationships.

3.3.1. Bayesian Linear Regression

Bayesian linear regression integrates model and data variability, prior beliefs and observed data to produce posterior distributions, inherently quantifying uncertainty in predictions. Rutherford et al. (2022) applied Bayesian linear regression to study brain structure changes, using a warped Bayesian framework with B-spline basis for age to capture non-linear data patterns. Their model estimates cortical thickness and subcortical volumes while accounting for model and data uncertainty.

The general mathematical formulation is:

$$\hat{y}, \sigma = w^T \Phi(x) + \epsilon$$

Here, σ is the estimated variance, w^T is the transposed weight vector learned by the model during training, $\Phi(x)$ is the B-spline basis expansion including age, sex, and scanner site, and $\epsilon = \nu(0, \beta)$ represents normally distributed noise with zero mean and precision β.

The study highlights the probabilistic nature of brain development across individuals, noting limitations such as data bias towards western societies and the UKBiobank, suggesting future research include diverse demographics for broader model applicability and consider random effects for site variation.

3.3.2. Gaussian Process Regression (GPR)

GPR is fundamentally a Bayesian approach. It places a prior over the space of possible functions and updates this prior with observed data to obtain a posterior distribution, and providing uncertainty estimations at every input point. GPR inherently models non-linear relationships without the need for predefined basis functions, thus is sometimes categorized as non-linear model.

Studies have demonstrated GPR's effectiveness in modeling various normative response variables like cortical thickness, brain activity, brain volume, and different brain regions' volumes (Ziegler et al., 2014; Wolfers et al., 2018; Marquand et al., 2016a; Gur et al., 2014; Wolfers et al., 2020).

Marquand et al. (2016a) used GPR to study the relationship between trait impulsivity, measured by delay discounting, and reward-related brain activity across regions. For each input covariate set, the model predicts a normative response variable like brain activity, providing a mean prediction ($\hat{y}$) and a variance estimate (σ), quantifying uncertainty. The mathematical model is:

$$\hat{y}, \sigma = f(x, \theta) + \epsilon$$

Where $f(x, \theta)$ represents the Gaussian process model, x the delay discounting, θ a parameter vector controlling function scale and covariate relevance, and ϵ denotes residuals. By combining predictions with confidence estimates, the authors created brain maps that illustrate the likelihood of normal brain activities. When paired with individual abnormality scores, these maps facilitate the identification of abnormal regions in individuums.

A limitation of GPR is its high computational cost, especially with large neuroimaging datasets, hindering scalability to clinical applications. Despite its detailed predictions and handling of uncertainty, GPR may overfit in less defined or noisier datasets.

3.3.3. Hierarchical Linear Modeling (HLM)

HLM is designed to analyze multi-level data, especially longitudinal studies with irregular measurements over time. It partitions outcome variance into fixed effects and random effects, providing growth trajectory estimates that capture universal trends and individual deviations.

Ordaz et al. (2013) used HLM version 6 and the unconditional growth model to study brain activity growth curves. They analyzed longitudinal fMRI data across brain regions, with age as a within-group predictor and sex and IQ as between-group predictors. However, HLM's applicability in voxelwise analyses across wide age ranges is limited by high variability and small effect sizes, impacting reliability and generalizability. Computational demands for large neuroimaging datasets further challenge the implementation and interpretation in developmental neuroscience.

4. Evaluating Output from Normative Models

This section explores how outputs from normative models enhance the identification and understanding of brain anomalies. By comparing predicted values ($\hat{y}$) with observed values (y), researchers can assess the extent of an individual's deviation from expected norms. Accurately determining this deviation and estimating prediction errors requires quantifying the uncertainty in model predictions, primarily measured by the variance of residuals within the reference group.

Approaches for utilizing predictions from normative models can be split into three categories:

1. Simple Regression Methods: Methods which use the residuals between true and predicted value to calculate individual deviations without incorporate uncertainty measurements.
2. Statistical Methods: Techniques that incorporate data variance or quantile bands to accommodate population diversity and specific patterns.
3. Bayesian Methods: These methods dynamically incorporate variance components of both data variability and model uncertainty into the calculation of individual deviations.

This section will detail each approach, highlighting their roles in refining predictions and enhancing the applicability of normative models in brain growth and development studies.

4.1. Simple Regression Methods

Calculating the residual between the predicted value ($\hat{y}$) and the actual measured value (y), expressed as $y - \hat{y}$, quantifies the divergence of each data point from its expected value. These methods rely on single-point estimations, providing a simple numerical measure of individual deviations from normative values and are commonly used in brain age studies to correlate deviations with cognitive performance or neuropsychiatric disorders (Cole & Franke, 2017; Gur et al., 2014).

For instance, Kessler et al. (2016) used this method to compute maturational deviation scores from a quadratic fit, identifying whether a participant's brain component expression deviates from age-expected levels, thereby linking deviations to cognitive outcomes like attention performance and ADHD diagnosis. Similarly, Erus et al. (2015) used the regression approach to evaluate brain maturation against normative data by using the residual between the normative output and the true brain maturation.

However, simple regression methods focus solely on residuals without accounting for overall dataset variability, limiting their ability to provide detailed insights into individual differences or support in-depth predictive analytics. While effective for establishing basic deviations and group-level correlations, these methods do not address prediction uncertainty, thereby restricting their application in precision medicine or personalized diagnostics.

4.2. Statistical Methods

Analytical frameworks effectively quantify individual deviations from typical brain morphologies, capturing population diversity and identifying age-specific or disease-specific changes. These advanced techniques focus on data distribution and employ statistical scores, namely z-scores or w-scores. These methods mainly use the standard deviation (σ) from the data to normalize and adjust for scale differences across predictions, mathematically expressed as $\frac{y-\hat{y}}{\sigma}$.

Bethlehem et al. (2020) used LOESS to calculate normative means and standard deviations for each age bin and brain region stratified by sex. They computed w-scores for individuals with autism to measure deviations in cortical thickness:

$$W_{\text{region}} = \frac{CT_{\text{region}} - \mu_{\text{norm region}}}{\sigma_{\text{norm region}}}$$

Where CT_{region} is the cortical thickness in a specific brain region, $\mu_{\text{norm region}}$ the mean cortical thickness, and $\sigma_{\text{norm region}}$ the standard deviation. Positive w-scores indicate higher cortical thickness, while negative scores suggest lower thickness than the norm. Mapping these scores highlighted regions where individuals with autism differ from typical patterns, revealing significant regional variations often obscured in broader analyses. Limitations include the exclusion of smaller age bins, potential biases from motion-affected data, and a cross-sectional design that longitudinal studies could improve.

Tahedl (2020) investigated methods to assess and validate cortical thinning, aiming to enhance the sensitivity and specificity of atrophy detection. Their approach utilized non-parametric methods to generate null distributions for cortical thickness data, which serve as reference points under the null hypothesis. The null hypothesis assumes no differences between the observed data and the expected normative values, providing a baseline for comparison. Standardization in their study is expressed as:

$$z_{\text{vertex}} = \frac{d_{\text{vertex}} - \mu_{\text{vertex}}}{\sigma_{\text{vertex}}}$$

Where d_{vertex} represents the cortical thickness at a specific vertex, while μ_{vertex} and σ_{vertex} are the mean and standard deviation of that vertex from the healthy control group, respectively. The use of null distributions is crucial as it enables the establishment of a statistical baseline to determine if observed differences are significant or simply due to random variation.

4.3. Bayesian Methods

This subsection explores the use of Bayesian statistical techniques to analyze outputs from normative models in neurological development. Unlike the previously introduced approaches, which provide static point estimates and use the standard deviations (σ_a) estimated from the data, Bayesian techniques dynamically estimate variance components and generate tailored predictions for each participant. These methods adjust for aleatoric uncertainty (σ_a^2), the inherent variability in the data, and epistemic uncertainty (σ_e^2), uncertainty in the model due to lack of knowledge. The generalized deviation formula for z-score calculation is:

$$z = \frac{y - \hat{y}}{\sqrt{\sigma_a^2 + \sigma_e^2}} \tag{1}$$

Therefore Bayesian analysis provides a flexible confidence interval that adapts to data sparsity or variability. Both uncertainties are estimated during the fitting of the normative model.

Rutherford et al. (2023) demonstrated the effectiveness of Bayesian normative modelling in identifying and quantifying brain abnormalities in schizophrenia compared to controls. They used Bayesian linear regression for surface area and functional connectivity, incorporating age, sex, data quality, and site as covariates. As shown in 1, z-scores were calculated to measure deviations while accounting for data and model uncertainty. The variance σ_d^2 reflects data variability due to noise, modelled as Gaussian noise with a mean of zero and precision β.

The study used these z-scores for multiple analyses, finding significant improvements in identifying subtle brain differences compared to traditional methods. Individual deviation maps summarized extreme deviations, revealing structural and functional anomalies across brain regions and networks. These deviations were not evident in raw data models, showcasing the normative model's ability to capture clinically relevant variations.

The authors used Support Vector Classification (SVC) and Principal Component Regression (PCR) in a downstream z-score analysis. SVC achieved an 87% classification accuracy in distinguishing schizophrenia from controls, surpassing traditional methods. PCR linked brain network activities to cognitive performance, predicting general cognitive abilities. The study's reliance on datasets excluding individuals with lower cognitive ranges and the lack of longitudinal data limit its generalizability.

Marquand et al. (2016a) used Gaussian Process Regression to predict brain activity based on input covariates, providing model variance estimates (σ_{ij}), which quantify uncertainty, for a subject i and brain location j, and variance learned form the normative distribution (σ^2_{nj}), corresponding to the data uncertainty. Using the z-score formula as in 1, the authors found significant deviations from the norm, highlighting atypical neural processing or pathology. The study linked deviations from normal behavior to ADHD symptoms, finding high hyperactivity scores associated with extreme functioning or different reward-related brain responses, correlated with hyperactivity but not inattention across a broad population range.

5. Comparative Analysis of Normative Modeling Approaches in Alzheimer's Disease Detection

This section compares a GAM and GAMLSS normative model implementation by Wachinger et al. (2024) with the publicly available Bayesian linear regression model by Marquand et al. (2016a). The goal is to analyze differences in quantitative model performance, qualitative normative growth curves, and the models' abilities to distinguish Alzheimer's disease patients from cognitively normal subjects.

5.1. Method

First, the PCNToolkit project was cloned from GitHub, and the Bayesian Linear Regression (BLR) model was trained on the public datasets HCP1200 and IXI to verify reproducibility. In a second step, the datasets from five different sites including only subjects with diagnoses Alzheimer's disease (AD), cognitively normal (CN), or mild cognitive impairment (MCI), were merged. From this merged dataset two training sets were created: Set A: Included 80% of all CN subjects for training, with the remaining 20% of healthy subjects used for testing the model's prediction accuracy for healthy individuals. Set B: Included only UK Biobank data for the training, again with a 80 - 20 split. The covariates for estimating the normative models were "age" and "gender."

Separate BLR models were trained on datasets A and B. Additionally, a third BLR model, referred to as the pretrained BLR, was refitted using the publicly available model lifespan_57K_82sites. This pre-trained model was selected for its inclusion of all cortical regions relevant to estimating Alzheimer's disease, as described by Wachinger et al. (2024). These models' parameters are optimized with Powell's method during training. Furhtermore, a sinarcsinh function is applied to handle non-Gaussian data and a cubic b-spline for non-linear adjustments.

The GAM and GAMLSS models were trained using the formulas provided by Wachinger et al. (2024) and dataset A. The GAM was trained with the "REML" method, and the estimation formula for the GAMLSS sigma value was adopted from their paper. The two not-pretrained BLR models were quantitatively compared using BIC, EV, and MSLL. In a downstream analysis, the not-pretrained model trained on dataset A, pre-trained BLR model, GAM, and GAMLSS models were evaluated using BIC, EV, MSLL, MSE, the percentage of z-scores lower than -2 per diagnosis, the correlation between z-score and diagnosis, and the visualization of the calculated growth curves.

5.2. Results

Training the BLR models on datasets A and B demonstrated no significant difference in overall performance. Both models showed similar abilities in recognizing the correlation between z-score and diagnosis. The primary difference observed was that the model trained on dataset B generally identified a lower percentage of abnormal z-scores across all categories. To refine the analysis, further evaluations focused exclusively on models trained on dataset A, to maintain consistency with the original PCNToolkit setup, which prefers having all sites included in the training data. After initial training, the models' performances were compared on the healthy hold-out test set, and then on the full dataset. The literature review showed coherence in evaluating the quantitative performance of normative models on mean-squared error (MSE) and Explained Variance (EV). Additionally, differences in scores for Bayesian Information Criterion (BIC), and Mean Standardized Log Loss (MSLL) are analyzed, all performance scores are visualized in Appendix A.

Explained variance measures how much of the variability in the data is captured by the model. Higher EV values indicate better model performance. Figure 2 shows that all models exhibit relatively low EV values in the left entorhinal region, suggesting complex patterns that are difficult to model. GAM and GAMLSS models were generally the most effective in explaining variance across different brain regions. Despite performing slightly lower than GAM and GAMLSS models, the not-pretrained BLR outperformed the pretrained BLR in all brain regions except the left entorhinal.

BIC assesses the balance between model fit and complexity, with lower values indicating better performance. All models exhibit positive BIC values in the entorhinal region, suggesting that the models did not fit well, suggesting either excessive complexity or insufficient capture of underlying patterns. The pretrained models consistently showed better performance and lower BIC scores compared to the not-pretrained models, likely due to better initial parameter estimates. The significantly lower BIC values for GAM and GAMLSS do not indicate worse performance; instead, they highlight the limitation of the BIC criterion in not being comparable across different model types. Detailed BIC values show that GAMLSS models performed better, indicating a better balance between model fit and complexity compared to GAM models.

MSLL measures the quality of probabilistic predictions, with lower values indicating better performance. GAM and GAMLSS models generally show lower MSLL values across all brain regions, signifying superior performance compared to BLR models. The pretrained BLR model excels in the left entorhinal region but lags in other regions, suggesting that pretraining helps capture complex patterns in this area while struggling with regions that are easier to model.

Changes in Explained Variance and Mean Squared Error upon including AD and MCI subjects were evaluated to assess the impact on model performance. This comparison provides insights into each model's robustness or susceptibility to outliers, Figures 6 and 7, visualize the findings respectively.

All models exhibited an increase in EV when evaluated on the entire dataset, with the BLR models, particularly the pretrained one, showing the highest increase. This increase suggests the models can explain the additional variability AD and MCI subjects introduce. However, a moderate increase in EV, as observed in the GAM and GAMLSS models, indicates better utility in distinguishing between CN and outlier subjects, demonstrating a balanced adaptation to the added complexity without being overly sensitive to the outliers.

Positive differences in MSE indicated worse performance with the full dataset. GAM and GAMLSS models showed substantial positive differences, particularly in the left entorhinal region. This behaviour is expected, as these models predict healthy values for the outlier subjects. The extent to which these predictions differ from the actual values provides insights into how abnormal an individual's value is. Thus, this behaviour is both expected and informative. The pretrained BLR models showed decreased performance only in the left entorhinal region, with improved performance elsewhere, possibly due to the exposure to diverse normal variances during pretraining, which may have mitigated overfitting.

The proportions of z-scores less than or equal to -2 for each diagnosis was evaluated, Figure 9 shows the behaviour averaged across all brain areas. This analysis is used to identify how performant the models are in distinguishing AD from CN subjects, expecting a lower amount of CN subjects having a z-score of -2 and most AD subjects having a z-score of -2 or lower. All models except GAMLSS align with the expected behaviour, showing high abnormal z-scores for AD, medium for MCI, and low for CN. The GAM and not-pretrained model performed consistently across all regions. The pretrained model showed lower numbers of abnormal z-scores for AD but also for MCI and CN, indicating potentially lower z-score magnitude or potential robustness to AD and MCI subjects.

To determine which model best identifies the correlation between diagnosis and z-score, both Spearman and Kendall correlation scores were computed, see 8. The results indicate that the not-pretrained model achieves the highest scores in both categories, followed by the GAMLSS model, the pretrained BLR, and then the GAM model.

The qualitative evaluation of the visualized growth curves is displayed in Appendix B

Normative growth curves plotted for different models show that GAM and GAMLSS models consistently indicate clear decreases in thickness with age. The BLR models add another layer by visualizing uncertainty around the percentile bands. For the pretrained model, this uncertainty is nearly zero, resulting in clean percentile curves. In contrast, the not-pretrained model exhibits wide bands around its curves, indicating high model uncertainty that could be reduced with more training data. The not-pretrained BLR model showed discrepancies for some regions, most pronoun in male plots, suggesting data deficiencies or overfitting. For instance, the graph for the left inferiorparietal region shows an increase in thickness with age, and all regions except the left fusiform display abnormal thickness values, exceeding physiological possibilities. Pretrained BLR models mirrored GAM and GAMLSS behaviors but with more variation, particularly in the left entorhinal region, suggesting potential accuracy in this area. Figures 10 - 17 illustrate the placement of AD subjects within the growth curves for each approach. Notably, only the not-pretrained model failed to categorize any AD patients as outliers, indicating potential incoherencies between the visualization of growth trajectories and quantitative performance.

5.3. Discussion

Despite BLR models trained on dataset A and B demonstrating similar capabilities in correlating z-scores with diagnoses, the lower identification of abnormal z-scores by the model trained on dataset B highlights potential nuances in data site variability. This observation suggests that including the dataset site into the training process might enhance the model's sensitivity to detecting abnormalities as a specific site-effect could be learned.

Explained Variance and Mean Standardized Log Loss analyses revealed that GAM and GAMLSS models performed best across all brain regions. The pretrained BLR model only outperformed the other model types in the left entorhinal cortex, suggesting that the pretrained data might have included vital information for modelling this specific region. This indicates that pretraining might enhances model performance for complex areas, additional pretraining on data may not be necessary for simpler regions.

The Bayesian Information Criterion analysis indicated that pretrained models have better initial parameter estimates, reducing model complexity. However, the limitation of comparing BIC across different model types was evident, underscoring the need for a different metric to compare the differences in model complexity vs. model fit of the two approaches.

Our findings on z-scores and their correlation with diagnoses suggest that the magnitude of z-scores must be considered when selecting models to identify abnormal values for targeted conditions. The GAM and non-pretrained models are comparable in the percentage of abnormal z-scores identified for each diagnosis. In contrast, the pretrained model identified a lower percentage of abnormal z-scores across cognitive statuses, indicating a lower overall z-score magnitude.

Overall, GAM, GAMLSS, and pretrained BLR models demonstrated better qualitative performance, while not-pretrained BLR models showed reliable quantitative performance. This underscores the importance of considering multiple metrics before selecting a model type for accurate brain development modeling.

5.4. BLR Uncertainty Estimation in Z-Score Calculation

The primary distinction between the z-score calculation of Wachinger et al. (2024) and Rutherford et al. (2023) lies in estimating the uncertainty, with Rutherford et al. (2023) incorporate both aleatoric and epistemic uncertainties in their analysis.

Aleatoric uncertainty, representing inherent data noise, is estimated via $\frac{1}{\beta}$. Comparing β values between the pretrained and not-pretrained models reveals a significantly higher β in the pretrained model. This is attributed to the pretrained model's extensive training on 57,000 CN subjects, providing a more accurate estimation of inherent noise.

Epistemic uncertainty, representing model uncertainty reducible with more data, is calculated using:

$$(\sigma_*^2)_d = \phi(x)^T A_d^{-1} \phi(x)$$

where $A = \Phi^T \Lambda_\beta \Phi + \Lambda_\alpha$. Here, $\phi(x)$ is the basis expansion of x, computed via the B-spline method from the scipy library ‘bspline‘. During training, hyperparameters, particularly A and β, are optimized using Powell's method.

Comparing epistemic uncertainty between the pretrained and not-pretrained models show significantly lower values for the pretrained model across all brain regions. This reflects the nature of epistemic uncertainty as reducible with more data, explaining how pretraining leads to lower model uncertainty. Regarding overall uncertainty, which is the square root of the sum of both uncertainties, epistemic uncertainty contributes more significantly in the not-pretrained models. In contrast, aleatoric uncertainty is the dominant factor in the pretrained models, almost nullifying the impact of epistemic uncertainty. A detailed visualization of these uncertainties can be found in Table 1.

The rationale for using the square root of the sum of the two uncertainties is grounded in probability theory. Assuming independence between aleatoric and epistemic uncertainties, their variances add directly. Consequently, the total uncertainty is modelled as:

$$\sqrt{\sigma_d^2 + (\sigma_*^2)_d}$$

Adding the square root independently would represent standard deviations and not calculate the combined uncertainty.

In the GAM approach, z-scores are calculated by dividing residuals by a constant standard deviation, assuming uniform variance across all data points. This means that the standard deviation is the same for every data point, reflecting an overall average variability.

In contrast, the GAMLSS approach calculates z-scores by dividing residuals by the model-predicted sigma value, representing the local standard deviation specific to each data point. This sigma value varies, reflecting the specific variability at each point.

Table 1 visualizes the consistently higher variance values in GAMLSS than those in GAM, indicating a more sensitive reflection of local differences in data variability.

The variance values alone do not provide direct insight into the similarity of z-scores, as they depend on the magnitude of predicted $\hat{y}$ values. For instance, the mean $\hat{y}$ value for the left entorhinal region is 0.715 for the not-pretrained model, 10.107 for the pretrained model, 3.320 for GAMLSS, and 3.321 for GAM.

Combining the knowledge about variance magnitude and similar $\hat{y}$ values between GAMLSS and GAM explains why GAMLSS does not identify z-scores lower than -2 for all patients, while GAM does. Therefore, while GAMLSS offers more precise z-scores by capturing local variability, it results in lower z-score magnitudes, suggesting the need to adjust thresholds for identifying abnormal values to maintain diagnostic accuracy.

The significant difference in $\hat{y}$ values between the pretrained and not-pretrained models highlights their different variance estimations and resulting z-scores, reflecting the impact of pretraining and warping mechanisms learned during model training.

In summary, z-scores classify an individual within the normative range, but understanding variance calculation and considering the magnitude of the predicted $\hat{y}$ values are crucial for interpreting these scores.

6. Conclusion

This paper reviewed and compared normative modeling approaches for estimating growth curves and calculating z-scores to analyze anatomical patterns. The analysis highlights the strengths and limitations of different modeling techniques in clinical and research settings.

Quantitative metrics show that model performance varies by metric, suggesting model selection should align with specific performance needs. Notably, the pretrained BLR model only excelled in the left entorhinal cortex, indicating the importance of region-specific training data and questioning the overall advantage of pretrained models.

Variance analysis reveals GAMLSS models offer more individualized residual normalization than linear methods like GAM, resulting in smaller z-scores and the need for adjusted thresholds to maintain diagnostic accuracy. Incorporating model and data uncertainty into z-score calculations provides a more individualized approach to understanding brain development. Aleatoric uncertainty is higher in pretrained models due to large dataset training, while epistemic uncertainty is lower, reflecting the impact of pretraining on model parameters.

Qualitative analysis shows GAM and GAMLSS produce nearly identical growth curves despite different z-score magnitudes. Non-pretrained models performed well quantitatively but showed unrealistic growth curves and high uncertainty. In contrast, pretrained models displayed more realistic growth curves with appropriate variability.

Combining quantitative and qualitative findings, GAMLSS is the best-performing normative model for brain development. Non-pretrained Bayesian regression models show significant quantitative improvements but lack qualitative accuracy. The need for more training data to derive meaningful growth curves for Bayesian models is highlighted. Future research should set appropriate thresholds for identifying abnormal z-scores and explore the performance issues in the left entorhinal cortex.

References

Bethlehem, R. A., Seidlitz, J., Romero-Garcia, R., Dumas, G., and Lombardo, M. V. Normative age modelling of cortical thickness in autistic males. *bioRxiv*, pp. 252593, 2018a.

Bethlehem, R. A., Seidlitz, J., Romero-Garcia, R., and Lombardo, M. V. Using normative age modelling to isolate subsets of individuals with autism expressing highly age-atypical cortical thickness features. *bioRxiv*, pp. 252593, 2018b.

Bethlehem, R. A., Seidlitz, J., Romero-Garcia, R., Trakoshis, S., Dumas, G., and Lombardo, M. V. A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Communications biology*, 3(1):486, 2020.

Bethlehem, R. A., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., et al. Brain charts for the human lifespan. *Nature*, 604(7906):525–533, 2022.

Brewer, J. B. Fully-automated volumetric mri with normative ranges: translation to clinical practice. *Behavioural neurology*, 21(1-2):21–28, 2009.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Cole, J. H. and Franke, K. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017.

Erus, G., Battapady, H., Satterthwaite, T. D., Hakonarson, H., Gur, R. E., Davatzikos, C., and Gur, R. C. Imaging patterns of brain development and their relationship to cognition. *Cerebral cortex*, 25(6):1676–1684, 2015.

Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P., and Thompson, P. M. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2): 67–77, 2010.

Gur, R. C., Calkins, M. E., Satterthwaite, T. D., Ruparel, K., Bilker, W. B., Moore, T. M., Savitt, A. P., Hakonarson, H., and Gur, R. E. Neurocognitive growth charting in psychosis spectrum youths. *JAMA psychiatry*, 71(4):366–374, 2014.

Huizinga, W., Poot, D. H., Vernooij, M. W., Roshchupkin, G. V., Bron, E. E., Ikram, M. A., Rueckert, D., Niessen, W. J., Klein, S., Initiative, A. D. N., et al. A spatiotemporal reference model of the aging brain. *NeuroImage*, 169:11–22, 2018.

Insel, T. R. Mental disorders in childhood: shifting the focus from behavioral symptoms to neurodevelopmental trajectories. *Jama*, 311(17):1727–1728, 2014.

Kessler, D., Angstadt, M., and Sripada, C. Growth charting of brain connectivity networks and the identification of attention impairment in youth. *JAMA psychiatry*, 73(5): 481–489, 2016.

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological psychiatry*, 80(7):552–561, 2016a.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):433–447, 2016b.

Ordaz, S. J., Foran, W., Velanova, K., and Luna, B. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. *Journal of Neuroscience*, 33(46):18109–18124, 2013.

Rutherford, S., Fraza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., et al. Charting brain growth and aging at high spatial precision. *elife*, 11:e72904, 2022.

Rutherford, S., Barkema, P., Tso, I. F., Sripada, C., Beckmann, C. F., Ruhe, H. G., and Marquand, A. F. Evidence for embracing normative modeling. *Elife*, 12:e85082, 2023.

Stasinopoulos, D. M. and Rigby, R. A. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46, 2008.

Tahedl, M. Towards individualized cortical thickness assessment for clinical routine. *Journal of translational medicine*, 18:1–12, 2020.

Wachinger, C., Hedderich, D., and Bongratz, F. Stochastic cortical self-reconstruction. *arXiv preprint arXiv:2403.06837*, 2024.

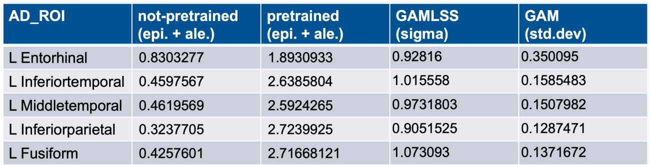
Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J. K., Ueland, T., Melle, I., Franke, B., et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA psychiatry*, 75(11):1146–1155, 2018.

Wolfers, T., Beckmann, C. F., Hoogman, M., Buitelaar, J. K., Franke, B., and Marquand, A. F. Individual differences v. the average patient: mapping the heterogeneity in adhd using normative models. *Psychological medicine*, 50(2): 314–323, 2020.

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., et al. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 4(6):567–578, 2019.

Ziegler, G., Ridgway, G. R., Dahnke, R., Gaser, C., Initiative, A. D. N., et al. Individualized gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. *NeuroImage*, 97: 333–348, 2014.

A. Quantitative Performance Visualizations



AD_ROI	not-pretrained (epi. + ale.)	pretrained (epi. + ale.)	GAMLSS (sigma)	GAM (std.dev)
L Entorhinal	0.8303277	1.8930933	0.92816	0.350095
L Inferiortemporal	0.4597567	2.6385804	1.015558	0.1585483
L Middletemporal	0.4619569	2.5924265	0.9731803	0.1507982
L Inferiorparietal	0.3237705	2.7239925	0.9051525	0.1287471
L Fusiform	0.4257601	2.71668121	1.073093	0.1371672

Figure 1. Variance Between Different Modelling Approaches

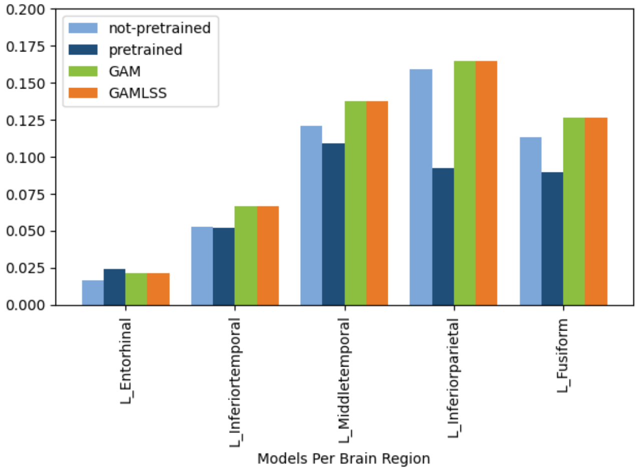


Figure 2. EV Healthy Testset

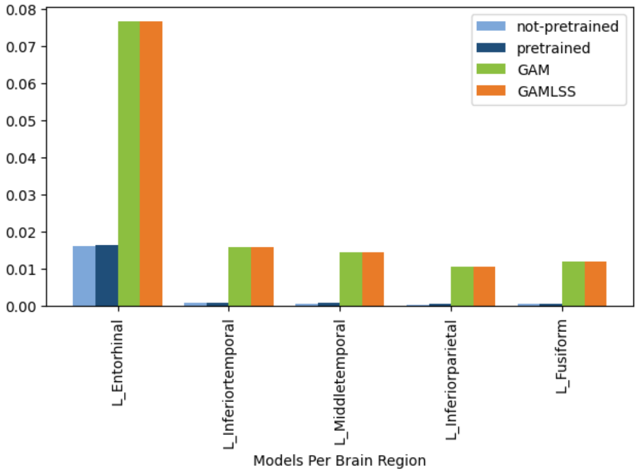


Figure 3. MSE Healthy Testset

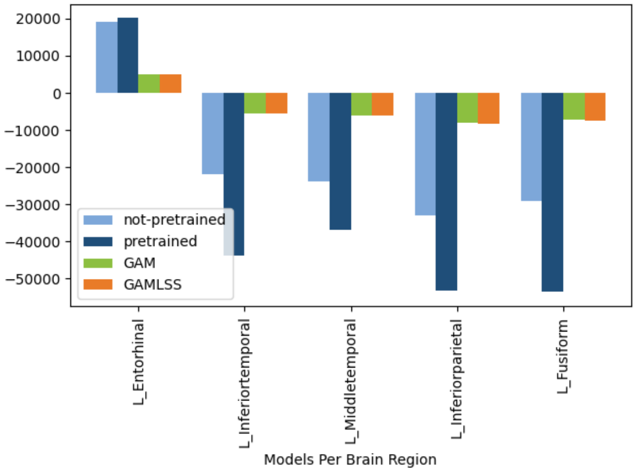


Figure 4. BIC Healthy Testset

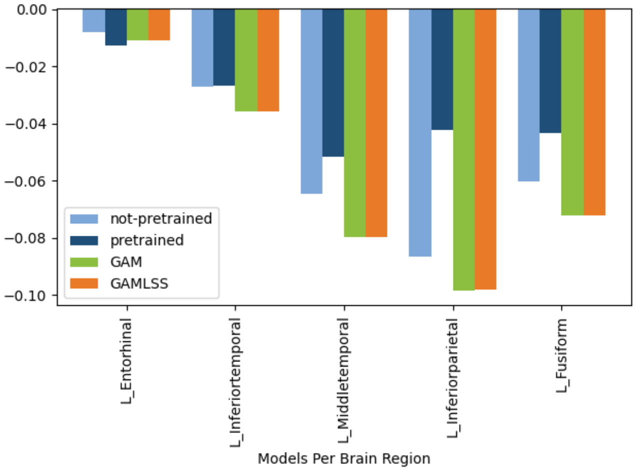


Figure 5. MSLL Healthy Testset

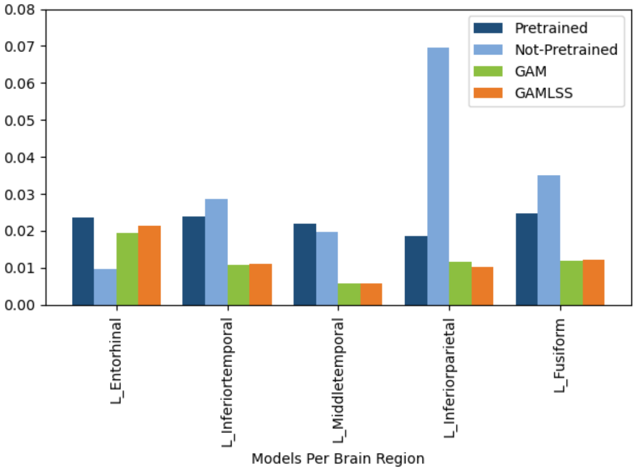


Figure 6. EV Differences All Data vs. Healthy Testset

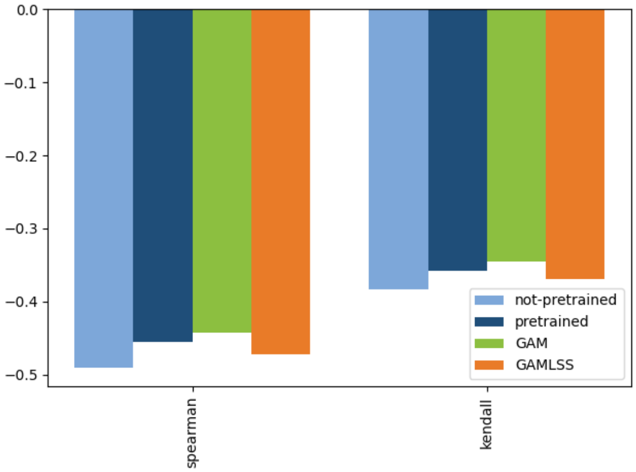


Figure 8. Correlation Diagnosis & Z-Score

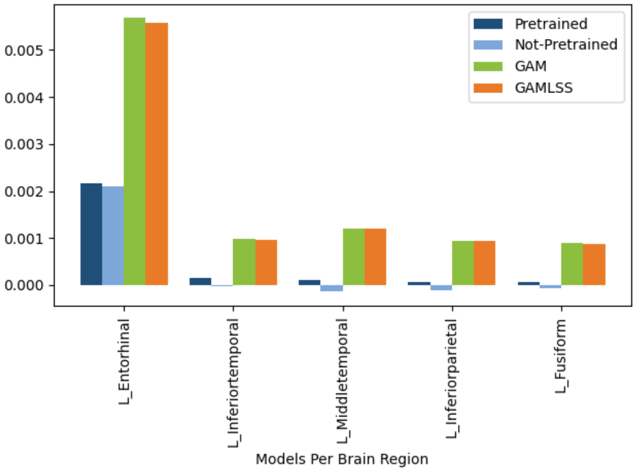


Figure 7. MSE Differences All Data vs. Healthy Testset

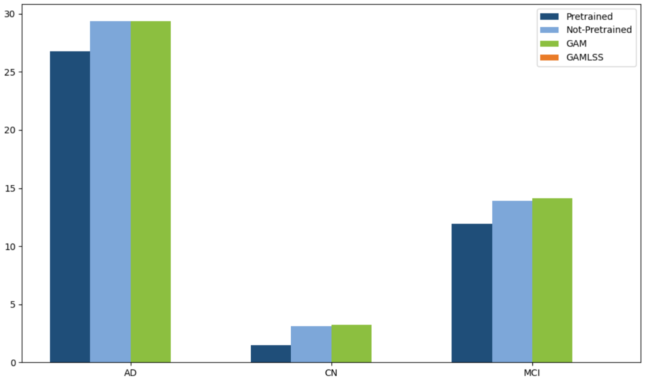


Figure 9. Abnormal Z-Score Percentage ROI

B. Qualitative Performance Visualization

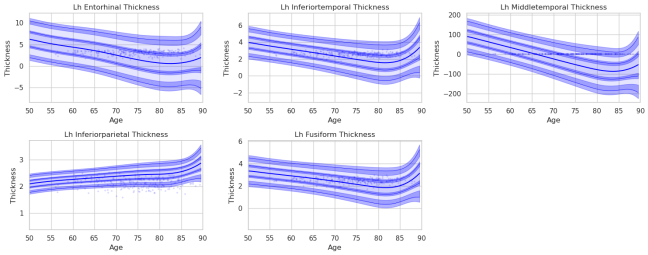


Figure 10. Not-Pretrained Normative Growth Curves Male

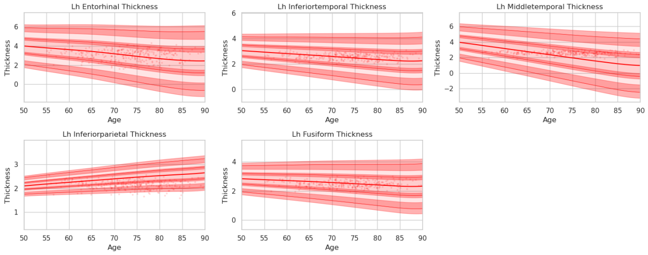


Figure 11. Not-Pretrained Normative Growth Curves Female

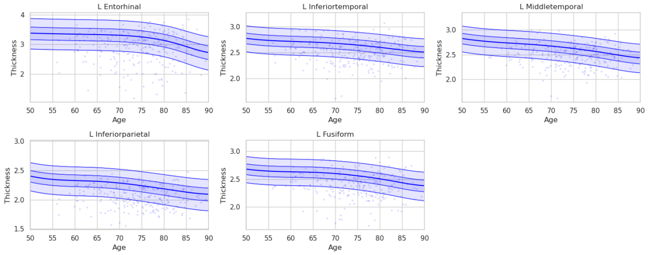


Figure 12. Pretrained Normative Growth Curves Male

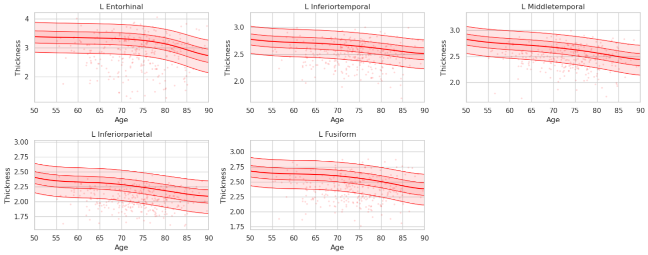


Figure 13. Pretrained Normative Growth Curves Female

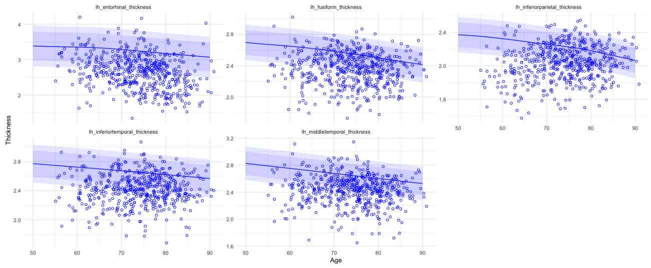


Figure 14. GAM Normative Growth Curves Male

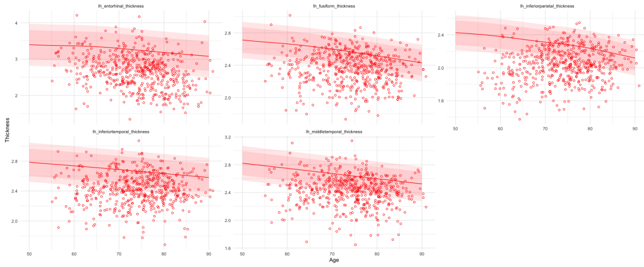


Figure 15. GAM Normative Growth Curves Female

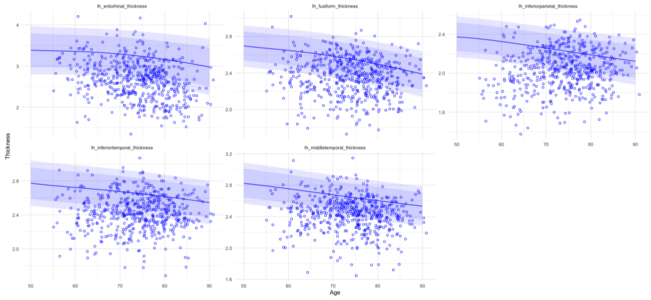


Figure 16. GAMLSS Normative Growth Curves Male

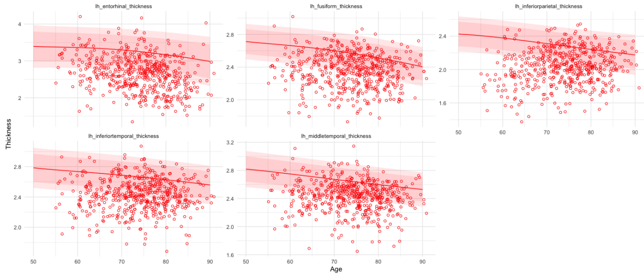


Figure 17. GAMLSS Normative Growth Curves Female