Using normative models pre-trained on cross-sectional data to evaluate longitudinal changes in neuroimaging data

Barbora Rehák Bučková[1,2,3], Charlotte Fraza[4], Rastislav Rehák[5], Marián Kolenič[3], Christian Beckmann[4], Filip Španiel[3], Andre Marquand[4†], and Jaroslav Hlinka[1,3†]

[1]Department of Complex Systems, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

[2]Department of Cybernetics, Czech Technical University in Prague, Prague, Czech Republic

[3]National Institute of Mental Health, Klecany, Czech Republic

[4]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

[5]CERGE-EI, a joint workplace of the Center for Economic Research and Graduate Education of Charles University, and the Economics Institute of the Czech Academy of Sciences, Prague, Czech Republic

[*]Corresponding author. Email: andre.marquand@donders.ru.nl, hlinka@cs.cas.cz

[†]These authors contributed equally to this work.

Abstract

Longitudinal neuroimaging studies offer valuable insight into intricate dynamics of brain development, ageing, and disease progression over time. However, prevailing analytical approaches rooted in our understanding of population variation are primarily tailored for cross-sectional studies. To fully harness the potential of longitudinal neuroimaging data, we have to develop and refine methodologies that are adapted to longitudinal designs, considering the complex interplay between population variation and individual dynamics of change.

We propose a framework based on normative modelling, which extends the use of pre-trained normative models to longitudinal studies. We introduce a quantitative metric termed the *"z-diff"* score, which serves as an indicator of the extent of change within an individual. Notably, our framework offers advantages such as flexibility in dataset size and ease of implementation, thanks to the reference model being trained on over 58,000 individuals.

To demonstrate our approach, we applied it to a longitudinal dataset of 98 patients diagnosed with early-stage schizophrenia who underwent MRI examinations shortly after diagnosis and one year later.

Compared to cross-sectional analyses, which showed global thinning of grey matter at the first visit, our method revealed a significant normalization of grey matter thickness in the frontal lobe over time. Furthermore, this result was not observed when using more traditional methods of longitudinal analysis, making our approach more sensitive to temporal changes.

Overall, our framework presents a flexible and effective methodology for analysing longitudinal neuroimaging data, providing insights into the progression of the disease that would otherwise be missed when using more traditional approaches.

1 Introduction

Longitudinal neuroimaging studies provide a unique opportunity to gain insight into the temporal dynamics of a disease, over and above the insights offered by cross-sectional studies. Efforts related to the acquisition of these datasets are non-trivial and require substantial time and funding while accounting for problems inherent to longitudinal studies such as the dropout of subjects between time points, standardization over multiple clinical centres, and changes in imaging technology over time. Considering this challenging task, it is of great consequence to have tools to effectively analyze them whilst also making use of more widely available cross-sectional data to refine inferences.

Despite the importance of having suitable tools for longitudinal data analysis, the development of appropriate methods are relatively scarce. Even though recent advances in the acquisition and publication of large neuroimaging datasets [1, 2] have significantly improved our understanding of population variation, the developed methodologies are largely focused on the cross-sectional nature of the data [3]. Indeed, these methods are essential for characterizing individual differences; however, the vital factor of longitudinal change is largely neglected. Consequently, efforts should be made to tailor the standing methods modelling population variation for a longitudinal context to advance our understanding of disease progression, which is currently modelled mostly by mixed-effects models [4].

Among the promising cross-sectional techniques for modelling population variation, which emerged in recent years, is normative modelling [5, 6]. This framework models each image-derived phenotype (IDP) (i.e., voxel intensity, regional average, or regional volume) independently as a function of clinical variables (e.g., age, sex, or scanning site) in a large healthy population. Subjects are subsequently compared to the healthy population, which enables us to evaluate individual differences, rather than just compare group differences between patients and controls [7, 8]. Application of these models has already provided valuable insights into the individual neuroanatomy of various diseases, such as Alzheimer's, schizophrenia, autism, and other neurological and mental disorders [9–12].

Longitudinal studies are conceptually well suited for normative modelling since they analyse individual trajectories over time. If adjusted appropriately, normative models could not only improve predictive accuracy but also identify patterns of

change, thereby enhancing our understanding of the disease. In contrast to traditional statistical methods that estimate the average change in the group, normative models could estimate individual deviations from healthy trajectories. This comprehensive approach would take into account both population heterogeneity and confounders, thus providing a more nuanced understanding of change over time.

Normative modelling is a relatively new area of research and thus, despite its potential, longitudinal normative models have not been extensively explored [6, 13]. Indeed, virtually all large-scale normative models released to date are estimated on cross-sectional data [6, 14] and a recent report [13] has provided empirical data to suggest that such cross-sectional models may underestimate the variance in longitudinal data [13]. However, from a theoretical perspective, it is very important to recognise that cross-sectional models describe group-level population variation across the lifespan, where such group level centiles are interpolated smoothly across time. It is well-known in the pediatric growth-charting literature (e.g. [15]) that centiles in such do not necessarily correspond to individual level trajectories, rather it is possible that individuals cross multiple centiles as they proceed through development, even in the absence of pathology. Crucially, classical growth charts, and current normative brain charts provide no information about how frequent such centile crossings are in general. In other words, they provide a *trajectory of distributions*, **not** a *distribution over trajectories*. There are different approaches to tackle this problem in the growth charting literature, including the estimation of 'thrive lines' that map centiles of constant velocity across the lifespan and can be used to declare 'failure to thrive' at the individual level see (e.g [15]) for details. Unfortunately, this approach requires densely sampled longitudinal neuroimaging data to estimate growth velocity, that are not available across the human lifespan at present. Therefore, in this work, we adopt a different approach based on estimates of the uncertainty in the centile estimates themselves together with the uncertainty with which a point is measured (e.g. bounded by the test-retest reliability, noise etc). By accounting for such variability, this provides a statistic to determine whether a centile crossing is large enough to be statistically different from the base level within the population.

We stress that our aim is not to build a longitudinal normative model *per se*. Considering the much greater availability of cross-sectional data relative to longitudinal data, we instead, leverage existing models constructed from densely sampled cross-sectional populations and provide methods for applying these to longitudinal

cohorts. We argue that although these models lack explicit intra-subject dynamics, they contain sufficient information to enable precise assessments of changes over time. Nevertheless, the inclusion of longitudinal data into existing models largely estimated from cross-sectional data is also an important goal and can be approached with hierarchical models [16], however, we do not tackle this problem here.

Our approach requires (i) a probabilistic framework to coherently manage uncertainty and (ii) cross-sectional models estimated on large reference cohort to accurately estimate population centiles. To this end, we utilise the Warped Bayesian Linear Regression normative model [7] as a basis for our work. Training these models requires significant amounts of data and computational resources, limiting their use for smaller research groups. However, the availability of pre-trained models has made them more accessible to researchers from a wider range of backgrounds, as reported by Rutherford et al. [14].

In summary, in this work, we propose a framework for using pre-trained normative models derived from cross-sectional data to evaluate longitudinal studies. We briefly present the existing model and derive a novel set of difference ('*z-diff*') scores for statistical evaluation of change between measurements. We then describe its implementation and showcase its practical application to an in-house longitudinal dataset of 98 patients in the early stages of schizophrenia who underwent fMRI examinations shortly after being diagnosed and one year after.

2 Methods

2.1 Model formulation

2.1.1 Original model for cross-sectional data

The original model [7] we use for pre-training is developed for a cross-sectional database $\mathbf{Y} = (y_{nd}) \in \mathbb{R}^{N \times D}$ and $\mathbf{X} = (x_{nm}) \in \mathbb{R}^{N \times M}$, where y_{nd} is the d-th IDP of the n-th subject and x_{nm} is the m-th covariate (e.g., age or sex) of the n-th subject.

The IDPs are treated separately, so we focus on a single IDP d. To simplify notation, this IDP-specific index is dropped and we denote $\mathbf{y} = (y^{(1)}, \ldots, y_N)^T$. The observations are assumed to be independent (across n). We model the distribution of $\varphi(y_n)$ conditional on exogenous variables $\mathbf{x}_n = (x_{n1}, \ldots, x_{nM})^T$, vector of parameters

$\mathbf{w}$, and a hyper-parameter β as

$$\varphi(y_n)|\mathbf{x}_n; \mathbf{w}; \beta \sim \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}), \tag{1}$$

where:

- $\varphi(y_n)$ is the original variable y_n transformed by a warping function φ, which is parametrized by hyper-parameters $\boldsymbol{\gamma}$ (e.g., composition of SinhArcsinh warping functions, which performed well in prior work [7]); this transformation is used to accommodate non-Gaussian errors in the original space of dependent variables;
- $\phi(\mathbf{x}_n) \in \mathbb{R}^K$ is a common B-spline basis expansion of the original independent variables $\mathbf{x}_n$ (specifically, cubic splines with three evenly spaced knot points [7]) to capture non-linear relationships, appended with site dummies to accommodate site level effects;
- β is a precision of measurements treated as a hyper-parameter (the framework in [7] allows also for site-specific precision terms, which we don't use here).

We write this as

$$\varphi(y_n) = \mathbf{w}^T \phi(\mathbf{x}_n) + \varepsilon_n, \ \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}), \tag{2}$$

where ε_n are independent from $\mathbf{x}_n$ and across n. We further denote $\boldsymbol{\Lambda}_\beta = \beta \mathbf{I} \in \mathbb{R}^{N \times N}$ and the design matrix $\boldsymbol{\Phi} = (\phi(\mathbf{x}_n)_k) \in \mathbb{R}^{N \times K}$.

The estimation of parameters $\mathbf{w}$ is performed by empirical Bayesian methods. In particular, prior about $\mathbf{w}$

$$\mathcal{N}(0, \boldsymbol{\Lambda}_\alpha^{-1}), \ \boldsymbol{\Lambda}_\alpha = \alpha \mathbf{I} \tag{3}$$

is combined with the likelihood function to derive the posterior

$$\mathbf{w}|\mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma} \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{A}^{-1}), \tag{4}$$

$$\mathbf{A} = \boldsymbol{\Phi}^T \boldsymbol{\Lambda}_\beta \boldsymbol{\Phi} + \boldsymbol{\Lambda}_\alpha, \tag{5}$$

$$\bar{\mathbf{w}} = \mathbf{A}^{-1} \boldsymbol{\Phi}^T \boldsymbol{\Lambda}_\beta \mathbf{y}. \tag{6}$$

The hyper-parameters $\alpha, \beta, \boldsymbol{\gamma}$ are estimated by maximizing the warped marginal log-likelihood.

The predictive distribution of $\varphi(y)$ for a subject with $\mathbf{x}$ is

$$\mathcal{N}(\bar{\mathbf{w}}^T\phi(\mathbf{x}), \phi(\mathbf{x})^T\mathbf{A}^{-1}\phi(\mathbf{x}) + \beta^{-1}). \tag{7}$$

Finally, the z-score characterizing the position of this subject within population is

$$z = \frac{\varphi(y) - \bar{\mathbf{w}}^T\phi(\mathbf{x})}{\sqrt{\phi(\mathbf{x})^T\mathbf{A}^{-1}\phi(\mathbf{x}) + \beta^{-1}}}, \tag{8}$$

where $\varphi(y)$ is the realized warped observation of IDP d for this subject.

Note that formulae (7) and (8) implicitly evaluate only (potentially new) subjects measured at sites already present in the original database $\mathbf{y}$, $\mathbf{\Phi}$. If we want to evaluate subjects measured at a new site, we will have to run an adaptation procedure to account for the effect of the new site. This adaptation procedure is described elsewhere and is readily accessible for use online (see [14]). In short, a sample of a reference (healthy) cohort measured on the same scanner as the population of interest is needed to accommodate a site-specific effect. Using the reference cohort, the parameters of the original pre-trained model are adapted and only then the sample of interest is analyzed (Figure 1).

In the following section, we develop a procedure that allows to extend the original cross-sectional framework pre-trained on database $\mathbf{y}$, $\mathbf{\Phi}$ to a new longitudinal dataset for evaluation of changes in regional brain thickness.

2.1.2 Adaptation to longitudinal data

We aim to adapt the original model [7], leveraging its pre-trained parameters, to design a z-score (for clarity further referred to as *z-diff* score) for a change between visits, based on which we could detect large changes in regional brain thickness.

To utilize the normative model of the healthy population pre-trained on cross-sectional data in the longitudinal setup, we have to make an assumption about the trajectory of healthy controls. The natural first approximate assumption is that a healthy subject does not deviate substantially from their position within the population as time progresses, so the observed position changes between the visits of a healthy control subject are assumed to stem from observation noise (due to technical or physiological factors), and are therefore constrained by the test-retest reliability of the measurement. Note that this does not imply that a healthy subject does not

change over time, but rather change that occurs follows approximately the centile of distribution at which the individual is placed. We acknowledge that this is a relatively strong assumption, but it is reasonable to assume that a subject's brain activity or structure will remain relatively stable over short time periods (years, but not necessarily decades), and any significant changes in the pattern compared to the normative reference database may indicate disease progression or response to intervention [12]. Also, it is very important to recognise that this model does not constrain a given subject to follow a population-level centile trajectory exactly, because the model includes an error component ($\xi^{(i)}$) that allows for individual level deviations from the population centile.

We connect to the cross-sectional model by invoking the above assumption through reinterpretation of the error term $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$. In particular, we decompose it to a subject-specific factor η and a measurement error $\xi^{(i)}$ in the i-th visit. While η is considered constant across the visits (of a healthy cohort subject) and captures the subject's position within the population, $\xi^{(i)}$ is specific for the i-th visit and captures a combination of movement, acquisition, processing noise, etc. Hence, the model for the i-th visit of a subject with given covariates $\mathbf{x}^{(i)}$ is

$$\begin{aligned} \varphi(y^{(i)}) &= \mathbf{w}^T \phi(\mathbf{x}^{(i)}) + \eta + \xi^{(i)} \\ \eta &\sim \mathcal{N}(0, \sigma_\eta^2) \\ \xi^{(i)} &\sim \mathcal{N}(0, \sigma_\xi^2) \\ \beta^{-1} &= \sigma_\eta^2 + \sigma_\xi^2 \end{aligned} \tag{9}$$

where η, $\xi^{(i)}$, and $\mathbf{x}^{(i)}$ are mutually independent for a given i, and $\xi^{(i)}$ and $\xi^{(j)}$ are independent for $i \neq j$. Note that we dropped the subject-specific index n introduced in (1). This change in notation should force the reader to distinguish between data used for training the original pre-trained model and a new set of longitudinal data that is going to be used for estimating the longitudinal change.

In our longitudinal data, we are interested in the change for a given individual across two visits. According to model (9), the difference in $\varphi(y)$ between visits 1 and 2, $\varphi(y^{(2)}) - \varphi(y^{(1)})$, for a subject with covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is given by

$$\varphi(y^{(2)}) - \varphi(y^{(1)}) = \mathbf{w}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + \xi^{(2)} - \xi^{(1)} \tag{10}$$

with $\xi^{(2)} - \xi^{(1)} \sim \mathcal{N}(0, 2\sigma_\xi^2)$. We use the posterior distribution of $\mathbf{w}$ (and hyperparameters $\alpha, \beta, \boldsymbol{\gamma}$) estimated on the original cross-sectional database $\mathbf{y}, \boldsymbol{\Phi}$ [7]. Therefore, the posterior predictive distribution for the difference $\varphi(y^{(2)}) - \varphi(y^{(1)})$ for our subject is (for more detailed derivation, please refer to the supplement)

$$\mathcal{N}\left(\bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})], [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2\right). \quad (11)$$

Finally, the z-score for the difference in $\varphi(y)$ between visits 1 and 2 is

$$z_{diff} = \frac{[\varphi(y^{(2)}) - \varphi(y^{(1)})] - \bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]}{\sqrt{[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2}}, \quad (12)$$

where $\varphi(y^{(2)}) - \varphi(y^{(1)})$ is the realized change in the warped observations of IDP d for this subject. Since this z_{diff} score is standard normal for the population of healthy controls, any large deviations may be used to detect suspicious changes.

The primary role of adaptation of the pre-trained cross-sectional model to longitudinal data is to account for the measurement noise variance σ_ξ^2. From the posterior predictive distribution (11), we have

$$\begin{aligned}
&\mathrm{E}\Big[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \mathrm{E}[\varphi(y^{(2)}) - \varphi(y^{(1)})|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}]\right)^2 \Big| \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}\Big] \\
&= [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2. \quad (13)
\end{aligned}$$

Hence, by the Law of Iterated Expectations (to integrate out $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$), we obtain

$$\begin{aligned}
&\mathrm{E}\Big[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]\right)^2 \\
&\quad - [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] \Big| \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}\Big] = 2\sigma_\xi^2. \quad (14)
\end{aligned}$$

Therefore, we estimate $2\sigma_\xi^2$ by the sample analogue of the left-hand side in (14). Specifically, we devote a subsample C of our controls to adaptation and we compute

$$\begin{aligned}
\widehat{2\sigma_\xi^2} = \frac{1}{|C|} \sum_{k \in C} \Big[&\left(\varphi(y_k^{(2)}) - \varphi(y_k^{(1)}) - \bar{\mathbf{w}}^T[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]\right)^2 \\
&- [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]\Big]. \quad (15)
\end{aligned}$$

Moreover, another useful feature of longitudinal data is that $[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]$ is

negligible (especially with stable covariates, like sex and age). Sex (typically) does not change across the two visits and age relatively little (in our target application) with respect to the full span of ageing. Consequently, $[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]$ in (15) is negligible in adult cohorts but must be treated with caution in developmental or ageing groups. Finally, it is apparent from (5) that $\mathbf{A}$ scales with a number of subjects and its inverse will be negligible for substantial training datasets, such as the one that was used for pre-training.

2.1.3 Implementation

To implement the method (Fig. 1), we used the **PCN toolkit**. Running the adaptation pipeline [14] returns all intermediate results, which is convenient for analysis.

2.2 Data

2.2.1 Early stages of schizophrenia patients

The clinical data used for the analysis were part of the Early Stages of Schizophrenia study [17]. We analysed data from 98 patients in the early stages of schizophrenia (38 females) and 67 controls (42 females) (Table 1). The inclusion criteria were as follows: The subjects were over 18 years of age and undergoing their first psychiatric hospitalisation. They were diagnosed with schizophrenia; or acute and transient psychotic disorders; and suffered from untreated psychosis for less than 24 months. Patients were medically treated upon admission, based on the recommendation of their physician. Patients suffering from psychotic mood disorders were excluded from the study.

Healthy controls over 18 years of age were recruited through advertisements unless: They had a personal history of any psychiatric disorder or had a positive family history of psychotic disorders in first- or second-degree relatives.

If a subject in either group (patient or control) had a history of neurological or cerebrovascular disorders or any MRI contraindications, they were excluded from the study.

The study was carried out in accordance with the latest version of the Declaration of Helsinki. The study design was reviewed and approved by the Research Ethics Board. Each participant received a complete description of the study and provided written informed consent.

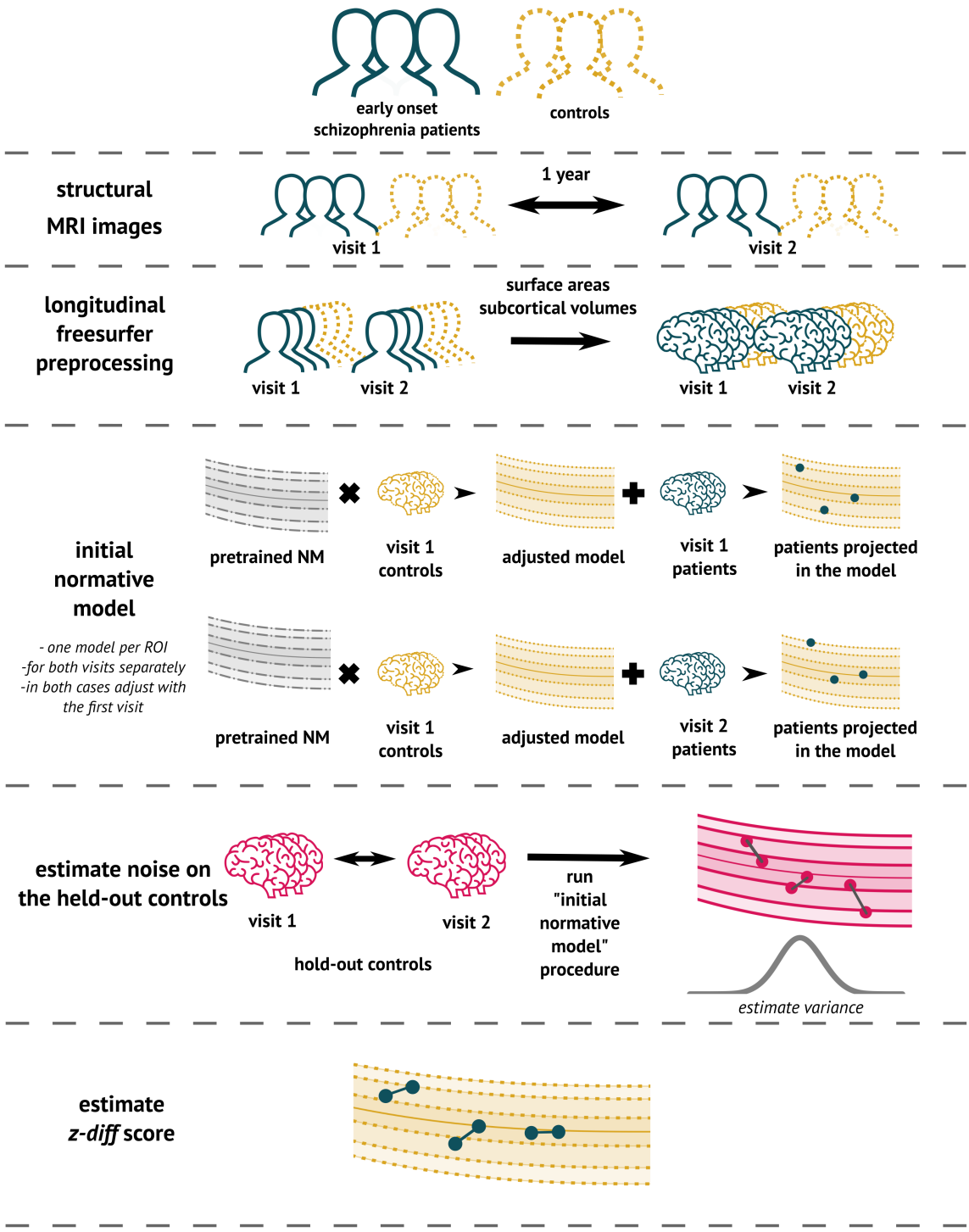


Figure 1: The overview of the analytical pipeline for our schizophrenia patients: First, data are preprocessed using Freesurfer's longitudinal pipeline. Subsequently, the pre-trained models are adjusted to a local sample of healthy controls. The site-specific measurement noise variance $\sigma^2_{\xi_s}$ in healthy subjects is estimated using held-out controls, and finally, the *z-diff* score is computed.

Data were acquired at the National Centre of Mental Health in Klecany, Czech Republic. The data were acquired at the National Institute of Mental Health using Siemens MAGNETOM Prisma 3T. The acquisition parameters of T1-weighted images using MPRAGE sequence were: 240 scans; slice thickness: 0.7 mm; repetition time: 2,400 ms; echo time: 2,34 ms; inversion time: 1000 ms; flip angle: 8°, and acquisition matrix: 320 mm $\times$ 320 mm.

2.3 Preprocessing and Analysis

All T1 images were preprocessed using the Freesurfer v.(7.2) `recon-all` pipeline. Preprocessing was performed twice, in cross-sectional and longitudinal [18] settings, to evaluate the effect of preprocessing. In line with [14], we performed a simple quality control procedure whereby we excluded all subjects having a rescaled Euler number greater than ten were labelled outliers and were not included in the analysis (Table 1) (see [14] and [16] for further details). The pre-trained model used for adaptation was the `lifespan_58K_82_sites` [14].

We evaluated the effect of preprocessing by running normative models for both visits and both types of preprocessing. We then tested for the difference in variance of the difference of the cross-sectional z-scores $z^{(2)} - z^{(1)}$ in held-out controls.

After preprocessing, patient data were projected into the adapted normative model (median Rho across all IDP was 0.3 and 0.26 for the first and the second visit, respectively – see Supp. Fig. 1). For each subject and visit, we obtained cross-sectional z-score, as well as the underlying values needed for it's computation, particularly $\phi(y)$, and $\bar{\mathbf{w}}^T\varphi(\mathbf{x})$. We conducted a cross-sectional analysis of the original z-scores to evaluate each measurement independently. We used the Wilcoxon test to determine whether the z-scores of the patients were significantly different from zero and corrected for multiple tests using the Benjamini-Hochberg FDR correction at the 5% level of significance.

Subsequently, following (12), we derived the *z-diff* scores of change between visits. We conducted two analyses, one to investigate the group-level effect, and another to link the *z-diff* to the changes in clinical scales.

At a group-level, we identified regions with *z-diff* scores significantly different from zero using the Wilcoxon test, accounting for multiple comparisons using the Benjamini-Hochberg FDR correction. Additionally, we performed a more traditional

	Patients	**Controls**
N (% females)	98 (39%)	67 (63%)
Age, median (min, max), years	27 (18, 46)	29 (18, 54)
Interval between visits, median (min, max), years	1.1 (0.9,2.7)	1.2 (0.9, 3)
Diagnosis (only for patients)		
Schizophrenia	53	
Brief psychotic disorder	45	
Length of disease, median (min, max), months	4 (1,21)	
Clinical scales (only for patients)	**Visit 1**	**Visit 2**
PANSS sum, median (min, max)	53 (30, 94)	44 (30, 84)
PANSS Positive Symptoms, median (min, max)	11 (7, 21)	8 (7, 26)
PANSS Negative Symptoms, median (min, max)	14.5 (7, 30)	11.5 (7, 24)
GAF, median (min, max)	70 (25, 100)	80.5 (40, 98)

Table 1: Clinical description of the dataset after quality control

longitudinal analysis. As all visits were approximately one-year apart, we conducted an analysis of covariance (ANCOVA). The ANCOVA model combines general linear model and ANOVA. Its purpose is to examine whether the means of a dependent variable (thickness in V2) are consistent across levels of a categorical independent variable (patients or controls) while accounting for the influences of other variables (age, gender, and thickness in V1). We conducted a separate test for each IDP and controlled the relevant p-values across tests using the FDR correction.

For linking the *z-diff* score to clinical change, we transformed the *z-diff* score across all IDPs using PCA to decrease the dimensionality of the data as well as to avoid fishing. We ran PCA with 10 components and using Spearman correlation related the scores with changes in the Positive and Negative Syndrome Scale (PANSS) and Global Assessment of Functioning (GAF) scale.

3 Results

3.1 Effect of preprocessing

After running the normative models in both visits, we visually observed a decrease in variance between the two visits in longitudinal preprocessing compared to cross-sectional (Figure 2). More specifically, we calculated the mean of the difference

between z-scores of V^2 and V^1 for each individual IDP, stratified by preprocessing and group, across all subjects. We then visualized the distribution of these means using a histogram (Figure 2C). Alternatively, we also computed the mean difference between z-scores of V^2 and V^1 across all IDPs for each subject, and plotted a histogram of these values. Based on these results, we continued the analysis using longitudinal preprocessing for all data, although we also compare our results to the results derived from cross-sectional preprocessing.

3.2 Cross-sectional results

At a group level, patients had significantly lower thicknesses in most areas compared to healthy populations. In particular, this difference was distinct even in the first visit, indicating structural changes prior to diagnosis (Figure 3).

3.3 Longitudinal results and patterns of change

A longitudinal analysis that evaluated the amount of structural change between the two visits showed a significant cortex normalization of several frontal areas, namely the right and left superior frontal sulcus, the right and left middle frontal sulcus, the right and left middle frontal gyrus, and the right superior frontal gyrus (Figure 4).

In terms of linking change in clinical scores with changes in *z-diff* scores, each of two scales were well correlated with different component. The first PCA component, which itself reflected the average change in global thickness across patients was correlated with the change in GAF score, whereas the second component significantly correlated with the change in PANSS score (see Fig. 5).

4 Discussion

Longitudinal neuroimaging studies allow us to assess the effectiveness of interventions and gain deeper insights into the fundamental mechanisms of underlying diseases. Despite the significant expansion of our knowledge regarding population variation through the availability of publicly accessible neuroimaging data, this knowledge, predominantly derived from cross-sectional observations, has not been adequately integrated into methods for evaluating longitudinal changes.

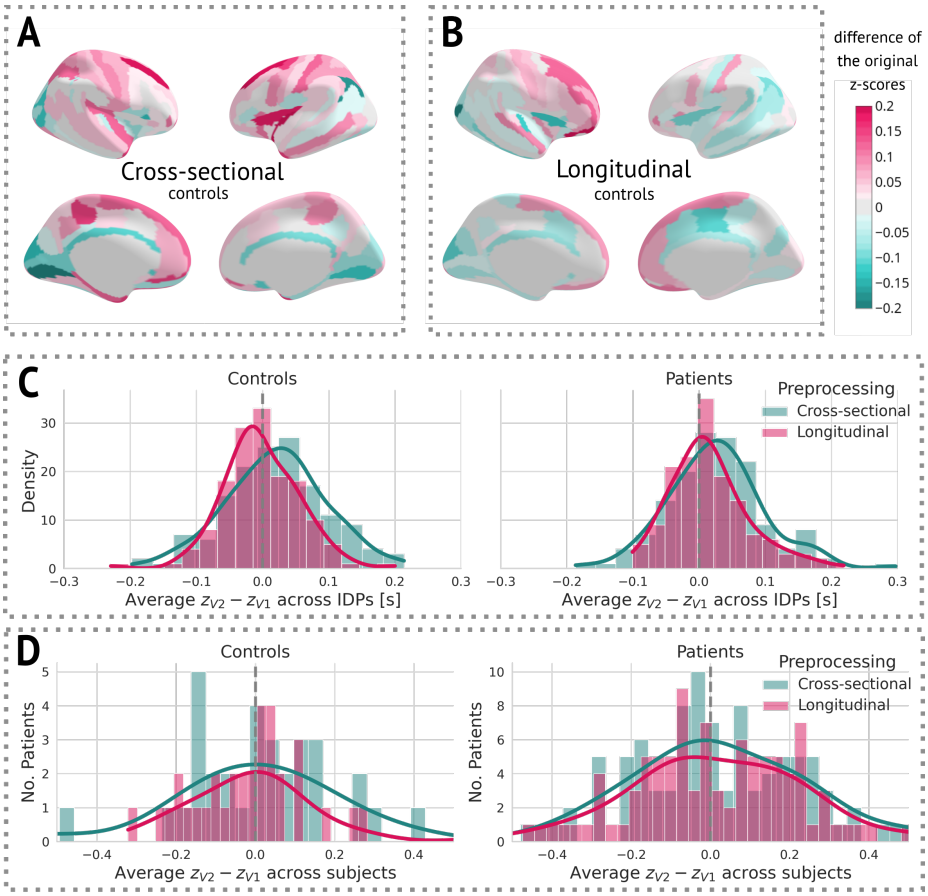


Figure 2: The effect of preprocessing across all subjects and IDPs: (A) Cross-sectional preprocessing: Heatmap of the difference of the original z-scores ($z^{(2)} - z^{(1)}$) on held-out controls. (B) Longitudinal preprocessing: Heatmap of the difference of the original z-scores ($z^{(2)} - z^{(1)}$) on held-out controls. (C) Histogram of the average ($z^{(2)} - z^{(1)}$) across all IDPs stratified by health status and preprocessing. (D) Histogram of the average ($z^{(2)} - z^{(1)}$) of each subject stratified by health status and preprocessing.

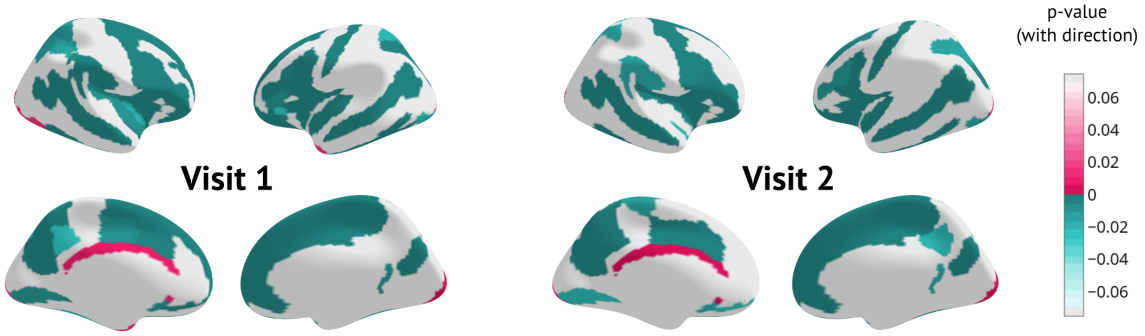


Figure 3: Cross-sectional results for each visit separately: p-values surviving Benjamini-Hochberg correction. The sign indicates the direction of change (negative means lower thickness in patients).

We propose an analytical framework that builds on normative modelling and generates unbiased features that quantify the degree of change between visits, whilst capitalising on information extracted from large cross-sectional cohorts.

4.1 Theoretical derivation

Our approach is rooted in the normative modelling method based on Bayesian regression [7], the pre-trained version of which recently became available [14]. We theoretically showed that the estimation of longitudinal changes using the results of the normative model is readily available based on a preexisting cross-sectional normative model and only requires a set of healthy controls on which the variance of healthy change might be estimated. We denoted the score obtained after running the procedure as a *z-diff* score, which quantifies the probability that such difference could have occurred between visits even in the absence of true brain change.

The core assumption of the method is that, on average, healthy controls do not deviate from their trajectory between visits. This assumption stems from the essential idea of normative modelling that, with respect to the reference population, the position of a healthy subject does not significantly deviate over time. To verify this assumption, we used the data of 33 healthy controls which were originally used for the site-specific adaptation (for more details, see the discussion part on implementation) and computed their *z-diff* scores. After averaging these scores across all subjects, the *z-diff* score of no region was statistically significant from zero (after FDR correction). However, as pointed out by a recent work studying the effect of cross-sectional nor-

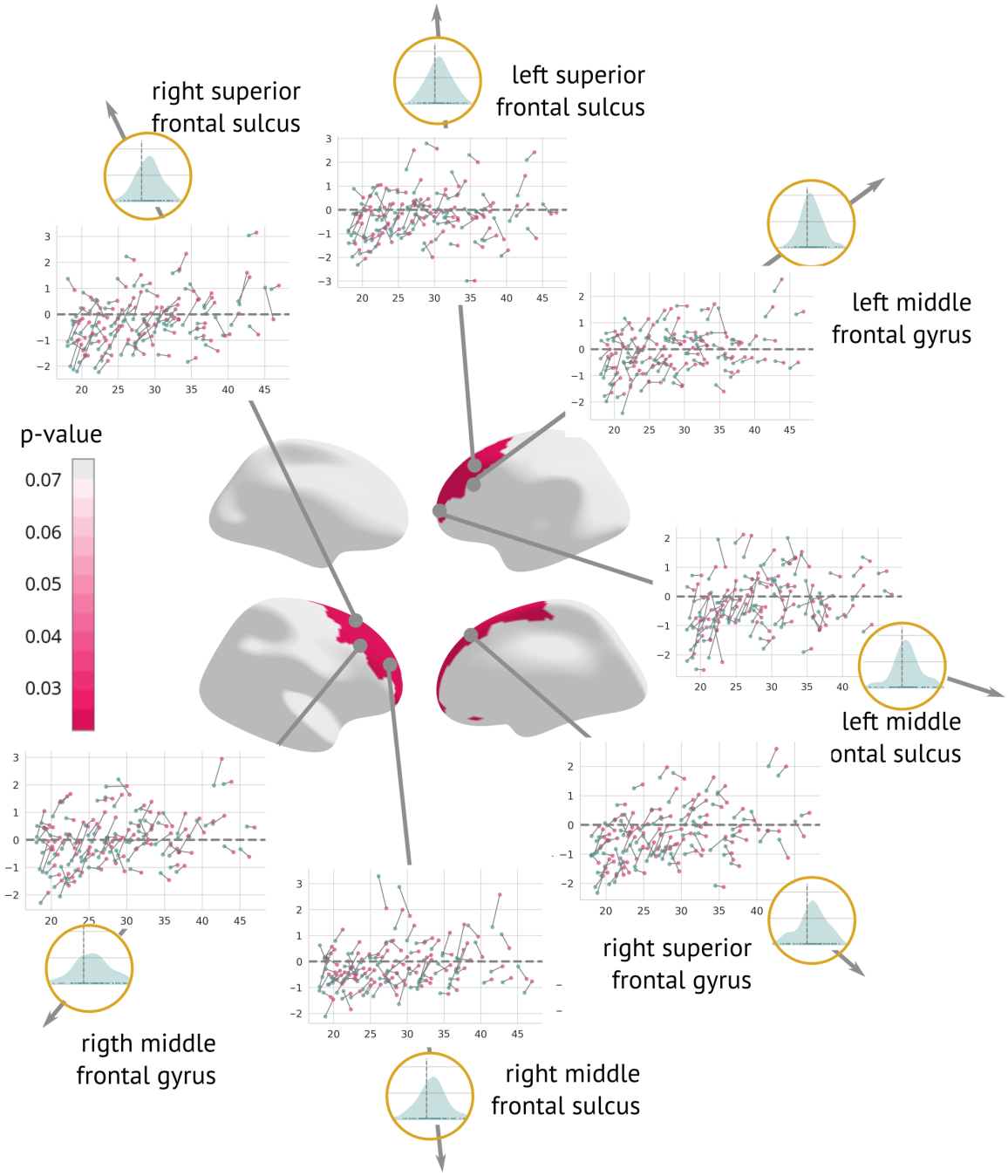


Figure 4: Regions significantly changed between the visits: Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of z-scores across all patients for both visits (the x-axis describes age, and the y-axis depicts the z-score. Blue dots represent the first and pink dots represent the second visit). The Grey dashed line highlights z=0. Histograms in the golden circles depict the distribution of the *z-diff* score.

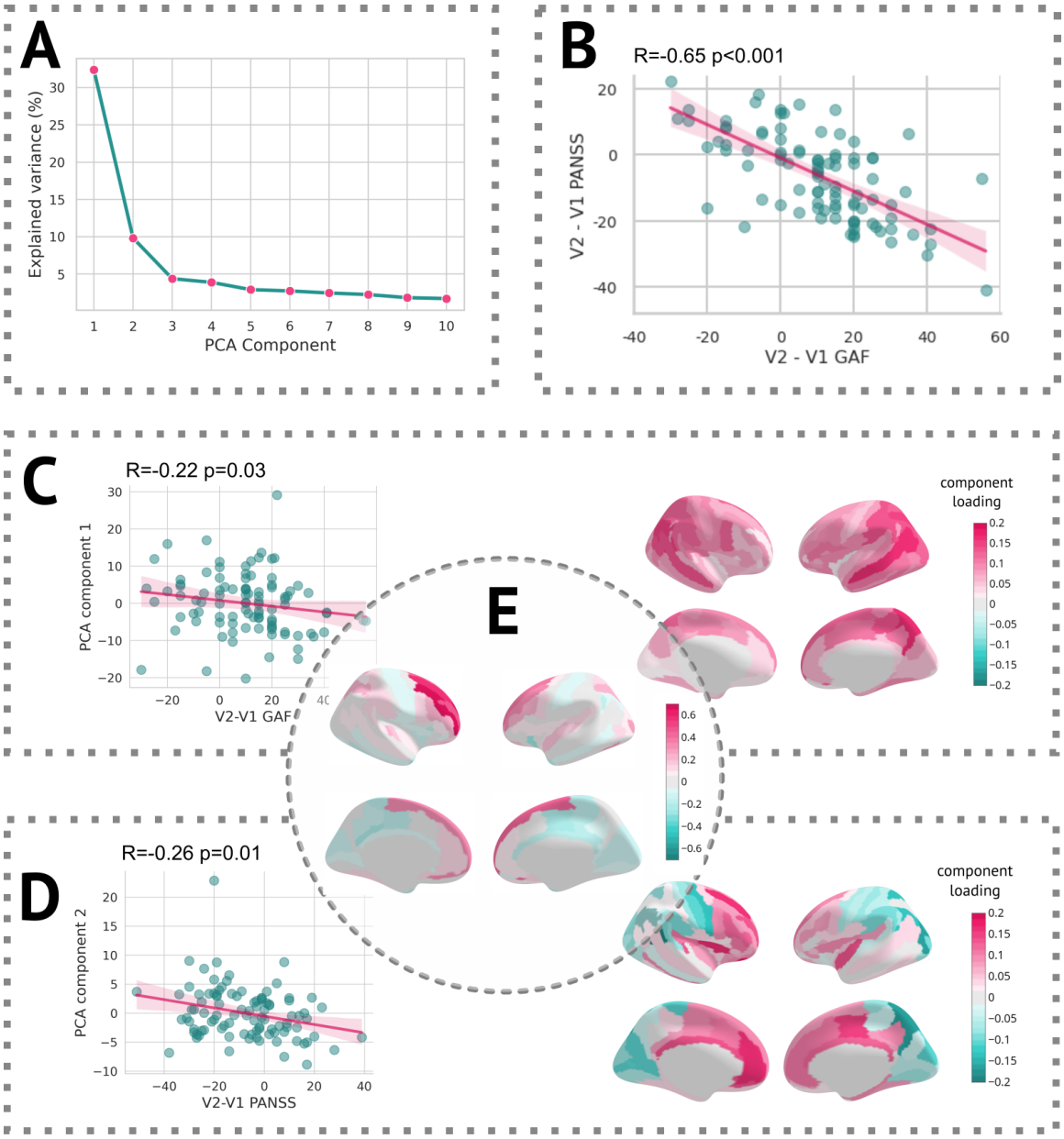


Figure 5: Results of the PCA analysis: (A) Scree plot of the explained variance of PCA components. (B) Scatterplot of change in the GAF scale vs. the change in the PANSS scale (C Left) Scatter plot of the first PCA component and difference in the GAF scale. (C Right) Heatmap of PCA loadings for the first component. (D Left) Scatter plot of the second PCA component and difference in the PANSS scale. (D Right) A Heatmap of PCA loadings for the second component. (E) Average *z-diff* score.

mative models on longitudinal predictions, the cross-sectionally derived population centiles *by design* lack information about longitudinal dynamics. Consequently, what may appear as a population-level trajectory does not necessarily align with individual subjects' actual trajectories.

To address this limitation, our model considers the variation in individual centiles. This is achieved by estimating the impact of noise and reliability, which manifests as apparent crossings of centiles observed in healthy controls. Naturally, by incorporating this element of uncertainty, the model's ability to detect subjects who experienced substantial changes in their trajectory over time decreases. As evident from the clinical findings, only a fraction of subjects were identified as having undergone significant changes (Supp. Fig. 2). However, at the group level, the significance of the observed changes persisted. Hence, while we adopt a cautious approach when assessing individual changes, the method still effectively identifies group-level changes.

Furthermore, unlike in [13], our approach does not aim to predict individual trajectories, but rather to quantify whether the observed changes over time exceed what would be expected by chance.

4.2 Implementation

At the implementation level, our model requires two stages of adaptation: site-specific adaptation, as presented in [14], and a second level where we compute the variance of healthy change (noise) in healthy controls. However, if the number of longitudinal controls is limited, the site-specific adaptation may be omitted. The purpose of site-specific adaptation is to generate unbiased cross-sectional z-scores that are zero-centred, with a variance of one for healthy controls. However, in the case of longitudinal analysis, the offset and normalization constant are irrelevant since they will be identical for both visits. Therefore, the estimation of healthy change is the only essential factor in producing the *z-diff* score. Note that in this scenario, the cross-sectional result should not be interpreted.

4.3 Clinical results

Examination of the preprocessed data showed the superiority of the longitudinal pipeline in terms of consistency over time. As our objective was to minimise the intra-individual variance, we preferred and primarily report the longitudinal approach in

further analyses. However, to assess the added benefit of the preprocessing, we also computed the core results (regions that significantly changed in time) for the cross-sectional data. The significant results were mostly consistent with a longitudinal pipeline: Six out of seven originally significant regions were still statistically significant (with the exception of the right middle frontal sulcus), and three other regions were labelled significant: the left superior frontal gyrus, the right inferior frontal sulcus, and the right medial or olfactory orbital sulcus (Supp. Fig. 3). Therefore, it is also possible to use cross-sectional preprocessing for longitudinal analysis; however, at a cost of increased between-visit variance and consequently decreased power (in comparison to the longitudinal preprocessing).

The observation of cortical normalization between the visits of early schizophrenia patients is to a degree counterintuitive and inconsistent with other works, which mostly report grey matter thinning. However, a meta-analysis of 50 longitudinal studies examining individuals with a heightened risk of psychosis revealed that 15 of the 19 studies indicated deviations in grey matter developmental trajectories between those with persistent symptoms and those whose symptoms resolved [19]. The authors propose that grey matter developmental trajectories may return to normal levels in individuals in the High-Risk Remitting group by early adulthood, whereas neurological irregularities may continue to advance in those whose symptoms do not resolve. Although our cohort had already received a diagnosis of schizophrenia, it is possible that early identification and treatment supported these compensatory mechanisms, as demonstrated by the normalization of grey matter thickness in frontal regions. Notably, the affected regions also increased in raw grey matter thickness (as measured in mm, see Supp. Fig. 4).

Furthermore, we observed significant correlations between the PCA components of the *z-diff* score and changes in clinical scales, as illustrated in Fig. 5. Notably, each clinical scale exhibited distinct associations with separate PCA components, despite substantial intercorrelations (Fig. 5 (B)).

The first PCA component, which predominantly captured global changes in grey matter thickness, displayed a negative correlation with improvements in the GAF score (Fig. 5 (C)). This unexpected inverse relationship would suggest that patients who demonstrated clinical improvement over time exhibited a more pronounced decrease in grey matter thickness, as quantified by the *z-diff* score. However, further investigation revealed that this correlation was primarily driven by the patients' GAF

scores in the initial visit. Specifically, the correlation between GAF scores at the first visit and the first PCA component yielded a coefficient of R = 0.19 (p = 0.06), whereas the correlation with scores at the second visit was R = -0.10 (p = 0.31). These findings suggest that lower GAF scores during the initial visit are predictive of subsequent grey matter thinning.

Conversely, the interpretation of the second PCA component, significantly correlated with changes in the PANSS score, was more straightforward (Fig. 5 (D)). The observed normalization of grey matter thickness in frontal areas was positively correlated with improvements in the PANSS scale, indicating that symptom amelioration was accompanied by the normalization of grey matter thickness in these regions.

Finally, we conducted an analysis of change using conventional statistical approaches to compare the results with normative modelling. Out of 148 areas tested by ANCOVA, 6 were statistically significant. However, after controlling for multiple comparisons, no IDP persisted. This result highlights the advantages of normative models and shows improved sensitivity of our method in comparison with more conventional approaches.

4.4 Limitations

Estimating the intra-subject variability is a complex task that might be affected by acquisition and physiological noise. Assumptions must be made about the longitudinal behaviour of healthy subjects. The former problem is unavoidable, whereas the latter might be addressed by constructing longitudinal normative models. However, the project necessary for such a task would have to map individuals across their lifespan consistently. The efforts to create such a dataset are already in progress through projects like the ABCD study [20], but much more data are still needed to construct a full-lifespan longitudinal model.

Additionally, our clinical results may be affected by selection bias, where subjects experiencing a worsening of their condition dropped out of the study, whereas patients with lower genetic risk or more effective treatment continued to participate.

4.5 Conclusion

In conclusion, we have successfully developed a method that utilizes pre-trained normative models to detect longitudinal changes in neuroimaging data. Our approach

offers a user-friendly implementation and has demonstrated its effectiveness through a comprehensive analysis. Specifically, we observed significant grey matter changes in the frontal lobe of schizophrenia patients over time, surpassing the sensitivity of conventional statistical approaches. This research represents a significant advancement in longitudinal neuroimaging analysis and holds great potential for further discoveries in neurodegenerative disorders.

Acknowledgements

This research was supported by the Czech Health Research Council (NU21-08-00432); the European Research Council (grant 'MENTALPRECISION', 10100118), the Wellcome Trust under an Innovator award ('BRAINCHART', 215698/Z/19/Z), the Ministry of Education, Youth and Sports (CZ.02.2.69/0.0/0.0/18_053/0017594); and the Czech Technical University Internal Grant Agency (SGS22/062/OHK3/1T/13).

References

[1] C. M. Williams, H. Peyre, R. Toro, and F. Ramus, "Neuroanatomical norms in the UK Biobank: The impact of allometric scaling, sex, and age," *Human Brain Mapping*, vol. 42, no. 14, pp. 4623–4642, 2021, ISSN: 1097-0193. DOI: 10.1002/hbm.25572.

[2] S. G. Heeringa and P. A. Berglund, *A Guide for Population-based Analysis of the Adolescent Brain Cognitive Development (ABCD) Study Baseline Data*, Feb. 2020. DOI: 10.1101/2020.02.10.942011.

[3] K. Franke and C. Gaser, "Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?" *Frontiers in Neurology*, vol. 10, 2019, ISSN: 1664-2295.

[4] C. Schuster, M. Elamin, O. Hardiman, and P. Bede, "Presymptomatic and longitudinal neuroimaging in neurodegeneration—from snapshots to motion picture: A systematic review," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 86, no. 10, pp. 1089–1096, Oct. 2015, ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp-2014-309888.

[5] A. F. Marquand, I. Rezek, J. Buitelaar, and C. F. Beckmann, "Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies," *Biological Psychiatry*, Obsessive-Compulsive Disorder, vol. 80, no. 7, pp. 552–561, Oct. 2016, ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2015.12.023.

[6] R. a. I. Bethlehem *et al.*, "Brain charts for the human lifespan," *Nature*, vol. 604, no. 7906, pp. 525–533, Apr. 2022, ISSN: 1476-4687. DOI: 10.1038/s41586-022-04554-y.

[7] C. J. Fraza, R. Dinga, C. F. Beckmann, and A. F. Marquand, "Warped Bayesian linear regression for normative modelling of big data," *NeuroImage*, vol. 245, p. 118 715, Dec. 2021, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2021.118715.

[8] M. Habes *et al.*, "The Brain Chart of Aging: Machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the iSTAGING consortium of 10,216 harmonized MR scans," *Alzheimer's & Dementia*, vol. 17, no. 1, pp. 89–102, 2021, ISSN: 1552-5279. DOI: 10.1002/alz.12178.

[9] W. H. L. Pinaya *et al.*, "Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multicohort study," *Scientific Reports*, vol. 11, no. 1, p. 15 746, Aug. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-95098-0.

[10] T. Wolfers *et al.*, "Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder," *Human Brain Mapping*, vol. 42, no. 8, pp. 2546–2555, 2021, ISSN: 1097-0193. DOI: 10.1002/hbm.25386.

[11] M. Zabihi *et al.*, "Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, The Bridging of Scales: Techniques for Translational Neuroscience, vol. 4, no. 6, pp. 567–578, Jun. 2019, ISSN: 2451-9022. DOI: 10.1016/j.bpsc.2018.11.013.

[12] A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann, "Conceptualizing mental disorders as deviations from normative

functioning," *Molecular Psychiatry*, vol. 24, no. 10, pp. 1415–1424, Oct. 2019, ISSN: 1476-5578. DOI: 10.1038/s41380-019-0441-1.

[13] M. A. Di Biase *et al.*, "Mapping human brain charts cross-sectionally and longitudinally," *Proceedings of the National Academy of Sciences*, vol. 120, no. 20, e2216798120, May 2023. DOI: 10.1073/pnas.2216798120.

[14] S. Rutherford *et al.*, "Charting Brain Growth and Aging at High Spatial Precision," Cold Spring Harbor Laboratory, Tech. Rep., Aug. 2021, ch. New Results, p. 2021.08.08.455487. DOI: 10.1101/2021.08.08.455487.

[15] T. Cole, "The development of growth references and growth charts," *Annals of Human Biology*, vol. 39, no. 5, pp. 382–394, Sep. 2012, ISSN: 0301-4460. DOI: 10.3109/03014460.2012.694475.

[16] S. M. Kia *et al.*, "Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression," *PLOS ONE*, vol. 17, no. 12, e0278776, 2022, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0278776.

[17] F. Spaniel *et al.*, "Altered Neural Correlate of the Self-Agency Experience in First-Episode Schizophrenia-Spectrum Patients: An fMRI Study," *Schizophrenia Bulletin*, vol. 42, no. 4, pp. 916–925, Jul. 2016, ISSN: 0586-7614. DOI: 10.1093/schbul/sbv188.

[18] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *Neuroimage*, vol. 61, no. 4, pp. 1402–1418, Jul. 2012, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2012.02.084.

[19] K. Merritt, P. Luque Laguna, A. Irfan, and A. S. David, "Longitudinal Structural MRI Findings in Individuals at Genetic and Clinical High Risk for Psychosis: A Systematic Review," *Frontiers in Psychiatry*, vol. 12, 2021, ISSN: 1664-0640.

[20] B. J. Casey *et al.*, "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites," *Developmental Cognitive Neuroscience*, The Adolescent Brain Cognitive Development (ABCD) Consortium: Rationale, Aims, and Assessment Strategy, vol. 32, pp. 43–54, Aug. 2018, ISSN: 1878-9293. DOI: 10.1016/j.dcn.2018.03.001.

Supplement

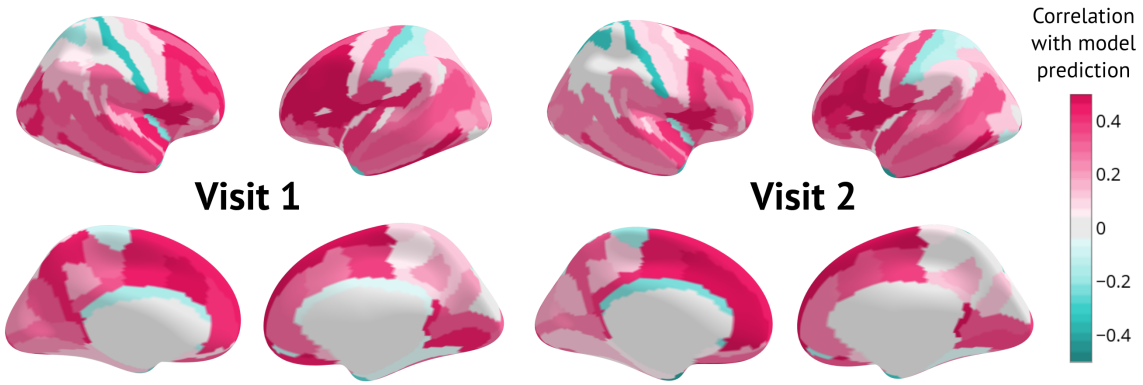
Posterior predictive distribution for difference between visits

Here we derive the posterior predictive distribution for the difference $\varphi(y^{(2)}) - \varphi(y^{(1)})$. The argument is standard. Denote $\Delta_{\mathbf{x}} = \phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})$ and $\Delta_y = \varphi(y^{(2)}) - \varphi(y^{(1)})$. Since $\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma} \sim \mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})$ and $\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w} \sim \mathcal{N}(\Delta_{\mathbf{x}}^T \mathbf{w}, 2\sigma_\xi^2)$, the posterior predictive density is

$$
\begin{aligned}
&f(\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) = \\
&\quad = \int f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \mathbf{w}, 2\sigma_\xi^2)}(\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w}) \cdot f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})}(\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) \, \mathrm{d}(\Delta_{\mathbf{x}}^T \mathbf{w}) \\
&\quad = \int f_{\mathcal{N}(0, 2\sigma_\xi^2)}(\Delta_y - \Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w}) \cdot f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})}(\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) \, \mathrm{d}(\Delta_{\mathbf{x}}^T \mathbf{w}).
\end{aligned}
$$

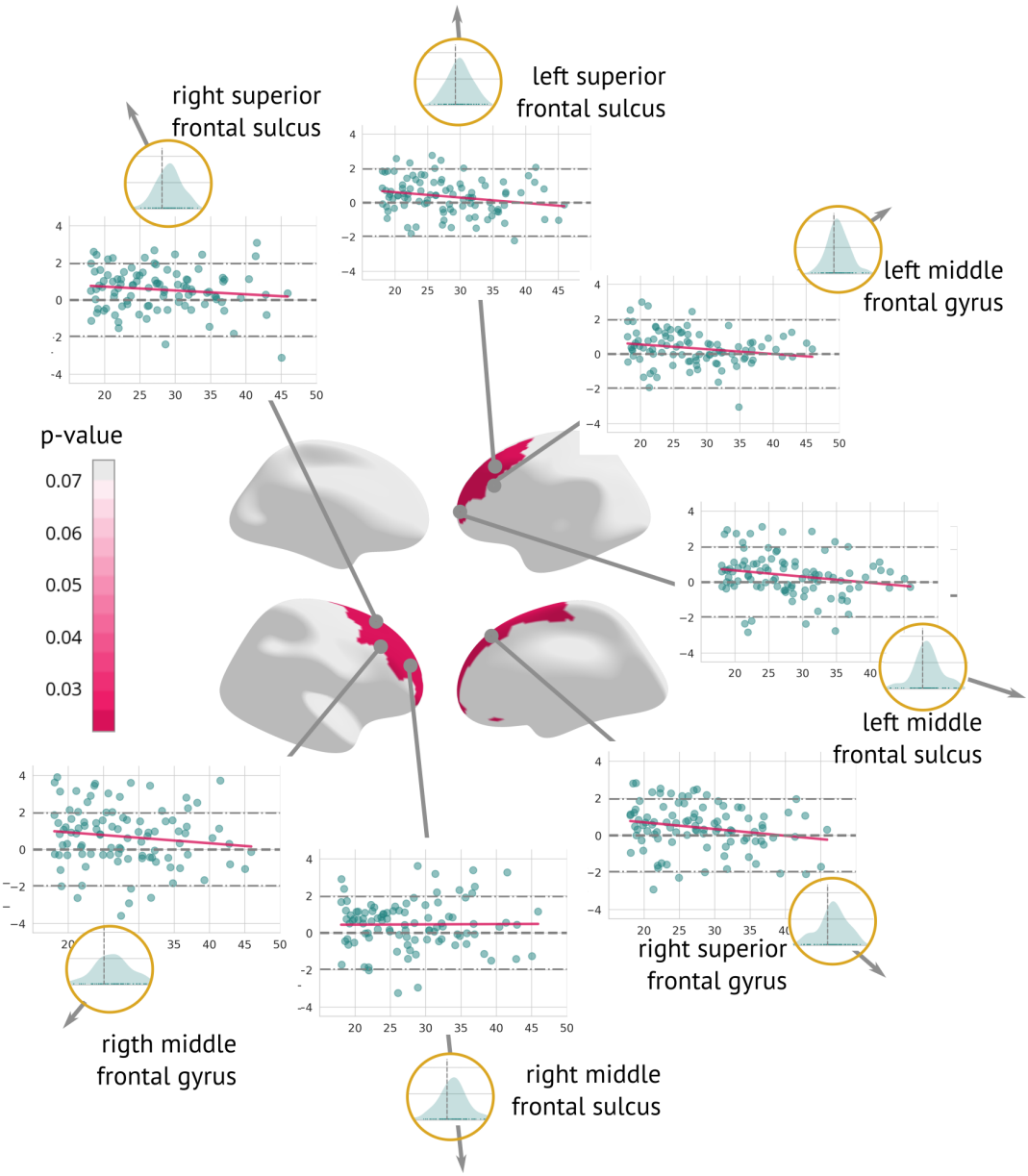
This has the familiar convolution form of the densities of $\mathcal{N}(0, 2\sigma_\xi^2)$ and $\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})$. It is known to produce the density of $\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}} + 2\sigma_\xi^2)$ (by completion to squares in the exponent).

Quality of fit across regions of interest

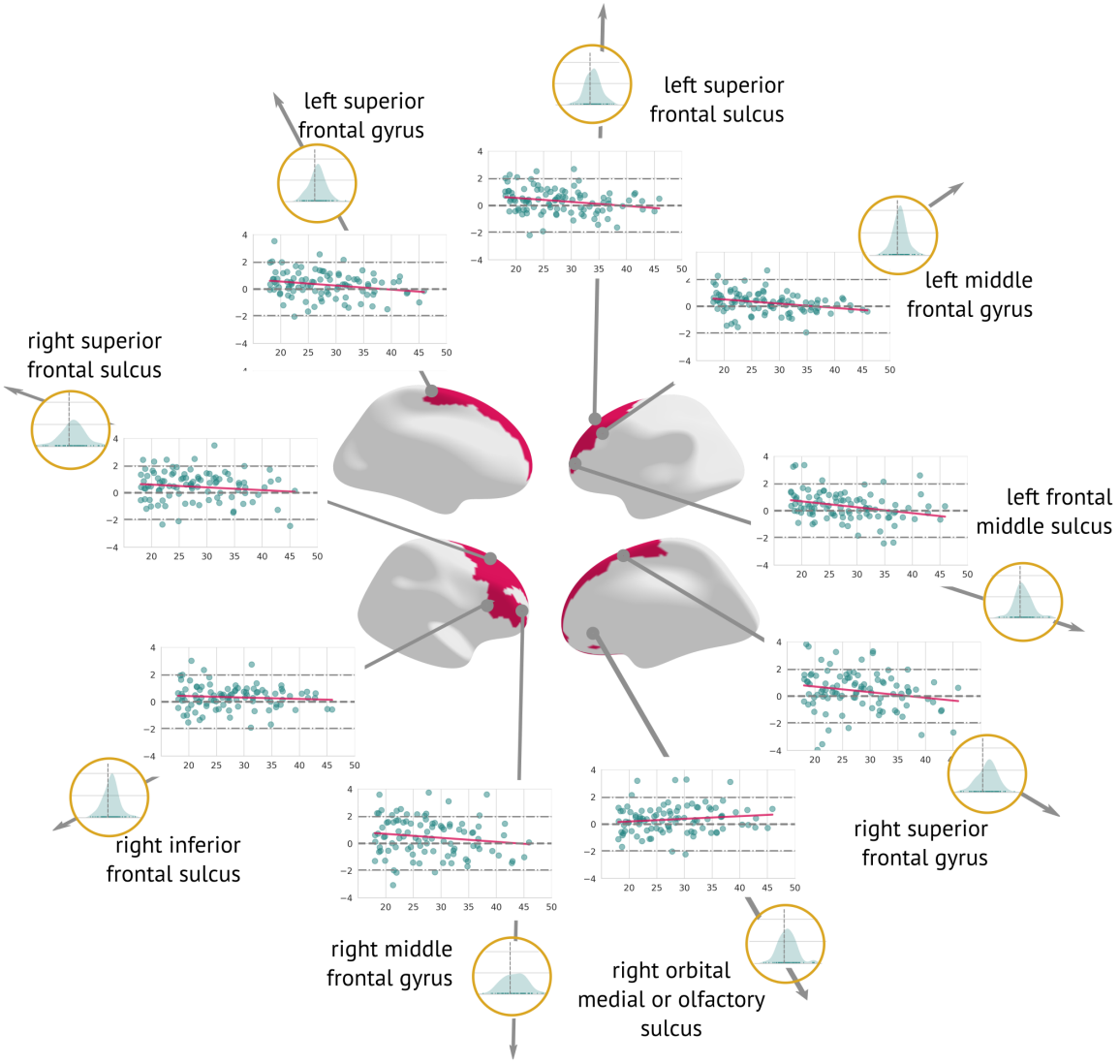


Supplementary Figure 1: **Quality of fit as measured by Rho for the first and the second visit.**

Comparison of preprocessing

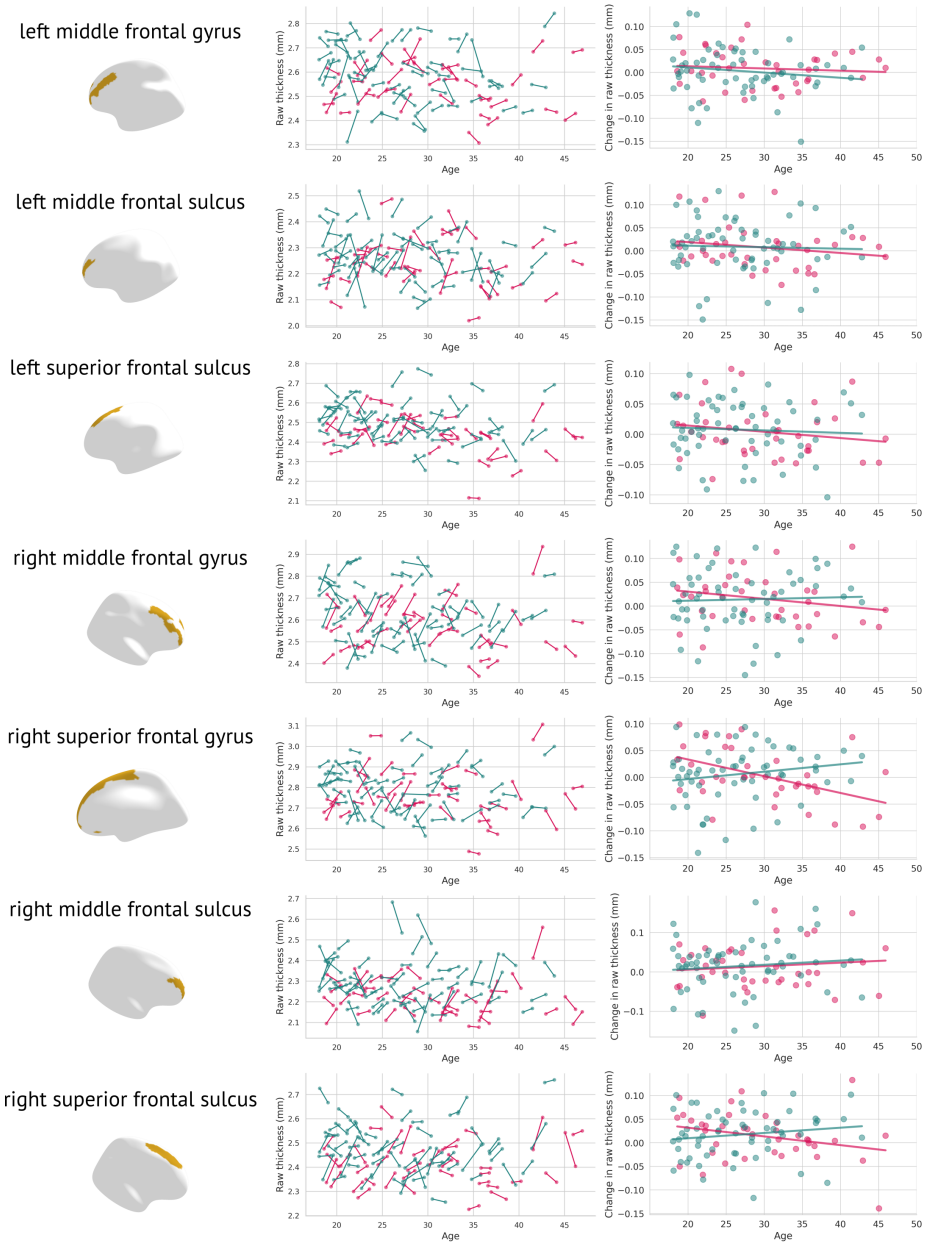


Supplementary Figure 2: Regions significantly changed between the visits (longitudinal preprocessing): Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of *z-diff* across all patients for both visits (the x-axis describes age, and the y-axis depicts the *z-diff*. Blue dots represent individual patients and the pink line shows a trend of *z-diff* change). The Grey dashed line highlights z=0. Histograms in the golden circles depict the distribution of the *z-diff* score.



Supplementary Figure 3: Regions significantly changed between the visits (cross-sectional preprocessing): Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of *z-diff* scores across all patients for both visits (the x-axis describes age, and the y-axis depicts the *z-diff* score. The grey dashed line highlights z=0. Histograms in the golden circles depict the distribution of the *z-diff* score.

Raw changes observed in significant regions



Supplementary Figure 4: Raw changes in grey matter thickness: Each significantly changed region is presented twice, once as a scatter plot containing the original grey matter thickness for both visits (left); females are plotted in pink, males in blue. The figure on the right depicts V2-V1 in raw thicknesses (separately for females – pink, and males – blue).