

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

Barbora Reháková Bučková, Charlotte Frazz, Rastislav Reháček, Marián Koleniň, Christian Beckmann, Filip Španiel, Andre Marquand , Jaroslav Hlinka 

Department of Complex Systems, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic • Department of Cybernetics, Czech Technical University in Prague, Prague, Czech Republic • National Institute of Mental Health, Klecany, Czech Republic • Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands • Max Planck Institute for Research on Collective Goods, Bonn, Germany • University of Cologne, Germany

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This paper addresses an **important** topic (normative trajectory modelling), seeking to provide a method aiming to accurately reflect the individual deviation of longitudinal/temporal change compared to the normal temporal change characterized based on a pre-trained population normative model. The evidence provided for the new methods is, however, **incomplete**, with the simulations validating the method needing to be extended.

<https://doi.org/10.7554/eLife.95823.2.sa2>

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.

We build on normative modelling framework, which enables the evaluation of an individual's position compared to a population standard. We extend this framework to evaluate an individual's longitudinal *change* compared to the longitudinal change reflected by the (population) standard dynamics. Thus, we exploit the existing normative models pre-trained on over 58,000 individuals and adapt the framework so that they can also be used in the evaluation of longitudinal studies. Specifically, we introduce a quantitative metric termed “z-*diff*” score, which serves as an indicator of a temporal change of an individual compared to a

population standard. Notably, our framework offers advantages such as flexibility in dataset size and ease of implementation.

To illustrate 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 normalisation 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 a 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. Consequently, it is crucial to have tools to effectively analyse them whilst also making use of more widely available cross-sectional data to refine inferences. Therefore, in this manuscript, we develop a novel method for evaluating longitudinal *changes* in a subject's neuroimaging data, building upon an existing normative modelling framework originally designed to assess a subject's *position* within a population. This adaptation allows us to track individual changes over time, providing a more dynamic understanding of neuroanatomical variations.

Normative modelling is a promising technique for modelling population variation in neuroimaging data [1, 2]. This framework models each image-derived phenotype (IDP) (e.g., voxel intensity, regional thickness, or regional volume) independently as a function of demographic or clinical variables (e.g., age, sex, scanning site) in a large healthy population. Subjects are subsequently compared to the normative model characterizing the healthy population, which enables us to evaluate the position of each individual, rather than just compare group differences between patients and controls [3, 4]. 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 [5–8].

Longitudinal data are conceptually well suited to extend standard 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 temporal change, thereby enhancing our understanding of the disease.

However, despite their potential, longitudinal normative models have not yet been systematically explored [2, 9]. Indeed, virtually all large-scale normative models released to date are estimated on cross-sectional data [2, 10] and a recent report [9] has provided empirical data to suggest that such cross-sectional models may underestimate the variance in longitudinal data [9]. 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., [11]) that centiles in such cross-sectional models 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 (e.g., see [11] 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 [12]; however, we do not tackle this problem here.

We derive a novel set of difference (“z-diff”) scores for statistical evaluation of longitudinal change between two measurements (the “diff” in the name refers to the temporal difference that we are evaluating as opposed to a one-time position evaluated by the simple z-score). We utilise the Warped Bayesian Linear Regression normative model [3] 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. [10]. We present a comprehensive theoretical analysis of our method, followed by numerical simulations and a 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

Here, we briefly present the original normative model [3], developed to be trained and used on cross-sectional data. In the following subsection, we take this model pre-trained on a large cross-sectional dataset, and extend it so that it can be used on longitudinal data.

The original model [3] is pre-trained on a cross-sectional dataset $\mathbf{Y} = (y_{nd}) \in \mathbb{R}^{N \times D}$, $\mathbf{X} = (x_{nm}) \in \mathbb{R}^{N \times M}$ of N subjects, for whom we observe D IDPs and M covariates (e.g., age or sex). Thus, y_{nd} is the d -th IDP of the n -th subject and x_{nm} is the m -th covariate of the n -th subject.

Since each IDP is treated separately, we focus on a fixed IDP d and drop this index for ease of exposition. To simplify notation, we denote $\mathbf{y} = (y_1, \dots, y_N)^T$ the column of observations of this fixed IDP across subjects. The observations are assumed to be independent (across n subjects). To model

the relationship between IDP y_n and covariates $\mathbf{x}_n = (x_{n1}, \dots, x_{nM})^T$, we want to exploit a normal linear regression model described in [3, 10]. However, we make a couple of adjustments first:

- To accommodate non-Gaussian errors in the original space of dependent variables, we transform the original variable y_n by a warping function $\phi(y_n)$, which is parametrised by hyper-parameters γ (see Section 2.3 in [3] for details).
- To capture non-linear relationships, we use a B-spline basis expansion of the original independent variables \mathbf{x}_n (see Section 2.3 in [3] for details). To accommodate site-level effects, we append it with site dummies. We denote the resulting transformation of \mathbf{x}_n as $\phi(\mathbf{x}_n) \in \mathbb{R}^K$.

We also treat the variance of measurements as a hyper-parameter and we denote it by σ^2 . Thus, we model the distribution of the transformed IDP $\phi(y_n)$ conditional on covariates \mathbf{x}_n , vector of parameters \mathbf{w} , and hyper-parameters σ^2 and γ as

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

where ε_n are independent from \mathbf{x}_n and across n . We further denote the design matrix $\Phi = (\phi(\mathbf{x}_n)_k) \in \mathbb{R}^{N \times K}$ ($\phi(\mathbf{x}_n)_k$ is the k -th element of vector $\phi(\mathbf{x}_n)$).

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

$$\mathcal{N}(0, \omega^2 \mathbf{I}) \quad (2)$$

is combined with the likelihood function to derive the posterior

$$\mathbf{w} | \mathbf{y}, \Phi; \omega^2, \sigma^2, \gamma \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{A}^{-1}), \quad (3)$$

$$\mathbf{A} = \sigma^{-2} \Phi^T \Phi + \omega^{-2} \mathbf{I}, \quad (4)$$

$$\bar{\mathbf{w}} = \sigma^{-2} \mathbf{A}^{-1} \Phi^T \mathbf{y}. \quad (5)$$

The hyper-parameters $\omega^2, \sigma^2, \gamma$ are estimated by maximising the warped marginal log-likelihood.

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

$$\mathcal{N}(\bar{\mathbf{w}}^T \phi(\mathbf{x}), \phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x}) + \sigma^2). \quad (6)$$

Hence, the z-score characterising the position of this subject within population is

$$z = \frac{\phi(y) - \bar{\mathbf{w}}^T \phi(\mathbf{x})}{\sqrt{\phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x}) + \sigma^2}}, \quad (7)$$

where $\phi(y)$ is the realised warped observation of IDP d for this subject. This score captures how surprising is the actual observation $\phi(y)$ relative to what one would expect for an average subject with the same characteristics $\bar{\mathbf{w}}^T \phi(\mathbf{x})$, and this deviation has to be compared to (normalized by) the variability stemming from the natural variability in the data (σ^2) and the modelling uncertainty ($\phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x})$).

In this form, the original models were fit on a large dataset consisting of 58,836 participants scanned across 82 sites. Specifically, cortical thickness and subcortical volumes were modelled, and the models were validated against a subset of 24,000 participants, the quality of which were checked manually [10].

Note that formulae (6) and (7) implicitly evaluate only (potentially new) subjects measured at sites already present in the original dataset \mathbf{y}, Φ . If we want to evaluate subjects measured at a new site, we will have to run an adaptation procedure to account for its effect. This adaptation procedure is described and readily accessible online in [10]. 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.

In the following section, we develop a procedure that allows us to extend the original cross-sectional framework pre-trained on dataset \mathbf{y}, Φ to evaluate a new longitudinal dataset for assessment of temporal changes.

2.1.2 Adaptation to longitudinal data

We adapt the original cross-sectional normative modelling framework [3] (reviewed in the previous section) to the evaluation of intra-subject longitudinal changes. Specifically, we design a score for a longitudinal change between visits (further referred to as *z-diff* score), based on which we can assess temporal changes in regional brain thickness and potentially detect any unusually pronounced deviations from normative trajectories.

We start by noticing that the original cross-sectional normative modelling framework [3] features an implicit assumption that pertains to the longitudinal view. Specifically, it assumes that had we randomly sampled the population at a different time (e.g., 5 years sooner or later), we would have gotten equivalent picture about the “norm” (up to randomness of the sampling, **Figure 1 A**). In other words, the parameters of the normative model would be the same irrespective of the time we sampled the population, including the case in which we would sample the same people again, just later (while appropriately compensating for the resulting under-representation of younger ages). We further assume comparability of people of any given age irrespective of their birth time (i.e., we assume independence of birth dates and trajectories, **Figure 1 B**). Together, these assumptions imply a form of stationarity (formally discussed in the next paragraph). These are indispensable assumptions for the practical usefulness of normative modelling, albeit one can see that in the real world they are not fulfilled perfectly, e.g., due to evolutionary dynamics, the ever-changing environment, or any changes in the distribution of those demographic variables that are not explicitly accounted for in the normative model.

Formally, we work with the process $\{(y_{n,t}, \mathbf{x}_{n,t})\}$ where n indexes a subject and t indexes age (to avoid technicalities, we assume discrete time). The minimal requirement imposed implicitly by the above assumptions is $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ for every age t . We further restrict our focus to the class of stationary Gaussian processes ε , i.e., processes such that $\varepsilon_{t_1}, \varepsilon_{t_2}, \dots, \varepsilon_{t_k}$ are jointly normal for any finite set of ages t_1, t_2, \dots, t_k and their joint distribution is invariant to any admissible time shift $t_1 + s, t_2 + s, \dots, t_k + s$. Furthermore, we focus on a specific process in this section for ease of exposition, and we discuss the general case in the next section.

The specific process we focus on in this section reflects the assumption that a healthy subject does not deviate substantially from their position within the population as they get older [8]; the observed position change between the visits stems from observation noise (due to technical or physiological factors) and is therefore constrained by the test-retest reliability of the measurement. Formally,

$$\varepsilon_{n,t} = \eta_n + \xi_{n,t},$$

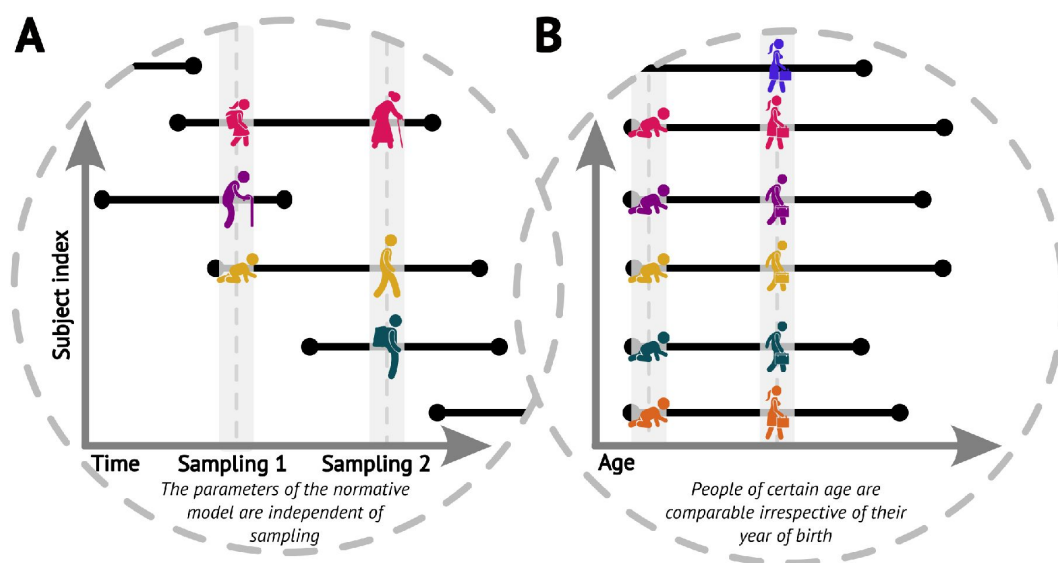


Figure 1

Visualisations of the core assumptions of normative modelling: (A) The parameters of the fitted normative model are independent of the time of sampling. (B) People of the same age are comparable irrespective of their year of birth (datasets sampled at different times can be combined)

where η is a subject-specific time-independent factor independent of the iid noise process ξ . Note that this does not imply that a healthy subject does not change over time, but rather that the change follows approximately the population centile at which the individual is placed. We generalise our method to other stationary Gaussian processes in the next section.

According to this model, the i -th visit of a healthy subject with given covariates $\mathbf{x}^{(i)}$ is generated by

$$\begin{aligned}\varphi(y^{(i)}) &= \mathbf{w}^T \phi(\mathbf{x}^{(i)}) + \eta + \xi^{(i)} \\ \eta &\sim \mathcal{N}(0, \sigma_\eta^2) \\ \xi^{(i)} &\sim \mathcal{N}(0, \sigma_\xi^2) \\ \sigma^2 &= \sigma_\eta^2 + \sigma_\xi^2\end{aligned}\tag{8}$$

where η , $\xi^{(i)}$, and $\mathbf{x}^{(i)}$ are mutually independent for a given i , and the measurement errors $\xi^{(i)}$ and $\xi^{(j)}$ are independent across visits $i \neq j$. Note that we dropped the subject-specific index n (and subsumed the age $t_n^{(i)}$ in the visit index i). This should remind the reader that the goal is just to evaluate longitudinal change of a given subject from our new longitudinal data, and not to re-estimate the parameters with these additional data. Nevertheless, to properly adapt the cross-sectional model, we will need to estimate one new parameter stemming from the further structure we impose on ε .

In our longitudinal data, we are interested in the change for a given individual across two visits. According to model (8), the difference in the transformed IDP between visits 1 and 2, $\phi(y^{(2)}) - \phi(y^{(1)})$, for a subject with covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is given by

$$\varphi(y^{(2)}) - \varphi(y^{(1)}) = \mathbf{w}^T [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + \xi^{(2)} - \xi^{(1)}\tag{9}$$

with $\xi^{(2)} - \xi^{(1)} \sim \mathcal{N}(0, 2\sigma_\xi^2)$. We use the posterior distribution of \mathbf{w} with hyper-parameters ω^2 , σ^2 , \mathbf{y} estimated on the original cross-sectional dataset \mathbf{y} , Φ (the estimates are available at <https://github.com/predictive-clinical-neuroscience/braincharts>). Therefore, the posterior predictive distribution for the difference $\phi(y^{(2)}) - \phi(y^{(1)})$ for our subject is (for more detailed derivation, please refer to the supplement)

$$\mathcal{N}(\bar{\mathbf{w}}^T [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})], [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2).\tag{10}$$

Hence, the z -score for the difference in the transformed IDP between visits 1 and 2 is

$$z\text{-diff} = \frac{[\varphi(y^{(2)}) - \varphi(y^{(1)})] - \bar{\mathbf{w}}^T [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]}{\sqrt{[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2}},\tag{11}$$

where $\phi(y^{(2)}) - \phi(y^{(1)})$ is the realised temporal change in the warped observations of the IDP 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 unusual temporal changes.

The primary role of adaptation of the (pre-trained) cross-sectional model to (new) longitudinal data is to account for the measurement noise variance σ_ξ^2 , thus taking care of the atemporal source of variability η . In other words, having an estimate of σ_ξ^2 in hand helps us to use the proper scaling. To arrive at an estimator of σ_ξ^2 , notice that from the posterior predictive distribution (10), we have (denoting the set of conditionals $\Omega = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \Phi; \omega^2, \sigma^2, \gamma\}$)

$$\begin{aligned} & \mathbb{E} \left[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \mathbb{E}[\varphi(y^{(2)}) - \varphi(y^{(1)}) | \Omega] \right)^2 \middle| \Omega \right] \\ &= [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2. \end{aligned} \quad (12)$$

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

$$\begin{aligned} & \mathbb{E} \left[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \bar{\mathbf{w}}^T [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] \right)^2 \right. \\ & \quad \left. - [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] \middle| \mathbf{y}, \Phi; \omega^2, \sigma^2, \gamma \right] = 2\sigma_\xi^2. \end{aligned} \quad (13)$$

Therefore, we estimate $2\sigma_\xi^2$ by the sample analogue of the left-hand side in (13). Specifically, we devote a subsample C of the controls from our (new) longitudinal data just to this estimation (i.e., the subjects from C will not be used in the evaluation) and we compute

$$\begin{aligned} \widehat{2\sigma_\xi^2} = \frac{1}{|C|} \sum_{k \in C} & \left[\left(\varphi(y_k^{(2)}) - \varphi(y_k^{(1)}) - \bar{\mathbf{w}}^T [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})] \right)^2 \right. \\ & \left. - [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})] \right], \end{aligned} \quad (14)$$

where $|C|$ denotes the number of subjects in subsample C .

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 (17) is negligible in adult cohorts but must be treated with caution in developmental or ageing groups. Finally, it is apparent from (4) that \mathbf{A} scales with the number of subjects, and its inverse will be negligible for substantial training datasets, such as the one that was used for pre-training.

To conclude this subsection, we caution against the naive use of the difference of the simple z -scores

$$\frac{\varphi(y^{(2)}) - \bar{\mathbf{w}}^T \phi(\mathbf{x}^{(2)})}{\sqrt{\phi(\mathbf{x}^{(2)})^T \mathbf{A}^{-1} \phi(\mathbf{x}^{(2)}) + \sigma^2}} - \frac{\varphi(y^{(1)}) - \bar{\mathbf{w}}^T \phi(\mathbf{x}^{(1)})}{\sqrt{\phi(\mathbf{x}^{(1)})^T \mathbf{A}^{-1} \phi(\mathbf{x}^{(1)}) + \sigma^2}} \quad (15)$$

instead of the z -diff score to evaluate the longitudinal change. The problem with such an approach is apparent by comparing it to the z -diff score (11): it does not properly account for the modelling uncertainty (instead of using the combined term $[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1} [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]$ to scale the difference of the numerators in (15), it scales the individual terms of the difference by their individual model uncertainty). More importantly, even if the modelling uncertainty is negligible, expression (15) does not properly scale the difference of the “residuals” because it incorrectly includes the common source of subject-level variability η (it uses $\sigma_\eta^2 + \sigma_\xi^2$ instead of $2\sigma_\xi^2$).

2.1.3 More general dynamics

The model we introduced in the previous section is an intuitive extension of the original model introduced in [section 2.1.1](#). However, the model operates with a seemingly strong (although reasonable) assumption that healthy subjects inherently follow their centiles. Due to the lack of large longitudinal data testing this assumption, in this section, we investigate the generalisation to other stationary Gaussian processes to illustrate the robustness of our method. As an example, we are able to deal with a stationary Gaussian AR(1) process $\varepsilon_t = \Phi\varepsilon_{t-1} + \xi_t$ with $|\Phi| < 1$, ξ iid $\mathcal{N}(0, (1 - \Phi^2)\sigma^2)$, and $\varepsilon^0 \sim \mathcal{N}(0, \sigma^2)$.

Importantly, our framework evaluates change only between two visits. Hence, we do not need to consider the full specification of the process ε , but only the time-dependence between the two visits that can arise under it. Formally, since we are in the class of stationary Gaussian processes, we only need to consider the autocorrelation between $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ $\rho \in [-1, 1]$. Just as an example, the stationary Gaussian AR(1) process introduced above would produce autocorrelation $\rho = \zeta^{T_2 - T_1}$, where $t_2 - t_1$ is the time between the two visits.

Considering this more general class of processes, this amounts to $\varepsilon^{(2)} - \varepsilon^{(1)} \sim \mathcal{N}(0, 2\sigma^2(1 - \rho))$. Going through the same derivations as before, we obtain the score for the evaluation of longitudinal change

$$z\text{-diff} = \frac{[\varphi(y^{(2)}) - \varphi(y^{(1)})] - \bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]}{\sqrt{[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma^2(1 - \rho)}} \quad (16)$$

and the estimator

$$2\sigma^2(1 - \rho) = \frac{1}{|C|} \sum_{k \in C} \left[\left(\varphi(y_k^{(2)}) - \varphi(y_k^{(1)}) - \bar{\mathbf{w}}^T[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})] \right)^2 - [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})] \right]. \quad (17)$$

These take the same form as in the specific case considered in the previous section. Hence, the method developed in the previous section does not depend on that particular assumption about the process ε and will still yield valid inferences even if the seemingly strong assumption of centile tracking is violated. In either case, we need one more free parameter to properly account for the potential non-iid dynamics of σ_ξ^2 in the previous section, ρ in this section. The only substantial difference is that while $2\sigma^2(1 - \rho)$ can be larger than $2\sigma^2$ (for $\rho < 0$), the process from the previous section leads to $2\sigma_\xi^2$ with values only lower than $2\sigma^2$. This could provide a test of the assumption about the process from the previous section: if our estimate $\widehat{2\sigma_\xi^2}$ is larger than the cross-sectional estimate $\widehat{2\sigma^2}$, then the assumption about ε in the previous section is not justified.

2.1.4 Simulation study

To formally evaluate the performance of the proposed method in making accurate inferences about the longitudinal changes, we conduct a simulation study. We imagine a practitioner who would use some lower and upper thresholds for the $z\text{-diff}$ score to detect unusual change. It is natural to choose the probability of dubbing a healthy control as unusual $\theta \in [0, 1]$, and use the $\frac{\theta}{2}$ and $1 - \frac{\theta}{2}$ quantiles of the standard normal distribution as the thresholds (denote them $q_{\frac{\theta}{2}}$ and $q_{1 - \frac{\theta}{2}}$ respectively). A subject with $z\text{-diff} < q_{\frac{\theta}{2}}$ Or $> q_{1 - \frac{\theta}{2}}$ is thus flagged as someone with unusual change. We would like to know how successfully this classification detects true changes, i.e., how

often it detects a patient with a disrupted trajectory. We capture the disruption by a process δ , i.e., the trajectory of a patient in our model would be $\phi(y_t) = \mathbf{w}^T \phi(\mathbf{x}_t) + \varepsilon_t + \delta_t$. We treat the realised change in the disruption between the two visits $\Delta := \delta(2) - \delta(1)$ as a fixed number to be detected.

For the simulation, we fix $\theta = 0.05$. For each combination of $\Delta \in [-4, 4]$ and $\rho \in (-1, 1)$, we generate a large number of patients with various age and gender, disruption Δ , and $(\varepsilon^{(2)}, \varepsilon^{(1)})$ from the bivariate normal distribution

$$\mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right). \quad (18)$$

Specifically, we produce $\varphi(y^{(1)}) = \bar{\mathbf{w}}^T \phi(\mathbf{x}^{(1)}) + \varepsilon^{(1)}$, $\varphi(y^{(2)}) = \bar{\mathbf{w}}^T \phi(\mathbf{x}^{(2)}) + \varepsilon^{(2)} + \Delta$ (the remaining parameters are the cross-sectional estimates). For each patient, we calculate the *z-diff* score and we look at the fraction of patients with a *z-diff* score surpassing the thresholds. The resulting simulation is depicted in **Figure 2**.

Two intuitive properties arise from this simulation: larger disruptions are easier to detect; positive autocorrelation in ε makes it easier to detect the disruptions, while negative autocorrelation makes it harder. Strong positive autocorrelation reflects a strong common component in ε across the two visits, which cancels out through subtracting the two visits, while strong negative autocorrelation indicates strong switching in ε across the two visits, which can be easily confounded with the true disruption Δ . Finally, if the true process for ε is as assumed in (8) (stable component plus noise), then higher noise σ_ε^2 corresponds to lower (but positive) ρ . Intuitively, we can see that more noise makes the detection of true disruptions more difficult.

2.1.5 Implementation

To implement the method (**Fig. 3**), we used the **PCN toolkit**. The exact steps of the analysis with detailed explanations are available in the online tutorial at PCNtoolkit-demo (<https://github.com/predictive-clinical-neuroscience/PCNtoolkit-demo>) in the tutorials section.

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 [13]. 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.

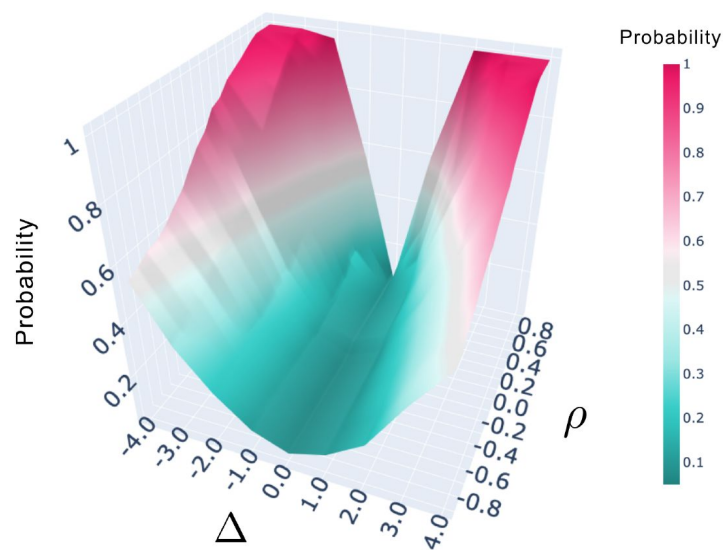


Figure 2

Probability of detecting a true disruption Δ for various values of autocorrelation ρ ($\theta = 0.05$)

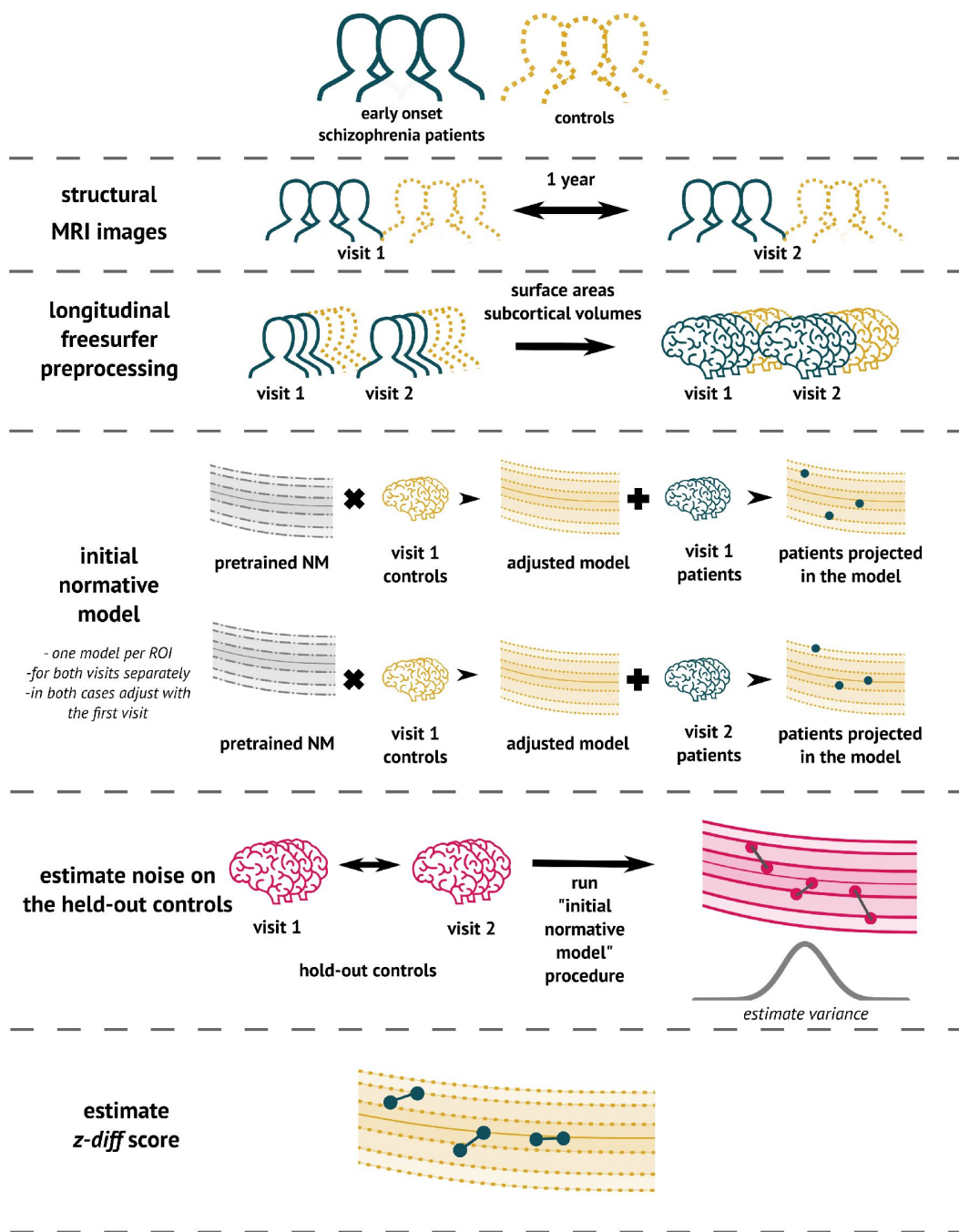


Figure 3

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 σ_{ξ}^2 in healthy subjects is estimated using held-out controls, and finally, the *z-diff* score is computed.

	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)
<i>Diagnosis (only for patients)</i>		
Schizophrenia	53	
Brief psychotic disorder	45	
Length of disease, median (min, max), months	4 (1, 21)	
<i>Clinical scales (only for patients)</i>		
	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

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.

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 × 320 mm.

2.3 Preprocessing and Analysis

Prior to normative modelling, all T1 images were preprocessed using the Freesurfer v.(7.2) recon-all pipeline. While in the context of longitudinal analysis the longitudinal Freesurfer preprocessing pipeline is appropriate, we additionally performed cross-sectional preprocessing [14]. The reason to conduct this analysis is threefold: First, the impact of preprocessing on the z-scores of normative models lacks prior investigation. Second, the training dataset of 58,000 subjects initially underwent cross-sectional preprocessing, introducing a methodological incongruity. Third, certain large-scale studies, constrained by computational resources, exclusively employ cross-sectional preprocessing. Understanding the consistency of results between the two approaches becomes crucial in such cases.

In line with [10], we performed a simple quality control procedure whereby all subjects having a rescaled Euler number greater than ten were labelled outliers and were not included in the analysis (Table 1) (see [10] and [12] for further details).

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**). The pre-trained model used for adaptation was the `lifespan_58K_82_sites` [10]. For each subject and visit, we obtained cross-sectional z-score, as well as the underlying values needed for its computation, particularly $\phi(y)$ and $\bar{w}^T \phi(x)$. We conducted a cross-sectional analysis of the original z-scores to evaluate each measurement independently. We then tested for the difference of the cross-sectional z-scores $z^{(2)} - z^{(1)}$ between the patients and held-out controls using Mann-Whitney U test and corrected for multiple tests using the Benjamini-Hochberg FDR correction at the 5% level of significance.

Subsequently, following (11), 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 longitudinal 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 longitudinal analysis. As all visits were approximately one-year apart, we conducted an analysis of covariance (ANCOVA). The ANCOVA model combines a general linear model and ANOVA. Its purpose is to examine whether the means of a dependent variable (thickness in visit 2) 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 visit 1). 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 longitudinal 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

3 Results

3.1 Effect of preprocessing

After obtaining cross-sectional z-scores for both types of preprocessing, we visually observed a decrease in variance between the two visits in longitudinal preprocessing compared to the cross-sectional one (**Figure 4** [↗](#)). More specifically, we calculated the mean of the difference between z-scores of visit 2 and visit 1 for each individual IDP, stratified by preprocessing and group, across all subjects. We then visualised the distribution of these means using a histogram (**Figure 4C** [↗](#)). Alternatively, we also computed the mean difference between z-scores of visit 2 and visit 1 across all IDPs for each subject, and plotted a histogram of these values. Note that this step was only done to estimate the effect of preprocessing on z-scores for further discussion. Its impact on the results is elaborated on in the discussion.

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 5** [↗](#)).

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 normalisation 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 6** [↗](#)).

In terms of linking longitudinal change in clinical scores with changes captured by *z-diff* scores, each of the two scales was 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. 7** [↗](#)).

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.

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.

Figure 4

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.

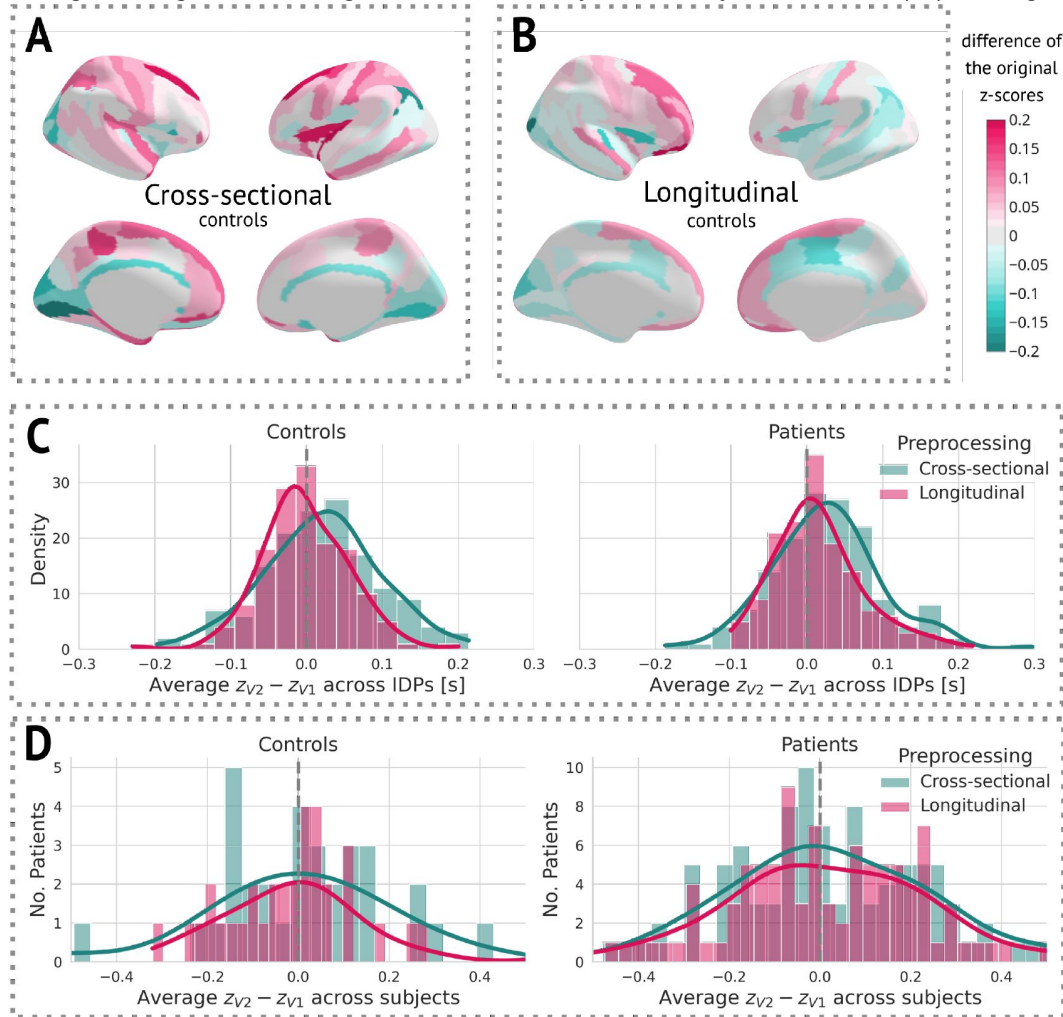
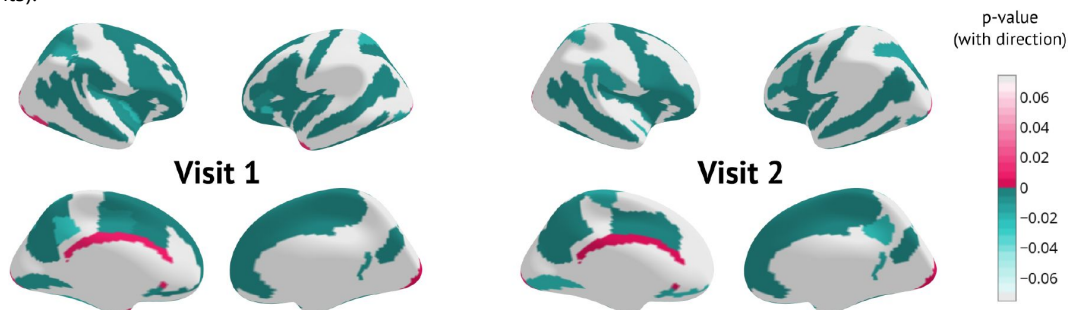


Figure 5

Cross-sectional results for each visit separately: p-values of Mann-Whitney U test between patients and held-out controls surviving Benjamini-Hochberg correction. The sign indicates the direction of change (negative means lower thickness in patients).



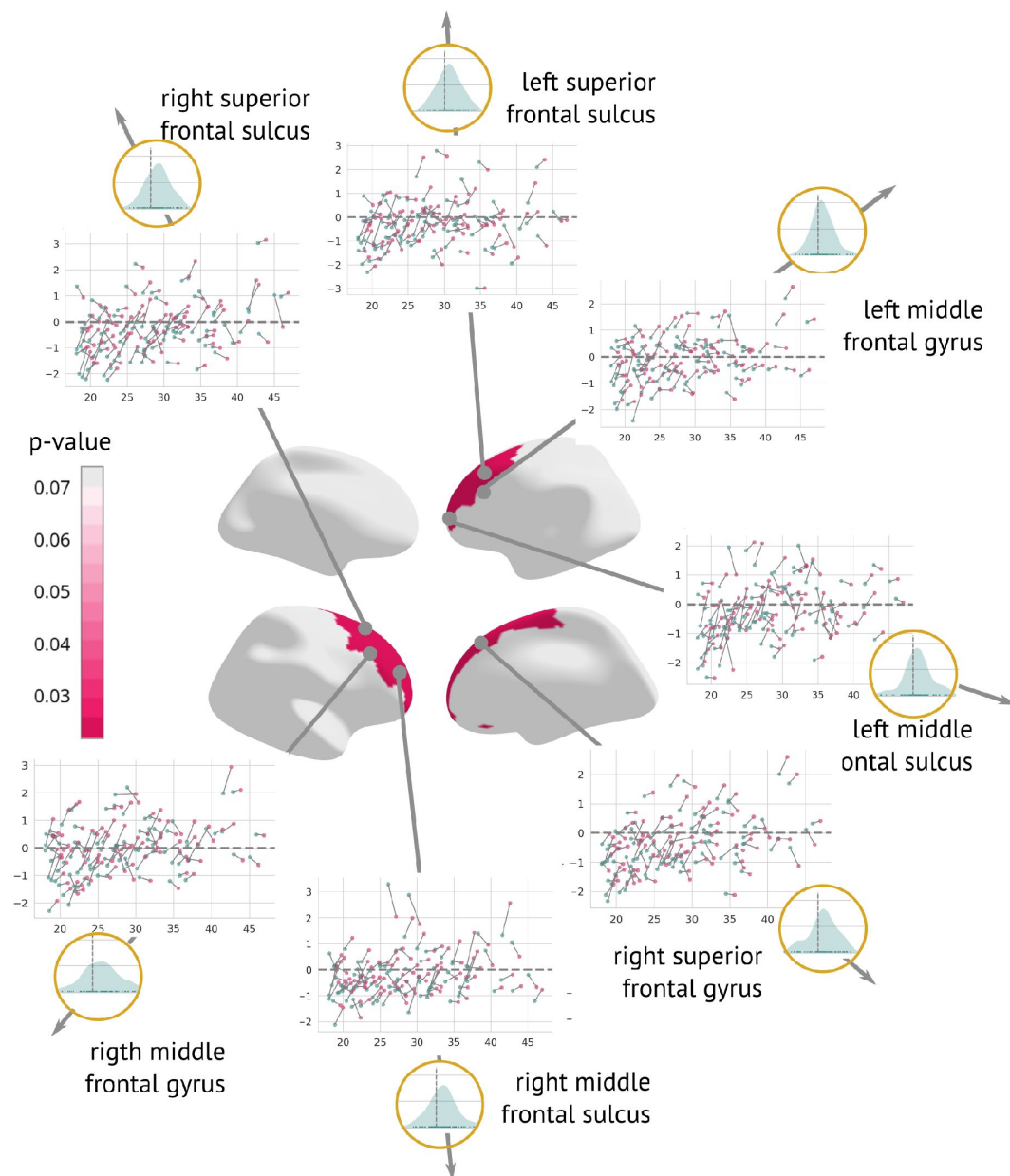


Figure 6

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 $z\text{-diff}$ score.

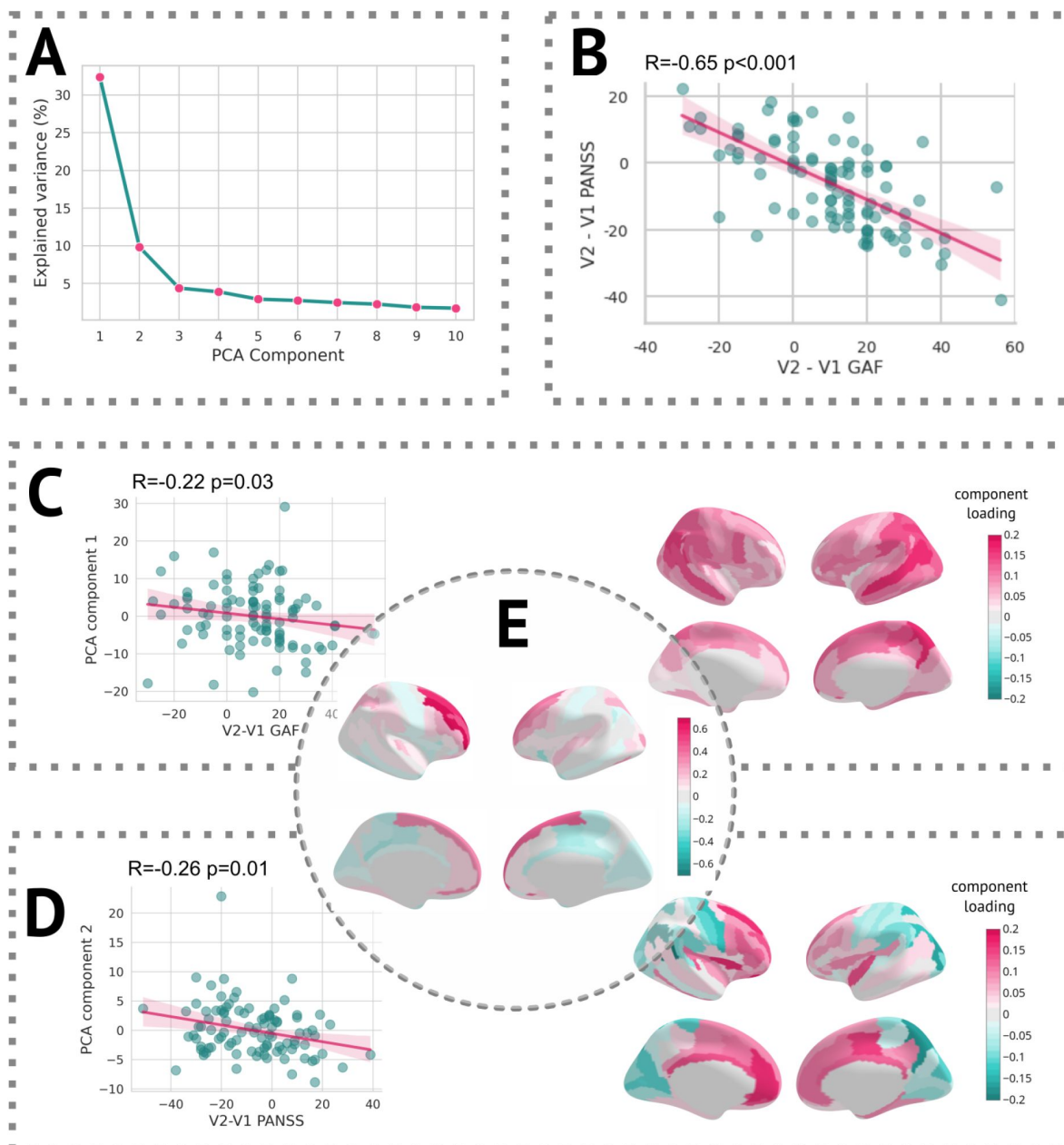


Figure 7

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.

4.1 Methodological contribution

Our approach is rooted in the normative modelling method based on Bayesian regression [3], the pre-trained version of which recently became available [10]. We showed that the estimation of longitudinal changes is 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 extent of change between visits beyond what one would expect in the healthy population.

To this end, our approach implies that in a group of healthy controls, we should observe only change that is consistent with the healthy population, i.e., zero average *z-diff* score. 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 [9] studying the effect of cross-sectional normative 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. Although it is important to keep this caveat in mind, it can be fully addressed only by proper longitudinal normative models, which is beyond the scope of this paper.

Instead, we argue that the population-level trajectory carries meaningful information about individual-level trajectories, and we allow for a flexible process of deviations between the two. By estimating the amplitude of the longitudinal change in healthy controls (adjusting for the population-level trajectory), we get an insight into this process. Naturally, if the healthy changes have a high amplitude (corresponding to low to negative ρ in section 2.1.4), it becomes more challenging to identify subjects who actually diverge from the “healthy” trajectory, i.e., the *z-diff* score becomes overly conservative. A potential reason for the high-amplitude residual process is substantial acquisition or processing noise. 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. Therefore, while the method adopts a cautious approach when assessing individual changes, it identifies effectively group-level changes. Note that this is not unique to our method, but is rather a general statistical feature.

Furthermore, unlike in [9], our approach does not aim to predict individual trajectories, but rather to quantify whether the observed changes over time exceed what would be expected.

4.2 Implementation

At the implementation level, our approach requires two stages of adaptation: site-specific adaptation, as presented in [10], and a second level where we compute the variance of healthy longitudinal 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-centered with a variance of one for healthy controls. However, in the case of longitudinal analysis, the offset and normalisation constant are irrelevant since they will be identical for both visits. Therefore, the estimation of healthy longitudinal 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 effect of preprocessing on z-scores showed that longitudinal pre-processing indeed decreases intra-subject variability compared to cross-sectional pre-processing. 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 normalisation between the visits of early schizophrenia patients is, to a degree, counterintuitive to the historical narrative, which mostly assumes grey matter thinning. There is now increasing evidence that: (i) trajectories of cortical thickness are highly variable across different individuals after the first psychotic episode and (ii) that individuals treated with second-generation antipsychotics and with careful clinical follow-up can show normalisation of cortical thickness atypicalities after the first episode [[15](#) [↗](#), [16](#) [↗](#)]. In [[15](#) [↗](#)], a cohort of 79 first-episode psychosis patients were longitudinally monitored with two follow-up, after a year and ten years. Although cross-sectionally, patients showed significantly lower (cross-sectional) z-scores at baseline (which is consistent with our findings), their proportion decreased over time, indicating an attenuation of differences over time. Canal et al. reported similar observation in larger cohort [[16](#) [↗](#)] of 357 people with first-episode psychosis followed over 10 year period. Notably, no changes in cortical thickness were observed within the first three years. Afterwards, the trajectories started diverging, with cortical thinning observed only in people who experienced worsening of negative symptoms on the expressivity dimension of Scale for the Assessment of Negative Symptoms.

Furthermore, 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 [[17](#) [↗](#)]. 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 normalisation 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** [↗](#)).

Additionally, we observed significant correlations between the PCA components of the *z-diff* score and longitudinal changes in clinical scales, as illustrated in **Fig. 7** [↗](#). Notably, each clinical scale exhibited distinct associations with separate PCA components, despite substantial intercorrelations (**Fig. 7 (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. 7 (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. 7 (D)** [↗](#)). The observed normalisation of grey matter thickness in frontal areas was positively correlated with improvements in the PANSS scale, indicating that symptom amelioration was accompanied by the normalisation of grey matter thickness in these regions.

Finally, we conducted an analysis of longitudinal 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 [[18](#) [↗](#)], but much more data are still needed to construct a full-lifespan longitudinal model.

Additionally, the *z-diff* score only quantifies the size of the change irrespective of the initial position (e.g. cross-sectional *z*-score being above or below 0). However, in subsequent analyses, it is possible to construct models that include both, the original (cross-sectional) position combined with the (longitudinal) change. Indeed, the non-random sampling of large cohort studies is a challenge for nearly all studies using such cohorts, and regardless of the statistical approach used.

Finally, 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

We have developed a method that utilises pre-trained normative models to detect unusual 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); Programme Johannes Amos Comenius ('BRADY' CZ.02.01.01/00/22 008/0004643); European Research Council (grant 'MENTALPRECISION', 10100118), the Wellcome Trust under an Innovator awards

Supplement

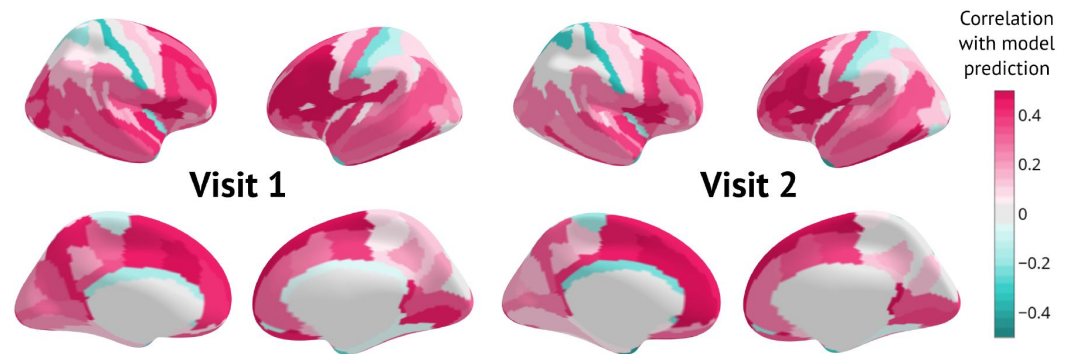
Posterior predictive distribution for difference between visits

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

$$\begin{aligned} f(\Delta_y | x^{(1)}, x^{(2)}; y, \Phi; \omega^2, \sigma^2, \gamma) &= \\ &= \int f_{\mathcal{N}(\Delta_x^T w, 2\sigma_\xi^2)}(\Delta_y | x^{(1)}, x^{(2)}; w) \cdot f_{\mathcal{N}(\Delta_x^T \bar{w}, \Delta_x^T A^{-1} \Delta_x)}(\Delta_x^T w | x^{(1)}, x^{(2)}; y, \Phi; \omega^2, \sigma^2, \gamma) d(\Delta_x^T w) \\ &= \int f_{\mathcal{N}(0, 2\sigma_\xi^2)}(\Delta_y - \Delta_x^T w | x^{(1)}, x^{(2)}; w) \cdot f_{\mathcal{N}(\Delta_x^T \bar{w}, \Delta_x^T A^{-1} \Delta_x)}(\Delta_x^T w | x^{(1)}, x^{(2)}; y, \Phi; \omega^2, \sigma^2, \gamma) d(\Delta_x^T w). \end{aligned}$$

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

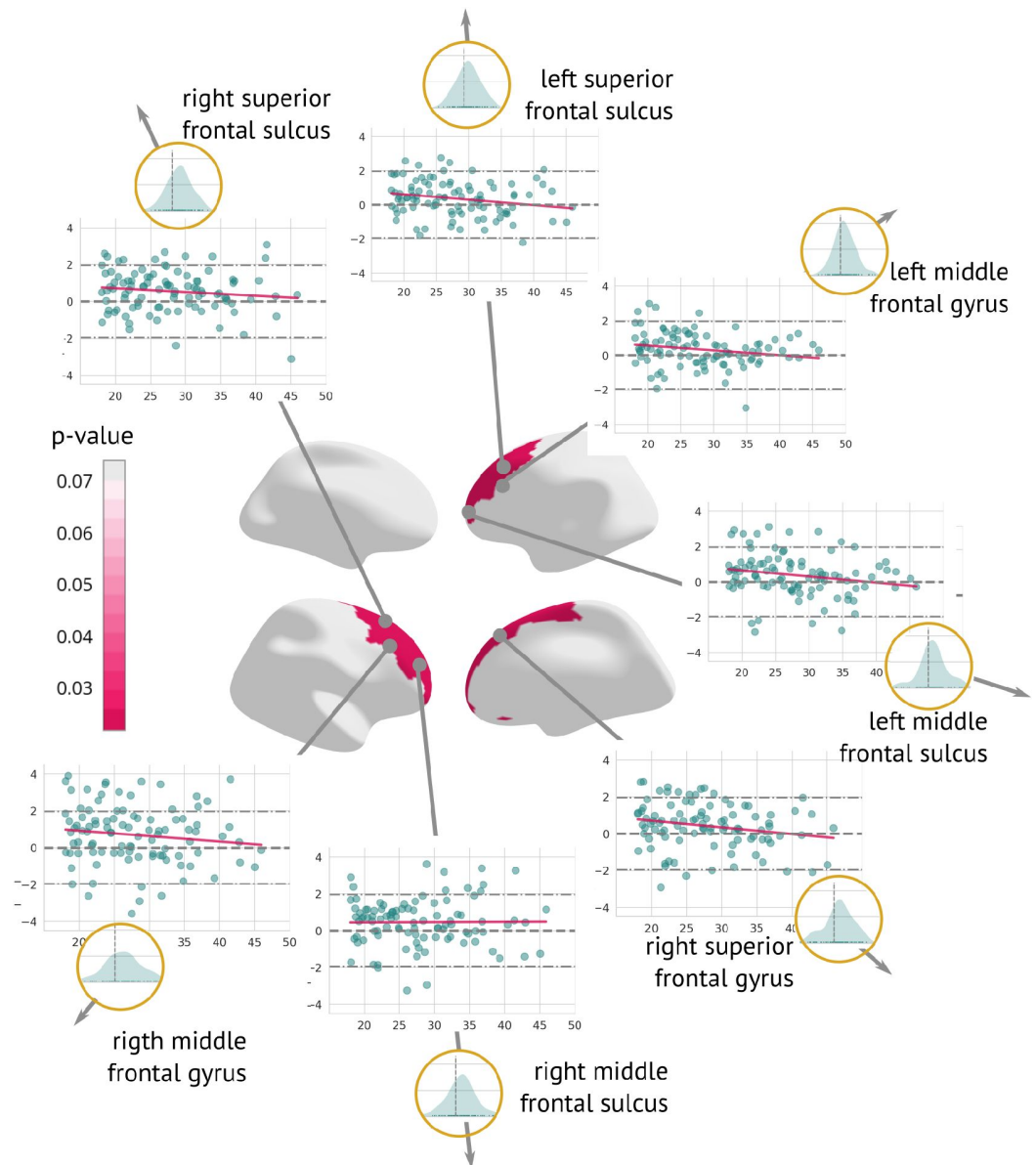
Quality of fit across regions of interest



Supplementary Figure 1

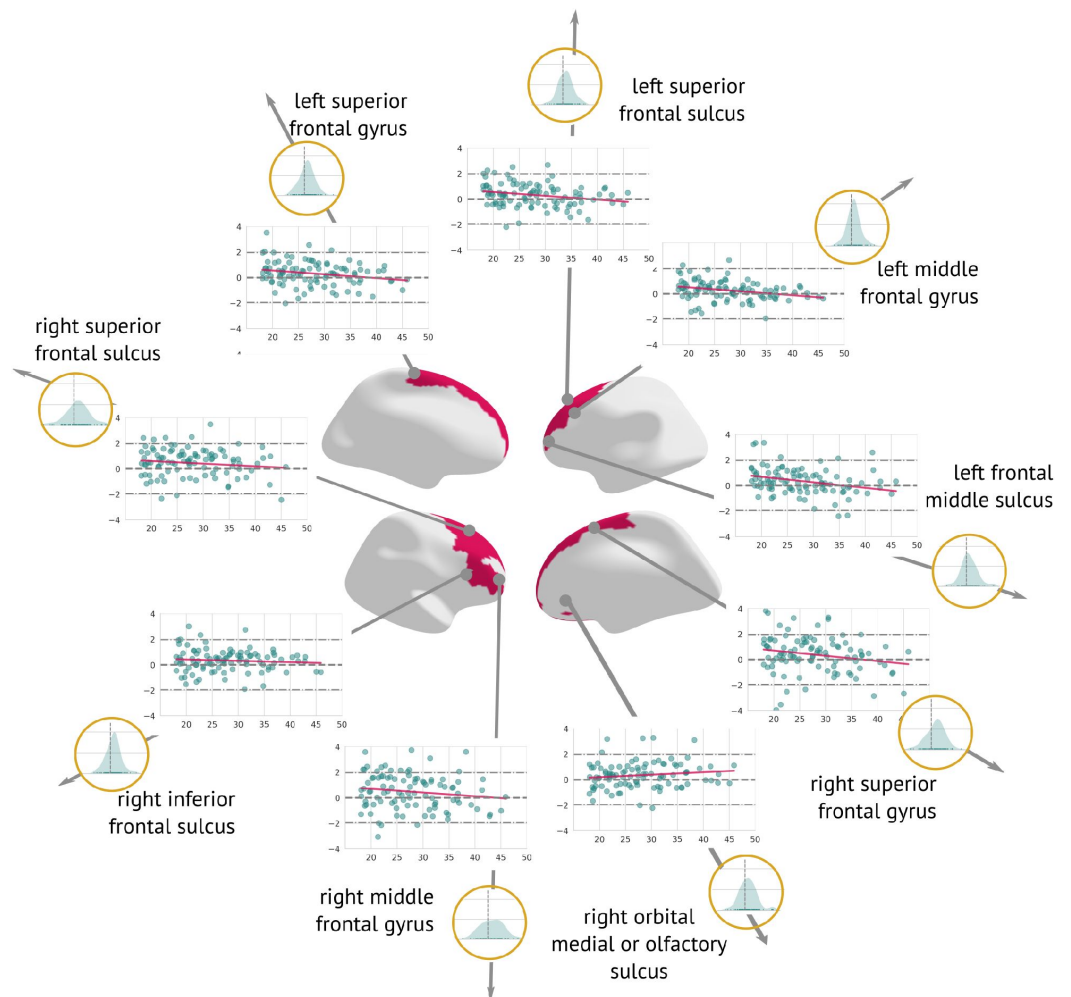
Quality of fit as measured by Rho for the first and the second visit.

Comparison of preprocessing



Supplementary Figure 2

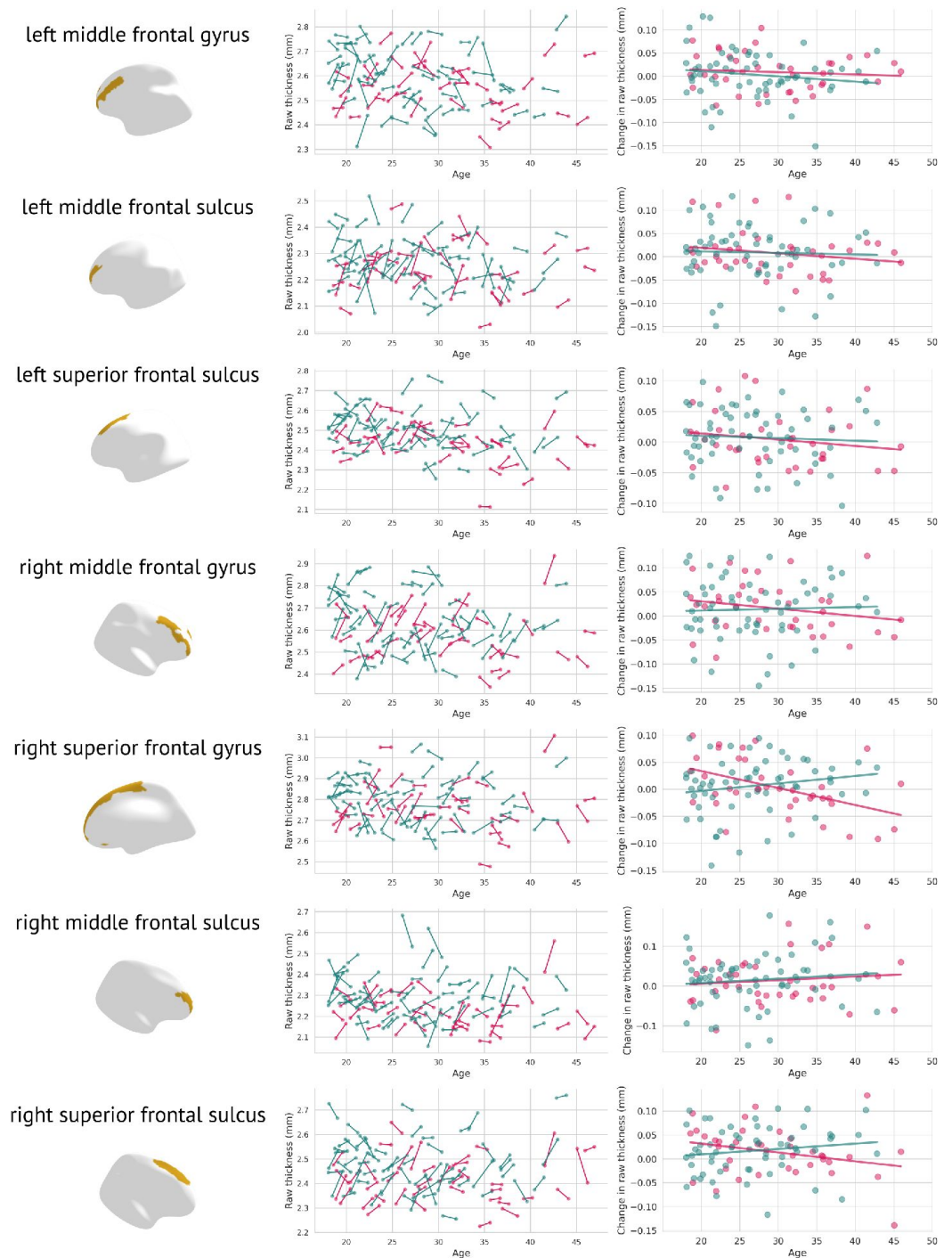
Regions significantly changed between the visits (longitudinal preprocessing): Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of $z\text{-diff}$ across all patients for both visits (the x-axis describes age, and the y-axis depicts the $z\text{-diff}$. Blue dots represent individual patients and the pink line shows a trend of $z\text{-diff}$ change). The Grey dashed line highlights $z=0$. Histograms in the golden circles depict the distribution of the $z\text{-diff}$ 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 visit 2 minus visit 1 in raw thicknesses (separately for females – pink, and males – blue).

References

- [1] Marquand A. F., Rezek I., Buitelaar J., Beckmann C. F. (2016) **Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies** *Biological Psychiatry, Obsessive-Compulsive Disorder* **80**:552–561 <https://doi.org/10.1016/j.biopsych.2015.12.023>
- [2] Bethlehem R. a. I., et al. (2022) **Brain charts for the human lifespan** *Nature* **604**:525–533 <https://doi.org/10.1038/s41586-022-04554-y>
- [3] Fraza C. J., Dinga R., Beckmann C. F., Marquand A. F. (2021) **Warped Bayesian linear regression for normative modelling of big data** *NeuroImage* **245** <https://doi.org/10.1016/j.neuroimage.2021.118715>
- [4] Habes M., et al. (2021) **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* **17**:89–102 <https://doi.org/10.1002/alz.12178>
- [5] Pinaya W. H. L., et al. (2021) **Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multicohort study** *Scientific Reports* **11** <https://doi.org/10.1038/s41598-021-95098-0>
- [6] Wolfers T., et al. (2021) **Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder** *Human Brain Mapping* **42**:2546–2555 <https://doi.org/10.1002/hbm.25386>
- [7] Zabihi M., et al. (2019) **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* **4**:567–578 <https://doi.org/10.1016/j.bpsc.2018.11.013>
- [8] Marquand A. F., Kia S. M., Zabihi M., Wolfers T., Buitelaar J. K., Beckmann C. F. (2019) **Conceptualizing mental disorders as deviations from normative functioning** *Molecular Psychiatry* **24**:1415–1424 <https://doi.org/10.1038/s41380-019-0441-1>
- [9] Di Biase M. A., et al. (2023) **Mapping human brain charts cross-sectionally and longitudinally** *Proceedings of the National Academy of Sciences* **120** <https://doi.org/10.1073/pnas.2216798120>
- [10] Rutherford S., et al. (2021) **Charting Brain Growth and Aging at High Spatial Precision** *bioRxiv* <https://doi.org/10.1101/2021.08.08.455487>
- [11] Cole T. (2012) **The development of growth references and growth charts** *Annals of Human Biology* **39**:382–394 <https://doi.org/10.3109/03014460.2012.694475>
- [12] Kia S. M., et al. (2022) **Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression** *PLOS ONE* **17** <https://doi.org/10.1371/journal.pone.0278776>

- [13] Spaniel F., et al. (2016) **Altered Neural Correlate of the Self-Agency Experience in First-Episode Schizophrenia-Spectrum Patients: An fMRI Study** *Schizophrenia Bulletin* **42**:916–925 <https://doi.org/10.1093/schbul/sbv188>
- [14] Reuter M., Schmansky N. J., Rosas H. D., Fischl B. (2012) **Within-subject template estimation for unbiased longitudinal image analysis** *Neuroimage* **61**:1402–1418 <https://doi.org/10.1016/j.neuroimage.2012.02.084>
- [15] Berthet P., et al. (2024) **A 10-year longitudinal study of brain cortical thickness in people with first-episode psychosis using normative models** *medRxiv* :2024–4
- [16] Canal-Rivero M., et al. (2023) **Longitudinal trajectories in negative symptoms and changes in brain cortical thickness: 10-year follow-up study** *The British Journal of Psychiatry* **223**:309–318
- [17] Merritt K., Luque Laguna P., Irfan A., David A. S. (2021) **Longitudinal Structural MRI Findings in Individuals at Genetic and Clinical High Risk for Psychosis: A Systematic Review** *Frontiers in Psychiatry* **12**
- [18] Casey B. J., et al. (2018) **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* **32**:43–54 <https://doi.org/10.1016/j.dcn.2018.03.001>

Editors

Reviewing Editor

Jason Lerch

University of Oxford, Oxford, United Kingdom

Senior Editor

Jonathan Roiser

University College London, London, United Kingdom

Reviewer #1 (Public review):

Summary:

In this manuscript, the authors provide a method aiming to accurately reflect the individual deviation of longitudinal/temporal change compared to the normal temporal change characterized based on pre-trained population normative model (i.e., a Bayesian linear regression normative model), which was built based on cross-sectional data. This manuscript aims at solving a recently identified problem of using normative models based on cross-sectional data to make inferences about longitudinal change.

Strengths:

The efforts of this work make a good contribution to addressing an important question of normative modeling. With the greater availability of cross-sectional studies for normative modeling than longitudinal studies, and the inappropriateness of making inferences about longitudinal subject-specific changes using these cross-sectional data-based normative models, it's meaningful to try to address this gap from the aspect of methodological development.

In the 1st revision, the authors added a simulation study to show how the performance of the classification based on z-diff scores relatively changes with different disruptions (and autocorrelation). Unfortunately, in my view this is insufficient as it only shows how the performance of using z-diff score relatively changes in different scenarios. I would suggest adding the comparison of performance to using the naïve difference in two simple z-scores to first show its better performance, which should also further highlight the inappropriate use of simple z-scores in inferring within-subject longitudinal changes. Additionally, Figure 1 is hard to read and obtain the actual values of the performance measure. I would suggest reducing it to several 2-dimensional figures. For example, for several fixed values of ρ , how the performance changes with different values of the true disruption (and also adding the comparison to the naïve method (difference in two z-scores)).

I would also suggest changing the title to reflect that the evaluation of "intra-subject" longitudinal change is the method's focus.

<https://doi.org/10.7554/eLife.95823.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Reviewer #1 (Public Review):

The models described are not fundamentally novel, essentially a random intercept model (with a warping function), and some flexible covariate effects using splines (i.e., additive models).

We respectfully but strongly disagree with the reviewer's assessment of the novelty of our work. The models referred to by the reviewer as "random intercept models ... and some flexible covariate effects" seem to relate to the estimation of normative models derived cross-sectionally as developed in and adopted from previous work, not to the work presented here. To be clear, the contributions of this work are: (i) a principled methodology to make statistical predictions for individual subjects in longitudinal studies based on a novel z-diff score, (ii) an approach to transfer information large scale normative models estimated on large scale cross-sectional data to longitudinal studies (iii) an extensive theoretical analysis of the properties of this approach and (iv) empirical evaluation on an unpublished psychosis dataset. Put simply, we provide the ability to estimate within subject *change* in normative models which until now only provide the ability to show a subject's *position* in the normative range at a given timepoint. With the exception of the reference [13] cited in the main text, we are not aware of any methods available that can achieve this. Based on this feedback combined with the feedback of the Reviewer 2, we now improved our introduction and clearly state our contribution right from the outset of the manuscript whilst also shortening the introduction to make it more concise. In this work, we are trying to be very transparent in showing to the reader that our method builds on a previously peer-reviewed model.

The assumption of constant quantiles is very strong, and limits the utility of the model to very short term data.

We now provide an extensive theoretical analysis of our approach (section 2.1.3), where we show that this assumption is actually not strictly necessary and that our approach yields valid inferences even under much milder assumptions. More specifically, we first provide a mathematical grounding for the assumption we made in the initial submission, then generalise our method to a wider class of residual processes and show that our original assumption of constant quantiles is not too restrictive. We also provide a simulation study to

show how the practitioner can evaluate the validity and implications of this assumption on a case-by-case basis. This generalisation is described in depth in section 2.1.3.

The schizophrenia example leads to a counter-intuitive normalization of trajectories, which leads to suspicions that this is driven by some artifact of the data modeling/imaging pipelines.

We understand that the observed normalisation effects might appear surprising. As we outlined in our provisional response, we would like to emphasise that there is increasing evidence that the old neurodegenerative view of psychosis is an oversimplification and that trajectories of cortical thickness are highly variable across different individuals after the first psychotic episode. More specifically, we have shown in an independent sample and with different methodology that individuals treated with second-generation antipsychotics and with careful clinical follow-up can show normalisation of cortical thickness atypicalities after the first episode (<https://www.medrxiv.org/content/10.1101/2024.04.19.24306008v2>, now accepted in Schizophrenia Bulletin). These results are well-aligned with the results we show in this manuscript. We now added remarks on this topic into the discussion. We would also like to re-emphasise that the data were processed with the utmost rigour using state of the art processing pipelines including quality control, which we have reported as transparently as possible. The confidence that the results are not ‘driven by some artifact of the data modeling/imaging pipelines’ is also supported by the fact that analysis of a group of healthy controls did not show any significant z-diffs (see Discussion section), neither frontally nor elsewhere. If the reviewer believes there are additional quality control checks that would further increase confidence in our findings, we would welcome the reviewer to provide specific details.

The method also assumes that the cross-sectional data is from a "healthy population" without describing what this population is (there is certainly every chance of ascertainment bias in large scale studies as well as small scale studies). This issue is completely elided over in the manuscript.

Indeed, we do not describe the cross-sectional population used for training the models, as these models were already trained and published with in-depth description of the datasets used for the training (<https://elifesciences.org/articles/72904>). We now make this more explicit in the section 2.1.1. of the manuscript (page 7), and also more explicitly acknowledge the possibility of ascertainment bias in the simulation section 2.1.4. However, we would like to emphasise that such ascertainment bias is not in any way specific to the analyses we report. In fact it is present in all studies that utilise large scale cohorts such as UK Biobank. Indeed, we are currently working on another manuscript to address this question in detail, but given the complexity of this problem and the fact that many publicly available legacy studies simply do not record sufficient demographic information, e.g. to assess racial bias properly, we believe that this is beyond the scope of the current work.

Reviewer #2 (Public Review):

The organization and clarity of this manuscript need enhancement for better comprehension and flow. For example, in the first few paragraphs of the introduction, the wording is quite vague. A lot of information was scattered and repeated in the latter part of the introduction, and the actual challenges/motivation of this work were not introduced until the 5th paragraph.

As noted above in our response to Reviewer 1, we significantly pruned the introduction, stating our objective in the first paragraph and elaborating on the topic later in the text. We hope that it is now less repetitive and easier to follow.

There are no simulation studies to evaluate whether the adjustment of the crosssectional normative model to longitudinal data can make accurate estimations and inferences regarding the longitudinal changes. Also, there are some assumptions involved in the modeling procedure, for example, the deviation of a healthy control from the population over time is purely caused by noise and constant variability of error/noise across x_n , and these seem to be quite strong assumptions. The presentation of this work's method development would be strengthened if the authors can conduct a formal simulation study to evaluate the method's performance when such assumptions are violated, and, ideally, propose some methods to check these assumptions before performing the analyses.

This comment encouraged us to zoom out from our original assumption and generalise our method to a wider class of residual processes (stationary Gaussian processes) in section 2.1.3. We now present a theoretical analysis of our model to show that our original assumption (of stable quantiles plus noise) is actually not necessary for valid inference in our method, which broadens the applicability of our method. Of course, we also discuss in what way the original assumption is restrictive and how it aligns with the more general dynamics. We also include a simulation study to evaluate the method's performance and elucidate the role of the more general dynamics in section 2.1.4.

The proposed "z-diff score" still falls in the common form of z-score to describe the individual deviation from the population/reference level, but now is just specifically used to quantify the deviation of individual temporal change from the population level. The authors need to further highlight the difference between the "z-score" and "z-diff score", ideally at its first mention, in case readers get confused (I was confused at first until I reached the latter part of the manuscript). The z-score can also be called a measure of "standardized difference" which kind of collides with what "z-diff" implies by its name.

We added the mention of the difference between z-score and z-diff score into the last paragraph of introduction.

Explaining that one component of the variance is related to the estimation of the model and the other is due to prediction would be helpful for non-statistical readers.

We now added an interpretation of the z-score in the original model below equation 7.

It would be easier for the non-statistical reader if the authors consistently used precision or variance for all variance parameters. Probably variance would be more accessible.

This was a very useful observation, we unified the notation and now only use variance.

The functions ψ were never explicitly described. This would be helpful to have in the supplement with a reference to that in the paper.

Indeed, while describing the original model we had to make choices about how to condense the necessary information from the original model so that we can build upon it. As the ϕ function is only used for data transformation in the original model, we did not further elaborate on it, however, we now refer to the specific section of the original paper of Fraza et al. 2021 where it is described more in detail (<https://www.sciencedirect.com/science/article/pii/S1053811921009873>).

What is the goal of equations (13) and (14)? The authors should clarify what the point of writing these equations is prior to showing the math. It seems like it is to obtain an

estimate of $\sigma_{\{k\}}^2$, which the reader only learns at the end.

We corrected the formatting.

What is the definition of "adaption" as used to describe equation (15)? In this equation, I think norm on subsample was not defined.

We added a more detailed description of the adaptation after equation 15.

"(the sandwich part with A)" - maybe call this an inner product so that it is not confused with a sandwich variance estimator. This is a bit unclear. Equation (8) does have the inner product involving A and β^{-1} does include variability of η . It seems like you mean that equation (8) incorrectly includes variability of η and does not have the right term vector component of the inner product involving A, but this needs clarifying.

We now changed the formulation to be less confusing and also explicitly clarified the caveat regarding the difference of z-scores.

One challenge with the z-diff score is that it does not account for whether a person sits above or below zero at the first time point. It might make it difficult to interpret the results, as the results for a particular pathology could change depending on what stage of the lifespan a person is in. I am not sure how the authors would address those challenges.

We agree with the outlined limitation in interpretation of overall trends when the position in the visit one is different between the subjects. However, this is a much broader challenge and is not specific to our approach. This effect is generally independent of the lifespan, but may further interact with the typical lifespan of disease. When the z scores are taken in the context of the cross-sectional normative models, it does make it possible to identify what the overall trend of an illness is across the lifespan, and individual patient's z-diffs not in line (with what would this typical group trajectory predicts) may e.g. correspond to early/late onset of their individual atrophy. We now make these considerations explicitly in the discussion section.

Reviewer #2 (Recommendations For The Authors):

Other minor suggestions to help improve the text:...

We thank Reviewer #2 for the list of minor suggestions to improve the text, which we all implemented in the manuscript.

<https://doi.org/10.7554/eLife.95823.2.sa0>