Longitudinal Reliability of Cross-Sectional Normative Models under Measurement Error

Wenxing Ji[1], Jianqiao Long[1], Qiwen Guan[1], Xinyu Li[2], Shouyong Jiang[3], Jichun Li[1(✉)], and Yujiang Wang[1(✉)]

[1] Newcastle University, Newcastle Upon Tyne, UK
jichun.li@ncl.ac.uk, Yujiang.Wang@newcastle.ac.uk
[2] University of Exeter, Exeter, UK
[3] Central South University, Changsha, China

Abstract. Longitudinal neuroimaging is essential for studying brain aging and disease progression. Most existing approaches track raw change over time, without reference to normative population trajectories. Cross-sectional normative models offer a principled alternative by contextualising an individual's brain against healthy population norms, but their application to longitudinal inference raises an underexplored question: how reliably can individual deviations from normative trajectories be distinguished from stochastic measurement error? We address this using the Longitudinal Deviation Index (LDI), which quantifies reliability as the ratio of true inter-individual residual slope variance to variance expected purely from measurement error. Leveraging a normative model pre-trained on over 6,800 healthy individuals, we derived theoretical LDI estimates from test–retest data and validated them against empirical longitudinal data from cognitively normal adults. Within an 8-year follow-up window ($LDI < 1$), individual deviations are indistinguishable from measurement noise, confounding valid longitudinal inference. Beyond 8 years ($LDI > 1$), cumulative longitudinal variations systematically surpass the noise floor, reflecting genuine biological progression. These results clarify the limits of longitudinal interpretation using cross-sectional normative models and provide a principled framework for noise-aware inference.

Keywords: Longitudinal reliability · LDI · Normative models · Measurement error.

1 Introduction

Longitudinal neuroimaging is central to understanding brain aging and disease progression, yet most existing approaches track raw change over time without reference to population norms [9, 21]. Normative models offer a principled alternative: by learning reference distributions from large healthy cohorts, they contextualise individual brain structure relative to expected age-related trajectories [13, 2], and have been applied broadly in neurological and psychiatric research [12, 16, 23].

Most normative models are built on cross-sectional data and, when applied longitudinally, implicitly assume that healthy individuals maintain a stable centile rank over time [6, 13]. Under this assumption, a subject's baseline centile defines their normative trajectory, and deviations from it are interpreted as biologically meaningful change [18]. However, longitudinal neuroimaging data are subject to non-negligible measurement variability from sources such as scanner drift, protocol changes, and preprocessing inconsistencies, even in the absence of true biological change. Conflating these noise-driven fluctuations with genuine structural progression is thus a real risk when applying cross-sectional normative models to repeated measures [7].

Because prior work has focused predominantly on cross-sectional accuracy or applied normative models longitudinally without quantifying reliability [5, 19], it remains unclear to what extent observed longitudinal deviations reflect genuine biological progression versus measurement noise [20]. We address this gap by introducing the Longitudinal Deviation Index (LDI), defined as the ratio of inter-individual variability in residual change to the variability expected purely from measurement error. Low LDI values indicate that residual slopes are dominated by noise; elevated values signal systematic longitudinal deviations reflecting biological progression not captured by the cross-sectional model.

2 Methods

2.1 Pre-trained Normative Model and Centile Trajectory

To assess longitudinal reliability, we used a pre-trained cross-sectional normative model based on our previous work [11], which was fitted for each log-transformed IDP independently using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS [4]) framework.

To apply the model to the new data, for each new subject i, the IDP Y_{ij} at age t_{ij} $(j = 1, \ldots, n)$ is assumed to stem from one of the scanning sites used during training, in order to predict model parameters and compute residuals. Subsequently, site-specific mean and variance parameters are estimated from the residuals of healthy control subjects at the new site, and these parameters are applied to calculate the calibrated IDP $Y_{ij}^{'}$. This procedure effectively removes site effects from the data and provides an expected, healthy distribution for each subject given their age and sex. A detailed description of this calibration procedure is provided in Section S2.2 of the supplementary material in [11].

Furthermore, under the fundamental assumption that healthy individuals maintain a relatively stable centile rank over time, we utilized normative centile curves to define the cross-sectional normative trajectory for each individual. Specifically, for subject i, the predicted IDP value Y_{ij}^{pred} at age t_{ij} is determined by evaluating the normative quantile function [17] conditioned on their baseline centile (at t_{i0}) and their new age t_{ij}. Consequently, the residual r_{ij} is explicitly defined as the deviation between the calibrated observation and this cross-sectional normative trajectory: $r_{ij} = Y_{ij}^{'} - Y_{ij}^{pred}$.

[4] https://cran.r-project.org/web/packages/gamlss/

2.2 Residual Trajectory and LDI

Assuming that changes in residuals follow a linear trend over time, the residual trajectory for each individual i is modeled as:

$$r_{ij} = a_i + b_i\, t_{ij} + \epsilon_{ij}, \tag{1}$$

where the intercept term a_i is set to 0, as the baseline residual is 0 by construction. The slope b_i represents the individual-specific deviation rate from the cross-sectional normative trajectory. We assume that population slopes b_i follow a normal distribution, $b_i \sim N(\delta, \sigma_b^2)$. Similarly, the measurement error ϵ_{ij} is normally distributed with mean 0 and variance σ_ϵ^2.

Furthermore, we define the LDI as the ratio of the true variance to the error variance for the residual slope b_i:

$$LDI = \frac{Var_{true}}{Var_{error}}. \tag{2}$$

2.3 Parameter Estimation

Assuming linear longitudinal change, applying linear regression to individuals with at least three observations allows us to estimate the true biological variance while isolating it from measurement error [4]. Consequently, the true (latent) variance can be estimated as the residual slope variance σ_b^2, given that inter-individual differences in the rate of deviation from the cross-sectional normative trajectory represent genuine biological variance.

The error variance quantifies the uncertainty associated with an individual's residual slope estimate b. Following ordinary least squares theory, this is defined as the ratio of the cross-sectional measurement error variance σ_ϵ^2 to the total sum of squared deviations of the time points (SST) [20, 22]:

$$Var_{error} = \frac{\sigma_\epsilon^2}{SST}. \tag{3}$$

In this study, the measurement error variance σ_ϵ^2 is derived from half of the variance of the test–retest difference [3]. The SST represents the spread of observation times:

$$SST = \sum_{j=1}^{n} (t_{ij} - \bar{t}_i)^{\,2}. \tag{4}$$

Under the assumption of approximately equal spacing, SST can be simplified relative to the total study duration t and the number of observations n [8, 15]:

$$SST_{\text{equidistant}} = \frac{t^2 n(n+1)}{12(n-1)}. \tag{5}$$

By substituting the estimated parameters (Eq. 3 and Eq. 5) into Equation 2, the LDI can be explicitly formulated as:

$$LDI = \frac{\sigma_b^2}{\sigma_\epsilon^2 \{t^2 n(n+1)/12(n-1)\}^{-1}}. \tag{6}$$

Equation 5 demonstrates that under equally spaced observations, SST grows quadratically with the follow-up duration t. Consequently, the slope error variance decreases proportionally to t^{-2}, implying that the LDI increases rapidly with longer follow-up durations.

2.4 Data

We utilized the OASIS-3 dataset [10], restricting our analysis to participants aged 42–95 who remained cognitively normal (Clinical Dementia Rating, CDR = 0 [14]) across all longitudinal visits. To estimate specific variance components, we derived three subsets (see Tab. 1): (1) a Test-Retest Subset ($n = 116$, interval < 90 days, mean = 15 days) from 11 scanners (sites) for calculating measurement error variance, with outliers > 5 median absolute deviations (MAD) from the mean within subjects removed; (2) Longitudinal Subset 1 ($n = 114$, ≥ 4 observations, spanning ≥ 4 years) from 15 scanners to estimate the true residual slope variance; and (3) Longitudinal Subset 2 ($n = 214$, ≥ 3 observations) from 15 scanners to empirically estimate the error variance of the residual slope.

Table 1. Descriptions for the test-retest subset and the longitudinal subsets. Age is reported in years, while follow-up time is in days (d) for the test-retest subset and years (y) for the longitudinal subsets. *Fem.* indicates the number of female participants; *Obs.* denotes the number of longitudinal imaging sessions per subject; *Site* specifies the number of scanners involved.

	N	Fem.	Obs.	Site	Age				Follow-up time			
					mean	SD	min	max	mean	SD	min	max
Test-Retest	116	66	2	11	66.23	8.13	42.70	84.72	15.11d	20.17d	1d	90d
Longitudinal												
1	114	70	4–9	15	67.36	8.29	45.96	92.44	8.86y	1.97y	4.14y	12.34y
2	214	132	3–9	15	68.09	8.39	45.96	92.44	7.48y	2.57y	1.78y	12.34y

We processed T1w data via FreeSurfer to extract regional IDP values across different modalities (cortical thickness, volume, and pial surface area). These data were subsequently mapped using the pre-trained cross-sectional normative model. To formally validate the underlying parametric assumptions, we conducted Kolmogorov-Smirnov tests across all 204 IDPs. The results confirm our assumptions: 100% of the features passed the normality check for measurement error, and 91.2% (186/204) exhibited no significant deviation from normality for the longitudinal residual slopes ($p \geq 0.05$).

Next, we evaluated follow-up durations (t) of 2 to 8 years (2-year intervals) across $n \in \{3, 5, 7, 9\}$ observations from a theoretical perspective. Due to data constraints, empirical validations were limited to $n \in \{3, 5\}$ over the same duration.

3 Results

3.1 Measurement Error is Age-independent across Cortical Regions

We quantified the measurement error using the test-retest data to assess its distribution and stability. Our analysis revealed consistency in error magnitude across morphometric measures, with entorhinal regions in cortical thickness exhibiting slightly lower error levels compared to volume and surface area (Fig. 1a). In addition, no significant linear relationship was found between measurement error and age in any region after correcting for multiple comparisons (FDR-corrected $p > 0.05$; Fig. 1b). This independence of measurement error from age was consistent across both the left and right hemispheres. These findings suggest that measurement error can be modeled as a stable, age-invariant parameter within the studied age range from 42.70 to 84.72 years, validating the use of a fixed error variance for subsequent LDI estimations.

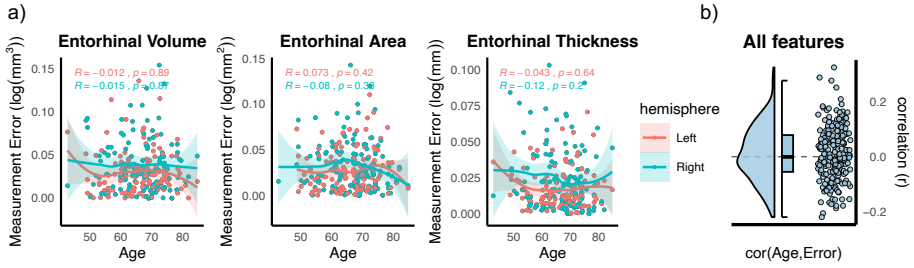

Fig. 1. Measurement error and age. **a)** Association between measurement absolute log error and age for entorhinal volume, area, and thickness in the test–retest dataset. Red = left hemisphere, green = right hemisphere. Data were fitted using loess models, with shaded ribbons indicating the corresponding confidence intervals. **b)** Pearson's correlations between measurement error and age across all features. Each point represents one feature, and the dashed line indicates a null correlation. No feature showed a significant association between measurement error and age after correction for multiple comparisons using the false discovery rate (FDR, $p < 0.05$).

3.2 LDI in Theoretical Analysis

Theoretically, LDI was derived from the estimated measurement error and residual slope variances across follow-up durations of 2–8 years with 3–9 equidistant observations (Eq. 6). A higher LDI indicates that genuine biological progression exceeds what the cross-sectional model captures. On average, for an 8-year follow-up, the LDI consistently exceeded 1 and increased alongside observation count (Fig. 2a). Conversely, for shorter periods (2–4 years), the biological signal was largely submerged by measurement error across nearly all frequencies. Notably, we observed comparable statistical power between 3 observations over 6

years and 9 observations over 4 years, which demonstrates that over short follow-up periods, true individual deviations from the expected normative trajectory are too subtle to overcome measurement error, rendering dense longitudinal sampling ineffective. In contrast, extending the temporal tracking window allows genuine biological deviations to accumulate and surpass the noise floor. Furthermore, LDI exhibited significant regional heterogeneity. For instance, under a 3-year follow-up with 8 observations, the LDI remained below 0.8 for bilateral transverse temporal volume and thickness, as well as for the right lingual pial area (Fig. 2b). In contrast, higher LDI levels were observed for the pial area and cortical thickness in the superior parietal and rostral middle frontal regions.

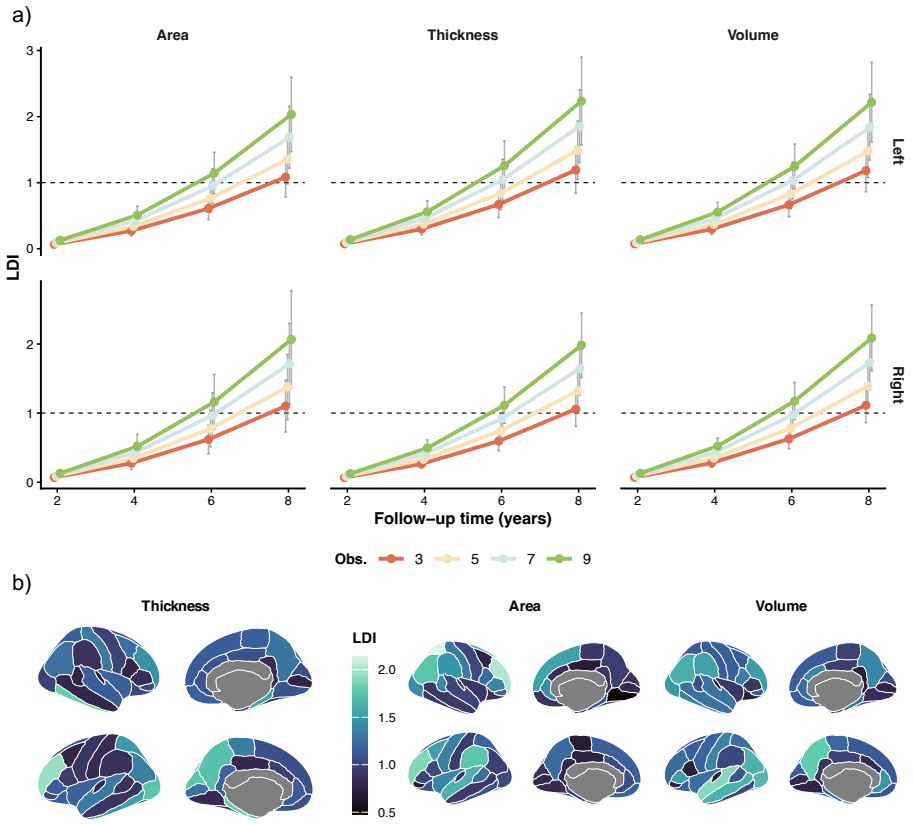

Fig. 2. LDI estimated theoretically. **a)** Trajectories of mean LDI across different brain regions over a 2 to 8 year follow-up period, evaluated under varying numbers of observations. Results are presented separately for the left (top row) and right (bottom row) hemispheres and different modalities. Error bars illustrate ±1 SD. **b)** Spatial maps of regional LDI at the 8-year follow-up based on 3 observations.

3.3 Empirical Estimation and Validation

Empirically, the residual slope variance was treated as a fixed parameter, while the error variance was estimated using the standard error (SE) of the subject-specific residual slope for each feature. Specifically, linear regression of the longitudinal residuals yields both the slope and its associated SE, which quantifies the estimation uncertainty. We modeled these SE values using a generalized linear mixed-effects model (GLMM) [1], including the total follow-up duration and the number of observations as fixed effects, alongside brain regions as a random intercept. Given that the SE underestimates the true error variance of the slope, we applied an observation-dependent empirical scaling factor to correct the predicted SE. This factor was derived from synthetic data: by adding random noise to known true slopes and refitting the data, we could empirically compute the true slope error variance because the true slope is known, see also [20]. As illustrated in Fig. 3, the relationship between the mean SE and the standard deviation (SD) of the true errors for the simulated slopes is reported for the number of observations. The magnitude of the deviation from the reference line (slope = 1) quantifies how much the SE underestimates the variance of the true slope error.

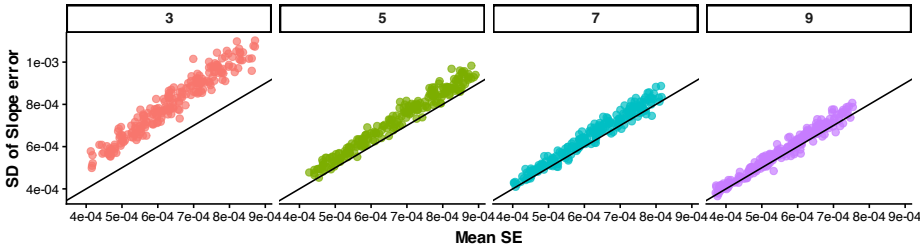

Fig. 3. Scatter plots of SE obtained via linear fitting versus the SD of slope errors across different observations in synthetic data. Each data point represents a simulation of a specific brain regional IDP over 500 iterations. The SE is reported as the mean.

Constrained by empirical data availability, we evaluated the empirical LDI as a function of follow-up duration exclusively for 3 and 5 observations (Fig. 4a). On average, the LDI for 5 observations was slightly higher than that for 3, a discrepancy likely driven by the smaller sample size available for the 5-observation cohort. Additionally, the difference between empirical and theoretical LDI remained minimal for follow-up periods of 2 to 6 years, increasing slightly at the 8-year mark (Fig. 4b). Across modalities, the mean theoretical-empirical LDI difference was near 0 (Fig. 4c), robustly validating the predictive accuracy of our theoretical framework. Furthermore, the LDI exhibited pronounced regional heterogeneity (Fig. 4d), corroborating our theoretical analysis. For instance, elevated LDI values were observed for the rostral middle frontal region.

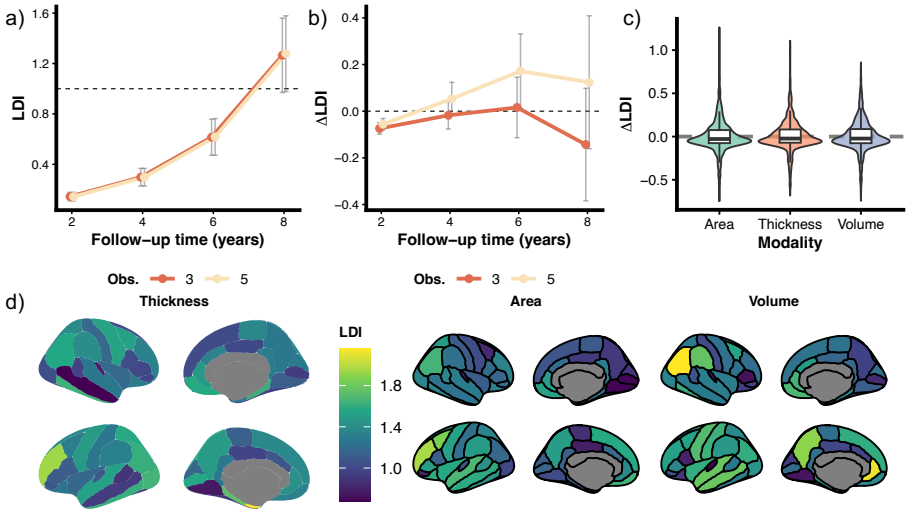

Fig. 4. LDI estimated empirically. **a)** Mean LDI across features as a function of total follow-up time and number of equidistant observations. Mean differences in LDI by **b)** follow-up time and number of observations, and **c)** modality, between the analytically derived and the empirical estimations of LDI. Positive ΔLDI indicates higher estimates for the analytical derivation of LDI. **d)** Empirically, spatial maps of the LDI for diverse brain structures at the 8-year follow-up across 3 observations.

4 Discussion and Conclusion

Our central finding is that cross-sectional normative models cannot support reliable individual-level longitudinal inference within an 8-year follow-up window. Below this threshold ($LDI < 1$), deviations from normative trajectories are statistically indistinguishable from stochastic measurement error, and residual changes cannot be confidently attributed to genuine biological progression. Beyond 8 years ($LDI > 1$), cumulative biological variation systematically exceeds the noise floor, and normative model residuals begin to carry interpretable longitudinal signals. This provides a principled basis for determining when cross-sectional normative models can be applied to longitudinal inference — a question largely unaddressed in prior work, where longitudinal change is more commonly tracked as raw deviation without normative reference [9, 21].

Regional LDI heterogeneity further qualifies these findings. Association cortices, including superior parietal and rostral middle frontal regions, exhibit higher LDI values, likely reflecting greater genuine inter-individual variability in longitudinal deviation rates (σ_b^2). These regions therefore reach reliable detection thresholds earlier. Conversely, bilateral transverse temporal and right lingual areas show persistently low LDI, suggesting individual trajectories there remain noise-dominated even at longer follow-ups.

This study has several limitations. First, limited empirical data prevented validation for 7 and 9 observations. Second, the evaluation was restricted to cognitively normal individuals; future work should characterise LDI in clinical cohorts such as Alzheimer's disease, where biological decline is expected to outpace the noise floor over shorter intervals. Third, a linear approximation was assumed for tractability, though biological changes are often non-linear. Fourth, the noise floor was established from 15-day test–retest data and may not capture longer-term physiological fluctuations. Fifth, LDI values are dataset-specific to OASIS-3; generalisability across larger, more heterogeneous cohorts should be formally validated.

In conclusion, the LDI offers a noise-aware framework for evaluating when longitudinal inference from cross-sectional normative models is statistically warranted. For follow-up windows below 8 years, residual deviations from normative trajectories should be interpreted cautiously, and study designs should explicitly account for measurement error rather than treating residual change as a direct proxy for biological progression.

Acknowledgments. This work was partly supported by the Innovate UK KTP fund (10159356) and the CACSUK research fund (RES/0550/7635) for Dr Jichun Li from Newcastle University. Yujiang Wang is supported by a UKRI Future Leaders Fellowships (MR/V026569/1). We also acknowledge the valuable contributions and discussions from Karoline Leiberg and Nida Alyas from the Computational Neurology, Neuroscience and Psychiatry (CNNP) Lab, Newcastle University, regarding the normative modelling framework and the conceptual development of this work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. Journal of statistical software **67**, 1–48 (2015)
2. Bethlehem, R.A., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., Adamson, C., Adler, S., Alexopoulos, G.S., Anagnostou, E., Areces-Gonzalez, A., et al.: Brain charts for the human lifespan. Nature **604**(7906), 525–533 (2022)
3. Bland, J.M., Altman, D.G.: Measurement error. BMJ: British medical journal **312**(7047), 1654 (1996)
4. Brandmaier, A.M., Lindenberger, U., McCormick, E.M.: Optimal two-time point longitudinal models for estimating individual-level change: Asymptotic insights and practical implications. Developmental Cognitive Neuroscience **70**, 101450 (2024)
5. Bucková, B.R., Fraza, C., Rehák, R., Kolenič, M., Beckmann, C.F., Španiel, F., Marquand, A.F., Hlinka, J.: Using normative models pre-trained on cross-sectional data to evaluate intra-individual longitudinal changes in neuroimaging data. Elife **13**, RP95823 (2025)
6. Cole, T.J.: The development of growth references and growth charts. Annals of human biology **39**(5), 382–394 (2012)

7. Di Biase, M.A., Tian, Y.E., Bethlehem, R.A., Seidlitz, J., Alexander-Bloch, A.F., Yeo, B.T., Zalesky, A.: Mapping human brain charts cross-sectionally and longitudinally. Proceedings of the National Academy of Sciences **120**(20), e2216798120 (2023)
8. Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: Applied longitudinal analysis. John Wiley & Sons (2012)
9. Gordon, B.A., Blazey, T.M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., Christensen, J., McDade, E., Wang, G., Xiong, C., et al.: Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant alzheimer's disease: a longitudinal study. The Lancet Neurology **17**(3), 241–250 (2018)
10. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medrxiv pp. 2019–12 (2019)
11. Little, B., Alyas, N., Surtees, A., Winston, G.P., Duncan, J.S., Cousins, D.A., Taylor, J.P., Taylor, P., Leiberg, K., Wang, Y.: Brain morphology normative modelling platform for abnormality and centile estimation: brain monocle. Imaging Neuroscience **3**, imag_a_00438 (2025)
12. Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F.: Conceptualizing mental disorders as deviations from normative functioning. Molecular psychiatry **24**(10), 1415–1424 (2019)
13. Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F.: Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biological psychiatry **80**(7), 552–561 (2016)
14. Morris, J.C.: The clinical dementia rating (cdr) current version and scoring rules. Neurology **43**(11), 2412–2412 (1993)
15. von Oertzen, T.: Power equivalence in structural equation modelling. British Journal of Mathematical and Statistical Psychology **63**(2), 257–272 (2010)
16. Pinaya, W.H., Scarpazza, C., Garcia-Dias, R., Vieira, S., Baecker, L., F da Costa, P., Redolfi, A., Frisoni, G.B., Pievani, M., Calhoun, V.D., et al.: Using normative modelling to detect disease progression in mild cognitive impairment and alzheimer's disease in a cross-sectional multi-cohort study. Scientific reports **11**(1), 15746 (2021)
17. Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z., De Bastiani, F.: Distributions for modeling location, scale, and shape: Using GAMLSS in R. Chapman and Hall/CRC (2019)
18. Rutherford, S., Barkema, P., Tso, I.F., Sripada, C., Beckmann, C.F., Ruhe, H.G., Marquand, A.F.: Evidence for embracing normative modeling. Elife **12**, e85082 (2023)
19. Rutherford, S., Fraza, C., Dinga, R., Kia, S.M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., et al.: Charting brain growth and aging at high spatial precision. elife **11** (2022)
20. Vidal-Piñeiro, D., Sørensen, Ø., Strømstad, M., Amlien, I.K., Anderson, M., Baaré, W.F., Bartrés-Faz, D., Brandmaier, A.M., Bråthen, A.C., Garrido, P., et al.: Reliability of structural brain change in cognitively healthy adult samples. Imaging Neuroscience **3**, imag_a_00547 (2025)
21. Vinci-Booher, S., Ren, X., Kay, K., Yu, C., Pestilli, F., Booth, J.R.: Dense longitudinal neuroimaging reveals individual brain change trajectories. Trends in cognitive sciences (2025)

22. Willett, J.B.: Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. Educational and psychological measurement **49**(3), 587–602 (1989)
23. Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., et al.: Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. Biological psychiatry: cognitive neuroscience and neuroimaging **4**(6), 567–578 (2019)