The practical impact of numerical variability on structural MRI measures of Parkinson's disease

Yohan Chatelain[1,2], Andrzej Sokołowski[2], Madeleine Sharp[3], Jean-Baptiste Poline[3], and Tristan Glatard[1,2,4,5]

[1]Krembil Centre for Neuroinformatics, The Centre for Addiction and Mental Health, Toronto, ON, Canada
[2]Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
[3]Department of Neurology and Neurosurgery, McGill University, Montreal, Canada
[4]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[5]Department of Computer Science, University of Toronto, Toronto, ON, Canada

Abstract

Numerical variability is rarely quantified in neuroimaging despite many biomarkers relying on subtle morphometric differences across individuals. We instrumented FreeSurfer, a widely used neuroimaging pipeline, to simulate numerical differences across computational environments, and used it to measure numerical variability in MRI analyses of Parkinson's disease patients and controls. In multiple cortical and subcortical regions, numerical variation reached nearly one-third of the population variability, altering statistical conclusions about group differences and clinical associations. To address this, we developed a practical tool that estimates the Numerical-Population Variability Ratio (NPVR), enabling researchers to assess the impact of numerical noise in existing studies. By applying this framework to thirteen previously published studies reporting MRI measures of Parkinson's disease, we quantified the probability of numerically induced false positives in the literature, highlighting a substantial impact of numerical variability on MRI measures of Parkinson's disease. These results underscore the importance of systematically evaluating numerical stability in neuroimaging.

1 Introduction

While Magnetic Resonance Imaging (MRI) remains a promising source of biomarkers for a variety of brain-related conditions, the reproducibility of MRI-derived measures across analytical conditions has been challenged in multiple contexts over the past few years. In structural MRI, cortical surface analyses have been shown to be substantially affected by software, parcellation, and quality control [3], in functional MRI, different research teams analyzing the data have found only moderately consistent results [4], and in diffusion MRI the variability among white matter bundle segmentation protocols using diffusion MRI was found to be comparable to variability across subjects [34]. The reliability of MRI-derived measures critically depends on a better understanding and characterization of the impacts of analytical variability, including data selection, analytical decisions, tool selection, computational infrastructure, and numerical state [20, 22].

Among these sources of analytical variability, numerical variability has been shown to have a measurable impact on MRI analyses [14, 12] but remains understudied, mainly due to the practical challenges of quantifying its effects. Numerical variability arises from rounding and truncation errors associated with the use of limited-precision numerical formats, such as the widespread IEEE-754 standard for floating-point arithmetic [27]. Numerical errors manifest slightly differently across computational platforms (hardware, operating systems, or library versions) and such differences sometimes accumulate and amplify across computational analyses, eventually leading to measurable differences in final outputs [33, 21, 8, 38, 6, 28]. Such issues occur particularly within high-dimensional optimization processes such as linear and non-linear image registration, or the training of deep learning models [13].

The implications of numerical variability for clinical measures are unknown. Previous studies focused on measuring its impact on image pre-processing and did not include any particular downstream analysis. The first aim of this study was to quantify the impact of numerical variability in structural

MRI analyses of Parkinson's disease (PD), where robust MRI measures of the disease have yet to be identified. We conducted typical cross-sectional and longitudinal analyses of structural MRI data of PD participants, measuring numerical variability through an experimental stochastic arithmetic approach.

Building on these observations, we developed an analytical framework and associated tools to rapidly assess the numerical quality of structural MRI analyses reported in the published literature, opening the possibility to conduct large-scale impact evaluations of numerical variability. By making numerical-variability evaluation accessible, our framework and tool enhance transparency, support peer review, and promote more reliable statistical inference in neuroimaging. Applying this framework to the PD literature, we obtained the first estimates of the numerical quality of MRI analyses in Parkinson's disease studies, highlighting a widespread impact of numerical variability on MRI measures of PD.

2 Results

2.1 Numerical variability alters statistical inference in MRI measures of PD

We assessed the impact of numerical variability on conclusions drawn from MRI analyses of Parkinson's disease, focusing on two common analyses: (1) volumetric group differences between PD subjects and Healthy Controls (HC), and (2) partial correlations between regional volumes and motor evaluation scores measured with the MDS-Unified Parkinson's Disease Rating Scale part 3 (UPDRS-III). For both, we conducted a cross-sectional analysis at baseline and a longitudinal analysis across two time points.

We selected T1-weighted MRI data from the Parkinson Progression Marker Initiative (PPMI) dataset [26], including participants with at least two usable visits separated by 0.9-2.0 years, and excluding participants with Mild Cognitive Impairment (MCI) or other neurological disorders. The final dataset included 112 PD participants without MCI (PD-non-MCI) and 89 HC participants (Table 2).

We processed all images for both time points using FreeSurfer's `recon-all`. To quantify the numerical variability, we introduced machine-level numerical noise via Monte Carlo Arithmetic (MCA) [29] to mimic realistic perturbations into this pipeline. MCA injects random, zero-mean perturbations into floating-point operations while preserving mathematical expectations. Injected perturbations are on the order of IEEE-754 machine precision. Such noise is representative of the noise introduced by typical variations in hardware (e.g., CPU models), software libraries (e.g., operating system updates), or parallelization (e.g., sequential vs multi-threaded executions). The validity of MCA to model realistic numerical variability in neuroimaging pipelines has been demonstrated in prior work for Operating Systems [33] and hardware variations [38]. We repeated the perturbed analyses, yielding 26 usable runs from which we estimated numerical variability. We verified the validity of the numerical perturbation approach by confirming that all unperturbed results belonged to the range defined by numerically perturbed results.

For both group comparisons and correlation analyses, statistical outcomes varied substantially across the 26 Monte Carlo Arithmetic (MCA) repetitions (Figures 1 and 2). For subcortical volumes (14 regions; Figure 1), p-values crossed the 0.05 threshold in 27% of all (regions, analysis) pairs, indicating frequent inconsistencies across repetitions. For cortical thickness (68 regions; Figure 2), 21% of (regions, analysis) pairs were similarly unstable. These fluctuations demonstrate that numerical noise alone can alter downstream statistical inference in structural MRI analyses of PD.

2.2 A practical model to quantify the impact of numerical variability

While the previous results demonstrate that numerical variability can alter statistical inference in MRI analyses of PD, routinely conducting computationally expensive Monte Carlo evaluations is impractical for most studies. To address this issue, we developed a closed-form analytical model that predicts numerical uncertainty using only standard summary statistics. The core of this framework is the Numerical-Population Variability Ratio (ν_{npv}), defined as the ratio between numerical variability (σ_{num}; Eq. 1) and population variability (σ_{pop}; Eq. 2):

$$\nu_{\text{npv}} = \frac{\sigma_{\text{num}}}{\sigma_{\text{pop}}}$$

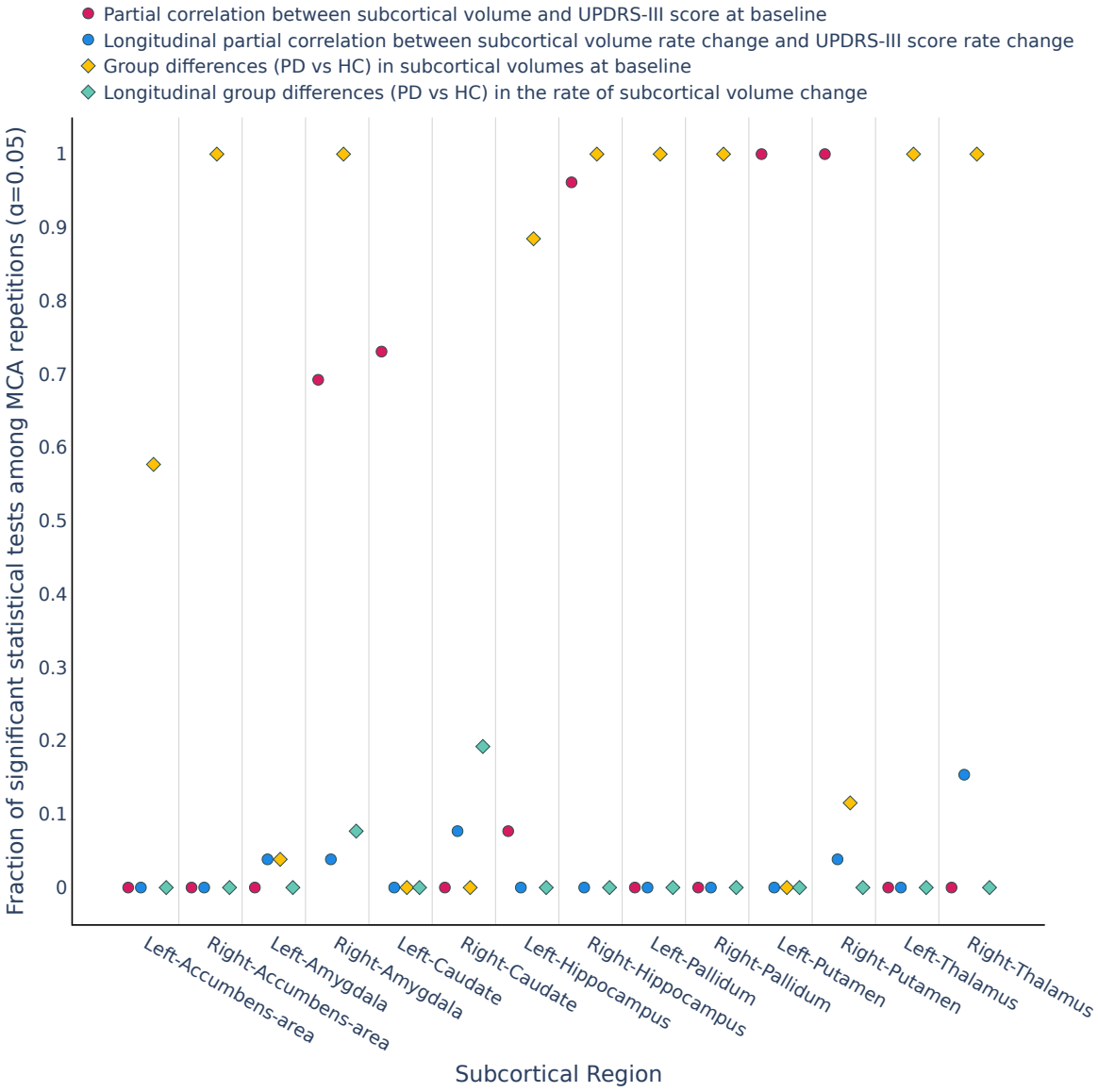


Figure 1: Proportion of significant tests ($p < 0.05$, uncorrected) for subcortical volumes across 26 numerical perturbations.

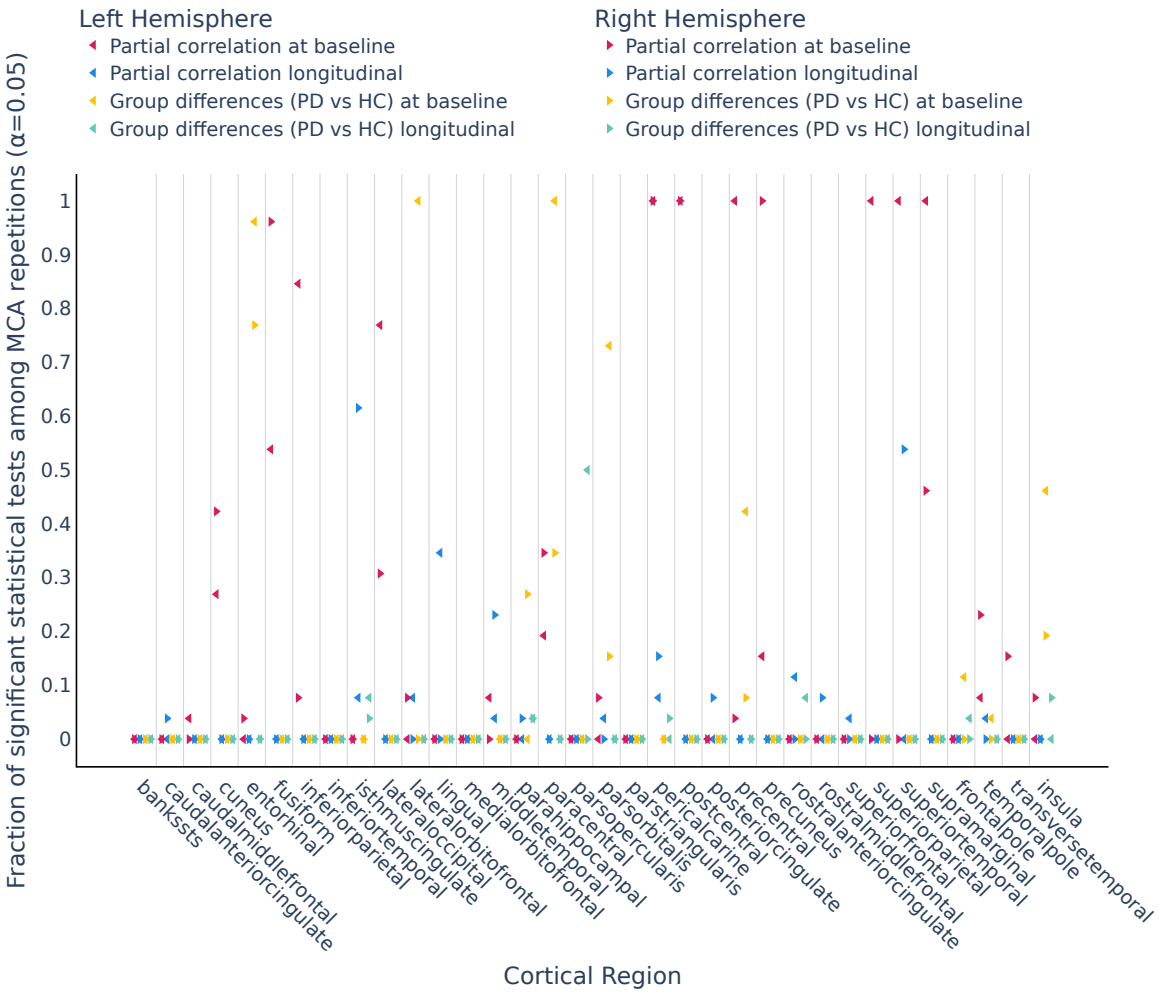


Figure 2: Proportion of significant tests ($p < 0.05$, uncorrected) for cortical thickness across 26 numerical perturbations.

The ν_{npv} metric standardizes the quantification of numerical instability, enabling direct comparisons across different brain regions, software pipelines, and cohorts. By propagating this ratio through standard statistical estimators using the delta-method (see Section 4.3.2), we derived approximations for the numerical uncertainty in common statistics (Table 1) and validated them through numerical simulations. This framework establishes a direct link between a pipeline's numerical instability (ν_{npv}), the study's sample size (n), and the resulting reliability of p-values and effect sizes. Crucially, because these formulas rely solely on summary statistics, they allow for the retrospective quality control of existing literature without requiring access to original raw data or costly re-computation.

Statistic	**Numerical standard deviation**	**Numerical p-value uncertainty**
Cohen's d	$\sigma_d \approx \nu_{\text{npv}} \frac{2}{\sqrt{n}}$	-
Two-sample t	$\sigma_t \approx \nu_{\text{npv}}$	$\sigma_p \approx 2 f_{t,df}(\lvert t \rvert) \nu_{\text{npv}}$
Partial correlation	$\sigma_r \geq \nu_{\text{npv}} \sqrt{\frac{(1-r^2)^3}{n-1}}$	$\sigma_p \geq 2 f_{t,df}(\lvert t \rvert) \sqrt{\frac{df}{n-1}} \nu_{\text{npv}}$
ANCOVA	$\sigma_F \approx 2\sqrt{F} \nu_{\text{npv}}$	$\sigma_p \approx 2\sqrt{F} f_{\mathcal{F}}(F; 1, df_2)\, \nu_{\text{npv}}$

Table 1: First-order numerical uncertainty of common statistical tests under Monte Carlo Arithmetic perturbations. Cohen's d formula assumes large and equal group sizes. $f_{t,df}$ and $f_{\mathcal{F}}(F; 1, df_2)$ denote the probability density functions of the Student's t-distribution with df degrees of freedom and the $\mathcal{F}$-distribution with $(1, df_2)$ degrees of freedom, respectively. The p-value approximation for the partial correlation uses $t = r(df/(1-r^2))^{1/2}$.

We applied this model to quantify the stability of the analyses reported above (Figure 3). In cross-sectional baseline analyses, the average ν_{npv} was 0.191 for the PD group and 0.176 for HC, with no significant difference between groups (permutation test, $p > 0.05$; Appendix Table 10). Applying the Cohen's d uncertainty formula (Table 1) to these ν_{npv} values reveals that suppressing numerical uncertainty to a negligible level ($\sigma_d \leq 0.01$) in cross-sectional studies would require approximately 1,340 participants.

In contrast, longitudinal analyses exhibited substantially higher instability, with average ν_{npv} values of 0.561 for PD and 0.549 for HC, although no significant difference was found between groups (permutation test, $p > 0.05$; Appendix Figure 10). This amplification is likely driven by catastrophic cancellation that occurs when subtracting nearly identical values between timepoints. Consequently, achieving the same level of numerical uncertainty ($\sigma_d \leq 0.01$) would require a sample size exceeding 12,000 participants, highlighting the critical impact of numerical noise on reliability.

To facilitate the use of this framework, we developed an open-source, interactive web interface available at yohanchatelain.github.io/brain_render (Appendix Figure 12). This tool allows researchers to upload standard summary statistics to instantly estimate the numerical uncertainty floor of their findings. The interface includes a 3D visualization engine that maps uncertainty onto cortical and subcortical atlases, allowing for interactive exploration of pipeline stability. To aid interpretation, the tool automatically flags regions where the reported effect size is indistinguishable from the estimated numerical noise, thereby identifying potentially spurious findings.

2.3 Impact of numerical variability on published findings

To assess the broader implications of numerical variability on MRI measures of Parkinson's disease, we applied numerical uncertainty propagation (Table 1) to re-evaluate findings from thirteen articles reporting MRI measures of Parkinson's disease, namely [5, 9, 10, 15, 23, 24, 25, 30, 32, 37, 36, 39, 40]. These studies were selected to illustrate the potential impact of numerical variability on reported outcomes based on available summary statistics.

For each p-value reported as significant in the original articles, we calculated the associated numerical variability using the formulas in Table 1. We then estimated the probability of a significance flip where a result transitions from significant to non-significant due to numerical error by modeling the p-value variability as a Beta distribution (see Section 4.4). We parametrized this distribution such that the reported p-value represents the mean and the calculated numerical variability represents the standard deviation. By computing the cumulative distribution function of this characterized Beta distribution at the original significance threshold, we derived the probability that the reported significant

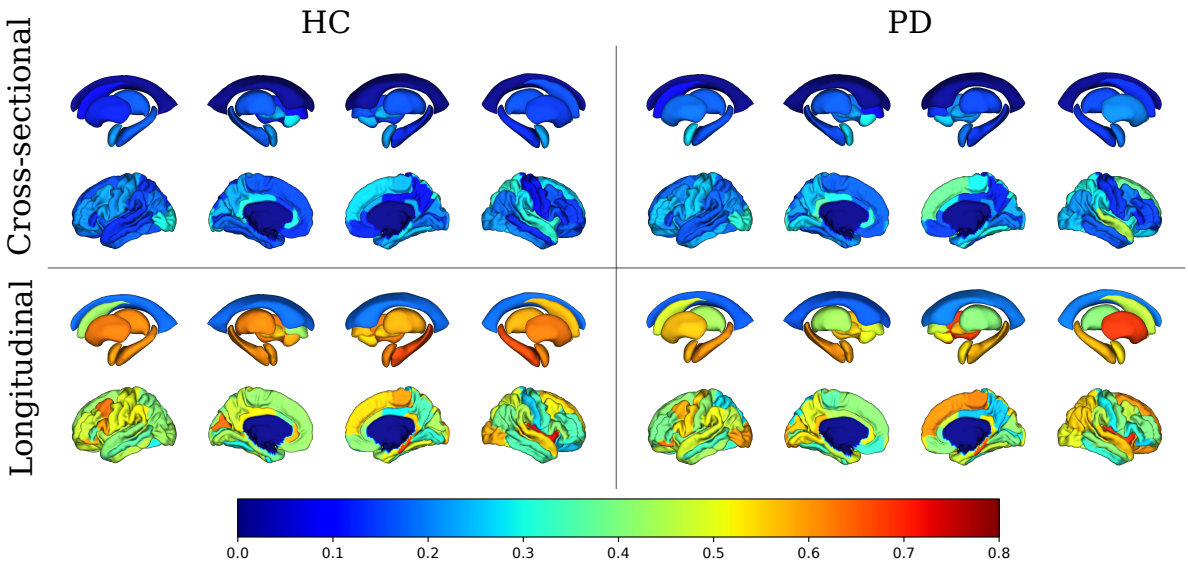


Figure 3: Numerical-Population Variability Ratio (ν_{npv}) for subcortical volumes (top row in each panel) and cortical thickness (bottom row in each panel) in healthy controls (HC) and Parkinson's disease (PD). Higher ν_{npv} values indicate higher computational uncertainty relative to inter-subject variability.

finding was actually a numerically induced misclassification.

Figure 4 shows the estimated probability of numerically induced misclassification as a function of the distance between reported p-values and the significance threshold, across multiple statistical tests. Each point corresponds to a significant result reported in the literature, with marker size reflecting sample size; the thirteen studies are anonymized to avoid singling out individual findings. The vertical axis shows the probability that numerical variability causes the result to cross the significance threshold, while the horizontal axis represents the distance to that threshold. Negative x-axis values correspond to reported significant results ($p < \alpha$), whereas positive values indicate reported non-significant results ($p > \alpha$). For negative distances, the misclassification probability quantifies the likelihood that a reported significant finding becomes non-significant (false positive) under numerical variability; for positive distances, it reflects the likelihood that a reported non-significant finding becomes significant (false negative).

Across statistical tests and analysis types, misclassification probability peaks at the significance threshold, indicating that findings near this boundary are most likely to change significance status under numerical variability. By contrast, results farther from the threshold exhibit lower misclassification probabilities. High misclassification rates persist near the threshold across sample sizes, demonstrating that even well-powered studies can remain numerically fragile. Misclassification probability therefore reflects the joint influence of distance to the significance threshold and the Numerical-Population Variability Ratio. These results reveal a widespread impact of numerical variability on reported MRI measures of Parkinson's disease and underscore the need for systematic numerical robustness assessments in neuroimaging studies.

3 Discussion

In this study, we demonstrate that numerical variability arising from floating-point computation constitutes a substantial source of variability in structural MRI analyses of Parkinson's disease, accounting for up to 90% of population variability. By systematically perturbing the FreeSurfer pipeline using Monte Carlo Arithmetic, we show that numerical noise alone can account for a non-negligible fraction of population variability in both cortical and subcortical measures. This variability is sufficient to alter downstream statistical inference, leading to frequent significance flips in group comparisons and clinical correlations, even when pipeline configurations and computational environments are held constant. These findings provide a concrete computational mechanism that helps explain persistent

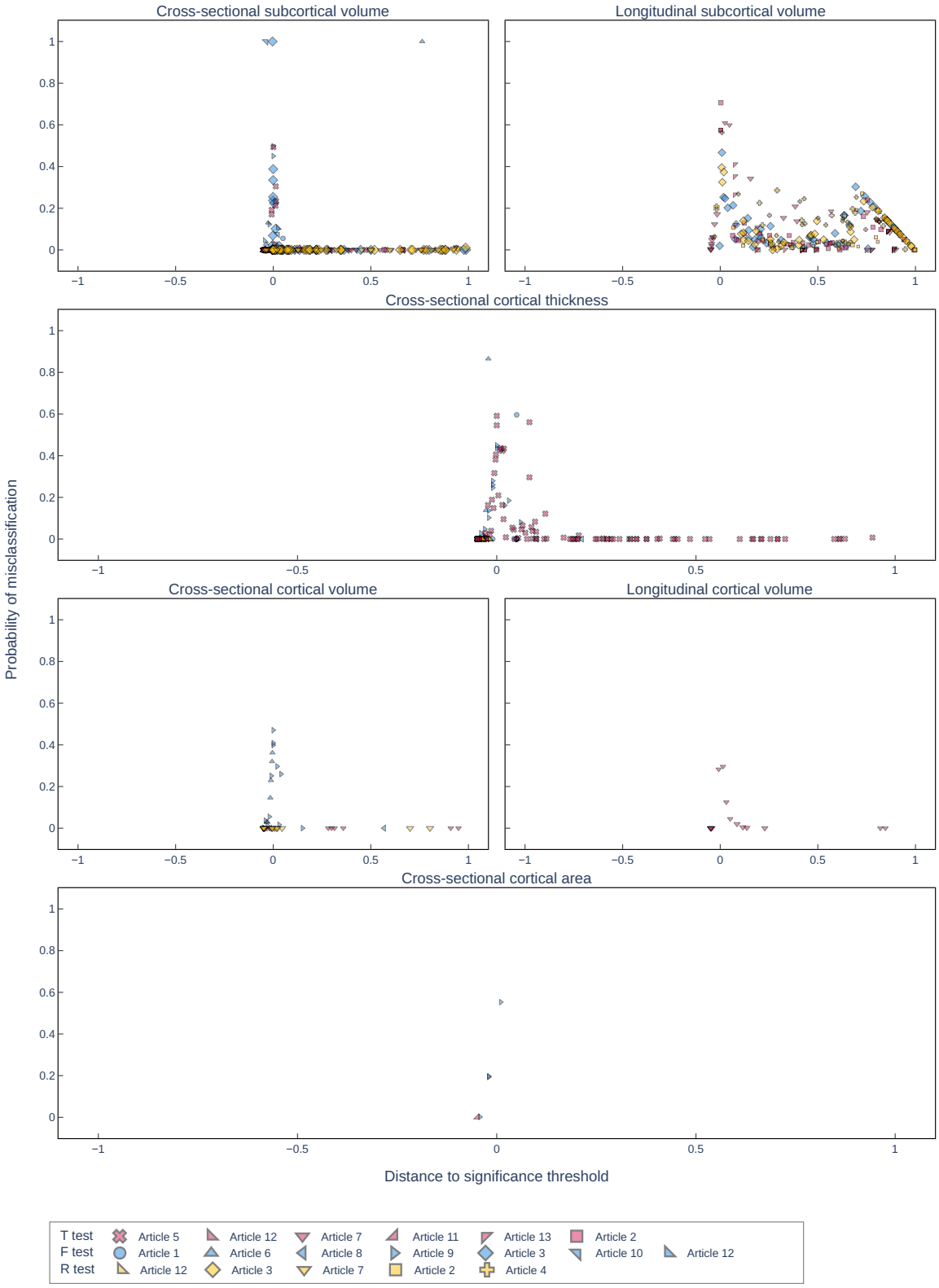


Figure 4: Probability of numerically-induced misclassifications in function of p-value distance to significance threshold across different statistics: T-values (T), F-values (F), and correlation coefficients (R). The marker's size reflects the sample size.

reproducibility challenges reported across clinical neuroimaging.

The Numerical-Population Variability Ratio (NPVR) introduced in this work formalizes numerical variability as a quantity directly comparable to population variability. By analytically propagating NPVR through common statistical estimators, we establish explicit relationships between computational instability, sample size, and uncertainty in effect sizes and p-values. This framework shows that numerical uncertainty does not vanish with increasing sample size in the same manner as sampling error, and that results near conventional significance thresholds are sensitive to numerical perturbations. Crucially, while applied here to numerical instability, this formalism is generalizable: the NPVR could model the propagation of other independent error sources, such as test-retest variability or instrumentation noise, offering a unified perspective on non-sampling errors.

Importantly, NPVR enables practical numerical quality evaluation without requiring access to raw imaging data or computationally expensive reprocessing. Applying this framework to previously published Parkinson's disease MRI studies revealed that a fraction of reported significant findings are susceptible to numerically induced instability, highlighting that numerical variability is not only a technical artifact but a substantive factor that can directly influence scientific conclusions. By making this assessment feasible from summary statistics alone, NPVR provides a scalable approach for retrospective evaluation of the existing literature. In this context, NPVR can be used by researchers and reviewers to estimate the potential contribution of numerical noise to observed false positive rates. Numerical validation demonstrated that these estimates tend to be conservative, thereby minimizing the risk of underestimating the impact of numerical variability.

Although our empirical analyses focused on FreeSurfer 7.3.1, the observed instabilities are not unique to this software. Prior work [28] has documented comparable numerical sensitivity in other widely used neuroimaging pipelines, including FMRIB Software Library [18] (FSL) and Advanced Normalization Tools [2] (ANTs), while suggesting lower sensitivity in Statistical Parametric Mapping [1] (SPM), potentially due to differences in optimization strategies and regularization. These findings indicate that numerical variability is a property of computational pipelines rather than a specific implementation artifact. The magnitude of the uncertainty then depends on algorithmic design choices, numerical precision, and optimization dynamics.

The increasing adoption of deep learning-based components in neuroimaging pipelines does not eliminate numerical variability but instead shifts its focus. The latest FreeSurfer release (v8) now integrates deep learning models such as FastSurfer [16] and Synthmorph [17] to replace classical segmentation and registration steps. While inference in trained models can be relatively stable [31], training itself introduces additional sources of variability, including stochastic optimization, weight initialization, and floating-point precision effects. Recent work [13] suggests that these factors can lead to different trained models under identical initial conditions, echoing the sensitivity to local minima observed in classical nonlinear optimization. Quantifying and controlling numerical variability across both classical and learning-based approaches remains therefore a critical challenge for the field.

We considered several factors that may influence the generalizability of our findings. The Parkinson's disease cohort analyzed here is relatively homogeneous in age and phenotype, which could reduce population variance and inflate NPVR estimates. However, we observed no significant differences in numerical variability between patients and healthy controls, supporting the interpretation that numerical instability is primarily a property of the computational pipeline rather than the clinical population. Extending NPVR measurements across diverse datasets, disease contexts, and software packages will be important to build a comprehensive map of computational reliability in neuroimaging, though our results already suggest that numerical variability is sufficiently large to warrant routine consideration.

More broadly, this work highlights that computational uncertainty should be treated as a serious component of uncertainty in neuroimaging, alongside biological variability and statistical sampling error. Floating-point rounding and truncation are only one contributor; algorithmic choices, preprocessing decisions, and data handling practices all interact with numerical precision to shape final results. Extending this quantification to the deep-learning training stage is equally important, given the field's central role in modern neuroimaging, and would support more robust and interpretable models. Systematic quantification of these effects is essential for the development of numerically robust software and for the reliable translation of neuroimaging biomarkers into clinical and personalized-medicine settings.

In conclusion, NPVR provides a principled, interpretable, and scalable framework to expose hidden numerical instability in neuroimaging analyses. By enabling routine evaluation of numerical uncer-

tainty, this approach strengthens transparency, supports more reliable inference, and offers a concrete path toward improving reproducibility in computational neuroscience.

4 Methods

We quantified the impact of numerical variability on structural MRI findings in Parkinson's disease (PD) using (1) stochastic perturbation experiments and (2) analytical uncertainty propagation. Empirically, we processed a longitudinal PPMI cohort using FreeSurfer instrumented with stochastic numerical noise to isolate run-to-run variability. Analytically, we derived closed-form approximations linking this pipeline instability (ν_{npv}) to uncertainty in effect sizes and test statistics, enabling retrospective quality control using only summary statistics.

The experimental workflow proceeded in four stages: (1) curation of a longitudinal dataset with two visits per participant; (2) repeated processing under Monte Carlo Arithmetic (MCA) perturbations via Fuzzy-libm; (3) rigorous quality control to distinguish numerical artifacts from technical failures; and (4) statistical evaluation of inference instability across repetitions.

4.1 Participants

Structural MRI data were obtained from the Parkinson's Progression Markers Initiative (PPMI; www.ppmi-info.org). The study included 201 participants: 112 individuals diagnosed with Parkinson's disease without mild cognitive impairment (PD-non-MCI) and 89 healthy controls (HC). All participants had two usable T1-weighted MRI scans acquired approximately 1.4 ± 0.5 years apart (0.9-2.0 years). Patients with mild cognitive impairment were excluded to minimize confounding effects of cognitive decline.

Inclusion criteria were: (i) diagnosis of idiopathic Parkinson's disease (PD-non-MCI) or healthy control status; (ii) availability of two high-quality T1-weighted scans at distinct visits; and (iii) absence of other neurological or psychiatric conditions. PD severity was quantified using the Unified Parkinson's Disease Rating Scale part III (UPDRS-III) in the OFF medication state at both baseline and follow-up visits.

All procedures were approved by the research ethics boards of participating PPMI sites, and written informed consent was obtained from all participants in accordance with the Declaration of Helsinki and was exempt from review by Concordia University's Research Ethics Unit. The PD and HC groups did not differ significantly in age, education, or sex distribution ($p > 0.05$; Table 2).

Cohort	**HC**	**PD-non-MCI**
n	89	112
Age (years)	60.7 ± 9.7	60.6 ± 8.9
Age range	30.6 - 79.8	39.2 - 78.3
Gender (male,%)	47 (52.8%)	74 (66.1%)
Education (years)	16.7 ± 3.4	16.0 ± 3.1
UPDRS-III OFF baseline	–	23.3 ± 10.0
UPDRS-III OFF follow-up	–	25.6 ± 11.2
Inter-visit interval (years)	1.4 ± 0.5	1.4 ± 0.6

Table 2: **Participant characteristics.** Values represent mean ± standard deviation. PD = Parkinson's disease; MCI = mild cognitive impairment; UPDRS = Unified Parkinson's Disease Rating Scale. The PD-non-MCI longitudinal subset corresponds to participants with available follow-up MRI and disease severity scores available.

4.2 Image acquisition and preprocessing

High-resolution T1-weighted MRI scans were obtained from the Parkinson's Progression Markers Initiative (PPMI). Data were acquired using standardized 3D MPRAGE protocols (TR = 2.3 s, TE = 2.98 ms, TI = 0.9 s, voxel size = 1 mm isotropic) across multiple sites. While the protocol was standardized, we accounted for minor acquisition variations inherent to the PPMI multisite design.

Structural images were processed using FreeSurfer 7.3.1 instrumented with Fuzzy-libm, a modified mathematical library that injects stochastic perturbations into floating-point operations to probe numerical stability (Section 4.3). To estimate numerical variability, each scan was processed 34 times.

We filtered the resulting outputs to exclude both technical failures (e.g., incomplete execution and FreeSurfer failures) and quality control (QC) failures.

For QC, we visually inspected nine representative slices (3 along each axis) per run to identify gross artifacts, such as missing brain tissue, blurring, or biologically implausible segmentation. Appendix Table 6 details the exclusion rates. To ensure statistical consistency across the cohort, we randomly subsampled the remaining runs to retain exactly 26 valid realisations per participant.

4.3 Numerical variability assessment

To quantify the numerical variability in the FreeSurfer pipeline, we employed Monte Carlo Arithmetic (MCA) [29]. MCA is a stochastic technique that simulates the propagation of rounding errors by introducing controlled random perturbations into floating-point operations. We utilized the Random Rounding (RR) mode, where the result of an arithmetic operation is perturbed by a zero-mean random noise scaled to the magnitude of the least significant bit. For any operation producing an exact result x, the perturbed result $\tilde{x}$ is modeled as $\tilde{x} = x + 2^{e_x - t}\xi$ where $e_x = \lfloor \log_2 |x| \rfloor$ is the exponent of x, t is the *virtual precision* parameter, and ξ is a random variable drawn uniformly from the interval $(-\frac{1}{2}, \frac{1}{2})$.

While comprehensive MCA instrumentation provides a rigorous bound on numerical error, it incurs a prohibitive computational cost (typically 100× to 1000× slowdown), rendering it intractable for large-scale neuroimaging cohorts. To address this, we used *Fuzzy-libm* [33], a lightweight implementation that restricts MCA instrumentation to elementary mathematical library functions (e.g., `exp`, `log`, `sin`, `cos`). By targeting these elementary functions, which are sources of divergence across operating systems [12], Fuzzy-libm significantly reduces computational overhead while effectively capturing the numerical variability relevant to cross-platform reproducibility [33, 38, 8].

Fuzzy-libm is deployed via a Docker container and uses the `LD_PRELOAD` mechanism to dynamically intercept calls to the standard system math library and redirect them to the instrumented version. The library is compiled using Verificarlo [7], an LLVM-based compiler that injects the MCA logic at compile time. Virtual precision parameters are set to match standard hardware precision ($t = 53$ bits for double precision and $t = 24$ bits for single precision), ensuring that the simulated variability remains representative of realistic machine-level precision errors.

4.3.1 Numerical-Population Variability Ratio (ν_{npv})

To quantify computational stability relative to population variation, we introduce the Numerical-Population Variability Ratio (ν_{npv}). For each brain region, ν_{npv} measures the ratio of measurement uncertainty arising from computational processes to natural inter-subject variation:

$$\nu_{\text{npv}} = \frac{\sigma_{\text{num}}}{\sigma_{\text{pop}}}$$

where σ_{num} represents numerical variability (measurement precision across MCA repetitions for individual subjects) and σ_{pop} represents population variability (inter-subject differences within each repetition). For each region of interest, measurements from k MCA repetitions across n subject-visit pairs form a data matrix $\mathcal{M}_{k\times n}$ with entries $x_i^{(r)}$, where $i = 1, \ldots, n$ indexes subject-visits and $r = 1, \ldots, k$ indexes repetitions. Let

$$\bar{x}_i = \frac{1}{k}\sum_{r=1}^{k} x_i^{(r)}, \qquad \bar{x}^{(r)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(r)}.$$

Numerical variability (within-subject, across repetitions):

$$\sigma_{\text{num}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{k-1}\sum_{r=1}^{k}\left(x_i^{(r)} - \bar{x}_i\right)^2\right]. \tag{1}$$

Population variability (within-repetition, across subjects):

$$\sigma_{\text{pop}}^2 = \frac{1}{k}\sum_{r=1}^{k}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^{(r)} - \bar{x}^{(r)}\right)^2\right]. \tag{2}$$

where $\bar{x}_i$ and $\bar{x}^{(r)}$ denote column and row means, respectively. Higher ν_{npv} values indicate regions where computational uncertainty approaches population variation.

4.3.2 Relationship between ν_{npv} and downstream statistical test uncertainty

To establish a quantitative link between a method's computational reproducibility and the reliability of group-level statistical inferences, we derived analytical expressions connecting numerical variability to the uncertainty of commonly used statistical tests (Cohen's d, t-tests, partial correlation, and ANCOVA). Our goal is to characterize how numerical noise propagates through the analytical pipeline to produce uncertainty in the reported statistics.

For each Monte Carlo Arithmetic (MCA) repetition r, we denote by $\tilde{\mathbf{x}}^{(r)} = (\tilde{x}_1^{(r)}, \ldots, \tilde{x}_N^{(r)})^\top$ the vector of perturbed measurements across N subjects. Each perturbed observation is modeled as the sum of the subject's biological value and a numerical error term:

$$\tilde{x}_i^{(r)} = x_i + \varepsilon_i^{(r)}, \qquad \mathbb{E}[\varepsilon_i^{(r)}] = 0, \qquad \mathrm{Cov}[\boldsymbol{\varepsilon}^{(r)}] = \Sigma_{\text{num}}.$$

Here, $\mathbf{x} = (x_1, \ldots, x_N)^\top$ represents the fixed, biological measurements, while $\boldsymbol{\varepsilon}^{(r)}$ captures the random numerical perturbations introduced during computation.

Assumptions. To isolate the contribution of numerical variability, we make three simplifying assumptions:

1. **Numerical error model.** The numerical perturbations are modeled as independent, zero-mean Gaussian random variables: $\varepsilon_i^{(r)} \sim \mathcal{N}(0, \sigma^2_{\text{num},i}), \boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\text{num}})$, where, under homoscedasticity, $\Sigma_{\text{num}} = \sigma^2_{\text{num}} I_N$.

2. **Population variability.** The between-subject variance $\sigma^2_{\text{pop}} = \mathrm{Var}(\{x_i\})$ dominates the numerical noise, i.e. $\sigma_{\text{num},i} \ll \sigma_{\text{pop}}$. This implies that the pooled empirical standard deviation of the observed data can be approximated by the biological one, $s_p \approx \sigma_{\text{pop}}$.

3. **Null-hypothesis scenario.** We condition on the observed biological measurements $\mathbf{x}$ and quantify variability only from numerical perturbations $\boldsymbol{\varepsilon}^{(r)}$. This isolates numerical variability effects from sampling variation; the resulting expressions remain accurate near the null and for small effect sizes. In this setting, the biological values $\mathbf{x} = (x_1, \ldots, x_N)^\top$ are treated as fixed, and all randomness arises from numerical perturbations $\boldsymbol{\varepsilon}^{(r)}$.

Under these assumptions, each downstream statistic $ds = f(\tilde{\mathbf{x}})$ can be linearized around the baseline $\mathbf{x}$ as $ds(\tilde{\mathbf{x}}) \approx ds(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^\top \boldsymbol{\varepsilon}$, allowing the numerical variance to be expressed through the delta method as

$$\mathrm{Var}_{\text{num}}[ds] \approx \nabla_{\mathbf{x}} f(\mathbf{x})^\top \Sigma_{\text{num}} \nabla_{\mathbf{x}} f(\mathbf{x}) = \sigma^2_{\text{num}} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2^2. \tag{3}$$

Table 1 summarizes the derived expressions for the numerical standard deviation of several common statistics.

Cohen's *d* Cohen's effect size quantifies the standardized difference between two sample means. For two independent groups G_1 and G_2 with sample sizes n_1 and n_2 ($df = n_1 + n_2 - 2$), we define:

$$d = \frac{\Delta}{s_p} = \frac{\bar{x}_1 - \bar{x}_2}{s_p}, \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{df}}. \tag{4}$$

The variance of d across Monte Carlo Arithmetic (MCA) repetitions, conditional on the fixed dataset $\mathbf{x}$, is given by:

$$\mathrm{Var}_{\text{num}}[d] = \mathrm{Var}\left[d(\tilde{\mathbf{x}}) \,\middle|\, \mathbf{x}\right].$$

Applying the multivariate delta method (Eq. 3) around the baseline $\mathbf{x}_0 = \mathbf{x}$, we obtain:

$$\mathrm{Var}_{\text{num}}[d] \approx \sigma^2_{\text{num}} \sum_{i=1}^{n} \left(\frac{\partial d}{\partial x_i}\right)^2. \tag{5}$$

For an observation $x_i \in G_g$ ($g \in \{1, 2\}$), the chain rule gives:

$$\frac{\partial d}{\partial x_i} = \frac{1}{s_p}\frac{\partial \Delta}{\partial x_i} - \frac{\Delta}{s_p^2}\frac{\partial s_p}{\partial x_i}, \quad \frac{\partial \Delta}{\partial x_i} = \pm\frac{1}{n_g}, \quad \frac{\partial s_p}{\partial x_i} = \frac{x_i - \bar{x}_g}{df\, s_p}$$

where the sign in $\partial\Delta/\partial x_i$ is positive for $g = 1$ and negative for $g = 2$. Substituting back into the expression for $\partial d/\partial x_i$, we have:

$$\frac{\partial d}{\partial x_i} = \pm\frac{1}{n_g s_p} - \frac{\Delta(x_i - \bar{x}_g)}{df\, s_p^3} \Rightarrow \left(\frac{\partial d}{\partial x_i}\right)^2 = \frac{1}{n_g^2 s_p^2} \pm \frac{2\Delta(x_i - \bar{x}_g)}{n_g df\, s_p^4} + \frac{\Delta^2 (x_i - \bar{x}_g)^2}{df^2 s_p^6}$$

and summing over all i in group G_g:

$$\sum_{i\in G_g}\left(\frac{\partial d}{\partial x_i}\right)^2 = \frac{1}{n_g s_p^2} \pm \frac{2\Delta}{n_g df\, s_p^4}\sum_{i\in G_g}(x_i - \bar{x}_g) + \frac{\Delta^2}{df^2 s_p^6}\sum_{i\in G_g}(x_i - \bar{x}_g)^2 = \frac{1}{n_g s_p^2} + \frac{\Delta^2}{df^2 s_p^6}\left((n_g - 1)s_g^2\right)$$

since $\sum_{i\in G_g}(X_i - \bar{X}_g) = 0$, so finally summing over both groups:

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{\partial d}{\partial X_i}\right)^2 &= \frac{1}{s_p^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{\Delta^2}{df^2 s_p^6}\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)\\ &= \frac{1}{s_p^2}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{df}\frac{\Delta^2}{s_p^2}\frac{1}{s_p^2}\frac{\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)}{df}\right)\\ &= \frac{1}{s_p^2}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{df}\right).\end{aligned}$$

Finally, assuming $s_p \approx \sigma_{\text{pop}}$, the population (between-subject) variance, Eq. (5) becomes:

$$\text{Var}_{\text{num}}[d] \approx \frac{\sigma_{\text{num}}^2}{\sigma_{\text{pop}}^2}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{df}\right). \tag{6}$$

For balanced groups ($n_1 = n_2 = n/2$) and large n, the d^2/df term is negligible:

$$\begin{aligned}\text{Var}_{\text{num}}[d] &\approx \frac{4}{n}\frac{\sigma_{\text{num}}^2}{\sigma_{\text{pop}}^2} = \frac{4}{n}\nu_{\text{npv}}^2\\ \sigma_{\text{num}}[d] &\approx \frac{2}{\sqrt{n}}\,\nu_{\text{npv}}\end{aligned}$$

where $\nu_{\text{npv}} = \sigma_{\text{num}}/\sigma_{\text{pop}}$ is the numerical-population variability ratio. This expression quantitatively links the numerical uncertainty captured by ν_{npv} to the variability of Cohen's d effect size, providing a practical measure of the stability of statistical inferences under finite-precision arithmetic.

Two-sample *t*-test statistic The pooled two-sample t statistic quantifies the standardized difference between two group means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where d is Cohen's d defined in Eq. (4). Defining $\omega_n = \frac{1}{n_1} + \frac{1}{n_2}$, the t statistic can be expressed as $t = d/\sqrt{\omega_n}$. From Eq. (6), the variance of d due to numerical perturbations propagates to the variance of t as:

$$\begin{aligned}\text{Var}_{\text{num}}[t] = \text{Var}\left[\frac{d}{\sqrt{\omega_n}}\right] &\approx \frac{1}{\omega_n}\nu_{\text{npv}}^2\left(\omega_n + \frac{d^2}{df}\right)\\ &= \nu_{\text{npv}}^2\left(1 + \frac{d^2}{df\,\omega_n}\right).\end{aligned} \tag{7}$$

To analyze the correction term, consider

$$df\,\omega_n = (n_1 + n_2 - 2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = (n_1 + n_2 - 2)\frac{n_1 + n_2}{n_1 n_2}$$

When $n_1 \gg n_2$, $df\,\omega_n \approx \frac{n_1-2}{n_2}$; symmetrically, when $n_2 \gg n_1$, $df\,\omega_n \approx \frac{n_2-2}{n_1}$. In both cases, $1/(df\,\omega_n) \to 0$, so the correction term $\frac{d^2}{df\,\omega_n}$ vanishes for unbalanced groups. When the groups are balanced ($n_1 = n_2 = n/2$), $df\omega_n = 4\big(1 - \frac{2}{n}\big) \to 4$ as $n \to \infty$, so that $\frac{d^2}{df\,\omega_n} \to d^2/4$. Substituting these results into Eq. (7) gives:

$$\mathrm{Var}_{\mathrm{num}}[t] \approx \nu_{\mathrm{npv}}^2\,(1+\epsilon),$$

where ϵ tends to 0 for strongly unbalanced groups and to $d^2/4$ for large balanced samples. For small effect sizes ($d^2 \ll 4$) and unbalanced groups, the correction term ϵ is negligible, yielding the simplified expression:

$$\begin{aligned}\mathrm{Var}_{\mathrm{num}}[t] &\approx \nu_{\mathrm{npv}}^2\\ \sigma_{\mathrm{num}}[t] &\approx \nu_{\mathrm{npv}}.\end{aligned}$$

The uncertainty of the corresponding p-values can then be derived. Let X be the random variable with $\mathbb{E}_{\mathrm{num}}[X] = t_0$ and $\mathrm{Var}_{\mathrm{num}}[X] = \nu_{\mathrm{npv}}^2$. Let f_t, F_t be the probability density and cumulative distribution functions of the Student t-distribution with df degrees of freedom. Applying the delta method to the two-sided p-value $p(X) = 2\,(1 - F_t(|X|))$ gives:

$$\begin{aligned}\mathrm{Var}_{\mathrm{num}}[p(X)] &= \mathrm{Var}_{\mathrm{num}}\left[2(1 - F_t(|X|))\right]\\ &\approx \left(-2f_t(|t_0|)\,\mathrm{sign}(t_0)\right)^2 \mathrm{Var}_{\mathrm{num}}[X]\\ &\approx 4(f_t(|t_0|))^2\nu_{\mathrm{npv}}^2,\end{aligned}$$

which gives the standard deviation of the numerical uncertainty in the p-value:

$$\sigma_{\mathrm{num}}[p(X)] = 2f_t(|t_0|)\,\nu_{\mathrm{npv}}.$$

ANCOVA group effect. Analysis of covariance (ANCOVA) evaluates group differences using a general linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\, \sigma^2 I),$$

where $\mathbf{y}$ is the vector of measurements across subjects, and $\mathbf{X}$ includes an intercept, diagnostic group (PD vs. HC), and covariates (e.g., age and sex). The adjusted group difference is expressed as the one-degree-of-freedom contrast $c^\top\boldsymbol{\beta}$, with contrast vector $c = [0,\,1,\,0,\,0]^\top$. Let $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ denote the ordinary least-squares (OLS) estimator and $\widehat{\sigma}_{\mathrm{res}}^2 = SS_{\mathrm{res}}/df_2$ the residual mean square, where $df_2 = n - \mathrm{rank}(\mathbf{X})$. The sum of squares associated with the group effect is:

$$SS_{\mathrm{group}} = \frac{(c^\top\widehat{\boldsymbol{\beta}})^2}{c^\top(\mathbf{X}^\top\mathbf{X})^{-1}c}, \qquad df_1 = 1.$$

The corresponding ANCOVA F statistic is given by:

$$F = \frac{MS_{\mathrm{group}}}{MS_{\mathrm{res}}} = \frac{(c^\top\widehat{\boldsymbol{\beta}})^2}{\widehat{\sigma}_{\mathrm{res}}^2\; c^\top(\mathbf{X}^\top\mathbf{X})^{-1}c}, \qquad F \sim \mathcal{F}(df_1 = 1,\, df_2).$$

Significance is evaluated using the upper-tail of the central $\mathcal{F}$ distribution under the null hypothesis:

$$p(X) = 1 - F_{\mathcal{F}}\left(X; df_1, df_2\right),$$

with $X \sim \mathcal{F}(df_1, df_2)$ and $F_{\mathcal{F}}$ the cumulative distribution function of the F distribution with (df_1, df_2) degrees of freedom. For $df_1 = 1$, the ANCOVA F statistic is equivalent to the two-sample t-test through $F = t^2$ (see [19, p. 403]). Then the uncertainty in the F statistic due to numerical noise follows directly from the uncertainty of t:

$$\mathrm{Var}_{\mathrm{num}}[F] = \mathrm{Var}_{\mathrm{num}}[t^2] \approx (2t)^2\mathrm{Var}_{\mathrm{num}}[t] = 4t^2\nu_{\mathrm{npv}}^2 = 4F\nu_{\mathrm{npv}}^2,$$

yielding

$$\sigma_{\mathrm{num}}[F] = 2\sqrt{F}\,\nu_{\mathrm{npv}}. \tag{8}$$

The uncertainty in the corresponding p-values can be obtained by the delta method. Let X be a random variable with $\mathbb{E}_{\text{num}}[X] = F_0$ and $\text{Var}_{\text{num}}[X] = 4F_0\nu_{\text{npv}}^2$ with $f_{\mathcal{F}}$, $F_{\mathcal{F}}$ the probability density and cumulative distribution functions. Applying the delta-method to the upper-tail p-value is $p(X) = 1 - F_{\mathcal{F}}(X)$ yields:

$$\begin{aligned}\text{Var}_{\text{num}}[p(X)] &= \text{Var}_{\text{num}}[1 - F_{\mathcal{F}}(X)] \\ &\approx f_{\mathcal{F}}(F_0)^2 \text{Var}_{\text{num}}[X] \\ &= 4F_0 f_{\mathcal{F}}(F_0)^2 \nu_{\text{npv}}^2,\end{aligned}$$

so that

$$\sigma_{\text{num}}[p(X)] = 2\sqrt{F_0}\, f_{\mathcal{F}}(F_0)\, \nu_{\text{npv}}. \tag{9}$$

Equations (8) and (9) show that numerical imprecision introduces a variance in the estimated F statistic and its corresponding p-value that scales linearly with the numerical-population variability ratio ν_{npv}, and proportionally to $\sqrt{F}$ for the group effect magnitude.

Partial correlation. Partial correlation measures the association between two variables (x, y) while controlling for the influence of one or more additional variables z. In our analysis, this corresponds to quantifying the relationship between regional brain measurements and UPDRS-III motor scores, controlling for age and sex. The sample partial correlation is defined as:

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}},$$

where r_{xy} denotes the Pearson correlation between variables x and y,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

To simplify notation, we set $a = r_{xy}$, $b = r_{xz}$, and $c = r_{yz}$ so that

$$R(a, b, c) = \frac{a - bc}{\sqrt{(1 - b^2)(1 - c^2)}} = \frac{a - bc}{D}, \qquad D = \sqrt{(1 - b^2)(1 - c^2)}.$$

Applying the delta method (Eq. 3) to the partial correlation, we have:

$$\text{Var}_{\text{num}}[R] \approx \sigma_{\text{num}}^2 \sum_{i=1}^{n} \left(\frac{\partial R}{\partial x_i}\right)^2, \tag{10}$$

Assuming only x is affected by numerical perturbations while y and z are fixed, the gradient with respect to each observation x_i is:

$$\frac{\partial R}{\partial x_i} = \frac{\partial R}{\partial a}\frac{\partial a}{\partial x_i} + \frac{\partial R}{\partial b}\frac{\partial b}{\partial x_i}.$$

The first-order partial derivatives are:

$$\frac{\partial R}{\partial a} = \frac{1}{D}, \qquad \frac{\partial R}{\partial b} = \frac{(1 - c^2)(ab - c)}{D^3}.$$

and the derivatives of the correlations with respect to x_i are (see Eq 22):

$$\frac{\partial a}{\partial x_i} = \frac{(v_i - a\, u_i)}{(n-1)s_x} = \frac{\alpha_i}{(n-1)s_x}, \qquad \frac{\partial b}{\partial x_i} = \frac{(w_i - b\, u_i)}{(n-1)s_x} = \frac{\beta_i}{(n-1)s_x}.$$

where $u_i = (x_i - \bar{x})/s_x$, $v_i = (y_i - \bar{y})/s_y$, and $w_i = (z_i - \bar{z})/s_z$ are standardized and centered observations of x, y, and z respectively. Then $\partial R/\partial x_i$ becomes:

$$\frac{\partial R}{\partial x_i} = \frac{1}{(n-1)s_x}\left[\frac{\alpha_i}{D} + \beta_i \frac{(1 - c^2)(ab - c)}{D^3}\right]$$

thus $(\partial R/\partial x_i)^2$ is:

$$\begin{aligned}\left(\frac{\partial R}{\partial x_i}\right)^2 &= \frac{1}{(n-1)^2 s_x^2}\left[\frac{\alpha_i^2}{D^2} + 2\frac{\alpha_i\beta_i(1-c^2)(ab-c)}{D^4} + \frac{\beta_i^2(1-c^2)^2(ab-c)^2}{D^6}\right] \\ &= \frac{1}{(n-1)^2 s_x^2}\left[\frac{\alpha_i^2}{(1-b^2)(1-c^2)} + 2\frac{\alpha_i\beta_i(1-c^2)(ab-c)}{(1-b^2)^2(1-c^2)^2} + \frac{\beta_i^2(1-c^2)^2(ab-c)^2}{(1-b^2)^3(1-c^2)^3}\right] \\ &= \frac{1}{(n-1)^2 s_x^2}\left[\frac{\alpha_i^2}{(1-b^2)(1-c^2)} + 2\frac{\alpha_i\beta_i(ab-c)}{(1-b^2)^2(1-c^2)} + \frac{\beta_i^2(ab-c)^2}{(1-b^2)^3(1-c^2)}\right].\end{aligned}$$

Using the correlation identities $\sum \alpha_i^2 = (n-1)(1-a^2)$, $\sum \beta_i^2 = (n-1)(1-b^2)$, and $\sum \alpha_i\beta_i = (n-1)(ab-c)$ (see proof in Appendix A.7), we sum over all i to obtain:

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{\partial R}{\partial x_i}\right)^2 &= \frac{1}{(n-1)s_x^2}\left[\frac{(1-a^2)}{(1-b^2)(1-c^2)} + 2\frac{(ab-c)^2}{(1-b^2)^2(1-c^2)} + \frac{(ab-c)^2}{(1-b^2)^2(1-c^2)}\right] \\ &= \frac{1}{(n-1)s_x^2}\left[\frac{(1-a^2)}{(1-b^2)(1-c^2)} + \frac{3(ab-c)^2}{(1-b^2)^2(1-c^2)}\right] \\ &= \frac{1}{(n-1)s_x^2}\left[\frac{(1-a^2)(1+3r_{yz,x}^2)}{(1-b^2)(1-c^2)}\right].\end{aligned}$$

Substituting back into Eq. (10) with $s_x^2 \simeq \sigma_{\text{pop}}^2$ gives:

$$\text{Var}_{\text{num}}[R] \approx \frac{\nu_{\text{npv}}^2}{(n-1)}\left[\frac{(1-a^2)(1+3r_{yz,x}^2)}{(1-b^2)(1-c^2)}\right]. \tag{11}$$

Since a, b, c are rarely reported in practice, we further simplify this expression by deriving the lower bound:

$$(1-R^2)^3 \leq \frac{(1-a^2)}{(1-b^2)(1-c^2)}, \tag{12}$$

First, note that the squared partial correlation is:

$$1-R^2 = \frac{(1-b^2)(1-c^2)-(a-bc)^2}{(1-b^2)(1-c^2)} = \frac{\Delta}{(1-b^2)(1-c^2)},$$

where

$$\Delta = (1-b^2)(1-c^2)-(a-bc)^2 = (1-a^2)(1-b^2)-(ac-b)^2 = (1-a^2)(1-c^2)-(ab-c)^2.$$

Each equality above follows from expanding both sides. Because every squared term is nonnegative, we obtain the three inequalities

$$\Delta \leq (1-a^2)(1-b^2), \qquad \Delta \leq (1-b^2)(1-c^2), \qquad \Delta \leq (1-c^2)(1-a^2). \tag{13}$$

Multiplying the first and third inequalities in (13) gives

$$\Delta^2 \leq (1-a^2)^2(1-b^2)(1-c^2),$$

and multiplying also by the middle one yields

$$\Delta^3 \leq (1-a^2)^2(1-b^2)^2(1-c^2)^2.$$

Since $0 \leq 1-a^2 \leq 1$, we have $(1-a^2)^2 \leq (1-a^2)$, so that

$$\Delta^3 \leq (1-a^2)(1-b^2)^2(1-c^2)^2.$$

Dividing both sides by $(1-b^2)^3(1-c^2)^3$ and substituting $\Delta = (1-b^2)(1-c^2)(1-R^2)$ yields

$$(1-R^2)^3 = \frac{\Delta^3}{(1-b^2)^3(1-c^2)^3} \leq \frac{1-a^2}{(1-b^2)(1-c^2)}.$$

This establishes the claimed bound (12). Since $3r_{yz,x}^2 + 1 \geq 1$ it follows immediately that

$$(1-R^2)^3 \leq \frac{(1-a^2)}{(1-b^2)(1-c^2)} \leq \frac{(1-a^2)(1+3r_{yz,x}^2)}{(1-b^2)(1-c^2)}$$

So, substituting into Eq. (11) gives the lower bound:

$$\nu_{\text{npv}}^2 \frac{(1-R^2)^3}{n-1} \lesssim \text{Var}_{\text{num}}[R]. \tag{14}$$

Taking the square root yields the standard deviation:

$$\nu_{\text{npv}} \sqrt{\frac{(1-r_{xy,z}^2)^3}{n-1}} \lesssim \sigma_{\text{num}}[R].$$

The two-sided significance of a partial correlation is computed from the t-statistic

$$t = R\sqrt{\frac{df}{1-R^2}}, \qquad df = n-k-2,$$

where k is the number of controlling variables. Let X be the random variable with $\mathbb{E}_{\text{num}}[X] = R_0$, $t_0^2 = R_0(df/(1-R_0^2))$ with $\text{Var}_{\text{num}}[X]$ bounded by Equation (14). Let f_t, F_t be the probability density and cumulative distribution functions of the Student t-distribution with df degrees of freedom. Applying the delta method to the two-sided p-value $p(X) = 2\,(1 - F_t(|X|))$ gives:

$$\text{Var}_{\text{num}}[p(X)] \approx \left(\frac{\partial p}{\partial t}\frac{\partial t}{\partial R}\right)^2 \text{Var}_{\text{num}}[X], \tag{15}$$

with the partial derivatives given by:

$$\frac{\partial p}{\partial t} = -2f_t(|t|)\,\text{sign}(t), \qquad \frac{\partial t}{\partial R} = \sqrt{\frac{df}{(1-R^2)^3}}. \tag{16}$$

Combining equations (15) and (16) yields

$$\text{Var}_{\text{num}}[p(X)] \geq 4f_t(|t_0|)^2 \frac{df}{(1-R_0^2)^3}\text{Var}_{\text{num}}[X].$$

Using Equation (14) to bound $\text{Var}_{\text{num}}[R]$, the dependence on $(1-R^2)^3$ cancels, leading to:

$$\sigma_{\text{num}}[p(X)] \geq 2f_t(|t_0|)\sqrt{\frac{df}{n-1}}\nu_{\text{npv}}.$$

4.4 Probability of false positives induced by numerical uncertainty

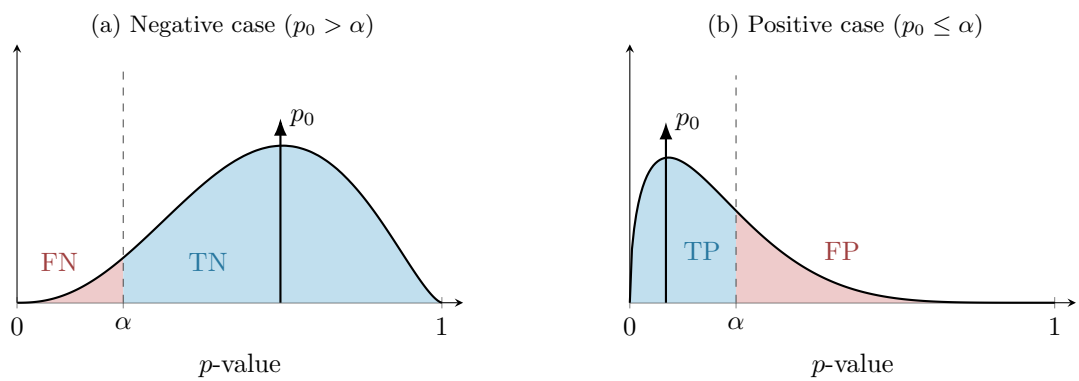
To quantify the probability of false positive findings arising from numerical variability, we model the computed p-value as a random variable following a Beta distribution, $p \sim \text{Beta}(a, b)$, which is suitable for modeling probabilities and proportions bounded on $[0, 1]$. The distribution is parameterized by shape parameters (a, b) determined from a target mean μ_p and variance σ_p^2 estimated using the uncertainty propagation formulae reported in Table 1. The parameters are given by:

$$a = \mu_p\left(\frac{\mu_p(1-\mu_p)}{\sigma_p^2} - 1\right), \quad b = (1-\mu_p)\left(\frac{\mu_p(1-\mu_p)}{\sigma_p^2} - 1\right).$$

This formulation defines a distribution of plausible p-values centered around the nominal value $p_0 = \mu_p$, with dispersion determined by numerical instability. Importantly, this model enables direct estimation of error rates caused by numerical perturbations relative to a fixed significance threshold α. Two distinct regimes are considered, depending on whether the nominal p-value p_0 lies above or below α (Fig. 5):

- Negative Case ($p_0 > \alpha$): The primary finding is non-significant, i.e., a true negative (TN). Here, we calculate the probability that numerical noise shifts the p-value below α, resulting in a false negative (FN) (instability leading to Type I-like errors).
- Positive Case ($p_0 \leq \alpha$): The primary finding is significant, i.e., a true positive (TP). We calculate the probability that numerical noise shifts the p-value above α, resulting in a false positive (FP) (instability leading to Type II-like errors).

The probabilities for these outcomes are obtained by computing the cumulative distribution function (CDF) of the Beta distribution up to the threshold α, as illustrated in Figure 5. This probabilistic modelisation provides an explicit link between numerical variability and classical Type I and Type II error rates, allowing us to isolate the contribution of numerical variability to spurious statistical significance. Numerical validation of this model is presented in Appendix Section E.



$$\mathbb{P}[\text{FN}] = \int_0^\alpha \text{Beta}(x; a, b)dx \qquad \mathbb{P}[\text{TP}] = \int_0^\alpha \text{Beta}(x; a, b)dx$$

$$\mathbb{P}[\text{TN}] = 1 - \mathbb{P}[\text{FN}] \qquad \mathbb{P}[\text{FP}] = 1 - \mathbb{P}[\text{TP}]$$

Figure 5: **Modeling inference stability using Beta distributions.** The panels illustrate the probability of significance flipping due to numerical uncertainty. The negative case ($p_0 > \alpha$, Fig. 5a), where the tail of the distribution crossing α represents the probability of a false negative finding (FN). The positive case ($p_0 \leq \alpha$, Fig. 5b), where the tail extending beyond α represents the probability of a false positive finding (FP).

5 Data Availability

The data that support the findings of this study are available from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the PPMI.

6 Code Availability

All MCA instrumentation scripts, FreeSurfer build instructions and analysis notebooks are available at `github.com/yohanchatelain/livingpark-numerical-variability`. Exact commit hashes are archived on Zenodo (DOI [to be added]) to ensure reproducibility.

7 Acknowledgements

The analyses were conducted on the Virtual Imaging Platform [11], which utilizes resources provided by the Biomed virtual organization within the European Grid Infrastructure (EGI). We extend our gratitude to Sorina Pop from CREATIS, Lyon, France, for her support. This work was funded by the Michael J. Fox Foundation for Parkinson's Research (MJFF- 021134; https://www.michaeljfox.org). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] John Ashburner. Computational anatomy with the spm software. Magnetic resonance imaging, 27(8):1163–1174, 2009.

[2] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). Insight j, 2(365):1–35, 2009.

[3] Nikhil Bhagwat, Amadou Barry, Erin W Dickie, Shawn T Brown, Gabriel A Devenyi, Koji Hatano, Elizabeth DuPre, Alain Dagher, Mallar Chakravarty, Celia MT Greenwood, et al. Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. GigaScience, 10(1):giaa155, 2021.

[4] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature, 582(7810):84–88, 2020.

[5] Marcos Hortes N Chagas, Vitor Tumas, Márcio A Pena-Pereira, João Paulo Machado-de Sousa, Antonio Carlos Dos Santos, Rafael Faria Sanches, Jaime EC Hallak, and José Alexandre S Crippa. Neuroimaging of major depression in parkinson's disease: Cortical thickness, cortical and subcortical volume, and spectroscopy findings. Journal of psychiatric research, 90:40–45, 2017.

[6] Yohan Chatelain, Loïc Tetrel, Christopher J Markiewicz, Mathias Goncalves, Gregory Kiar, Oscar Esteban, Pierre Bellec, and Tristan Glatard. A numerical variability approach to results stability tests and its application to neuroimaging. IEEE Transactions on Computers, 2024.

[7] Christophe Denis, Pablo de Oliveira Castro, and Eric Petit. Verificarlo: checking floating point accuracy through monte carlo arithmetic. In 2016 IEEE 23nd Symposium on Computer Arithmetic (ARITH), 2016.

[8] Morgane Des Ligneris, Axel Bonnet, Yohan Chatelain, Tristan Glatard, Michaël Sdika, Gaël Vila, Valentine Wargnier-Dauchelle, Sorina Pop, and Carole Frindel. Reproducibility of tumor segmentation outcomes with a deep learning model. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2023.

[9] Anna I Garcia-Diaz, B Segura, Hugo César Baggio, MJ Marti, F Valldeoriola, Yaroslau Compta, P Vendrell, N Bargallo, Eduardo Tolosa, and C Junque. Structural mri correlates of the mmse and pentagon copying test in parkinson's disease. Parkinsonism & related disorders, 20(12):1405–1410, 2014.

[10] Niels JHM Gerrits, Anita C van Loenhoud, Stan F van den Berg, Henk W Berendse, Elisabeth MJ Foncke, Martin Klein, Diederick Stoffers, Ysbrand D van der Werf, and Odile A van den Heuvel. Cortical thickness, surface area and subcortical volume differentially contribute to cognitive heterogeneity in parkinson's disease. PloS one, 11(2):e0148852, 2016.

[11] Tristan Glatard, Carole Lartizien, Bernard Gibaud, Rafael Ferreira Da Silva, Germain Forestier, Frédéric Cervenansky, Martino Alessandrini, Hugues Benoit-Cattin, Olivier Bernard, Sorina Camarasu-Pop, et al. A virtual imaging platform for multi-modality medical image simulation. IEEE transactions on medical imaging, 32(1):110–118, 2012.

[12] Tristan Glatard, Lindsay B Lewis, Rafael Ferreira da Silva, Reza Adalat, Natacha Beck, Claude Lepage, Pierre Rioux, Marc-Etienne Rousseau, Tarek Sherif, Ewa Deelman, et al. Reproducibility of neuroimaging analyses across operating systems. Frontiers in neuroinformatics, 9:12, 2015.

[13] Inés Gonzalez-Pepe, Vinuyan Sivakolunthu, Yohan Chatelain, and Tristan Glatard. Uncertain but useful: Leveraging cnn variability into data augmentation. arXiv preprint arXiv:2509.05238, 2025.

[14] Ed HBM Gronenschild, Petra Habets, Heidi IL Jacobs, Ron Mengelers, Nico Rozendaal, Jim Van Os, and Machteld Marcelis. The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. PloS one, 7(6):e38234, 2012.

[15] Alexandru Hanganu, Christophe Bedetti, Clotilde Degroot, Beatriz Mejia-Constain, Anne-Louise Lafontaine, Valerie Soland, Sylvain Chouinard, Marie-Andree Bruneau, Samira Mellah, Sylvie Belleville, et al. Mild cognitive impairment is linked with faster rate of cortical thinning in patients with parkinson's disease longitudinally. Brain, 137(4):1120–1129, 2014.

[16] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. NeuroImage, 219:117012, 2020.

[17] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. IEEE transactions on medical imaging, 41(3):543–558, 2021.

[18] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. Neuroimage, 62(2):782–790, 2012.

[19] Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. Continuous univariate distributions, volume 2, volume 2. John wiley & sons, 1995.

[20] David N Kennedy, Sanu A Abraham, Julianna F Bates, Albert Crowley, Satrajit Ghosh, Tom Gillespie, Mathias Goncalves, Jeffrey S Grethe, Yaroslav O Halchenko, Michael Hanke, et al. Everything matters: the repronim perspective on reproducible neuroimaging. Frontiers in neuroinformatics, 13:1, 2019.

[21] Gregory Kiar, Yohan Chatelain, Pablo de Oliveira Castro, Eric Petit, Ariel Rokem, Gaël Varoquaux, Bratislav Misic, Alan C Evans, and Tristan Glatard. Numerical uncertainty in analytical pipelines lead to impactful variability in brain networks. PloS one, 16(11):e0250755, 2021.

[22] Gregory Kiar, Jeanette A Mumford, Ting Xu, Joshua T Vogelstein, Tristan Glatard, and Michael P Milham. Why experimental variation in neuroimaging should be embraced. Nature communications, 15(1):9411, 2024.

[23] Mechelle M Lewis, Guangwei Du, Eun-Young Lee, Zeinab Nasralah, Nicholas W Sterling, Lijun Zhang, Daymond Wagner, Lan Kong, Alexander I Tröster, Martin Styner, et al. The pattern of gray matter atrophy in parkinson's disease differs in cortical and subcortical regions. Journal of neurology, 263(1):68–75, 2016.

[24] Jianyu Li, Yuanchao Zhang, Zitong Huang, Yihan Jiang, Zhanbing Ren, Daihong Liu, Jiuquan Zhang, Roberta La Piana, and Yifan Chen. Cortical and subcortical morphological alterations in motor subtypes of parkinson's disease. npj Parkinson's Disease, 8(1):167, 2022.

[25] E Mak, N Bergsland, MG Dwyer, R Zivadinov, and NJAAJN Kandiah. Subcortical atrophy is associated with cognitive impairment in mild parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. American Journal of Neuroradiology, 35(12):2257–2264, 2014.

[26] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). Progress in neurobiology, 95(4):629–635, 2011.

[27] Peter Markstein. The new ieee-754 standard for floating point arithmetic. Technical report, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2008. Lecture / tutorial on the IEEE-754 revision.

[28] Niusha Mirhakimi, Yohan Chatelain, Jean-Baptiste Poline, and Tristan Glatard. Numerical uncertainty in linear registration: An experimental study. arXiv preprint arXiv:2508.00781, 2025.

[29] Douglass Stott Parker. Monte Carlo arithmetic: exploiting randomness in floating-point arithmetic. Citeseer, 1997.

[30] Clelia Pellicano, Flavia Niccolini, Kit Wu, Sean S O'Sullivan, Andrew D Lawrence, Andrew J Lees, Paola Piccini, and Marios Politis. Morphometric changes in the reward system of parkinson's disease patients with impulse control disorders. Journal of neurology, 262(12):2653–2661, 2015.

[31] Inés Gonzalez Pepe, Vinuyan Sivakolunthu, Hae Lang Park, Yohan Chatelain, and Tristan Glatard. Numerical uncertainty of convolutional neural networks inference for structural brain mri analysis. In International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, pages 64–73. Springer, 2023.

[32] Andrius Radziunas, Vytenis Pranas Deltuva, Arimantas Tamasauskas, Rymante Gleizniene, Aiste Pranckeviciene, Kestutis Petrikonis, and Adomas Bunevicius. Brain mri morphometric analysis in parkinson's disease patients with sleep disturbances. BMC neurology, 18(1):88, 2018.

[33] Ali Salari, Yohan Chatelain, Gregory Kiar, and Tristan Glatard. Accurate simulation of operating system updates in neuroimaging using monte-carlo arithmetic. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3, pages 14–23. Springer, 2021.

[34] Kurt G Schilling, François Rheault, Laurent Petit, Colin B Hansen, Vishwesh Nath, Fang-Cheng Yeh, Gabriel Girard, Muhamed Barakovic, Jonathan Rafael-Patino, Thomas Yu, et al. Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset? Neuroimage, 243:118502, 2021.

[35] Devan Sohier, Pablo De Oliveira Castro, François Févotte, Bruno Lathuilière, Eric Petit, and Olivier Jamond. Confidence intervals for stochastic arithmetic. ACM Transactions on Mathematical Software (TOMS), 47(2):1–33, 2021.

[36] Andrzej Sokołowski, Nikhil Bhagwat, Yohan Chatelain, Mathieu Dugré, Alexandru Hanganu, Oury Monchi, Brent McPherson, Michelle Wang, Jean-Baptiste Poline, Madeleine Sharp, et al. Longitudinal brain structure changes in parkinson's disease: A replication study. Plos one, 19(1):e0295069, 2024.

[37] Andrzej Sokołowski, Nikhil Bhagwat, Dimitrios Kirbizakis, Yohan Chatelain, Mathieu Dugré, Jean-Baptiste Poline, Madeleine Sharp, and Tristan Glatard. The impact of freesurfer versions on structural neuroimaging analyses of parkinson's disease. bioRxiv, pages 2024–11, 2024.

[38] Gaël Vila, Emmanuel Medernach, Ines Gonzalez Pepe, Axel Bonnet, Yohan Chatelain, Michaël Sdika, Tristan Glatard, and Sorina Camarasu Pop. The impact of hardware variability on applications packaged with docker and guix: A case study in neuroimaging. In Proceedings of the 2nd ACM Conference on Reproducibility and Replicability, pages 75–84, 2024.

[39] Heather Wilson, Flavia Niccolini, Clelia Pellicano, and Marios Politis. Cortical thinning across parkinson's disease stages and clinical correlates. Journal of the neurological sciences, 398:31–38, 2019.

[40] Wenyi Yang, Xueqin Bai, Xiaojun Guan, Cheng Zhou, Tao Guo, Jingjing Wu, Xiaojun Xu, Minming Zhang, Baorong Zhang, Jiali Pu, et al. The longitudinal volumetric and shape changes of subcortical nuclei in parkinson's disease. Scientific Reports, 14(1):7494, 2024.

A Partial derivatives of sample statistics

We derive below the partial derivatives of common sample statistics for a dataset $x = \{x_1, x_2, \ldots, x_n\}$ with respect to an individual observation x_i, where n denotes the sample size. The Kronecker delta δ_{ij} equals 1 when $i = j$ and 0 otherwise.

A.1 Sample Mean

The partial derivative of the sample mean with respect to x_i is constant:

$$\frac{\partial \overline{x}}{\partial x_i} = \frac{1}{n}. \tag{17}$$

Proof. The sample mean is defined as

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j.$$

Taking the partial derivative with respect to x_i gives

$$\frac{\partial \overline{x}}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial x_j}{\partial x_i} = \frac{1}{n} \sum_{j=1}^{n} \delta_{ij} = \frac{1}{n}.$$

□

A.2 Sample Variance

The partial derivative of the sample variance with respect to x_i is

$$\frac{\partial s^2}{\partial x_i} = \frac{2(x_i - \overline{x})}{n-1}. \tag{18}$$

Proof. The sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2.$$

Differentiating with respect to x_i yields

$$\begin{aligned}
\frac{\partial s^2}{\partial x_i} &= \frac{1}{n-1} \sum_{j=1}^{n} \frac{\partial}{\partial x_i} (x_j - \overline{x})^2 \\
&= \frac{2}{n-1} \sum_{j=1}^{n} (x_j - \overline{x}) \frac{\partial (x_j - \overline{x})}{\partial x_i} \\
&= \frac{2}{n-1} \sum_{j=1}^{n} (x_j - \overline{x}) \left(\delta_{ij} - \frac{1}{n} \right) \\
&= \frac{2}{n-1} \left[(x_i - \overline{x}) \left(1 - \frac{1}{n} \right) - \frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (x_j - \overline{x}) \right].
\end{aligned}$$

Since $\sum_{j=1}^{n} (x_j - \overline{x}) = 0$ then $\sum_{\substack{j=1 \\ j \neq i}}^{n} (x_j - \overline{x}) = -(x_i - \overline{x})$ so the second term simplifies, giving

$$\frac{\partial s^2}{\partial x_i} = \frac{2(x_i - \overline{x})}{n-1}.$$

□

A.3 Sample Standard Deviation

The partial derivative of the sample standard deviation with respect to x_i is

$$\frac{\partial s}{\partial x_i} = \frac{x_i - \overline{x}}{(n-1)s}. \tag{19}$$

Proof. Given that $s = \sqrt{s^2}$, the derivative follows directly from the chain rule:

$$\begin{aligned}
\frac{\partial s^2}{\partial x_i} &= \frac{2(x_i - \overline{x})}{n-1}, \\
2s\frac{\partial s}{\partial x_i} &= \frac{2(x_i - \overline{x})}{n-1}, \\
\frac{\partial s}{\partial x_i} &= \frac{x_i - \overline{x}}{(n-1)s}.
\end{aligned}$$

□

A.4 Pooled Standard Deviation

The pooled standard deviation is a weighted average of the variances of two groups $|G_1| = n_1$ and $|G_2| = n_2$ with $df = n_1 + n_2 - 2$. Its partial derivative with respect to x_i, $i \in G_g$ is given by

$$\frac{\partial s_p}{\partial x_i} = \frac{x_i - \overline{x}_g}{df\, s_p}. \tag{20}$$

Proof. Let x be partitioned into two groups G_1 and G_2 with sizes n_1 and n_2, $\overline{x}_1$, $\overline{x}_2$ be the sample means and s_1, s_2 be the sample standard deviations of groups G_1 and G_2, respectively. Let $df = n_1 + n_2 - 2$ then the pooled standard deviation s_p is defined as

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{df}}.$$

Differentiating s_p with respect to x_i in group G_g gives:

$$\begin{aligned}
\frac{\partial s_p}{\partial x_i} &= \frac{\partial}{\partial x_i}\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{df}\right]^{\frac{1}{2}} \\
&= \frac{1}{2s_p}\frac{1}{df}\frac{\partial}{\partial x_i}\left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right] \\
&= \frac{1}{df}\frac{1}{2s_p}2(x_i - \bar{x}_g) \\
\frac{\partial s_p}{\partial x_i} &= \frac{x_i - \bar{x}_g}{df\, s_p}
\end{aligned}$$

□

A.5 Sample Covariance

The partial derivative of the sample covariance with respect to x_i is

$$\frac{\partial s_{xy}}{\partial x_i} = \frac{y_i - \overline{y}}{(n-1)}. \tag{21}$$

Proof. The sample covariance between two variables x and y is defined as

$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}).$$

Taking the partial derivative with respect to x_i gives

$$\begin{aligned}
\frac{\partial s_{xy}}{\partial x_i} &= \frac{1}{n-1}\sum_{j=1}^{n}\frac{\partial}{\partial x_i}\left[(x_j-\overline{x})(y_j-\overline{y})\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})\frac{\partial(x_j-\overline{x})}{\partial x_i} \\
&= \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})\left(\delta_{ij}-\frac{1}{n}\right) \\
&= \frac{1}{n-1}\left[(y_i-\overline{y})\left(1-\frac{1}{n}\right)-\frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(y_j-\overline{y})\right].
\end{aligned}$$

Since $\sum_{j=1}^{n}(y_j-\overline{y})=0$ then $\sum_{\substack{j=1\\ j\neq i}}^{n}(y_j-\overline{y})=-(y_i-\overline{y})$ so the second term simplifies, giving

$$\frac{\partial s_{xy}}{\partial x_i}=\frac{y_i-\overline{y}}{(n-1)}.$$

□

A.6 Pearson correlation coefficient

The partial derivative of $r_{x,y}$ with respect to an individual observation x_i is given by

$$\frac{\partial r_{x,y}}{\partial x_i}=\frac{1}{(n-1)s_x}\left(\frac{y_i-\overline{y}}{s_y}-\frac{r_{x,y}}{s_x}\frac{x_i-\overline{x}}{s_x}\right). \tag{22}$$

Proof. The Pearson correlation coefficient r between two variables x and y is defined as

$$r_{x,y}=\frac{s_{xy}}{s_x s_y},$$

using the quotient rule, we differentiate $r(x,y)$ with respect to x_i:

$$\begin{aligned}
\frac{\partial r_{x,y}}{\partial x_i} &= \frac{1}{s_x^2 s_y^2}\left[\frac{\partial s_{xy}}{\partial x_i}\cdot s_x s_y - s_{xy}\cdot\frac{\partial s_x s_y}{\partial x_i}\right] \\
&= \frac{1}{s_x s_y}\frac{\partial s_{xy}}{\partial x_i}-\frac{s_{xy}}{s_x^2 s_y}\frac{\partial s_x}{\partial x_i} \\
&= \frac{1}{s_x s_y}\frac{\partial s_{xy}}{\partial x_i}-\frac{r_{x,y}}{s_x}\frac{\partial s_x}{\partial x_i}.
\end{aligned}$$

Substituting the partial derivatives of the sample covariance (Eq. 21) and standard deviation (Eq. 19) we obtain

$$\begin{aligned}
\frac{\partial r_{x,y}}{\partial x_i} &= \frac{1}{s_x s_y}\cdot\frac{y_i-\overline{y}}{(n-1)}-\frac{r_{x,y}}{s_x}\cdot\frac{x_i-\overline{x}}{(n-1)s_x} \\
&= \frac{1}{s_x s_y}\cdot\frac{y_i-\overline{y}}{(n-1)}-\frac{r_{x,y}}{s_x^2}\cdot\frac{x_i-\overline{x}}{(n-1)} \\
&= \frac{1}{(n-1)s_x}\left(\frac{y_i-\overline{y}}{s_y}-\frac{r_{x,y}}{s_x}\frac{x_i-\overline{x}}{s_x}\right).
\end{aligned}$$

□

A.7 Correlation identities

Let $\tilde{x}_i = (x_i - \overline{x})/s_x$, $\tilde{y}_i = (y_i - \overline{y})/s_y$ and $\tilde{z}_i = (z_i - \overline{z})/s_z$ be the standardized variables. The following identities hold:

$$\sum_{i=1}^{n} (\tilde{x}_i - r_{xy}\tilde{y}_i)^2 = (n-1)(1 - r_{xy}^2) \tag{23}$$

$$\sum_{i=1}^{n} (\tilde{y}_i - r_{xy}\tilde{x}_i)(\tilde{z}_i - r_{xz}\tilde{x}_i) = (n-1)(r_{xy}r_{xz} - r_{yz}) \tag{24}$$

Proof. Note that $(n-1)s_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$ and $(n-1)s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ then by using the definitions of standardized variables, we have for the first identity:

$$\begin{aligned}
\sum_{i=1}^{n} (\tilde{x}_i - r_{xy}\tilde{y}_i)^2 &= \sum_{i=1}^{n} \left(\tilde{x}_i^2 - 2r_{xy}\tilde{x}_i\tilde{y}_i + r_{xy}^2\tilde{y}_i^2\right) \\
&= \sum_{i=1}^{n} \left[\frac{(x_i - \overline{x})^2}{s_x^2} - 2r_{xy}\frac{(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} + r_{xy}^2\frac{(y_i - \overline{y})^2}{s_y^2}\right] \\
&= \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{2r_{xy}}{s_x s_y}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \frac{r_{xy}^2}{s_y^2}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= (n-1) - 2r_{xy}(n-1)r_{xy} + r_{xy}^2(n-1) \\
&= (n-1)(1 - 2r_{xy}^2 + r_{xy}^2) \\
&= (n-1)(1 - r_{xy}^2).
\end{aligned}$$

and for the second identity:

$$\begin{aligned}
\sum_{i=1}^{n} (\tilde{y}_i - r_{xy}\tilde{x}_i)(\tilde{z}_i - r_{xz}\tilde{x}_i) &= \sum_{i=1}^{n} \left(\tilde{y}_i\tilde{z}_i - r_{xz}\tilde{y}_i\tilde{x}_i - r_{xy}\tilde{x}_i\tilde{z}_i + r_{xy}r_{xz}\tilde{x}_i^2\right) \\
&= (n-1)r_{yz} - r_{xz}(n-1)r_{xy} - r_{xy}(n-1)r_{xz} + r_{xy}r_{xz}(n-1) \\
&= (n-1)(r_{yz} - 2r_{xy}r_{xz} + r_{xy}r_{xz}) \\
&= (n-1)(r_{xy}r_{xz} - r_{yz}).
\end{aligned}$$

□

B Cross-sectional numerical variability quantification at baseline visit

As a complementary component of our analysis, we quantified cross-sectional numerical variability for cortical and subcortical measures at baseline. Numerical variability was assessed using two metrics: (1) the number of significant digits [35] ($p_s = 0.95$, confidence $1 - \alpha_s = 0.95$), calculated using the `significantdigits` package[1] (version 0.4.0); and (2) the extended Sørensen-Dice coefficient, which measures the spatial overlap of segmentation masks across n repetitions. The extended Dice coefficient is defined as:

$$\mathrm{Dice}(A_1, A_2, \ldots, A_n) = \frac{n\,|\bigcap_{i=1}^{n} A_i|}{\sum_{i=1}^{n} |A_i|}, \tag{25}$$

where A_i represents the segmentation for the i-th Monte Carlo sample.

FreeSurfer 7.3.1 demonstrated limited numerical precision across all anatomical measures. Cortical metrics showed averages of 1.61 ± 0.20 significant digits for thickness, 1.33 ± 0.23 for surface area, and 1.33 ± 0.23 for cortical volume (Figure 7). Subcortical volumes exhibited similar instability, with an average of 1.33 ± 0.22 significant digits (Figure 8). These results indicate that, on average, FreeSurfer measurements are precise to only one decimal place, with specific instances exhibiting complete loss

[1] github.com/verificarlo/significantdigits

of significant digits. While variability was observed across all metrics, cortical thickness displayed the highest relative precision (range: $1.22 - 1.93$ digits), whereas surface area ($0.82 - 1.72$) and cortical volume ($0.80 - 1.72$) were notably less stable. Table 3 and Table 4 provide a detailed breakdown of within-subject significant digits and standard-deviation average across all subjects for each cortical region and metric. Table 5 presents the corresponding data for subcortical regions.

To assess spatial stability, we evaluated the volumetric overlap using the extended Sørensen-Dice coefficient. The analysis revealed substantial inter-subject and regional variability (Figure 6). Cortical regions showed marked instability, with Dice coefficients ranging from 0.00 to 0.91 (mean 0.75 ± 0.11), indicating that for some regions, numerical noise could lead to a complete lack of spatial overlap. Subcortical structures appeared comparatively more stable, ranging from 0.18 to 0.94 (mean 0.82 ± 0.08). Overall, subcortical segmentations demonstrated higher spatial robustness compared to cortical parcellations.

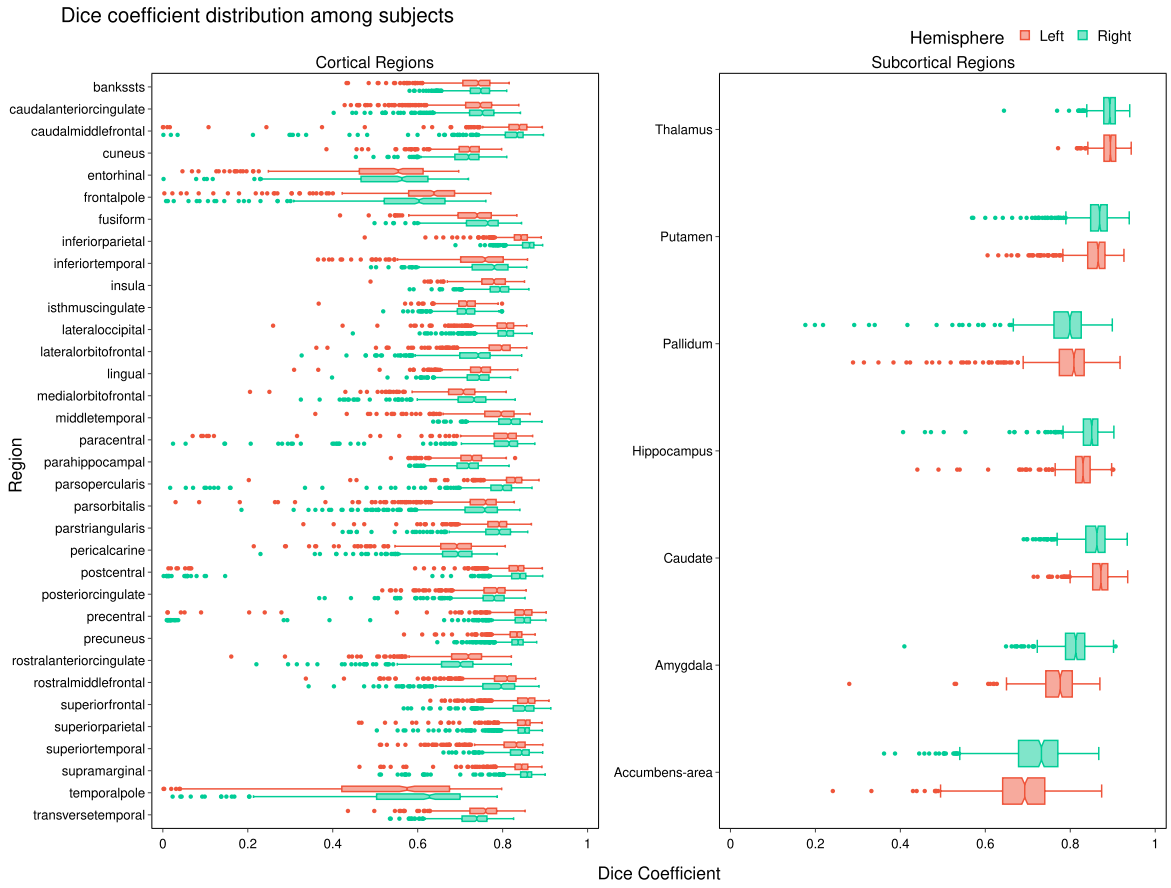


Figure 6: Numerical variability of segmentation. The distribution of extended Sørensen-Dice coefficients, highlighting the variability in spatial overlap for cortical and subcortical segmentations due to numerical variability. Lower values indicate higher instability.

Table 3: Within-subject significant digits averaged across all subjects.

Region	**cortical thickness**		**surface area**		**cortical volume**	
	lh	rh	lh	rh	lh	rh
banksssts	1.65 ± 0.16	1.69 ± 0.13	1.15 ± 0.18	1.21 ± 0.13	1.08 ± 0.17	1.14 ± 0.13
caudalanteriorcingulate	1.38 ± 0.14	1.40 ± 0.14	1.14 ± 0.22	1.19 ± 0.18	1.14 ± 0.24	1.21 ± 0.20
caudalmiddlefrontal	1.77 ± 0.18	1.77 ± 0.19	1.40 ± 0.21	1.31 ± 0.23	1.40 ± 0.22	1.30 ± 0.23
cuneus	1.52 ± 0.19	1.54 ± 0.19	1.34 ± 0.14	1.33 ± 0.14	1.32 ± 0.14	1.28 ± 0.15
entorhinal	1.22 ± 0.23	1.22 ± 0.23	0.82 ± 0.19	0.87 ± 0.18	0.80 ± 0.19	0.81 ± 0.18

Continued on next page

Table 3: Within-subject significant digits averaged across all subjects. (Continued)

Region	**cortical thickness**		**surface area**		**cortical volume**	
	lh	rh	lh	rh	lh	rh
fusiform	1.66 ± 0.17	1.71 ± 0.16	1.41 ± 0.18	1.43 ± 0.19	1.33 ± 0.18	1.37 ± 0.20
inferiorparietal	1.81 ± 0.15	1.82 ± 0.13	1.53 ± 0.18	1.59 ± 0.20	1.50 ± 0.17	1.56 ± 0.17
inferiortemporal	1.66 ± 0.17	1.70 ± 0.16	1.37 ± 0.25	1.38 ± 0.21	1.37 ± 0.23	1.41 ± 0.19
isthmuscingulate	1.46 ± 0.12	1.43 ± 0.13	1.27 ± 0.15	1.24 ± 0.15	1.27 ± 0.14	1.27 ± 0.15
lateraloccipital	1.75 ± 0.18	1.77 ± 0.17	1.58 ± 0.15	1.57 ± 0.16	1.49 ± 0.16	1.50 ± 0.15
lateralorbitofrontal	1.65 ± 0.17	1.51 ± 0.15	1.44 ± 0.23	0.95 ± 0.13	1.51 ± 0.16	1.12 ± 0.14
lingual	1.54 ± 0.22	1.52 ± 0.21	1.47 ± 0.18	1.46 ± 0.17	1.50 ± 0.18	1.49 ± 0.18
medialorbitofrontal	1.50 ± 0.15	1.53 ± 0.15	1.09 ± 0.16	1.15 ± 0.14	1.15 ± 0.17	1.21 ± 0.13
middletemporal	1.74 ± 0.16	1.81 ± 0.14	1.42 ± 0.23	1.52 ± 0.19	1.44 ± 0.21	1.55 ± 0.18
parahippocampal	1.54 ± 0.14	1.56 ± 0.12	1.13 ± 0.13	1.09 ± 0.13	1.11 ± 0.13	1.07 ± 0.13
paracentral	1.59 ± 0.22	1.60 ± 0.22	1.40 ± 0.17	1.40 ± 0.19	1.36 ± 0.18	1.36 ± 0.20
parsopercularis	1.74 ± 0.17	1.71 ± 0.16	1.38 ± 0.19	1.30 ± 0.18	1.38 ± 0.19	1.30 ± 0.20
parsorbitalis	1.53 ± 0.20	1.51 ± 0.20	1.21 ± 0.14	1.21 ± 0.18	1.19 ± 0.16	1.22 ± 0.18
parstriangularis	1.68 ± 0.17	1.63 ± 0.19	1.33 ± 0.16	1.30 ± 0.22	1.30 ± 0.16	1.28 ± 0.21
pericalcarine	1.33 ± 0.21	1.30 ± 0.22	1.23 ± 0.20	1.21 ± 0.22	1.18 ± 0.17	1.18 ± 0.17
postcentral	1.84 ± 0.24	1.81 ± 0.26	1.68 ± 0.23	1.69 ± 0.28	1.64 ± 0.20	1.63 ± 0.24
posteriorcingulate	1.57 ± 0.13	1.56 ± 0.14	1.37 ± 0.20	1.35 ± 0.21	1.39 ± 0.19	1.39 ± 0.22
precentral	1.79 ± 0.26	1.76 ± 0.28	1.71 ± 0.24	1.64 ± 0.27	1.72 ± 0.22	1.66 ± 0.28
precuneus	1.83 ± 0.13	1.84 ± 0.13	1.65 ± 0.21	1.66 ± 0.21	1.61 ± 0.18	1.62 ± 0.19
rostralanteriorcingulate	1.34 ± 0.14	1.39 ± 0.15	1.00 ± 0.16	1.07 ± 0.17	1.11 ± 0.19	1.11 ± 0.18
rostralmiddlefrontal	1.77 ± 0.19	1.74 ± 0.19	1.44 ± 0.24	1.41 ± 0.28	1.49 ± 0.21	1.48 ± 0.25
superiorfrontal	1.87 ± 0.17	1.85 ± 0.18	1.61 ± 0.23	1.56 ± 0.27	1.64 ± 0.21	1.62 ± 0.25
superiorparietal	1.92 ± 0.18	1.93 ± 0.17	1.72 ± 0.24	1.65 ± 0.28	1.66 ± 0.22	1.60 ± 0.26
superiortemporal	1.83 ± 0.17	1.85 ± 0.15	1.57 ± 0.22	1.58 ± 0.18	1.52 ± 0.21	1.57 ± 0.18
supramarginal	1.83 ± 0.16	1.85 ± 0.15	1.57 ± 0.22	1.59 ± 0.26	1.56 ± 0.20	1.56 ± 0.24
frontalpole	1.26 ± 0.23	1.23 ± 0.20	0.94 ± 0.11	0.91 ± 0.11	0.88 ± 0.17	0.87 ± 0.14
temporalpole	1.24 ± 0.26	1.28 ± 0.25	0.94 ± 0.16	0.99 ± 0.19	0.86 ± 0.20	0.91 ± 0.22
transversetemporal	1.47 ± 0.20	1.46 ± 0.18	1.17 ± 0.13	1.13 ± 0.11	1.20 ± 0.15	1.15 ± 0.13
insula	1.47 ± 0.16	1.42 ± 0.14	1.13 ± 0.18	1.00 ± 0.18	1.29 ± 0.16	1.19 ± 0.19

Table 4: Within-subject standard-deviation average across all subjects for cortical metrics.

Region	**cortical thickness (mm)**		**surface area (mm^2)**		**cortical volume (mm^3)**	
	lh	rh	lh	rh	lh	rh
bankssts	0.02 ± 0.01	0.02 ± 0.01	28.65 ± 15.97	21.73 ± 8.68	77.25 ± 37.44	59.87 ± 20.45
caudalanteriorcingulate	0.04 ± 0.01	0.04 ± 0.01	19.98 ± 13.83	21.01 ± 14.96	51.33 ± 37.32	51.67 ± 41.74
caudalmiddlefrontal	0.02 ± 0.01	0.02 ± 0.01	38.58 ± 36.77	46.65 ± 44.68	104.41 ± 108.02	124.11 ± 112.10
cuneus	0.02 ± 0.01	0.02 ± 0.01	28.45 ± 11.50	31.25 ± 15.67	60.72 ± 25.52	74.77 ± 34.16
entorhinal	0.08 ± 0.05	0.08 ± 0.05	27.41 ± 16.67	22.37 ± 11.70	125.48 ± 71.07	115.94 ± 57.21
fusiform	0.02 ± 0.01	0.02 ± 0.01	50.70 ± 25.16	47.86 ± 28.19	182.92 ± 92.31	170.22 ± 103.05
inferiorparietal	0.01 ± 0.01	0.01 ± 0.01	53.01 ± 29.19	59.90 ± 50.62	145.66 ± 72.95	159.55 ± 110.14
inferiortemporal	0.02 ± 0.01	0.02 ± 0.01	64.73 ± 42.27	58.75 ± 34.04	198.15 ± 127.44	168.38 ± 84.67
isthmuscingulate	0.03 ± 0.01	0.03 ± 0.01	23.74 ± 11.07	23.35 ± 13.99	57.43 ± 29.59	53.05 ± 34.34
lateraloccipital	0.02 ± 0.01	0.02 ± 0.01	53.82 ± 24.63	56.35 ± 28.61	156.83 ± 66.16	160.98 ± 76.00
lateralorbitofrontal	0.02 ± 0.01	0.03 ± 0.01	43.31 ± 30.16	117.14 ± 33.75	92.60 ± 56.29	217.89 ± 69.06
lingual	0.03 ± 0.01	0.03 ± 0.01	44.26 ± 22.65	46.73 ± 23.96	89.19 ± 46.24	95.82 ± 49.65
medialorbitofrontal	0.03 ± 0.01	0.03 ± 0.01	66.04 ± 24.11	58.06 ± 19.00	147.37 ± 57.84	134.52 ± 42.26
middletemporal	0.02 ± 0.01	0.02 ± 0.01	53.01 ± 34.97	44.87 ± 28.36	165.49 ± 108.52	135.26 ± 77.98
parahippocampal	0.03 ± 0.01	0.03 ± 0.01	19.55 ± 8.42	20.45 ± 7.81	64.22 ± 25.29	65.43 ± 24.59

Continued on next page

Table 4: Within-subject standard-deviation average across all subjects for cortical metrics. (Continued)

Region	cortical thickness (mm)		surface area (mm^2)		cortical volume (mm^3)	
	lh	rh	lh	rh	lh	rh
paracentral	0.03 ± 0.02	0.03 ± 0.01	22.94 ± 12.98	26.94 ± 19.80	63.71 ± 40.74	73.88 ± 56.66
parsopercularis	0.02 ± 0.01	0.02 ± 0.01	28.65 ± 28.77	29.46 ± 26.82	80.67 ± 92.87	82.38 ± 89.16
parsorbitalis	0.03 ± 0.02	0.03 ± 0.02	17.82 ± 9.77	21.41 ± 10.66	60.63 ± 45.20	68.18 ± 36.64
parstriangularis	0.02 ± 0.01	0.02 ± 0.01	25.67 ± 14.65	34.86 ± 37.79	71.73 ± 45.49	96.87 ± 102.22
pericalcarine	0.03 ± 0.02	0.04 ± 0.02	36.04 ± 20.18	42.02 ± 24.82	59.64 ± 29.98	68.61 ± 34.89
postcentral	0.01 ± 0.02	0.02 ± 0.02	43.47 ± 67.12	45.98 ± 83.10	100.26 ± 121.35	104.53 ± 156.51
posteriorcingulate	0.02 ± 0.01	0.02 ± 0.01	21.93 ± 13.05	24.39 ± 19.52	52.42 ± 33.33	56.27 ± 52.59
precentral	0.02 ± 0.02	0.02 ± 0.02	46.92 ± 53.54	57.46 ± 70.35	118.04 ± 157.21	148.21 ± 233.10
precuneus	0.01 ± 0.01	0.01 ± 0.00	38.04 ± 42.87	38.95 ± 40.96	100.91 ± 111.15	102.24 ± 96.62
rostralanteriorcingulate	0.05 ± 0.02	0.04 ± 0.02	34.80 ± 15.03	22.00 ± 10.59	81.04 ± 41.59	61.95 ± 33.93
rostralmiddlefrontal	0.02 ± 0.01	0.02 ± 0.01	92.87 ± 96.23	108.40 ± 132.97	213.81 ± 259.58	252.00 ± 358.20
superiorfrontal	0.01 ± 0.01	0.01 ± 0.01	85.23 ± 86.47	98.14 ± 120.75	223.91 ± 234.89	243.75 ± 304.56
superiorparietal	0.01 ± 0.01	0.01 ± 0.01	49.49 ± 80.81	62.89 ± 96.86	132.77 ± 207.97	161.39 ± 235.01
superiortemporal	0.02 ± 0.01	0.01 ± 0.01	47.70 ± 33.64	41.38 ± 23.84	156.30 ± 101.85	129.01 ± 78.70
supramarginal	0.01 ± 0.01	0.01 ± 0.01	50.87 ± 58.82	50.06 ± 83.24	136.23 ± 168.28	133.99 ± 207.69
frontalpole	0.07 ± 0.04	0.07 ± 0.04	12.99 ± 4.02	16.42 ± 4.47	56.49 ± 32.17	67.84 ± 28.93
temporalpole	0.09 ± 0.05	0.08 ± 0.05	25.08 ± 10.71	22.16 ± 11.78	154.60 ± 79.32	138.28 ± 78.33
transversetemporal	0.03 ± 0.02	0.03 ± 0.02	12.73 ± 5.33	9.98 ± 3.33	29.55 ± 12.34	24.91 ± 8.79
insula	0.04 ± 0.02	0.04 ± 0.01	73.45 ± 30.66	95.70 ± 37.63	146.49 ± 64.11	183.39 ± 81.47

Table 5: Within-subject significant digits averaged across all subjects for subcortical volumes.

Region	Significant digits	Standard deviation (mm^3)
Left-Thalamus	1.42 ± 0.21	120.08 ± 69.61
Left-Caudate	1.57 ± 0.20	38.83 ± 25.11
Left-Putamen	1.49 ± 0.22	65.88 ± 46.39
Left-Pallidum	1.25 ± 0.19	47.81 ± 25.09
Left-Hippocampus	1.48 ± 0.17	56.23 ± 41.03
Left-Amygdala	1.13 ± 0.16	48.71 ± 20.04
Left-Accumbens-area	0.88 ± 0.16	24.20 ± 8.80
Right-Thalamus	1.42 ± 0.20	118.92 ± 68.76
Right-Caudate	1.51 ± 0.24	49.37 ± 42.71
Right-Putamen	1.51 ± 0.25	68.07 ± 70.23
Right-Pallidum	1.22 ± 0.19	49.11 ± 30.50
Right-Hippocampus	1.55 ± 0.18	48.59 ± 28.98
Right-Amygdala	1.23 ± 0.17	42.21 ± 18.68
Right-Accumbens-area	0.99 ± 0.15	20.50 ± 7.72

C Numerical-Population Variability Ratio (ν_{npv})

C.1 ν_{npv} maps

Figure 9 shows the ν_{npv} maps for cortical surface area and volume, for each region. The color scale indicates the ν_{npv} value, with warmer colors indicating higher ν_{npv} values. The maps provide a visual representation of the variability in the ν_{npv} values across different cortical regions, highlighting regions with higher or lower ν_{npv} values.

Table 6: Summary of executions failure and excluded subjects. To standardize the sample, we keep 26 repetitions per subject/visits pair. Subject/visit pairs with less than 26 repetitions were excluded which is 12 subjects.

Stage	**Number of rejected repetitions**	**Total number of repetitions**
Cluster failure	1246 (5.80%)	21488
FreeSurfer failure	68 (0.33%)	21488
QC failure	319 (1.48%)	21488
Total	1633 (7.60%)	21488

C.2 Consistency results

We evaluated the stability of statistical inferences by quantifying the frequency of significance flipping across 26 MCA repetitions (Table 7). The analysis revealed that longitudinal processing introduces considerably more instability than baseline cross-sectional analysis, particularly for cortical metrics. Notably, partial correlation analysis exhibited higher susceptibility to instability in cortical volume (53% of regions unstable) and area (38% unstable) compared to ANCOVA.

To investigate whether longitudinal processing amplifies numerical instability, we compared the dispersion of the test statistics (F-values for ANCOVA, correlation coefficients r for partial correlation) in longitudinal versus cross-sectional pipelines. We used a one-sided Ansari-Bradley test on mean-centered distributions to test the specific hypothesis that longitudinal processing has significantly higher numerical variance compared to the baseline (Tables 8 and 9).

For subcortical structures (Table 8), partial correlation yielded significantly higher variance in the longitudinal setting for nearly all regions (13/14), whereas ANCOVA remained largely stable, with only the caudate showing significant variance inflation. This trend was mirrored in cortical regions (Table 9); partial correlation analysis consistently demonstrated significantly greater variance longitudinally across volume, thickness, and surface area. In contrast, ANCOVA showed significant variance differences in far fewer regions.

Table 7: Variance variability in subcortical structures (Ansari-Bradley test). Comparison of cross-sectional vs. longitudinal variances of test statistics. Significant results (*) indicate that longitudinal variance is significantly greater than baseline variance ($p_{\text{FDR}} < 0.05$, one-sided). W: test statistic.

Metric	**ANCOVA**		**Partial correlation**	
	Baseline	**Longitudinal**	**Baseline**	**Longitudinal**
Cortical Area (68 regions)	18 (27%)	36 (53%)	4 (6%)	26 (38%)
Cortical Thickness (68 regions)	11 (16%)	9 (13%)	19 (28%)	17 (25%)
Cortical Volume (68 regions)	14 (20%)	20 (29%)	8 (12%)	36 (53%)
Subcortical Volume (14 regions)	4 (29%)	2 (14%)	4 (29%)	5 (36%)

Table 9: Ansari-Bradley test results for cross-sectional vs. longitudinal variances in cortical regions. * indicates FDR-corrected significance ($p < 0.05$). lh/rh = left/right hemisphere. ACC = anterior cingulate cortex, MF = middle frontal. W = statistic, p = p-value.

Region	Cortical Volume				Cortical Thickness				Cortical Area			
	ANCOVA		Partial Corr		ANCOVA		Partial Corr		ANCOVA		Partial Corr	
	W	p	W	p	W	p	W	p	W	p	W	p
bankssts (lh)	447	1.6e-4*	486	4.1e-8*	454	4.8e-5*	433	1.2e-3*	455	4.1e-5*	481	1.6e-7*
bankssts (rh)	403	2.9e-2	466	5.0e-6*	469	2.7e-6*	449	1.1e-4*	447	1.6e-4*	461	1.3e-5*
caudalACC (lh)	446	1.8e-4*	487	3.1e-8*	322	0.86	475	6.9e-7*	350	0.52	480	2.0e-7*
caudalACC (rh)	327	0.81	513	1.1e-12*	501	3.1e-10*	453	5.8e-5*	304	0.96	500	4.6e-10*
caudalMF (lh)	437	6.8e-4*	420	5.6e-3*	479	2.6e-7*	413	1.2e-2	362	0.35	461	1.3e-5*
caudalMF (rh)	439	5.2e-4*	479	2.6e-7*	352	0.49	445	2.1e-4*	390	8.0e-2	489	1.7e-8*

Continued on next page

Table 8: =

Variance variability in cortical regions. One-sided Ansari-Bradley test comparing cross-sectional vs. longitudinal variances of test statistics. ($*p_{\text{FDR}} < 0.05$). L/R = Left/Right. W = statistic, p = p-value.

Region	**ANCOVA**		**Partial Corr**	
	W	**p**	**W**	**p**
L-Thalamus	248	1.00	465	6.1e-6*
L-Caudate	501	3.1e-10*	459	2.0e-5*
L-Putamen	327	0.81	491	9.6e-9*
L-Pallidum	264	1.00	449	1.1e-4*
L-Hippocampus	314	0.91	428	2.2e-3*
L-Amygdala	261	1.00	476	5.5e-7*
L-Accumbens	265	1.00	442	3.3e-4*
R-Thalamus	281	1.00	441	3.9e-4*
R-Caudate	485	5.5e-8*	478	3.4e-7*
R-Putamen	294	0.98	489	1.7e-8*
R-Pallidum	212	1.00	469	2.7e-6*
R-Hippocampus	335	0.73	493	5.1e-9*
R-Amygdala	316	0.90	408	1.9e-2*
R-Accumbens	214	1.00	418	7.0e-3*

Table 9: Ansari-Bradley test results for cross-sectional vs. longitudinal variances in cortical regions. * indicates FDR-corrected significance ($p < 0.05$). lh/rh = left/right hemisphere. ACC = anterior cingulate cortex, MF = middle frontal. W = statistic, p = p-value. (Continued)

Region	Cortical Volume				Cortical Thickness				Cortical Area			
	ANCOVA		Partial Corr		ANCOVA		Partial Corr		ANCOVA		Partial Corr	
	W	p	W	p	W	p	W	p	W	p	W	p
cuneus (lh)	441	3.9e-4*	471	1.7e-6*	507	2.6e-11*	454	4.8e-5*	486	4.1e-8*	483	9.4e-8*
cuneus (rh)	387	9.8e-2	481	1.6e-7*	474	8.7e-7*	461	1.3e-5*	497	1.4e-9*	487	3.1e-8*
entorhinal (lh)	359	0.39	409	1.7e-2	297	0.98	396	5.2e-2	474	8.7e-7*	434	1.0e-3*
entorhinal (rh)	372	0.23	428	2.2e-3*	246	1.00	396	5.2e-2	505	6.2e-11*	455	4.1e-5*
fusiform (lh)	477	4.3e-7*	447	1.6e-4*	421	5.0e-3*	419	6.3e-3*	421	5.0e-3*	455	4.1e-5*
fusiform (rh)	457	2.8e-5*	460	1.6e-5*	423	4.0e-3*	437	6.8e-4*	475	6.9e-7*	454	4.8e-5*
inferiorparietal (lh)	477	4.3e-7*	503	1.4e-10*	347	0.56	458	2.4e-5*	487	3.1e-8*	471	1.7e-6*
inferiorparietal (rh)	369	0.26	427	2.5e-3*	460	1.6e-5*	383	0.13	409	1.7e-2	455	4.1e-5*
inferiortemporal (lh)	437	6.8e-4*	444	2.5e-4*	356	0.44	461	1.3e-5*	472	1.4e-6*	444	2.5e-4*
inferiortemporal (rh)	335	0.73	406	2.3e-2	326	0.82	457	2.8e-5*	375	0.20	456	3.4e-5*
isthmuscingulate (lh)	432	1.3e-3*	504	9.5e-11*	515	3.1e-13*	462	1.1e-5*	413	1.2e-2	505	6.2e-11*
isthmuscingulate (rh)	494	3.7e-9*	469	2.7e-6*	489	1.7e-8*	455	4.1e-5*	513	1.1e-12*	474	8.7e-7*
lateraloccipital (lh)	494	3.7e-9*	470	2.1e-6*	430	1.7e-3*	440	4.5e-4*	475	6.9e-7*	460	1.6e-5*
lateraloccipital (rh)	472	1.4e-6*	454	4.8e-5*	498	9.5e-10*	412	1.3e-2	471	1.7e-6*	461	1.3e-5*
lateralorbitofrontal (lh)	265	1.00	414	1.1e-2	184	1.00	464	7.5e-6*	294	0.98	440	4.5e-4*
lateralorbitofrontal (rh)	271	1.00	478	3.4e-7*	422	4.5e-3*	429	2.0e-3*	258	1.00	468	3.3e-6*
lingual (lh)	250	1.00	463	9.1e-6*	494	3.7e-9*	468	3.3e-6*	453	5.8e-5*	497	1.4e-9*
lingual (rh)	435	9.0e-4*	460	1.6e-5*	480	2.0e-7*	467	4.1e-6*	477	4.3e-7*	459	2.0e-5*
medialorbitofrontal (lh)	459	2.0e-5*	467	4.1e-6*	479	2.6e-7*	410	1.6e-2	396	5.2e-2	469	2.7e-6*
medialorbitofrontal (rh)	395	5.6e-2	484	7.2e-8*	300	0.97	430	1.7e-3*	405	2.5e-2	459	2.0e-5*
middletemporal (lh)	488	2.3e-8*	487	3.1e-8*	252	1.00	419	6.3e-3*	457	2.8e-5*	464	7.5e-6*
middletemporal (rh)	352	0.49	463	9.1e-6*	396	5.2e-2	428	2.2e-3*	427	2.5e-3*	450	9.5e-5*
parahippocampal (lh)	502	2.1e-10*	476	5.5e-7*	462	1.1e-5*	437	6.8e-4*	425	3.2e-3*	463	9.1e-6*
parahippocampal (rh)	473	1.1e-6*	468	3.3e-6*	349	0.54	458	2.4e-5*	415	9.6e-3*	459	2.0e-5*
paracentral (lh)	314	0.91	424	3.6e-3*	204	1.00	425	3.2e-3*	491	9.6e-9*	438	6.0e-4*
paracentral (rh)	331	0.77	484	7.2e-8*	280	1.00	441	3.9e-4*	330	0.78	485	5.5e-8*
parsopercularis (lh)	362	0.35	461	1.3e-5*	494	3.7e-9*	418	7.0e-3*	459	2.0e-5*	472	1.4e-6*

Continued on next page

Table 9: Ansari-Bradley test results for cross-sectional vs. longitudinal variances in cortical regions. * indicates FDR-corrected significance ($p < 0.05$). lh/rh = left/right hemisphere. ACC = anterior cingulate cortex, MF = middle frontal. W = statistic, p = p-value. (Continued)

Region	Cortical Volume				Cortical Thickness				Cortical Area			
	ANCOVA		Partial Corr		ANCOVA		Partial Corr		ANCOVA		Partial Corr	
	W	p	W	p	W	p	W	p	W	p	W	p
parsopercularis (rh)	388	9.1e-2	482	1.2e-7*	305	0.96	477	4.3e-7*	447	1.6e-4*	466	5.0e-6*
parsorbitalis (lh)	293	0.98	490	1.3e-8*	230	1.00	462	1.1e-5*	419	6.3e-3*	462	1.1e-5*
parsorbitalis (rh)	295	0.98	455	4.1e-5*	336	0.71	427	2.5e-3*	269	1.00	471	1.7e-6*
parstriangularis (lh)	339	0.68	507	2.6e-11*	401	3.5e-2	419	6.3e-3*	419	6.3e-3*	504	9.5e-11*
parstriangularis (rh)	404	2.7e-2	455	4.1e-5*	288	0.99	394	6.0e-2	362	0.35	463	9.1e-6*
pericalcarine (lh)	442	3.3e-4*	465	6.1e-6*	441	3.9e-4*	470	2.1e-6*	460	1.6e-5*	509	9.9e-12*
pericalcarine (rh)	390	8.0e-2	449	1.1e-4*	480	2.0e-7*	451	8.1e-5*	443	2.9e-4*	492	7.0e-9*
postcentral (lh)	329	0.79	460	1.6e-5*	331	0.77	423	4.0e-3*	372	0.23	477	4.3e-7*
postcentral (rh)	350	0.52	449	1.1e-4*	269	1.00	426	2.8e-3*	514	6.1e-13*	387	9.8e-2
posteriorcingulate (lh)	317	0.90	488	2.3e-8*	487	3.1e-8*	436	7.9e-4*	328	0.80	465	6.1e-6*
posteriorcingulate (rh)	315	0.91	478	3.4e-7*	480	2.0e-7*	450	9.5e-5*	356	0.44	485	5.5e-8*
precentral (lh)	284	0.99	457	2.8e-5*	239	1.00	372	0.23	444	2.5e-4*	424	3.6e-3*
precentral (rh)	325	0.83	420	5.6e-3*	312	0.93	387	9.8e-2	470	2.1e-6*	488	2.3e-8*
precuneus (lh)	240	1.00	408	1.9e-2	251	1.00	483	9.4e-8*	410	1.6e-2	430	1.7e-3*
precuneus (rh)	288	0.99	414	1.1e-2	244	1.00	465	6.1e-6*	363	0.34	457	2.8e-5*
rostralACC (lh)	269	1.00	436	7.9e-4*	505	6.2e-11*	467	4.1e-6*	353	0.48	460	1.6e-5*
rostralACC (rh)	275	1.00	501	3.1e-10*	481	1.6e-7*	448	1.3e-4*	275	1.00	448	1.3e-4*
rostralmiddlefrontal (lh)	370	0.25	482	1.2e-7*	276	1.00	430	1.7e-3*	283	0.99	489	1.7e-8*
rostralmiddlefrontal (rh)	318	0.89	418	7.0e-3*	319	0.88	431	1.5e-3*	219	1.00	442	3.3e-4*
superiorfrontal (lh)	370	0.25	443	2.9e-4*	456	3.4e-5*	464	7.5e-6*	481	1.6e-7*	435	9.0e-4*
superiorfrontal (rh)	392	6.9e-2	432	1.3e-3*	271	1.00	467	4.1e-6*	376	0.19	480	2.0e-7*
superiorparietal (lh)	369	0.26	333	0.75	369	0.26	440	4.5e-4*	489	1.7e-8*	386	0.10
superiorparietal (rh)	506	4.0e-11*	430	1.7e-3*	437	6.8e-4*	463	9.1e-6*	456	3.4e-5*	430	1.7e-3*
superiortemporal (lh)	413	1.2e-2	484	7.2e-8*	427	2.5e-3*	441	3.9e-4*	446	1.8e-4*	481	1.6e-7*
superiortemporal (rh)	401	3.5e-2	452	6.8e-5*	364	0.32	469	2.7e-6*	425	3.2e-3*	470	2.1e-6*
supramarginal (lh)	466	5.0e-6*	466	5.0e-6*	274	1.00	488	2.3e-8*	509	9.9e-12*	481	1.6e-7*
supramarginal (rh)	394	6.0e-2	422	4.5e-3*	483	9.4e-8*	412	1.3e-2	376	0.19	432	1.3e-3*
frontalpole (lh)	385	0.11	445	2.1e-4*	319	0.88	453	5.8e-5*	347	0.56	428	2.2e-3*
frontalpole (rh)	284	0.99	410	1.6e-2	391	7.5e-2	451	8.1e-5*	218	1.00	399	4.1e-2
temporalpole (lh)	356	0.44	379	0.16	308	0.94	384	0.12	385	0.11	436	7.9e-4*
temporalpole (rh)	252	1.00	344	0.61	421	5.0e-3*	403	2.9e-2	225	1.00	396	5.2e-2
transversetemporal (lh)	406	2.3e-2	467	4.1e-6*	367	0.29	421	5.0e-3*	445	2.1e-4*	482	1.2e-7*
transversetemporal (rh)	324	0.84	417	7.8e-3*	477	4.3e-7*	442	3.3e-4*	347	0.56	435	9.0e-4*
insula (lh)	249	1.00	470	2.1e-6*	220	1.00	457	2.8e-5*	263	1.00	438	6.0e-4*
insula (rh)	267	1.00	444	2.5e-4*	401	3.5e-2	439	5.2e-4*	235	1.00	410	1.6e-2

C.2.1 Consistency of statistical tests

For both group comparisons and correlation analyses, statistical outcomes varied substantially across the 26 Monte Carlo Arithmetic (MCA) repetitions (Figures 10 and 11). For cortical area (68 regions; Figure 10), significance flipped in 31% of all (regions, analysis) pairs, indicating frequent inconsistencies across repetitions. For cortical volume (68 regions; Figure 11), 29% of pairs were similarly unstable. These fluctuations demonstrate that numerical noise alone can alter downstream statistical inference in structural MRI analyses of PD.

Table 10: Permutation test comparison of ν_{npv} between HC and PD. No significance detected after Bonferroni correction ($\alpha = 0.05/8$). Values are observed differences (PD−HC). 95% CIs from percentile bootstrap.

Metric	Observed Δ	95% CI	p-value	Study
area	0.005	[-0.002, 0.013]	0.734	cross-sectional
thickness	0.027	[0.020, 0.034]	0.033	cross-sectional
volume	0.013	[0.004, 0.022]	0.371	cross-sectional
subcortical volume	0.014	[0.004, 0.025]	0.646	cross-sectional
area	0.021	[-0.002, 0.044]	0.482	longitudinal
thickness	0.002	[-0.014, 0.020]	0.919	longitudinal
volume	0.025	[0.002, 0.049]	0.370	longitudinal
subcortical volume	-0.027	[-0.047,-0.006]	0.489	longitudinal

C.3 Interactive web tool

D Distribution of statistical test coefficients

Figures 13, 15 and 16 show the distribution of partial correlation coefficients and F-statistics for subcortical volume measures, cortical thickness, cortical areas and cortical volumes, across all subjects and regions. Red triangles indicate the unperturbed (IEEE-754) run for reference. For all analyses, the unperturbed (IEEE-754) result is included in the range of numerically perturbed results, which supports the validity of the numerical perturbation.

E Numerical validation of flip significance simulations

To assess the validity of the analytical uncertainty framework used to model numerically induced significance instability, we compared simulated estimates derived from closed-form uncertainty expressions with empirical estimates obtained from Monte Carlo Arithmetic (MCA) repetitions. Formally, the proportion of significant tests (positive prediction rate, PPR) is defined as:

$$\text{PPR} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{26}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

Figure 18 compares standard deviations of p-values estimated directly from 26 MCA runs with those predicted by the analytical uncertainty formulas (Table 1). Each point corresponds to a cortical or subcortical brain region. Results show close agreement between sampled and simulated estimates, particularly in the cross-sectional setting. In the longitudinal analyses, simulated standard deviations tend to be slightly larger than sampled values, indicating a conservative bias of the analytical model.

Figure 19 evaluates how this analytical uncertainty propagates to downstream inference by comparing the simulated positive prediction rate (PPR; Eq. 26) with the empirical proportion of significant tests observed across MCA repetitions. Cross-sectional results closely follow the identity line, while longitudinal results show a moderate overestimation of PPR, consistent with the conservative overestimation of p-value variability observed in Fig. 18.

Finally, Figure 20 assesses the Beta distribution model used to translate sampled p-value variability into PPR estimates. Here, PPR is computed using MCA-derived standard deviations propagated through the Beta model (Eq. 4.4). Both cross-sectional and longitudinal results align closely with the identity line, supporting the suitability of the Beta modeling framework for capturing numerically induced variability in statistical significance.

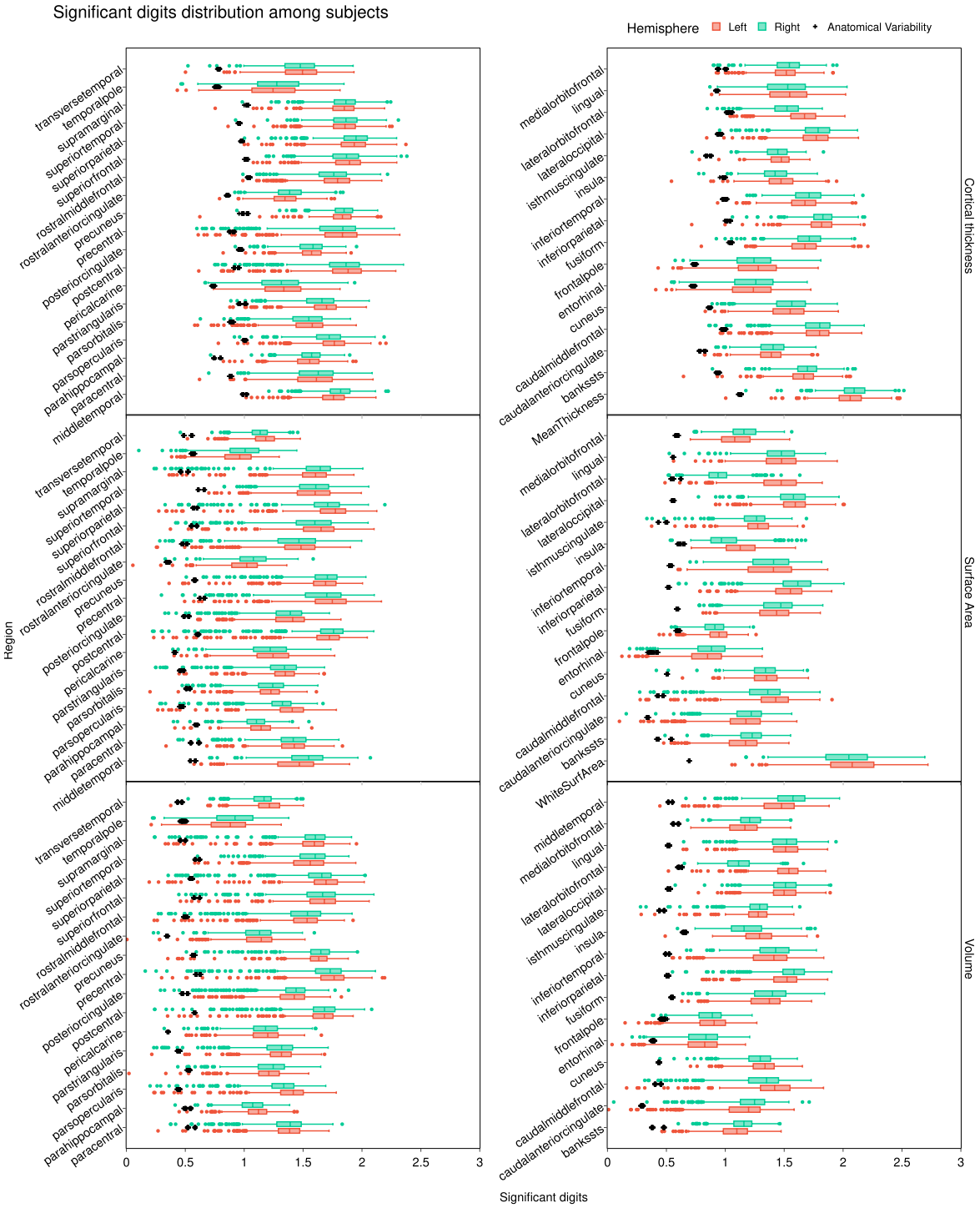


Figure 7: Numerical precision of cortical measures. The number of significant digits for cortical thickness, surface area, and volume across regions. Lower values indicate higher instability. Significant digits for population variability (black cross) remain consistently lower than numerical variability on average, indicating that numerical variability does not dominate population variability.

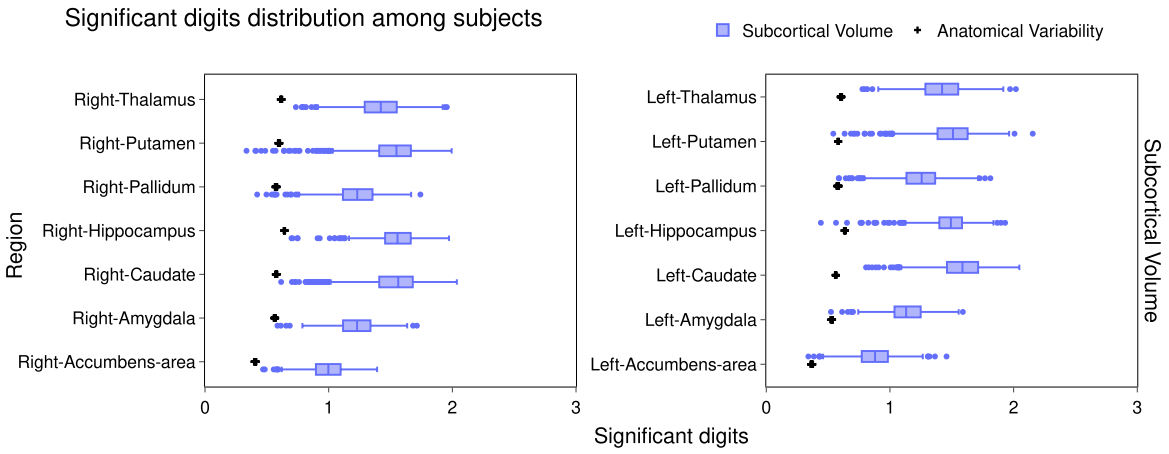


Figure 8: Numerical precision of subcortical volumes. The number of significant digits estimated for each subcortical region, showing comparable instability to cortical volume measures. Significant digits for population variability (black cross) remain consistently lower than numerical variability on average, indicating that numerical variability does not dominate population variability.

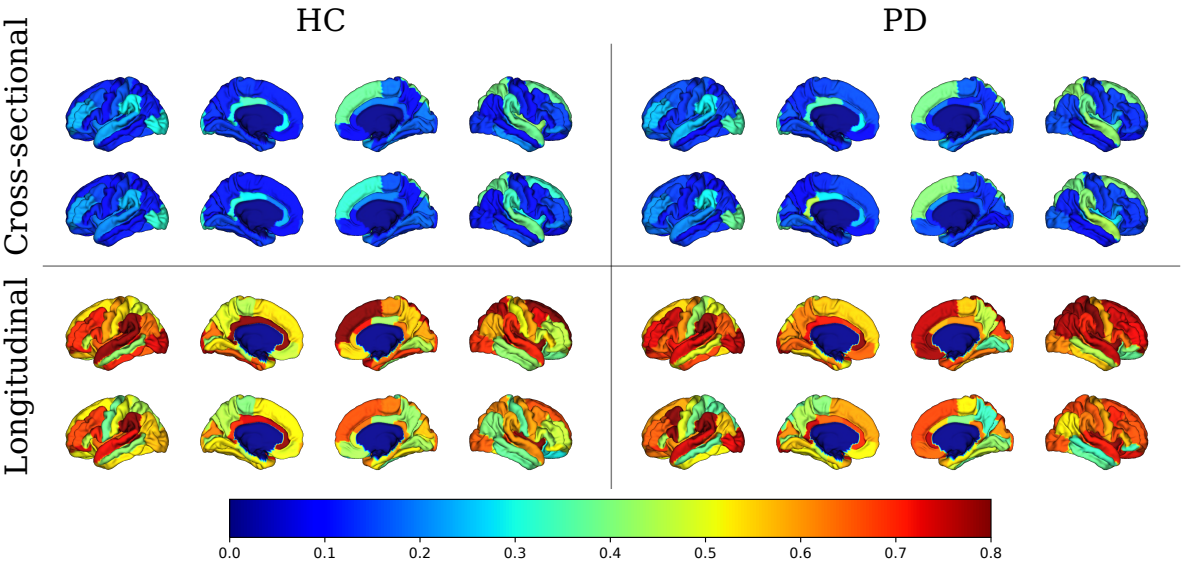


Figure 9: Numerical-Population Variability Ratio (ν_{npv}) for cortical surface area (top row in each panel) area and cortical volume (bottom row) in HC and PD. Panels show ν_{npv} maps at baseline for PD and HC, longitudinally for HC and PD. Higher ν_{npv} values indicate greater computational uncertainty relative to inter-subject anatomical variability. Warmer colors denote higher ν_{npv} values.

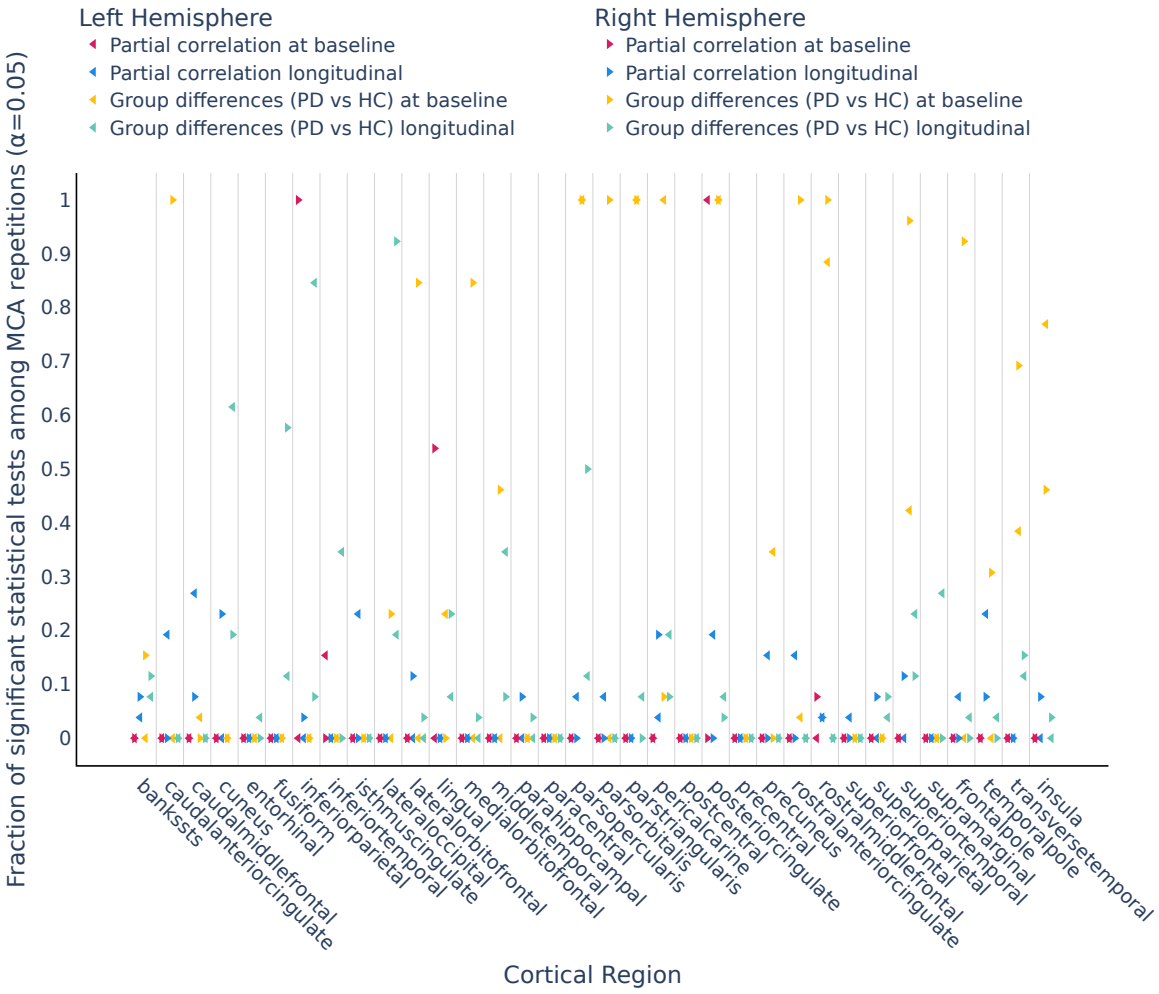


Figure 10: Consistency of statistical tests for cortical area across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency than others.

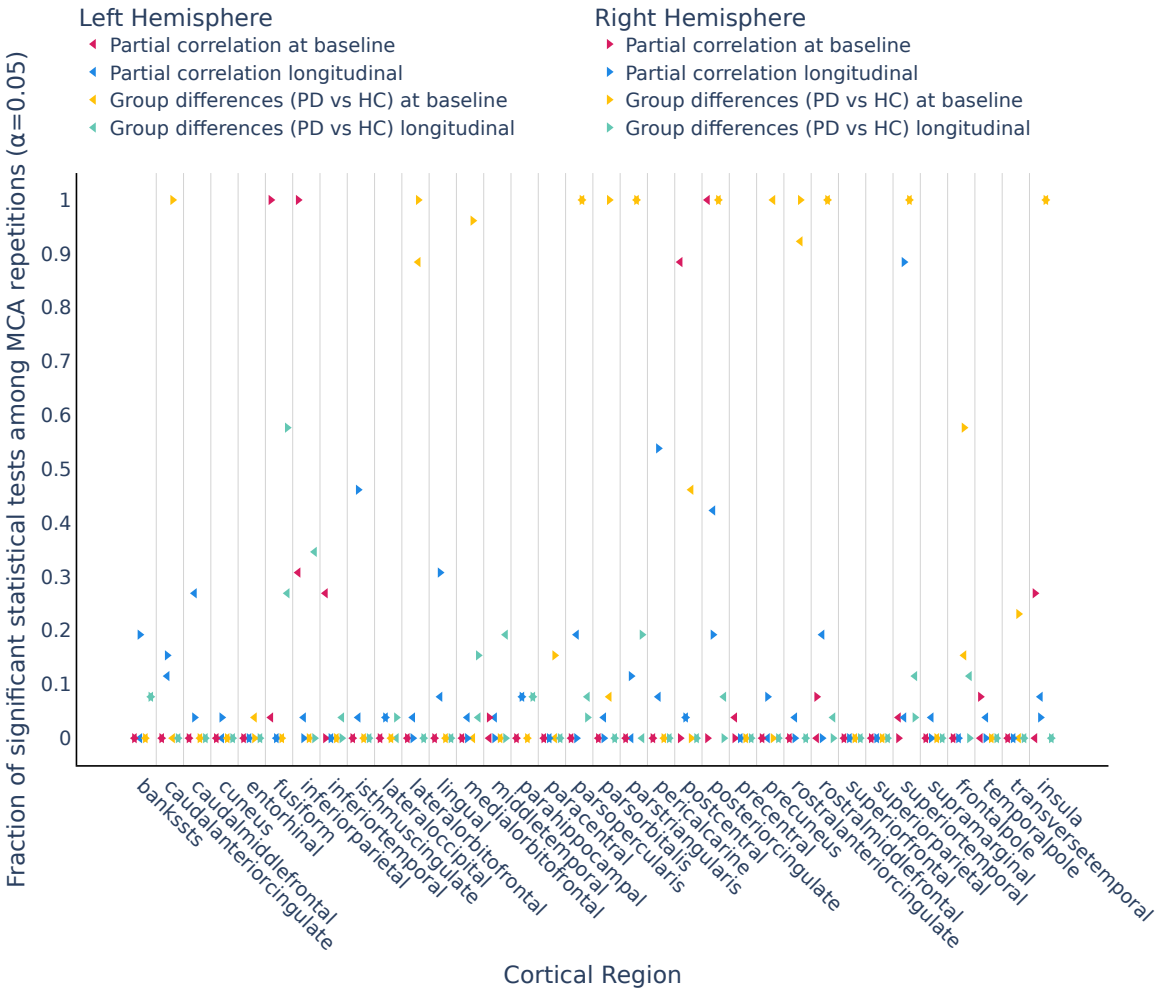


Figure 11: Consistency of statistical tests for cortical volume across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency than others.

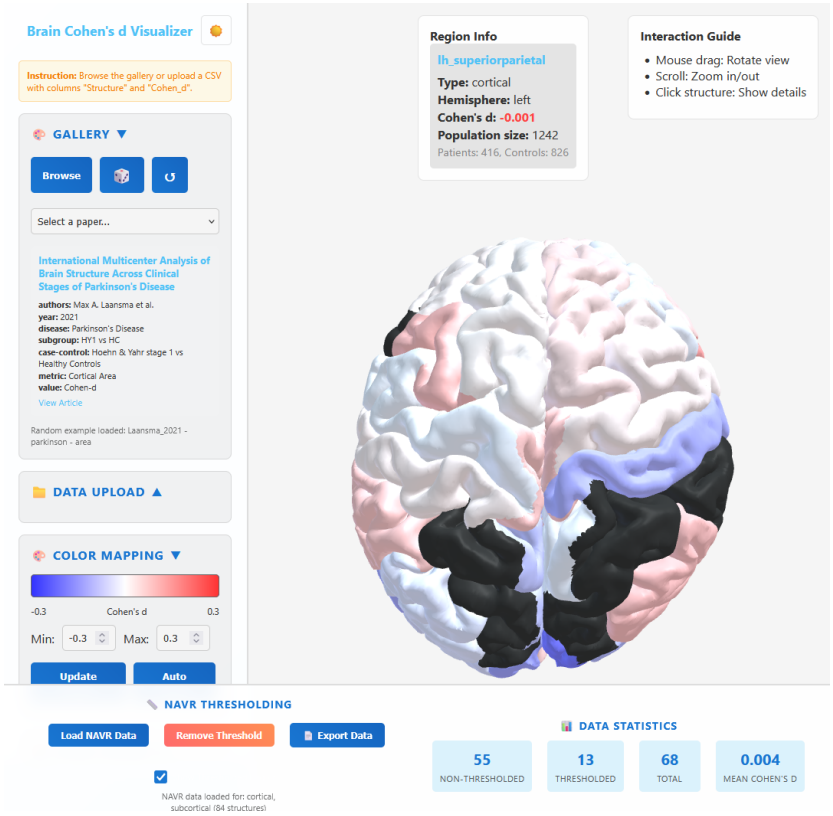


Figure 12: Interactive web tool for estimating NPVR and assessing numerical variability in neuroimaging studies. Users can input summary statistics to obtain NPVR values and visualize the impact of numerical variability on effect size estimates. The tool is available at yohanchatelain.github.io/brain_render.

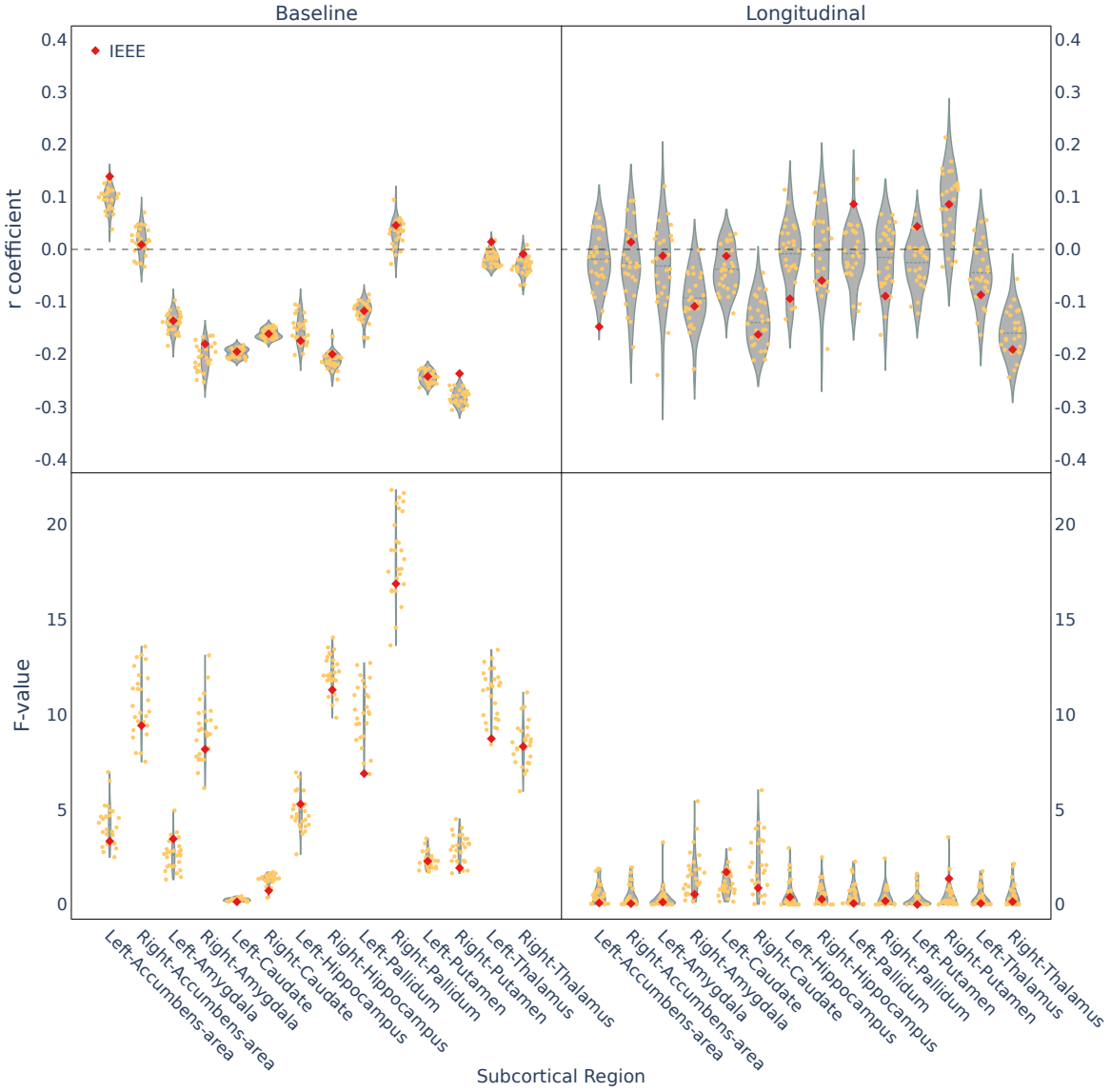
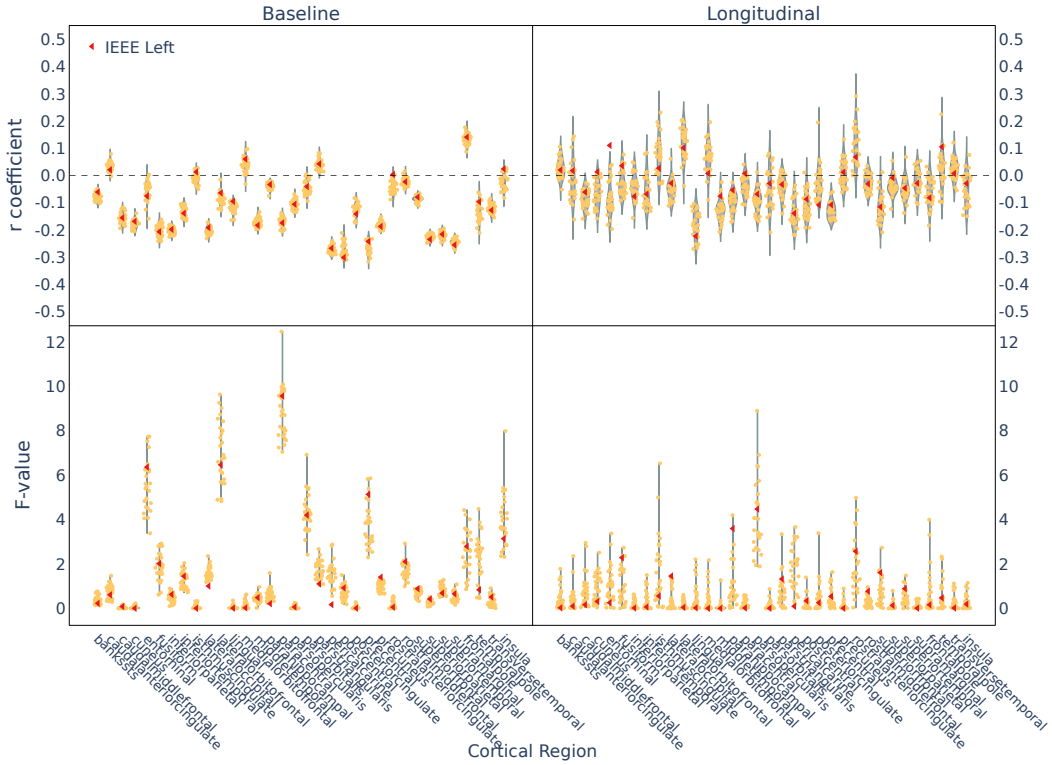
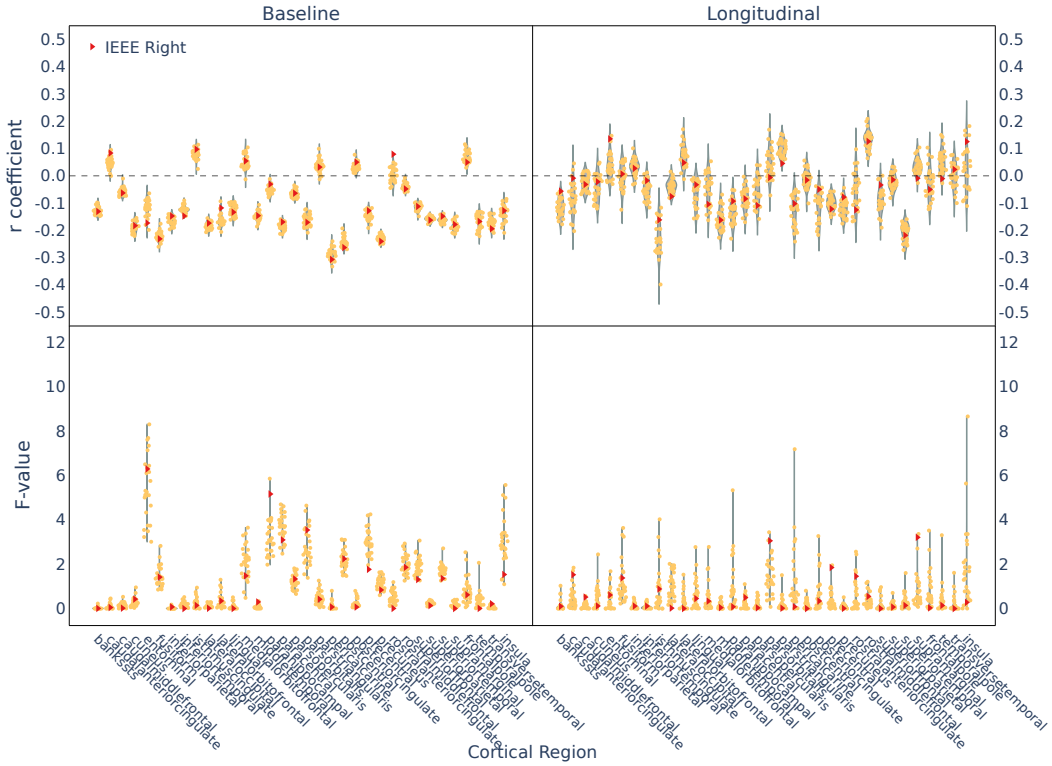


Figure 13: Distribution of partial correlation coefficients (r-values) and F-statistics from ANCOVA across MCA repetitions for subcortical volume measures. Red dots represent the IEEE-754 unperturbed results.

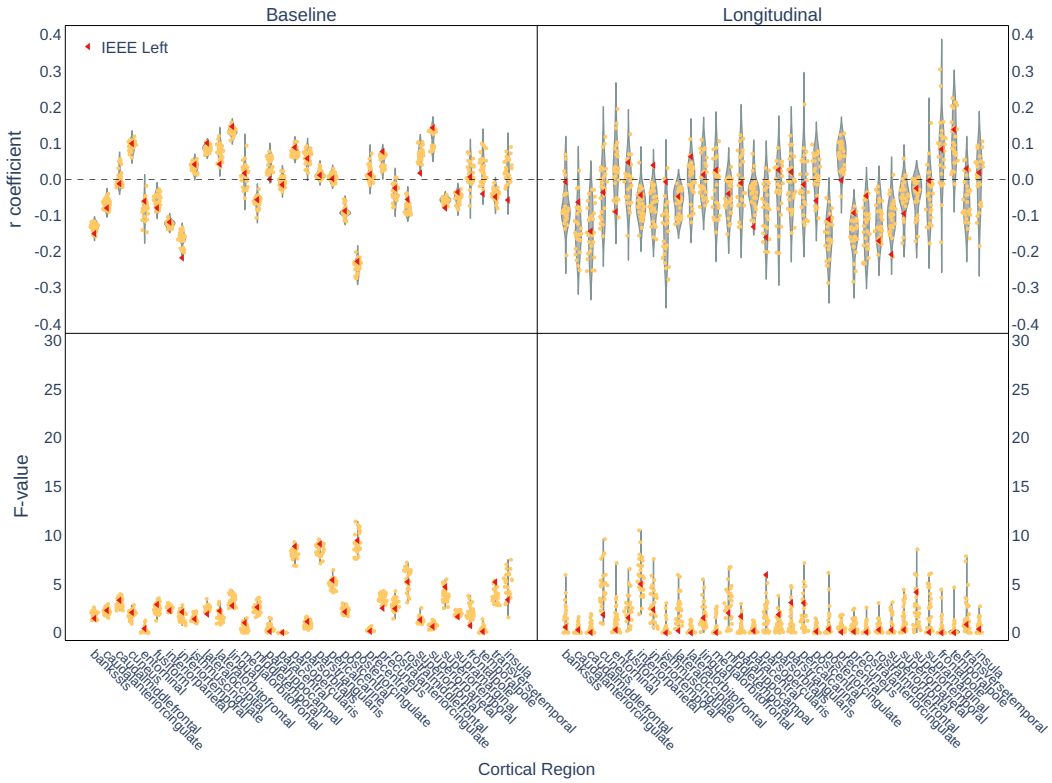


(a) Left hemisphere

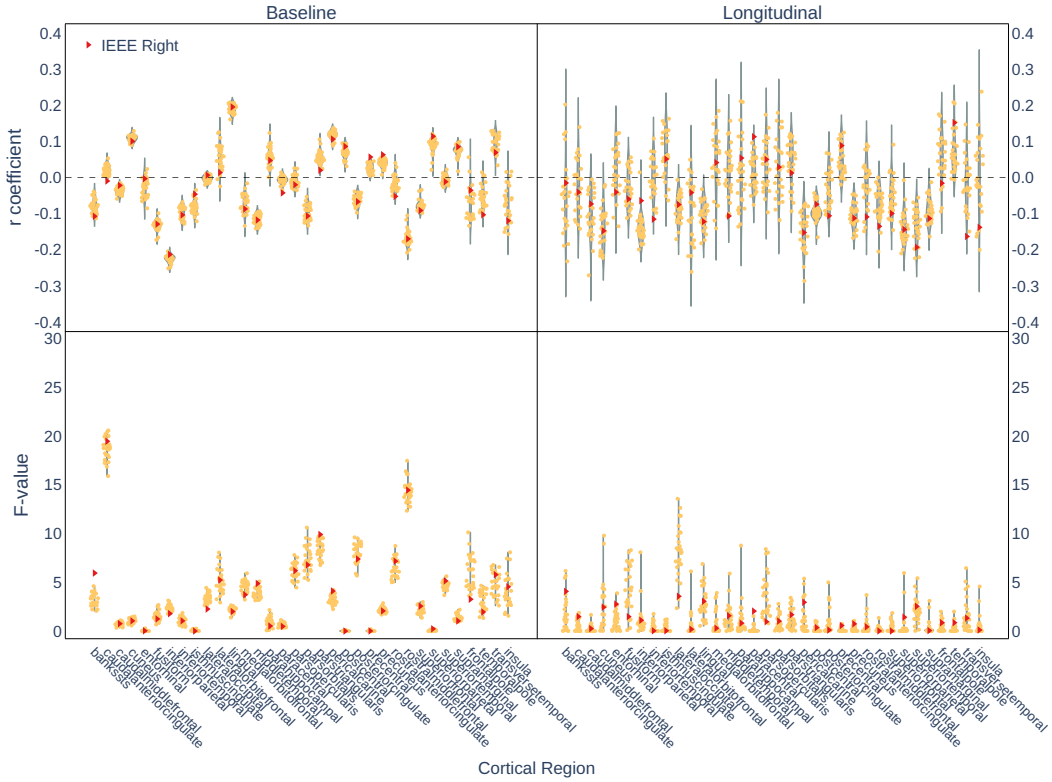


(b) Right hemisphere

Figure 14: Distribution of partial correlation coefficients for cortical thickness across all subjects and regions. Red triangles indicate the IEEE-754 run for reference.

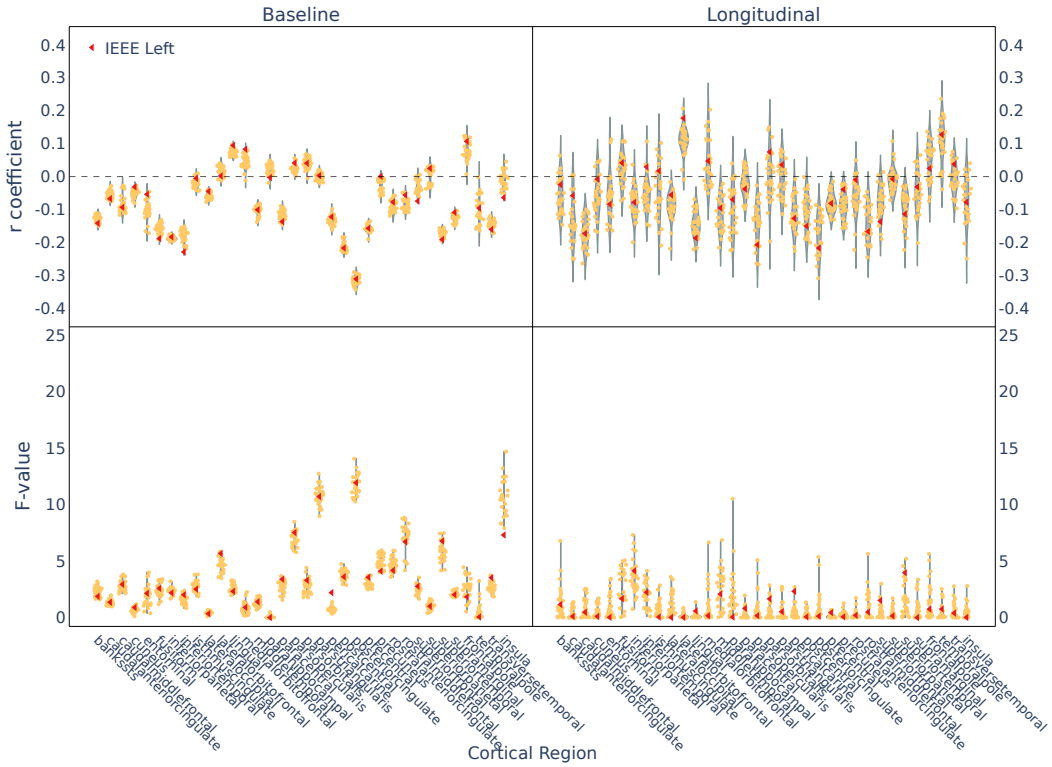


(a) Left hemisphere

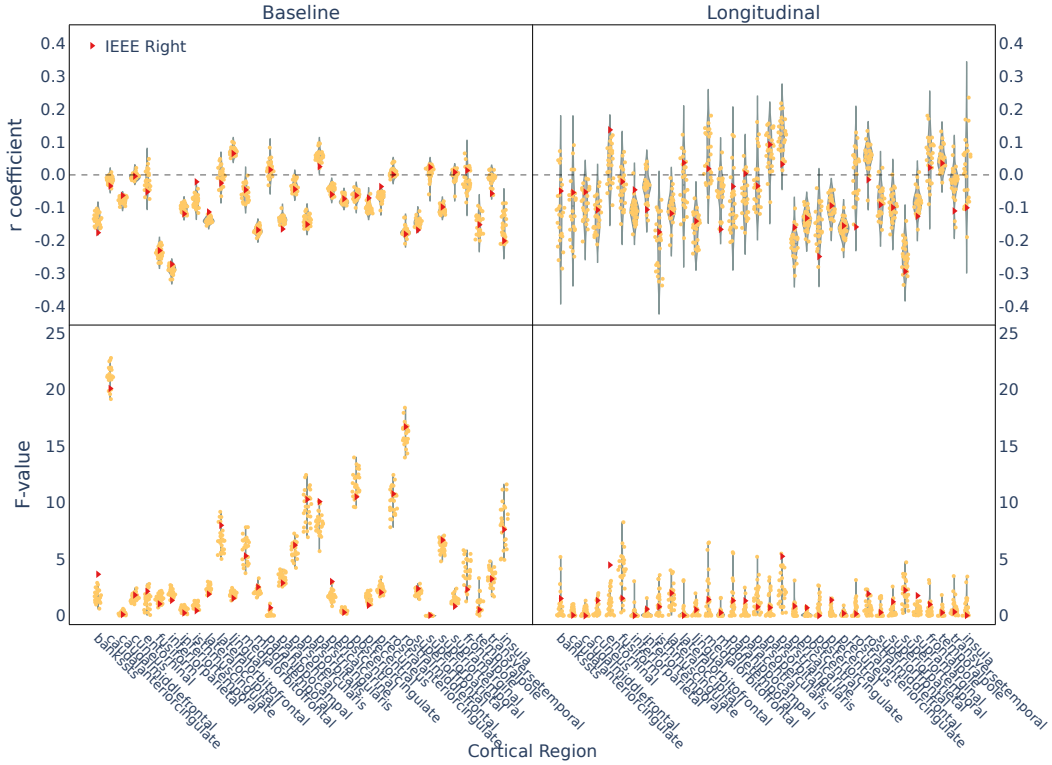


(b) Right hemisphere

Figure 15: Distribution of partial correlation coefficients for cortical surface area across all subjects and regions. Red triangles indicate the IEEE-754 run for reference.



(a) Left hemisphere



(b) Right hemisphere

Figure 16: Distribution of partial correlation coefficients for cortical volume across all subjects and regions. Red triangles indicate the IEEE-754 run for reference. The distribution shows the variability in the coefficients, with some regions exhibiting higher consistency than others.

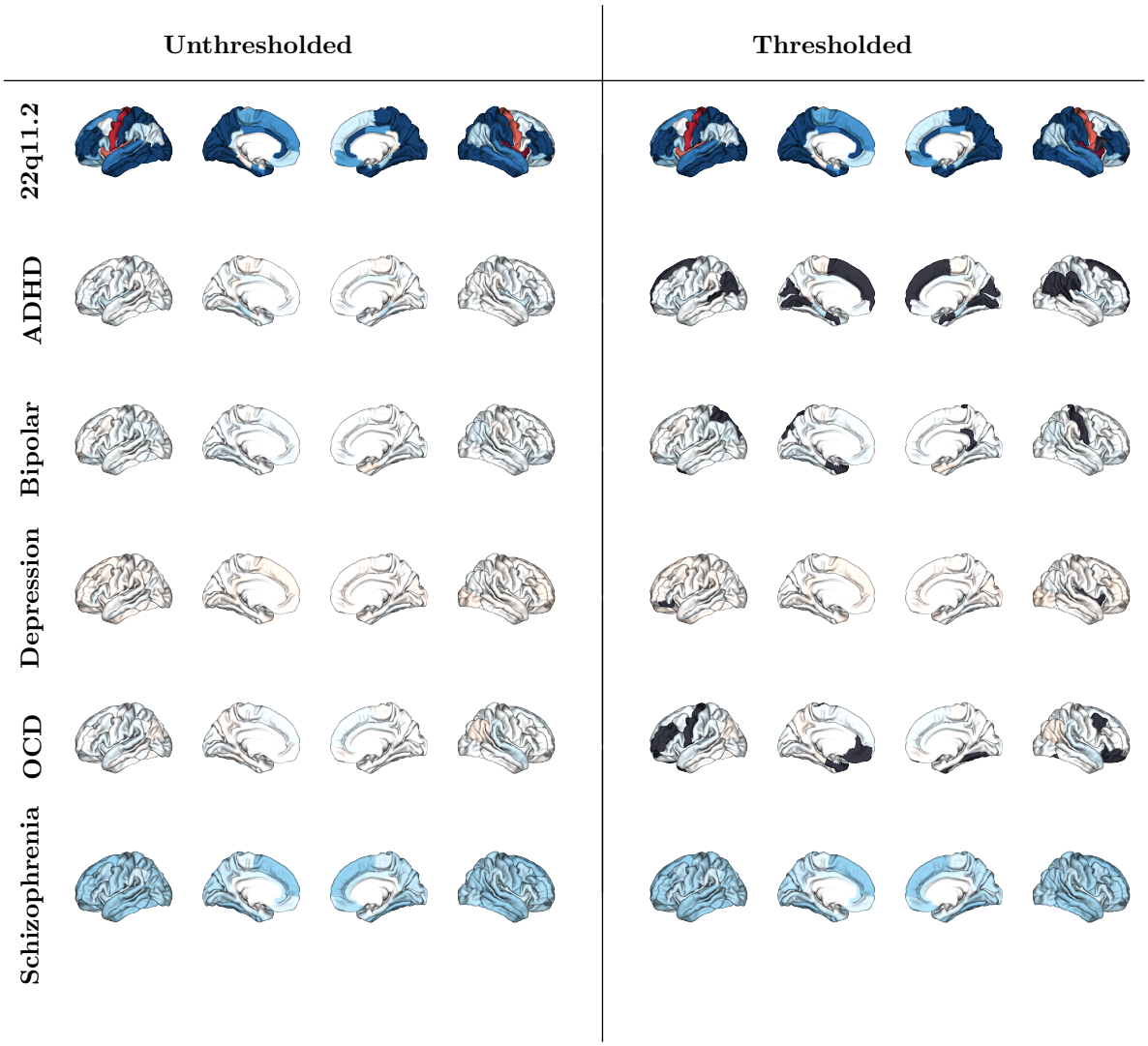


Figure 17: ENIGMA cortical area Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{npv} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.

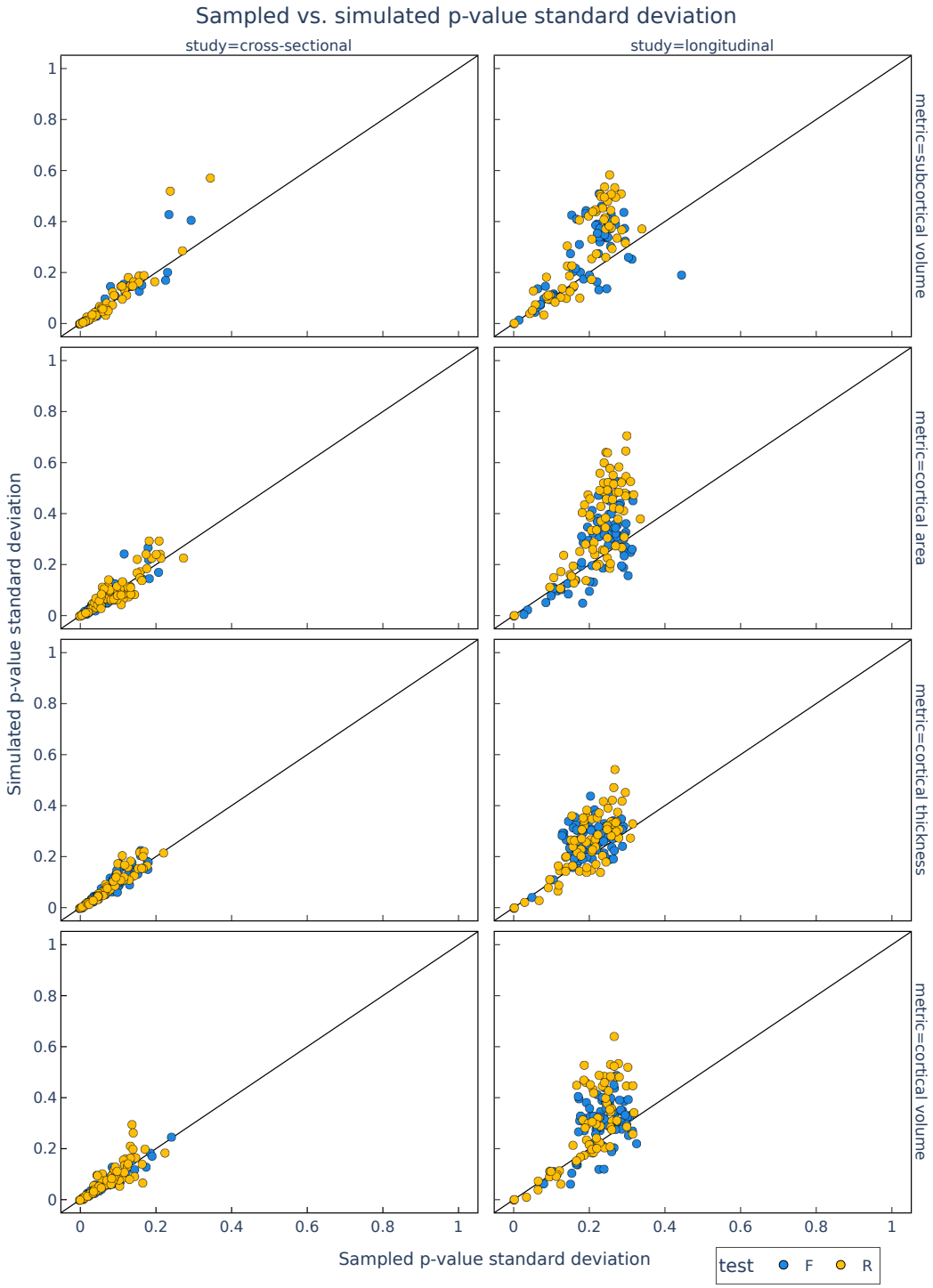


Figure 18: Comparison of standard deviation estimates of p-values obtained from MCA sampling and analytical simulations. Sampled estimates are computed from 26 MCA repetitions, while simulated estimates are derived from the uncertainty propagation formulas reported in Table 1. Each point corresponds to a cortical or subcortical brain region. The solid line denotes the identity line. Results show close agreement between sampled and simulated estimates, with cross-sectional analyses exhibiting tighter clustering than longitudinal analyses. In the longitudinal setting, simulated standard deviations tend to slightly overestimate sampled variability, indicating conservative behavior of the analytical uncertainty model.

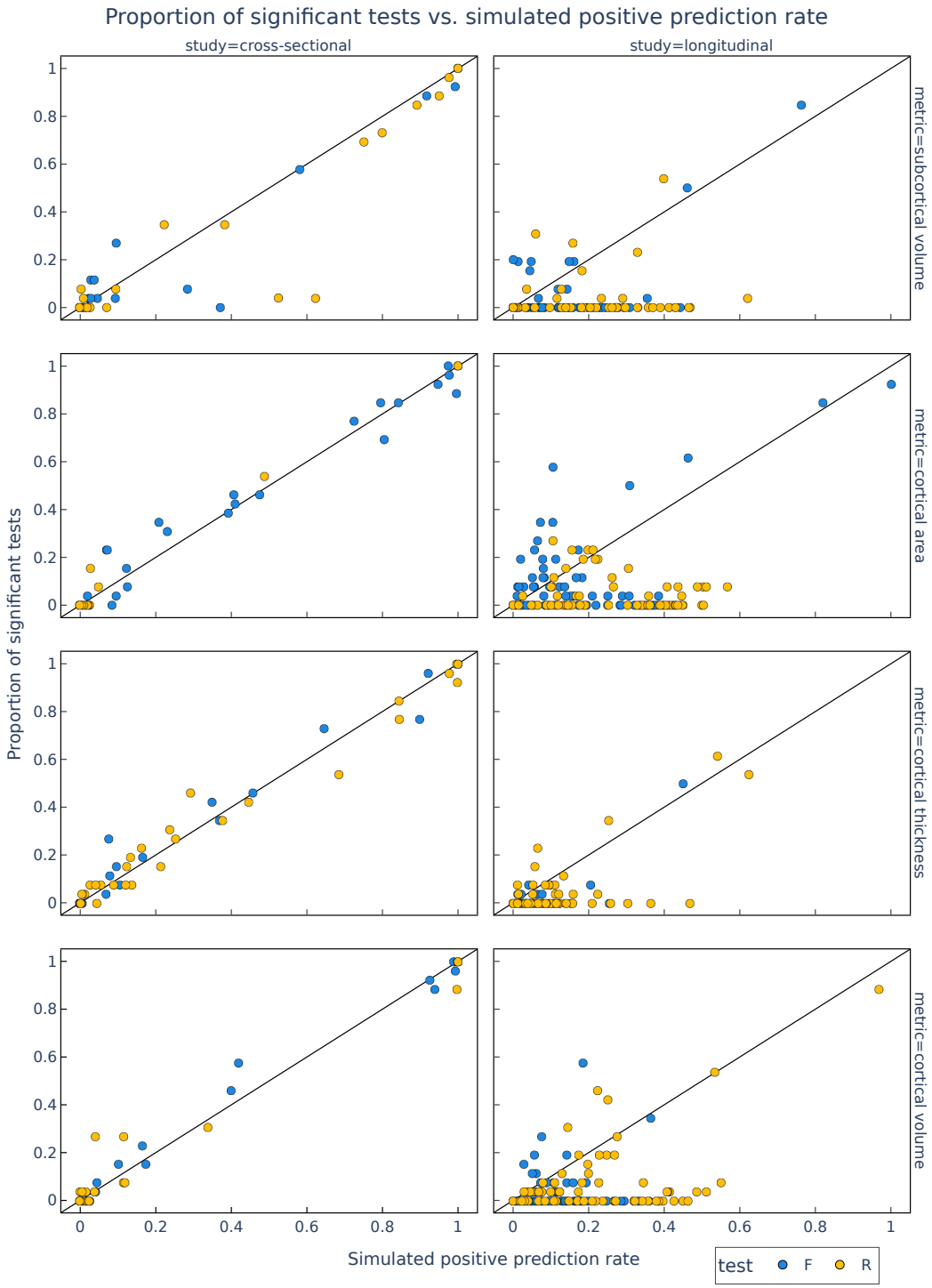


Figure 19: Comparison between simulated positive prediction rate (PPR; Eq. 26) and the empirical proportion of significant tests (Section 2.1). Simulated PPR values are computed using analytical uncertainty estimates derived from Table 1. Each point represents a cortical or subcortical brain region, and the solid line indicates the identity line. Simulated PPR closely matches empirical significance rates in the cross-sectional setting, while longitudinal analyses show a tendency toward overestimation, consistent with the corresponding overestimation of p-value variability.

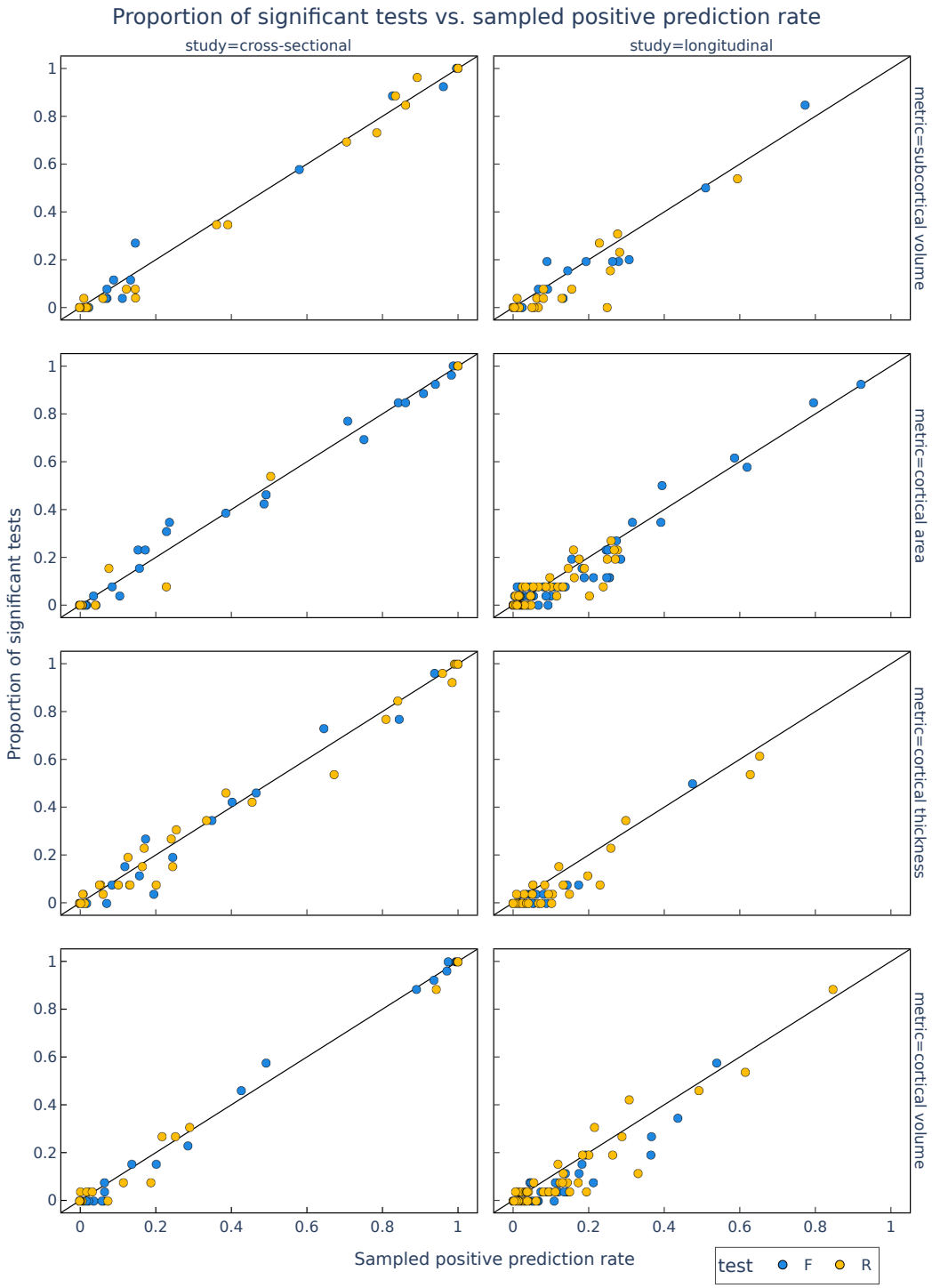


Figure 20: Comparison between sampled positive prediction rate (PPR; Eq. 26) and the empirical proportion of significant tests (Section 2.1). Sampled PPR values are computed from MCA-derived p-value standard deviations and propagated through the Beta distribution model (Eq. 4.4). Each point corresponds to a cortical or subcortical brain region, and the solid line denotes the identity line. Both cross-sectional and longitudinal results align closely with the identity line, supporting the suitability of the Beta modeling framework for capturing numerically induced variability in statistical significance.