

Numerical variability in structural MRI measurements on Parkinson’s disease

Yohan Chatelain, Andrzej Sokołowski, Madeleine Sharp, Jean-Baptiste Poline, Tristan Glatard

July 22, 2025

Abstract

Numerical variability, computational differences arising from floating-point arithmetic, poses a critical yet overlooked threat to reproducibility in clinical neuroimaging. We systematically quantify the impact of numerical variability in FreeSurfer, a widely-used MRI analysis tool, by analyzing how it processes identical brain scans under slightly different computational conditions—mimicking real-world variations across computers and software versions. To quantify this problem, we developed a metric that reveals when computational noise drowns out real biological signals. In Parkinson’s disease patients, we found that up to half the measured brain differences could be computational artifacts rather than disease effects. This computational instability can flip research conclusions: the same data analyzed on different computers can yield opposite results about whether a brain region shows disease-related changes. When we applied our framework to landmark brain imaging studies, including ENIGMA consortium findings, many reported effects fell below the computational noise threshold, particularly in smaller studies. By quantifying and addressing numerical uncertainty, our study presents essential guidelines to enhance computational reliability and reproducibility in clinical neuroscience.

1 Introduction

Modern science is built on a foundation of computational analysis. From climate modeling to genomics and neuroscience, complex software pipelines are indispensable for transforming raw data into discovery. While researchers are vigilant against software version changes and methodological errors, a more fundamental challenge is often overlooked: the numerical variability of the computations themselves. Seemingly deterministic algorithms can produce slightly different results on each run due to factors like floating-point arithmetic and parallelization, introducing a layer of computational uncertainty that is rarely quantified [BNHC⁺20, GHJ⁺12, BBD⁺21].

This challenge is particularly acute in clinical research, where the goal is to detect subtle biological signals of disease. In neuroimaging, for instance, metrics such as cortical thickness are used as potential biomarkers to track neurological disorders like Parkinson’s disease (PD) [HPZ⁺23]. If the computational noise in measuring these biomarkers is on the same order of magnitude as the subtle brain changes caused by the disease, any resulting scientific conclusions become inherently fragile. This may contribute to the well-documented “reproducibility crisis” where promising findings fail to translate into reliable clinical tools.

Here, we dissect the impact of within-software numerical variability on the reliability of clinical neuroimaging research. We use Monte Carlo Arithmetic to systematically assess FreeSurfer, a cornerstone software package, and develop a novel framework—the Numerical-Anatomical Variability Ratio (ν_{nav})—to directly compare the magnitude of computational uncertainty to biological variability. Using data from a major PD cohort, we (1) quantify how numerical instability impacts group comparisons and brain-behavior correlations, (2) derive the theoretical relationship between ν_{nav} and the reliability of statistical effect sizes, and (3) apply this framework to re-evaluate the robustness of landmark findings in the field.

2 Impacts of numerical variability

Numerical variability can have significant consequences for scientific research, particularly in fields like neuroimaging where precise measurements are crucial. In this section, we explore the various ways in which numerical variability can impact research outcomes. We first demonstrate how numerical instability can lead to fluctuating scientific conclusions. Next, we introduce a framework to quantify the impact of numerical variability. Finally, we apply this framework to re-evaluate landmark studies in the field, revealing widespread potential for unreliable effect sizes.

2.1 Numerical instability leads to fluctuating scientific conclusions

Numerical instability is a pervasive issue in scientific computing, particularly in neuroimaging, where complex algorithms process large datasets to extract biologically relevant features.

The culprit is floating-point arithmetic—the way computers handle real numbers. Just as $1 \div 3$ cannot be perfectly represented in decimal (0.333...), most calculations produce tiny rounding errors. In a pipeline with millions of calculations, these errors snowball into meaningful differences. Although the IEEE-754 standard provides a consistent framework for representing and manipulating floating-point numbers, their non-associativity leads to variability in results when computations are performed in different orders—determined by the underlying hardware and software forming a so-called computational environment.

To measure this instability, we processed each brain scan 26 times while introducing controlled computational perturbations—simulating the natural variations that occur across different computers. We found that the numerical variability significantly impacted the reproducibility of our findings. In many cases, the same analysis performed on slightly different numerical states produced divergent results, calling into question the reliability of the original conclusions.

We assessed numerical variability in FreeSurfer 7.3.1 using Monte Carlo Arithmetic (MCA) [Par97] to simulate how well behave the software under different execution environments (e.g. different hardware configurations, operating systems, scientific computational libraries), more particularly the Fuzzy-libm extension [SCKG21] that applies random perturbations to mathematical library functions's output. MCA has demonstrated its ability to simulate numerical variability in several scientific domains, including neuroimaging [KdOCR⁺20, DLBC⁺23, CTM⁺24, VMGP⁺24]. We executed 26 `recon-all` runs for each subject's MRI data from the Parkinson's Progression Markers Initiative (PPMI), simulating 26 numerical states.

For the same brain scans, statistical tests gave different answers depending on computational conditions. A brain region showing 'significant' disease effects in one run showed no effect in another—purely due to rounding errors.

We first tested the statistical significance of group differences between Parkinson's disease (PD) patients and healthy controls (HC), and their correlations with clinical measures (UPDRS scores), across the 26 MCA repetitions. Similar to the NARPS study [BNHC⁺20] that highlighted the impact of analytical variability in neuroimaging, we observed substantial fluctuations in the results due to numerical variability. We used a similar layout to NARPS paper to report the results, with the proportion of statistically significant tests ($p < 0.05$) varied substantially for both cortical thickness (Figure 2) and subcortical volumes (Figure 1). For many brain regions, the significance of a finding was inconsistent, appearing in some computational runs but not others. Ratios near 0.5 indicate maximal uncertainty, where a reported result could be attributed to a lucky roll of the computational dice. This variability in statistical significance suggests that the conclusions drawn from these analyses are not robust and may be influenced by the specific numerical state of the computation. This is specific to brain regions, some regions showing high consistency across numerical states, while others exhibit substantial variability. This instability is particularly concerning when finding biomarkers for the Parkinson's disease progression, as it suggests that the biological signals of interest may be obscured by computational noise. Sokolowski et al. [SBK⁺24] highlighted similar issues when playing with different FreeSurfer versions, showing that the results can vary significantly between versions 5, 6 and 7. Biomarkers are crucial for understanding disease mechanisms and tracking progression. Having reliable biomarkers is essential to better understand the nature of the disease which remains not well understood [BOK21].

This instability is also reflected in the effect sizes themselves. Partial correlation coefficients and F-statistics from ANCOVA analyses showed substantial spread around the unperturbed IEEE-754

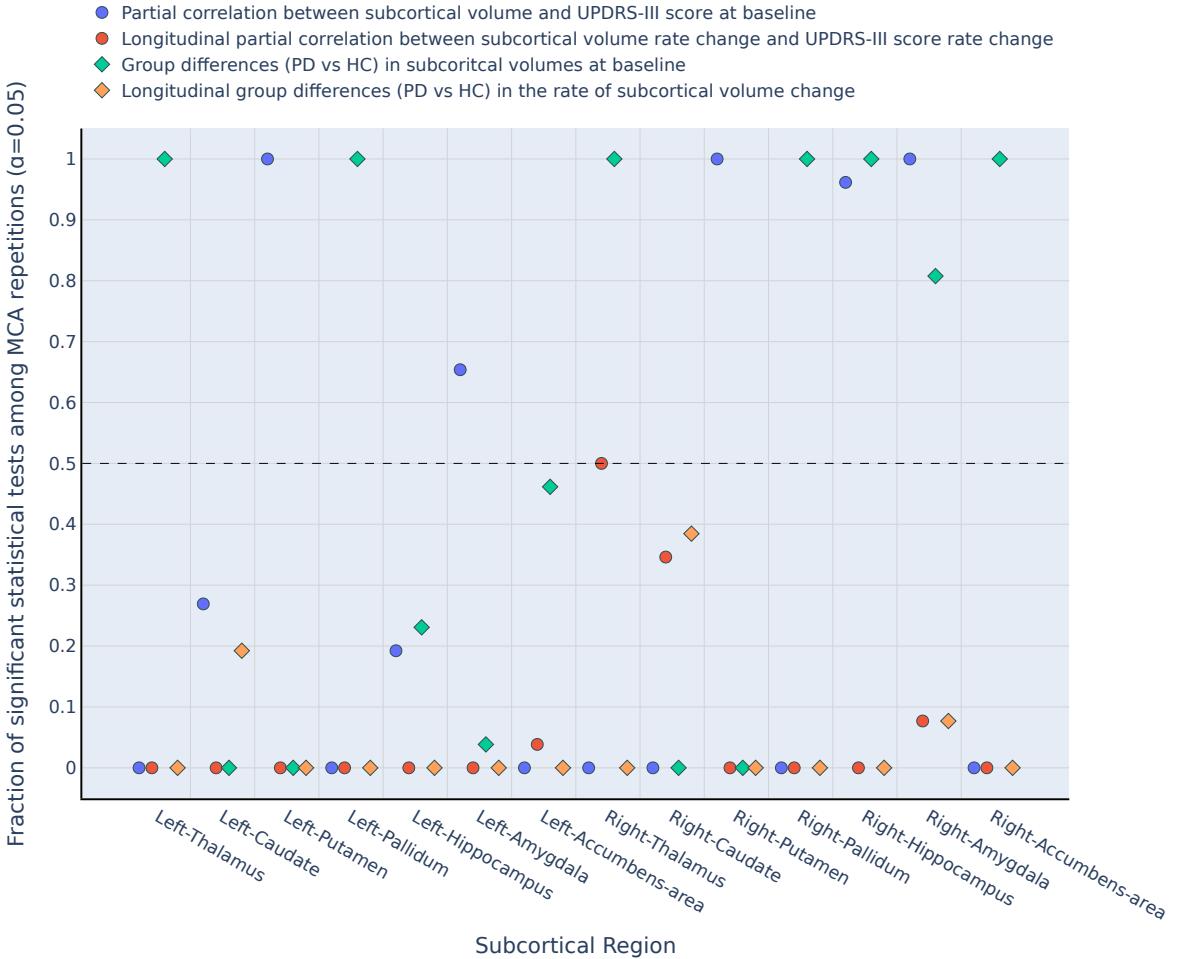


Figure 1: Proportion of significant tests ($p < 0.05$) for subcortical volumes across 26 numerical perturbations. measures.

results (red markers in Figure 3 and 4). This demonstrates that numerical variability affects not only the binary outcome of statistical significance but also the magnitude of the estimated effect, further complicating the interpretation of results. We can notice that the unperturbed results (red dots) are contained within the distribution of coefficients and F-statistics, which indicates that our methodology is sound, but the variability in the results highlights the need for caution in interpretation. In particular, when interpreting results for possible clinical applications, it is important to consider the potential impact of numerical variability on the robustness of findings. Yellow dots around red ones can be seen as many potential results that could be obtained by running the same analysis under computational environments.

2.2 A framework to quantify the impact of numerical variability

How much computational noise is too much? We developed the Numerical-Anatomical Variability Ratio (ν_{nav}), a metric that answers this question by comparing computational uncertainty to real biological differences between people. We built this tool to be broadly applicable to any neuroimaging analysis, allowing researchers to easily assess the impact of numerical variability on their findings. Our metric works like a signal-to-noise ratio in reverse: when $\nu_{\text{nav}} = 0.5$, half the observed 'biological variation' is actually computational noise. When ν_{nav} approaches 1, we're measuring noise, not brains. ν_{nav} is defined as:

$$\nu_{\text{nav}} = \frac{\sigma_{\text{num}}}{\sigma_{\text{anat}}}$$

where σ_{num} is the numerical variability (see 1) and σ_{anat} is the anatomical variability (see 2).

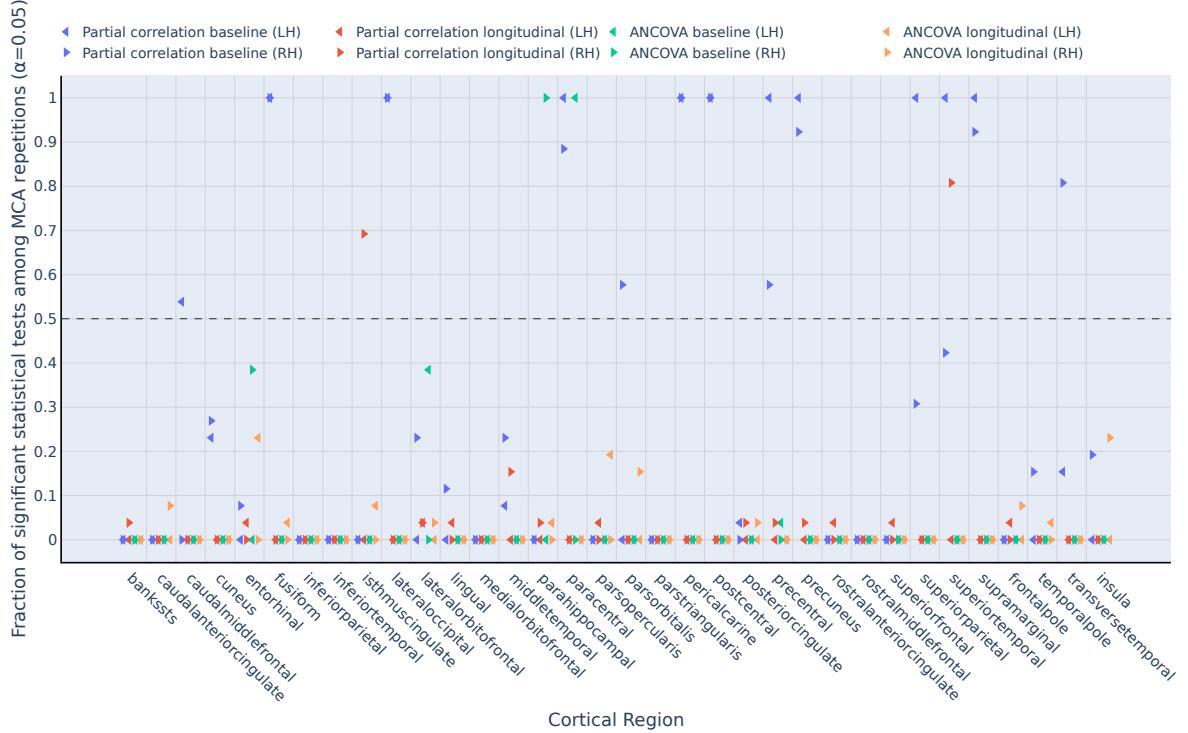


Figure 2: Proportion of significant tests ($p < 0.05$) for cortical thickness across 26 numerical perturbations. measures.

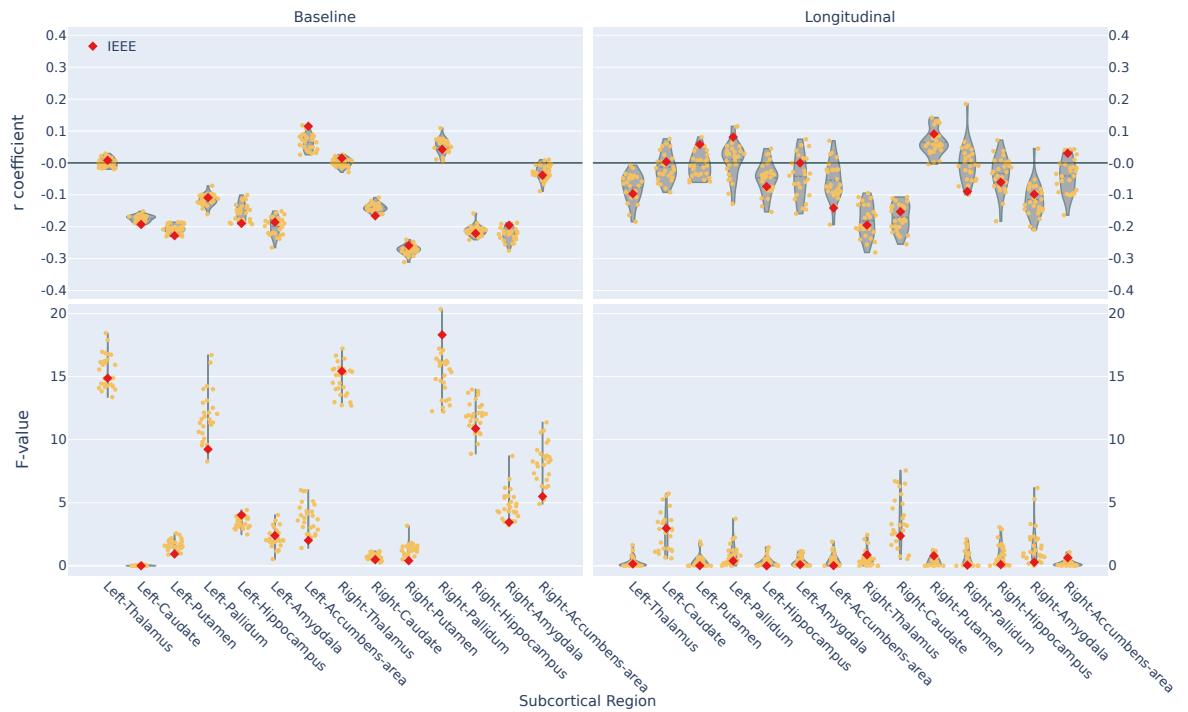


Figure 3: Distribution of partial correlation coefficients (r-values) and F-statistics from ANCOVA across MCA repetitions for subcortical volume measures. Red dots represent the IEEE results. The top row shows r-values, while the bottom row shows F-values. The left column represents baseline analysis, and the right column represents longitudinal analysis.



Figure 4: Distribution of partial correlation coefficients for cortical thickness 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.

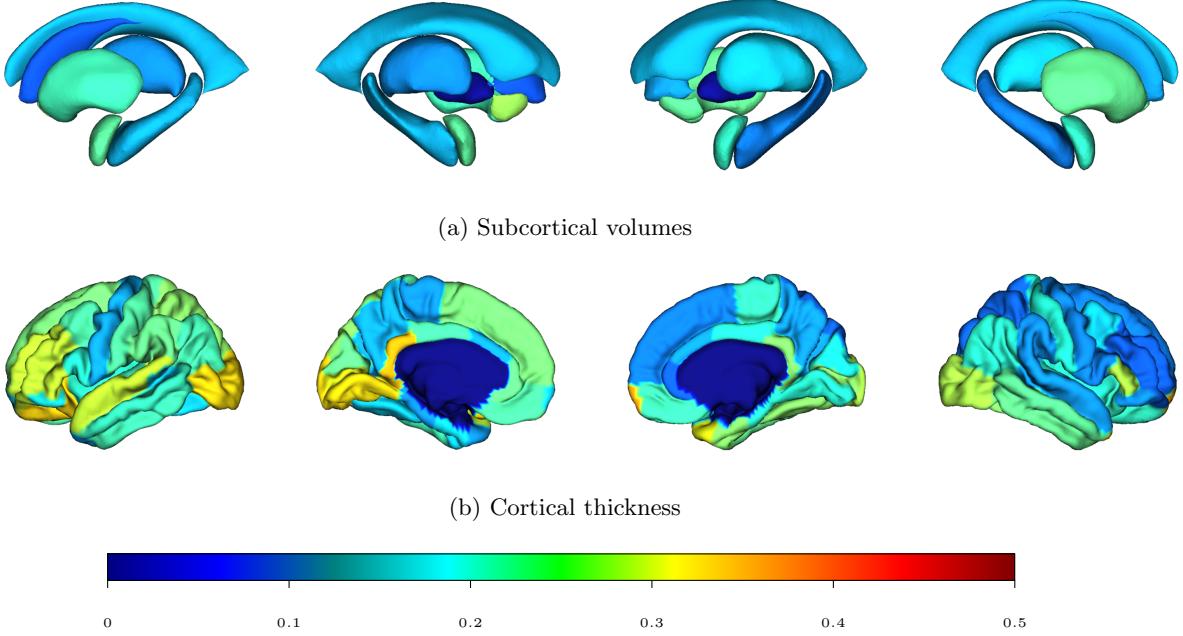
Having a tool to quantify numerical variability is crucial for several reasons. First, it allows researchers to assess the reliability of their findings and reveal meaningful signals from noise. Second, it provides a framework for comparing the impact of the numerical variability versus the anatomical variability, enabling more informed decisions about study design and data interpretation. Finally, it allows for the rapid assessment of numerical variability in existing neuroimaging studies, facilitating the identification of potentially unreliable results. Indeed, measuring numerical variability can be a time-consuming process, especially when dealing with large datasets and complex analyses. By providing an analytical model of ν_{nav} , we enable researchers to apply this tool to any neuroimaging paper or existing results, significantly speeding up the process of assessing numerical precision. We developed an online tool¹ that allows researchers to threshold their results based on ν_{nav} values, providing a practical way to identify potentially unreliable findings.

Figures 5b and 5a present the ν_{nav} for cortical thickness and subcortical volumes. These brain maps visualize the regions where the problem of numerical variability is most severe. Regions with high ν_{nav} values (e.g., approaching 0.5, shown in warmer colors) indicate areas where up to half of the observed variability is due to computational noise rather than true anatomical differences, severely compromising the ability to detect genuine biological effects.

We derived the theoretical relationship between ν_{nav} and the uncertainty in Cohen's d, a standard measure of effect size:

$$\sigma_d = \frac{2}{\sqrt{N}} \nu_{\text{nav}}$$

where σ_d is the standard deviation of Cohen's d and N is the total sample size. This formula provides a direct way to translate the numerical stability of a measurement into its impact on statistical power and reliability, offering a practical tool for researchers to gauge the robustness of their findings.



(c) Numerical-Anatomical Variability Ratio (ν_{nav}) for subcortical volumes 5a and cortical thickness 5b across regions and groups. Higher ν_{nav} values indicate greater computational uncertainty relative to biological variation. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav} values.

¹https://yohanchatelain.github.io/brain_render/

2.3 Re-evaluating landmark studies reveals widespread potential for unreliable effect sizes

The true test of our discovery: what happens when we apply this metric to the field’s most influential findings? ENIGMA [TSM⁺14]—the world’s largest neuroimaging consortium—has shaped our understanding of psychiatric disorders by analyzing tens of thousands of brain scans. When we applied our computational noise threshold to their published brain maps, large regions went black: the reported effects were smaller than measurement noise.

This finding raises important questions about the reliability of existing neuroimaging literature. Are we overestimating the effects of certain brain regions? Are there systematic biases in how we report and interpret these findings?

Figures 6 and 7 present the powerful results of this analysis. The left panels show the original Cohen’s d maps for various disorders, while the right panels show the same maps after thresholding. Regions colored in black indicate where the reported effect size is smaller than the underlying computational noise of the measurement tool. While many of the core, highly significant findings of ENIGMA remain robust (as expected from their large sample sizes), a considerable number of secondary findings are shown to be potentially unreliable. This analysis demonstrates that our framework can be used as a critical tool to evaluate the robustness of findings in the literature and guide future research toward more reliable effects.

3 Discussion

Our analysis of FreeSurfer reveals that its measurements possess a numerical precision of approximately one to two significant digits. This computational variability is not trivial; in some brain regions, it accounts for up to half of the observed anatomical variance between healthy individuals and patients with Parkinson’s disease. Such a low signal-to-noise ratio has direct consequences for scientific reliability. We demonstrate that for the exact same dataset, the statistical significance of clinical correlations and group differences can appear or vanish simply due to numerical errors in the underlying computation.

These findings offer a potential explanation for some of the reproducibility challenges in clinical neuroimaging. To move beyond simply identifying the problem, we introduced the Numerical-Anatomical Variability Ratio (ν_{nav}) as a framework to quantify the impact of this instability. By establishing a direct theoretical link between ν_{nav} and the uncertainty of Cohen’s d effect sizes, we provide a practical tool for researchers. Our re-analysis of landmark ENIGMA findings illustrates this utility. While the large sample sizes of the consortium provide robustness to its primary findings, our framework can flag secondary effect sizes that are smaller than the computational noise floor, suggesting they may not be reliable, particularly in the context of the smaller sample sizes typical of exploratory studies.

While we focused on FreeSurfer and Parkinson’s disease as a test case, the problem is universal. Floating-point arithmetic affects all software, and preliminary tests suggest other neuroimaging packages suffer similar instabilities. The ν_{nav} values we report are specific to FreeSurfer 7.3.1, and future software versions may exhibit improved numerical stability. Furthermore, our analysis used a cohort that was relatively homogeneous in age, which may decrease the total anatomical variance and thus inflate the ν_{nav} ratio. We therefore propose that the ν_{nav} framework is not a final judgment on any particular finding, but rather a methodology that should be applied across diverse software, datasets, and disease models to build a more complete picture of computational reliability. The fuzzy-libm library used here is a powerful tool for probing floating-point instability, but future work should also investigate other sources of variability, such as algorithmic choices and data handling practices.

In conclusion, our findings emphasize the need to treat computational uncertainty with the same rigor as statistical uncertainty. The ν_{nav} framework offers a practical step in this direction, providing a means to assess and compare the numerical reliability of neuroimaging methodologies. By quantifying and accounting for computational noise, we can enhance the search for subtle neurological changes and build more robust and reproducible models of human health and disease.

4 Methods

4.1 Numerical variability assessment

We employed Monte Carlo Arithmetic (MCA) [Par97] to quantify numerical instability in FreeSurfer computations. MCA introduces controlled random perturbations into floating-point operations, simulating rounding errors that occur across different computational environments. This stochastic approach enables systematic assessment of result stability by measuring variation across multiple runs of identical analyses.

We used Fuzzy-libm [SCKG21], which extends MCA to mathematical library functions (`exp`, `log`, `sin`, `cos`) through Verificarlo [DdOCP16], an LLVM-based compiler. Virtual precision parameters were set to 53 bits for double precision and 24 bits for single precision to simulate realistic machine-level precision errors.

4.2 Participants

We analyzed data from the Parkinson’s Progression Markers Initiative (PPMI), a multi-site longitudinal study. From 316 initial participants, we selected 125 Parkinson’s disease patients without mild cognitive impairment (PD-non-MCI) and 106 healthy controls (HC) with complete longitudinal T1-weighted MRI data. PD-MCI patients were excluded to avoid confounding effects of cognitive impairment.

Inclusion criteria required: (1) primary PD diagnosis or healthy control status, (2) availability of two visits with T1-weighted scans, and (3) absence of other neurological diagnoses. PD severity was assessed using the Unified Parkinson’s Disease Rating Scale (UPDRS). The study received ethics approval from participating institutions, and all participants provided written informed consent (Table 1).

PD and HC groups showed no significant age differences ($p > 0.05$) but differed in education ($t = -2.05$, $p = 0.04$) and sex distribution ($\chi^2 = 4.15$, $p = 0.04$). The longitudinal cohort showed no significant demographic differences between groups (Table 1).

Cohort	HC	PD-non-MCI
n	90	118
Age (y)	60.7 ± 9.7	61.1 ± 9.2
Age range	30.6 – 79.8	39.2 – 78.3
Gender (male, %)	48 (53.3%)	77 (65.3%)
Education (y)	16.7 ± 3.3	16.2 ± 2.9
UPDRS III OFF baseline	–	23.6 ± 10.3
UPDRS III OFF follow-up	–	25.6 ± 11.2
Duration T2 - T1 (y)	1.4 ± 0.5	1.4 ± 0.6

Table 1: **Abbreviations:** MCI = Mild Cognitive Impairment; UPDRS = Unified Parkinson’s Disease Rating Scale; PD = Parkinson’s disease. Values are expressed as mean \pm standard deviation. PD-non-MCI longitudinal sample is a subsample of the PD-non-MCI original sample that had longitudinal data and disease severity scores available.

4.3 Image acquisition and preprocessing

T1-weighted MRI images were obtained from PPMI that uses standardized acquisition parameters: repetition time = 2.3 s, echo time = 2.98 s, inversion time = 0.9 s, slice thickness = 1 mm, number of slices = 192, field of view = 256 mm, and matrix size = 256×256 . However, since PPMI is a multisite project there may be slight differences in the sites’ setup.

We processed images using FreeSurfer 7.3.1 instrumented with Fuzzy-libm to introduce controlled numerical perturbations. Each participant underwent 34 `recon-all` executions, extracting cortical thickness, surface area, and volumes. After quality control and exclusion of failed runs, we randomly selected 26 successful repetitions per subject to ensure balanced datasets for statistical analysis.

Longitudinal processing followed the standard FreeSurfer stream [RSRF12]: cross-sectional processing of both timepoints, followed by creation of an unbiased within-subject template [RF11] using

robust registration [RRF10]. Downstream analyses used unperturbed FreeSurfer to prevent additional numerical perturbations.

4.4 Numerical Variability Assessment

We assessed FreeSurfer 7.3.1 numerical stability in cross-sectional and longitudinal contexts using the Numerical-Anatomical Variability Ratio (ν_{nav}) and its relationship to statistical effect sizes.

4.4.1 Numerical-Anatomical Variability Ratio (ν_{nav})

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

$$\nu_{\text{nav}} = \frac{\sigma_{\text{num}}}{\sigma_{\text{anat}}}$$

where σ_{num} represents numerical variability (measurement precision across MCA repetitions for individual subjects) and σ_{anat} represents anatomical variability (inter-subject differences within each repetition).

For each region of interest, measurements from n MCA repetitions across m subject-visit pairs form a data matrix $\mathcal{M}_{n \times m}$, where element $x_{i,j}$ represents the measurement for subject j in repetition i .

Numerical variability quantifies intra-subject measurement consistency:

$$\sigma_{\text{num}}^2 = \frac{1}{m} \sum_{j=1}^m \left[\frac{1}{n-1} \sum_{i=1}^n (x_{i,j} - \bar{x}_{\cdot,j})^2 \right] \quad (1)$$

Anatomical variability captures inter-subject differences:

$$\sigma_{\text{anat}}^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{m-1} \sum_{j=1}^m (x_{i,j} - \bar{x}_{i,\cdot})^2 \right] \quad (2)$$

where $\bar{x}_{\cdot,j}$ and $\bar{x}_{i,\cdot}$ denote column and row means, respectively. Higher ν_{nav} values indicate regions where computational uncertainty approaches or exceeds biological variation, potentially compromising the detection of true anatomical differences.

4.4.2 Relationship between ν_{nav} and Effect Size Uncertainty

We derived the theoretical relationship between ν_{nav} and Cohen's d variability to quantify how measurement uncertainty affects statistical effect sizes in group comparisons.

For a balanced two-group design with total sample size N , each observation decomposes as $X_{ij} = \mu_i + \varepsilon_{ij}^{(\text{anat})} + \varepsilon_{ij}^{(\text{num})}$, where μ_i represents the true group mean, $\varepsilon_{ij}^{(\text{anat})} \sim \mathcal{N}(0, \sigma_{\text{anat}}^2)$ captures anatomical variation, and $\varepsilon_{ij}^{(\text{num})} \sim \mathcal{N}(0, \sigma_{\text{num}}^2)$ represents numerical uncertainty.

The standard deviation of Cohen's d attributable to measurement error is:

$$\sigma_d = \frac{2}{\sqrt{N}} \nu_{\text{nav}} \quad (3)$$

This relationship emerges from error propagation analysis. The difference in group means has variance $\text{Var}(\bar{X}_1 - \bar{X}_2) = 4(\sigma_{\text{anat}}^2 + \sigma_{\text{num}}^2)/N$, with the numerical component contributing $4\sigma_{\text{num}}^2/N$. Since Cohen's d normalizes by the pooled standard deviation $\sqrt{\sigma_{\text{anat}}^2 + \sigma_{\text{num}}^2}$, the measurement error contribution becomes $\sigma_d = (2\sigma_{\text{num}}/\sqrt{N})/\sigma_{\text{anat}} = (2/\sqrt{N})\nu_{\text{nav}}$.

This formula indicates that regions with $\nu_{\text{nav}} = 0.1$ contribute approximately $0.2/\sqrt{N}$ uncertainty to Cohen's d, while regions with $\nu_{\text{nav}} = 1.0$ contribute $2/\sqrt{N}$ uncertainty. The relationship provides a direct link between computational stability (ν_{nav}) and statistical reliability in neuroimaging studies.

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

The code used to conduct the analyses is available at [URL to be added upon publication].

7 Acknowledgements

The analyses were conducted on the Virtual Imaging Platform [[GLG⁺12](#)], 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.

References

- [BBD⁺21] 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.
- [BNHC⁺20] 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.
- [BOK21] Bastiaan R Bloem, Michael S Okun, and Christine Klein. Parkinson’s disease. *The Lancet*, 397(10291):2284–2303, 2021.
- [CTM⁺24] 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.
- [DdOCP16] 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.
- [DLBC⁺23] 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.
- [GHJ⁺12] 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.
- [GLG⁺12] 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.

- [HPZ⁺23] Elizabeth Haddad, Fabrizio Pizzagalli, Alyssa H Zhu, Ravi R Bhatt, Tasfiya Islam, Iyad Ba Gari, Daniel Dixon, Sophia I Thomopoulos, Paul M Thompson, and Neda Jahanshad. Multisite test-retest reliability and compatibility of brain metrics derived from freesurfer versions 7.1, 6.0, and 5.3. *Human Brain Mapping*, 44(4):1515–1532, 2023.
- [KdOCR⁺20] Gregory Kiar, Pablo de Oliveira Castro, Pierre Rioux, Eric Petit, Shawn T Brown, Alan C Evans, and Tristan Glatard. Comparing perturbation models for evaluating stability of neuroimaging pipelines. *The International Journal of High Performance Computing Applications*, 34(5):491–501, 2020.
- [Par97] Douglass Stott Parker. *Monte Carlo arithmetic: exploiting randomness in floating-point arithmetic*. Citeseer, 1997.
- [RF11] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *Neuroimage*, 57(1):19–21, 2011.
- [RRF10] Martin Reuter, H Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4):1181–1196, 2010.
- [RSRF12] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.
- [SBK⁺24] 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.
- [SCF⁺21] 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.
- [SCKG21] 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.
- [TSM⁺14] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, 8:153–182, 2014.
- [VMGP⁺24] 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.

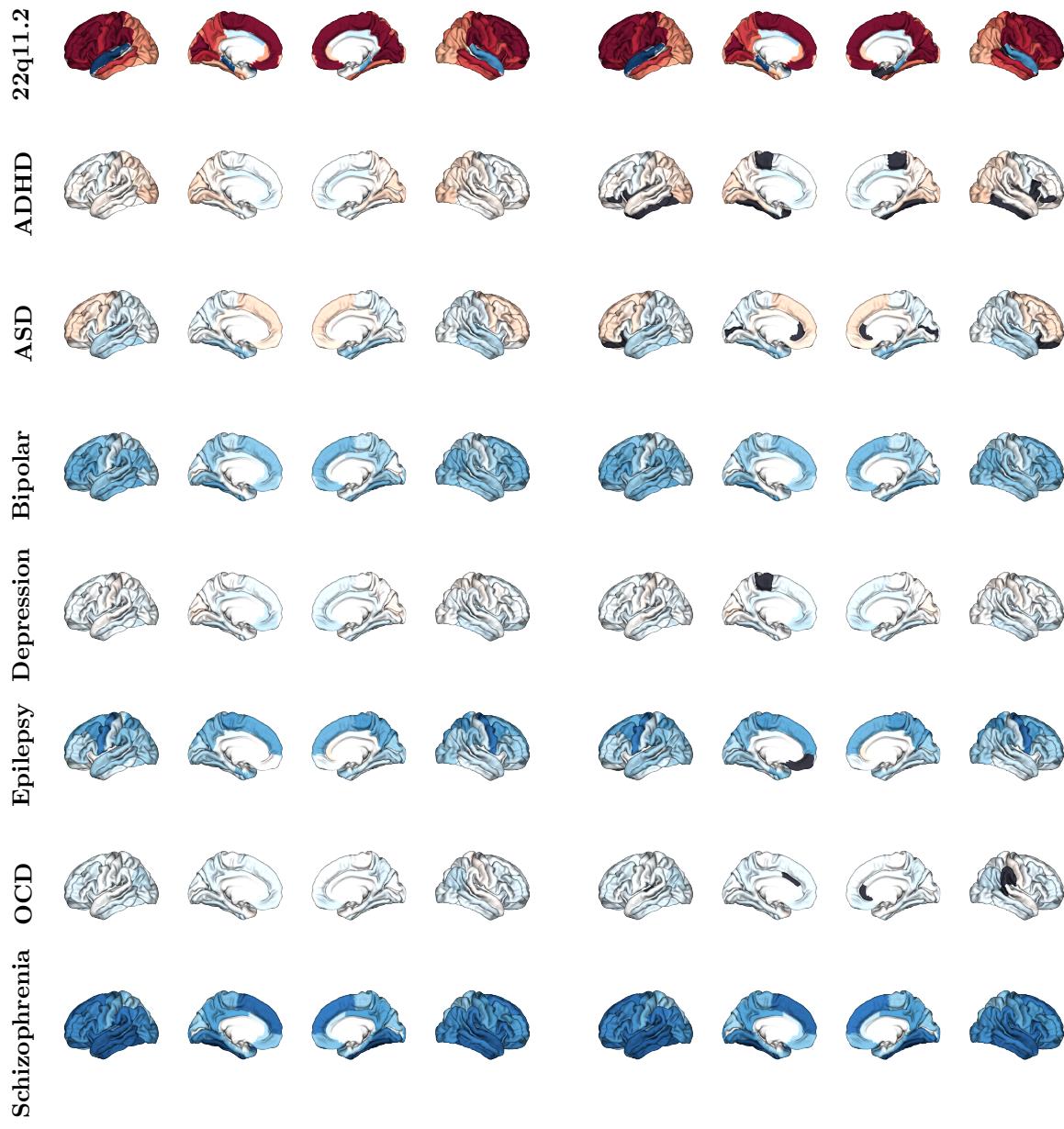


Figure 6: ENIGMA cortical thickness Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.

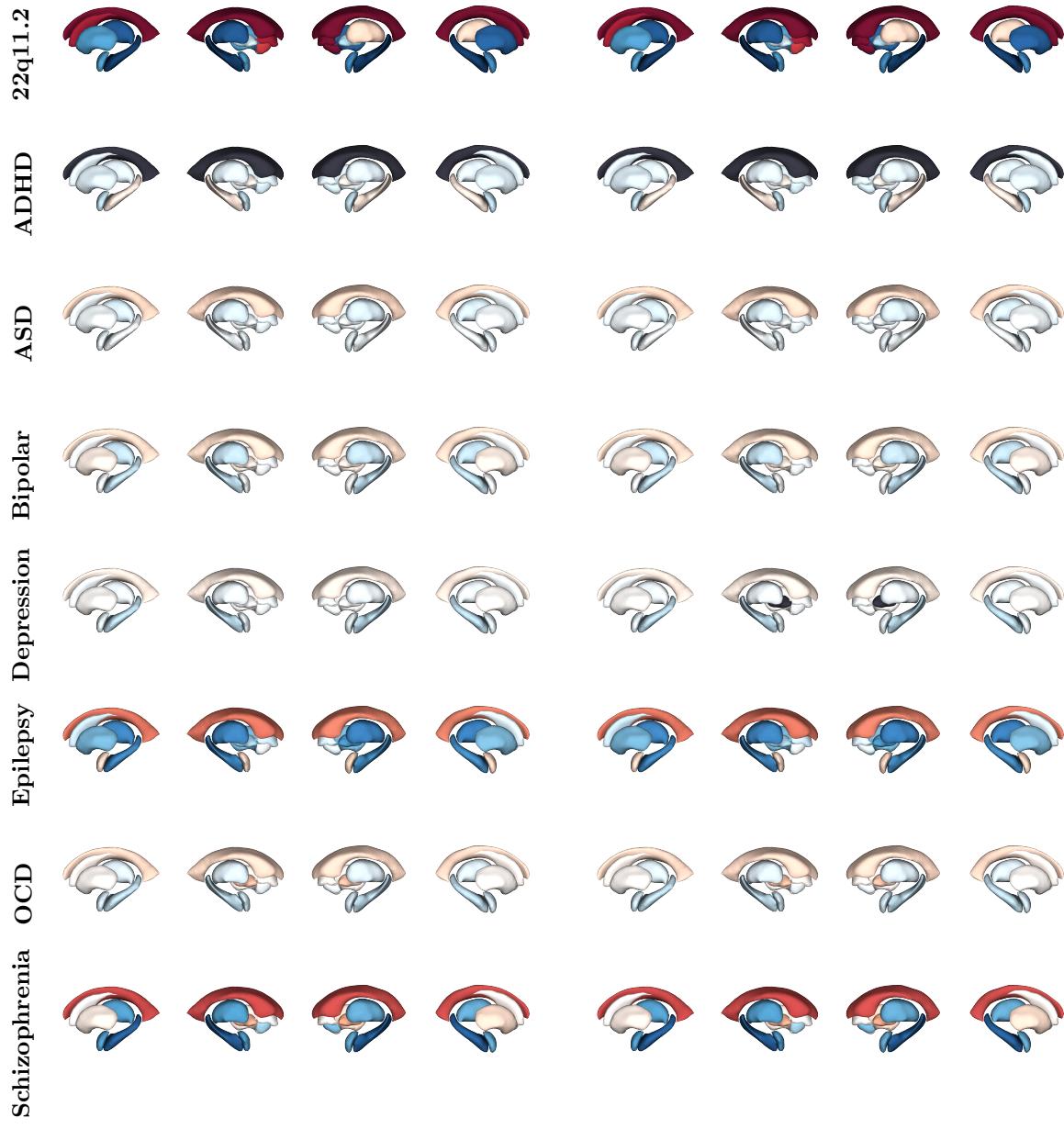


Figure 7: ENIGMA subcortical volume Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.

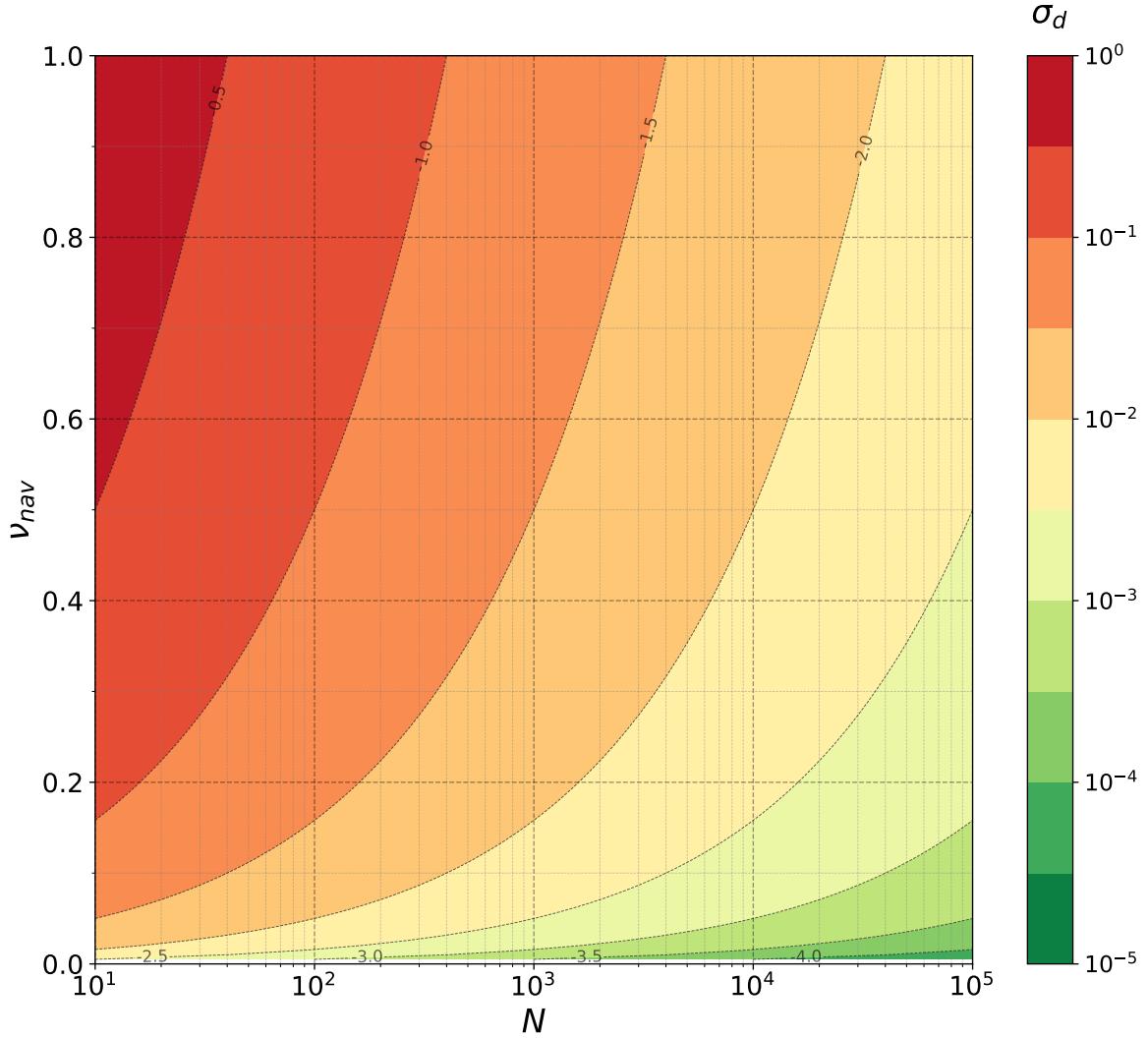


Figure 8: Relationship between ν_{nav} and population sample size N for predicting the uncertainty in Cohen's d effect size estimation. The contour lines represent different ν_{nav} values, showing how numerical variability scales with sample size. With a typical ν_{nav} value of 0.2, to maintain reliable effect size estimates $\sigma_d \leq 0.01$, the plot suggests to use $N \geq 1500$.

A Formula

A.1 Significant digits formula

We compute the number of significant bits \hat{s} with probability $p_s = 0.95$ and confidence $1 - \alpha_s = 0.95$ using the `significantdigits` package² (version 0.4.0). `significantdigits` implements the Centered Normality Hypothesis approach described in [SCF⁺21]:

$$\hat{s}_i = -\log_2 \left| \frac{\hat{\sigma}_i}{\hat{\mu}_i} \right| - \delta(n, \alpha_s, p_s),$$

where $\hat{\sigma}_i$ and $\hat{\mu}_i$ are the average and standard deviation over the repetitions, and

$$\delta(n, \alpha_s, p_s) = \log_2 \left(\sqrt{\frac{n-1}{\chi^2_{1-\alpha_s/2}}} \Phi^{-1} \left(\frac{p_s+1}{2} \right) \right) \quad (4)$$

is a penalty term for estimating \hat{s}_i with probability p_s and confidence level $1 - \alpha_s$ for a sample size n . Φ^{-1} is the inverse cumulative distribution of the standard normal distribution and χ^2 is the Chi-2 distribution with $n-1$ degrees of freedom.

A.2 Extended Sørensen-Dice coefficient

The extended Sørensen-Dice coefficient is a measure of overlap between multiple sets, defined as follows:

$$\text{Dice}(A_1, A_2, \dots, A_n) = \frac{n |\bigcap_{i=1}^n A_i|}{\sum_{i=1}^n |A_i|}$$

B Cross-sectional Analysis

As a side result, the cross-sectional analysis measures the impact of numerical variability in FreeSurfer version 7.3.1 on the PPMI (Parkinson’s Progression Markers Initiative) cohort. This involves comparing the estimation of structural MRI measures, including cortical and subcortical volumes, cortical thickness, and surface area. The goal is to assess the stability of these key metrics and quantify the numerical variability.

FreeSurfer 7.3.1 showed limited numerical precision across all cortical measures: 1.61 ± 0.20 significant digits for cortical thickness, 1.33 ± 0.23 for surface area, and 1.33 ± 0.23 for cortical volume (Figures 10). Subcortical volumes have a similar precision with 1.33 ± 0.22 significant digits on average (Figure 11). These values indicate measurements are typically precise to only one decimal place, with some instances showing complete precision loss. Regional consistency was observed within each metric type, with cortical thickness showing the highest precision (range: 1.22 – 1.93 digits) compared to surface area (0.82 – 1.72 digits) and cortical volume (0.80 – 1.72 digits). Subcortical volumes exhibited the highest precision (range: 0.88 – 1.57 digits), with a mean of 1.33 ± 0.22 significant digits.

To measure the structural overlap, we evaluated using the extended Sørensen-Dice coefficient: Dice coefficients revealed substantial inter-subject variability, particularly in temporal pole regions (Figure 9). We also observed that the Dice coefficient varies across regions, with some regions showing higher variability than others with cortical volume (0.00 – 0.91) with a mean of 0.75 ± 0.11 and subcortical volume (0.18 – 0.94) with a mean of 0.82 ± 0.08 . Finally, we noticed that subcortical volume measurements are more stable than cortical volume.

²<https://github.com/verificarlo/significantdigits>

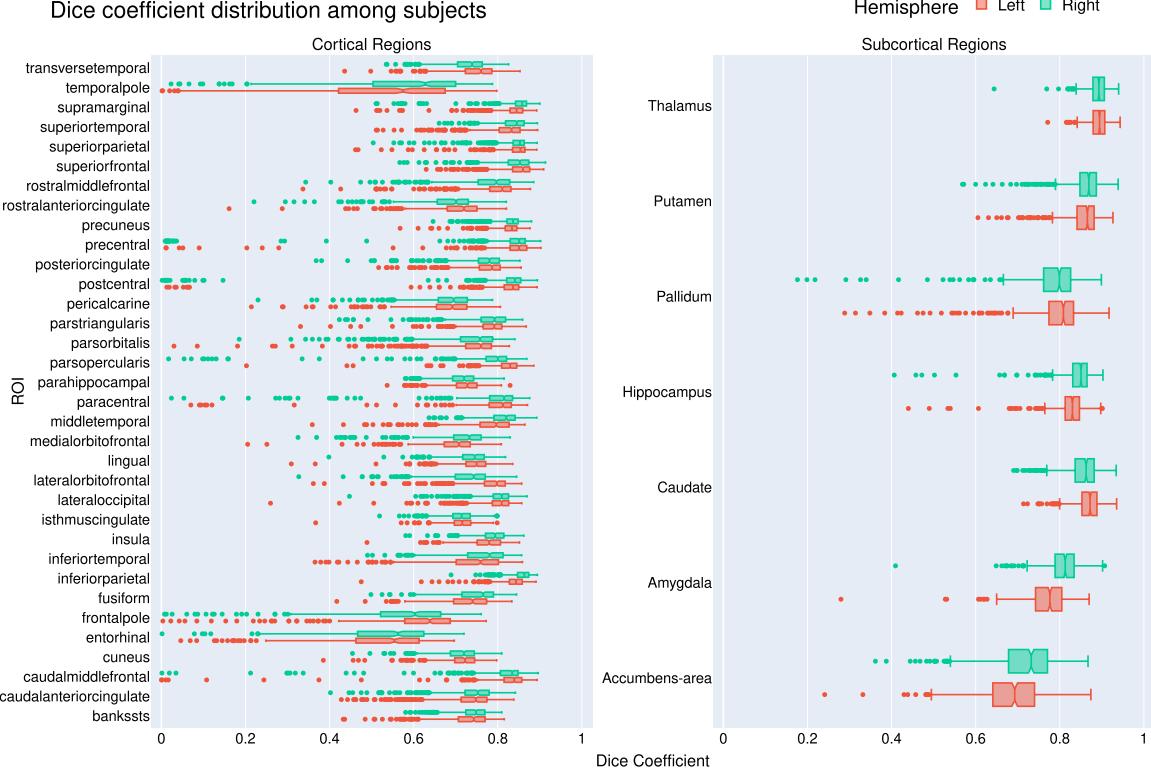


Figure 9: Dice coefficient.

B.1 Significant digits average across all subjects

Table 2: Significant digits average across all subjects.

Region	cortical thickness		surface area		cortical volume	
	lh	rh	lh	rh	lh	rh
bankssts	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
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
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

Continued on next page

Significant digits distribution among subjects

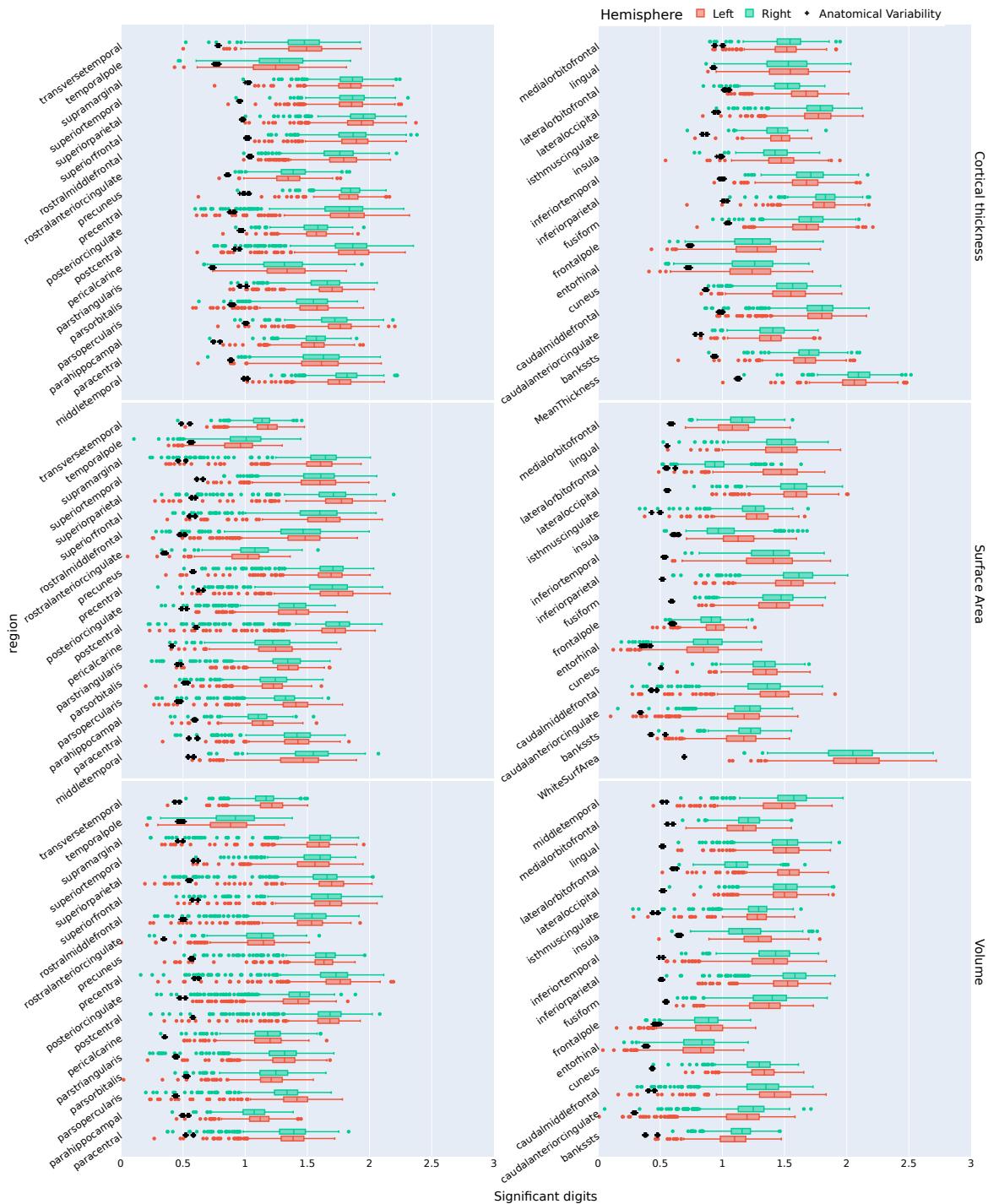


Figure 10: Number of significant digits for each cortical region and metric.

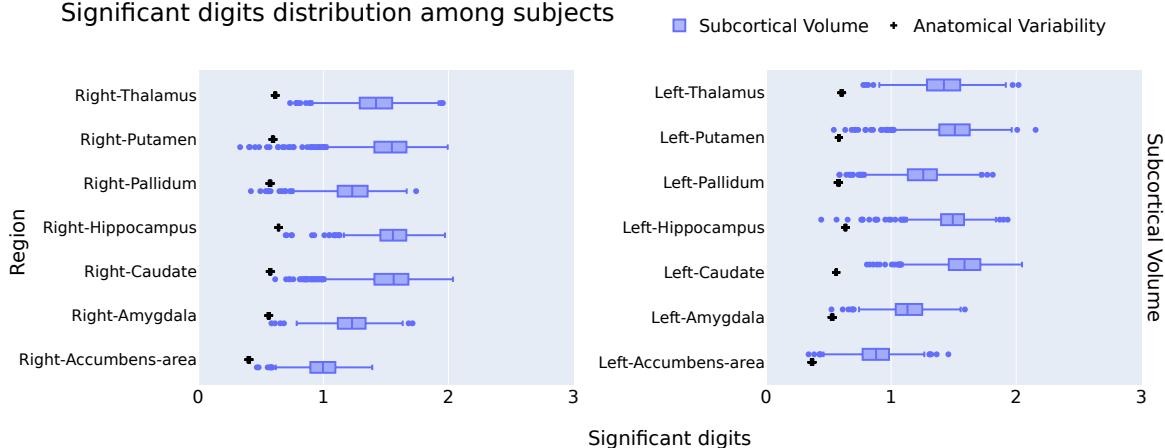


Figure 11: Number of significant digits of subcortical volume for each subcortical region.

Table 2: Significant digits average across all subjects. (Continued)

Region	cortical thickness		surface area		cortical volume	
	lh	rh	lh	rh	lh	rh
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 3: 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

Continued on next page

Table 3: 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
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
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 4: Significant digits and standard-deviation average 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

Table 5: 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		

Status	Cohort	HC	PD-non-MCI	PD-MCI
Before QC	n	106	181	29
	Age (y)	60.6 ± 10.2	61.7 ± 9.6	67.7 ± 7.7
	Age range	30.6 – 84.3	36.3 – 83.3	49.9 – 80.5
	Gender (male, %)	58 (54.7%)	119 (65.7%)	–
	Education (y)	16.6 ± 3.3	15.9 ± 2.9	–
After QC	n	103	175	27
	Age (y)	60.7 ± 10.3	61.4 ± 9.5	67.8 ± 7.9
	Age range	30.6 – 84.3	36.3 – 79.9	49.9 – 80.5
	Gender (male, %)	57 (55.3%)	114 (65.1%)	20 (74.1%)
	Education (y)	16.6 ± 3.3	15.9 ± 2.9	15.0 ± 3.5
After MCI exclusion	n	103	121	–
	Age (y)	60.7 ± 10.3	60.7 ± 9.1	–
	Age range	30.6 – 84.3	39.2 – 78.3	–
	Gender (male, %)	57 (55.3%)	80 (66.1%)	–
	Education (y)	16.6 ± 3.3	16.1 ± 3.0	–
	UPDRS III OFF baseline	–	23.4 ± 10.1	–
	UPDRS III OFF follow-up	–	25.8 ± 11.1	–
	Duration T2 - T1 (y)	1.4 ± 0.5	1.4 ± 0.7	–

Abbreviations: MCI = Mild Cognitive Impairment; UPDRS = Unified Parkinson’s Disease Rating Scale; PD = Parkinson’s disease. Descriptive statistics before and after quality control (QC). Values are expressed as mean ± standard deviation. PD-non-MCI longitudinal sample is a subsample of the PD-non-MCI original sample that had longitudinal data and disease severity scores available.

C Numerical-Anatomical Variability Ratio (ν_{nav})

C.1 ν_{nav} maps

Figures 12 and 13 show the ν_{nav} maps for cortical surface area and volume, respectively. The maps show the average ν_{nav} values across all subjects for each cortical region. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav} values. The maps provide a visual representation of the variability in the ν_{nav} values across different cortical regions, highlighting regions with higher or lower ν_{nav} values.

C.2 Consistency results

C.2.1 Consistency of statistical tests

Figures 14 and 15 show the consistency of statistical tests for cortical area and volume, respectively, across all subjects and regions. The plots show 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. The red triangles indicate the IEEE-754 run for reference.

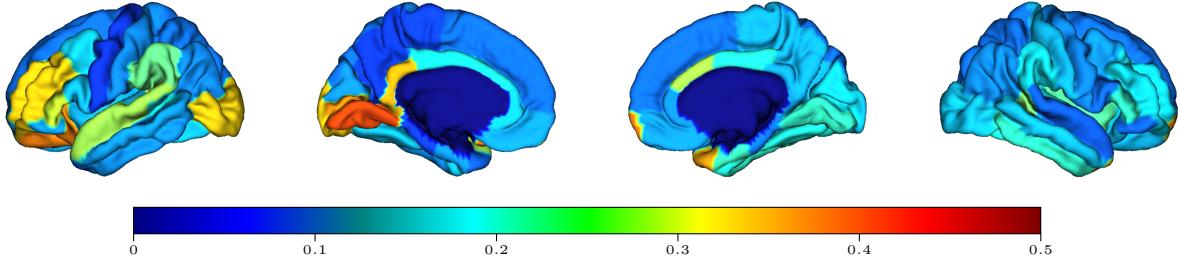


Figure 12: Numerical-Anatomical Variability Ratio (ν_{nav}) for cortical surface area across regions and groups. Higher ν_{nav} values indicate greater computational uncertainty relative to biological variation. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav} values.

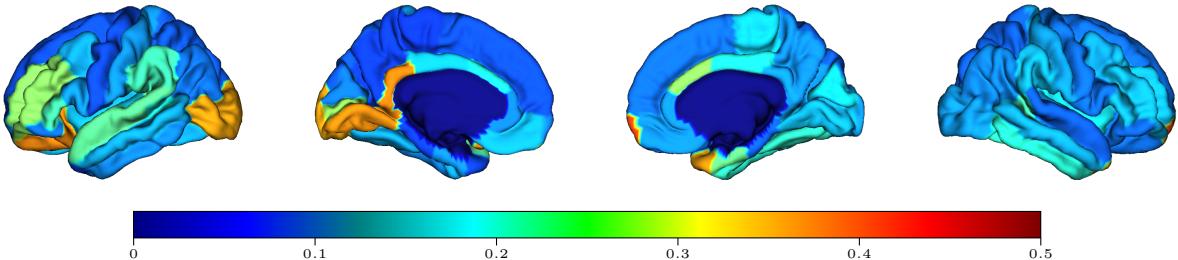


Figure 13: Numerical-Anatomical Variability Ratio (ν_{nav}) for cortical volume across regions and groups. Higher ν_{nav} values indicate greater computational uncertainty relative to biological variation. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav} values.

C.2.2 Distribution of statistical tests coefficients

Figures 16 and 17 show the distribution of partial correlation coefficients for cortical area and volume, respectively, across all subjects and regions. The 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.

C.2.3 Thresholding existing Cohen's d values from the literature

We applied a thresholding approach to the Cohen's d values reported in the literature to identify the most relevant findings for our analysis. This involved setting a minimum effect size threshold, below which results were considered non-significant or uninformative. The threshold was determined based on the distribution of Cohen's d values across studies, with a focus on retaining only those effects that were robust and consistent.

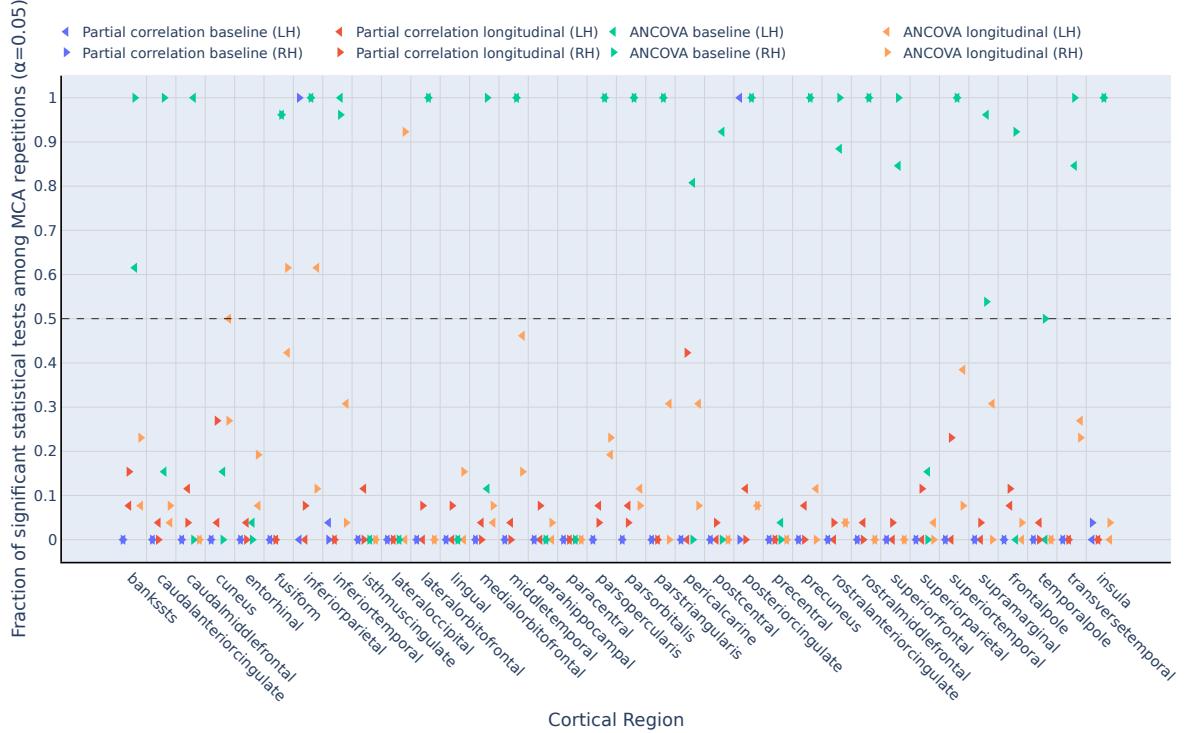


Figure 14: Consistency of statistical tests for cortical area across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency than others.

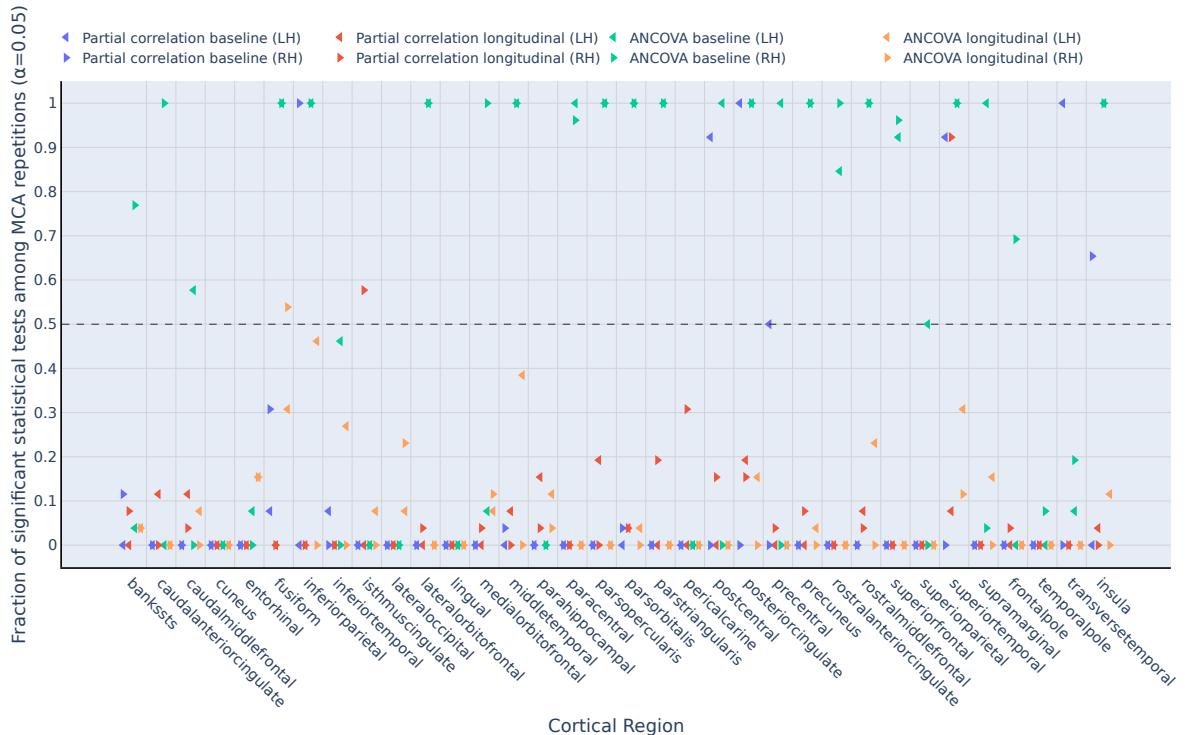


Figure 15: Consistency of statistical tests for cortical volume across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency than others.

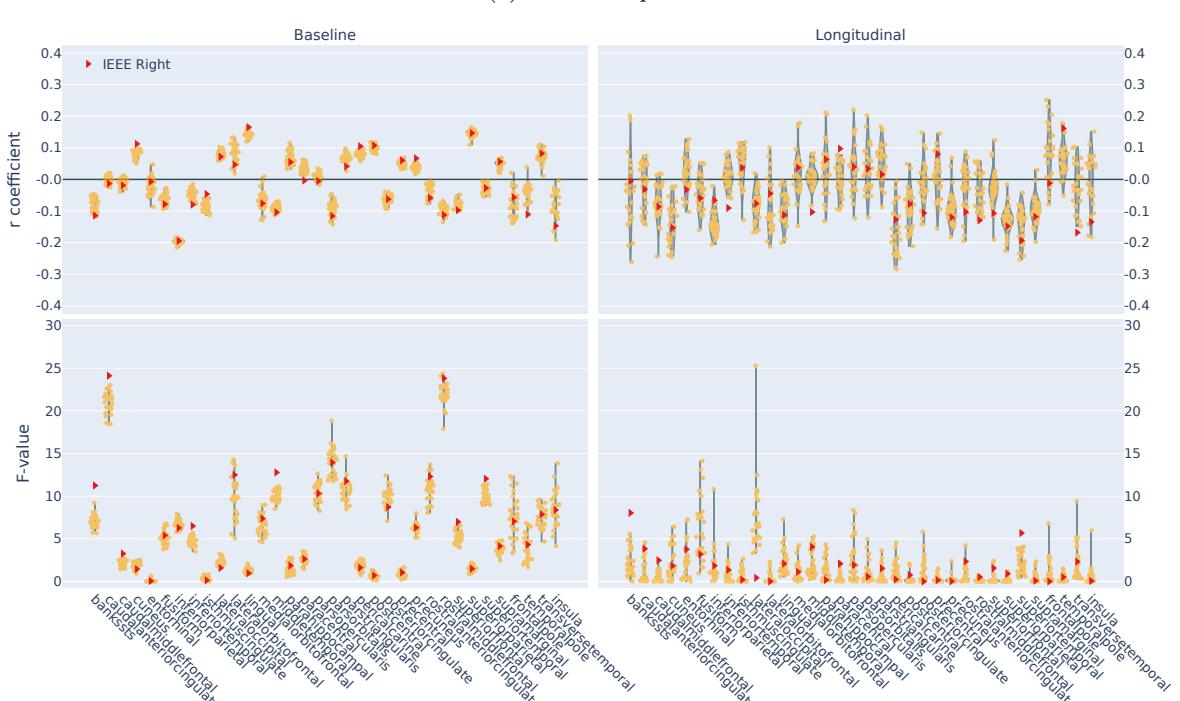
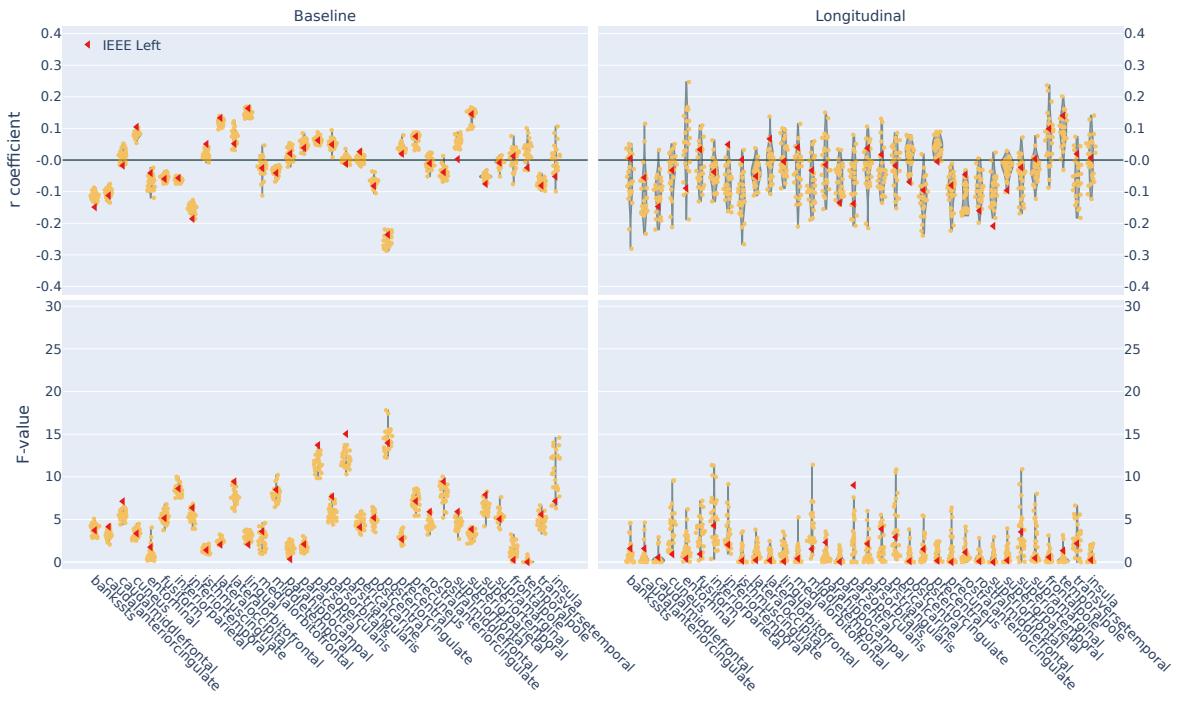
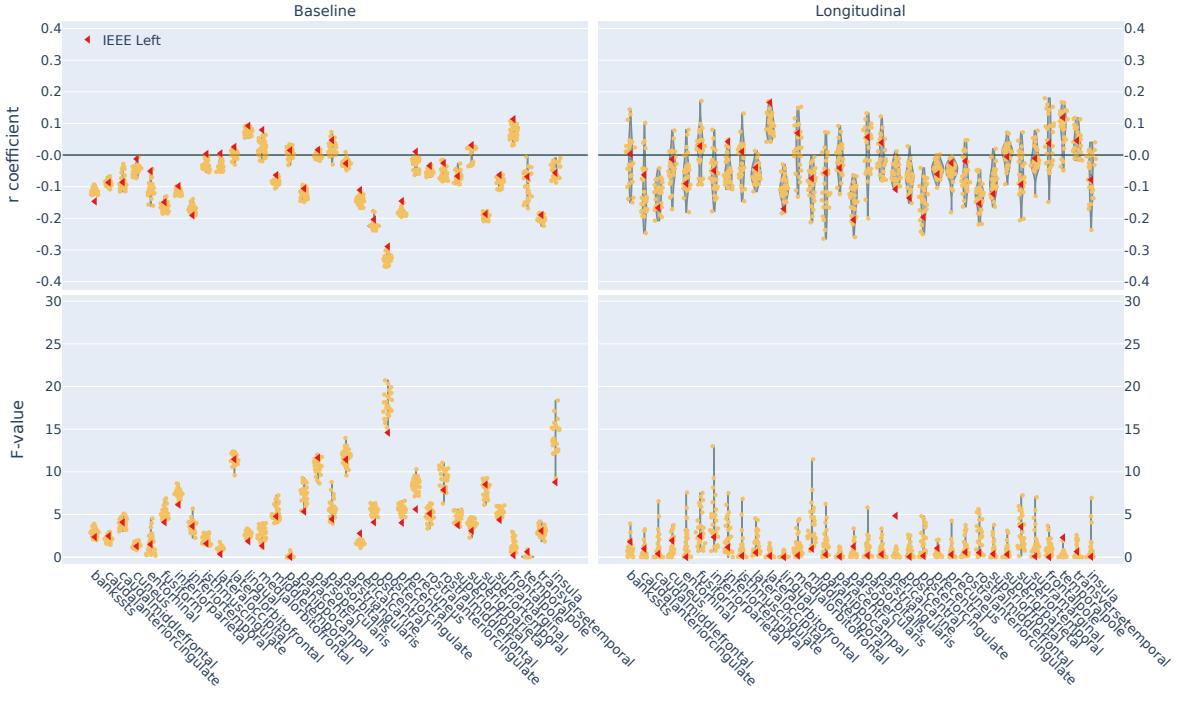
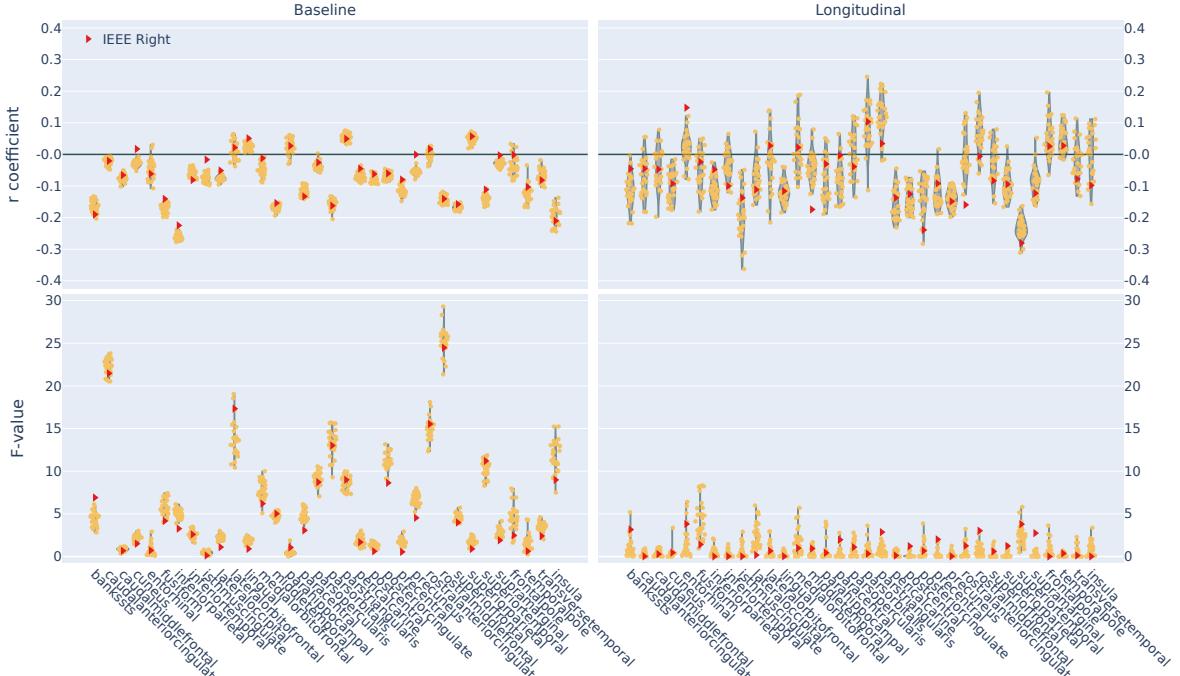


Figure 16: Distribution of partial correlation coefficients for cortical area across all subjects and regions. Red triangles indicate the IEEE-754 run for reference. The distribution shows the variability in the coefficients, with some regions exhibiting higher consistency than others.



(a) Left hemisphere



(b) Right hemisphere

Figure 17: 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.

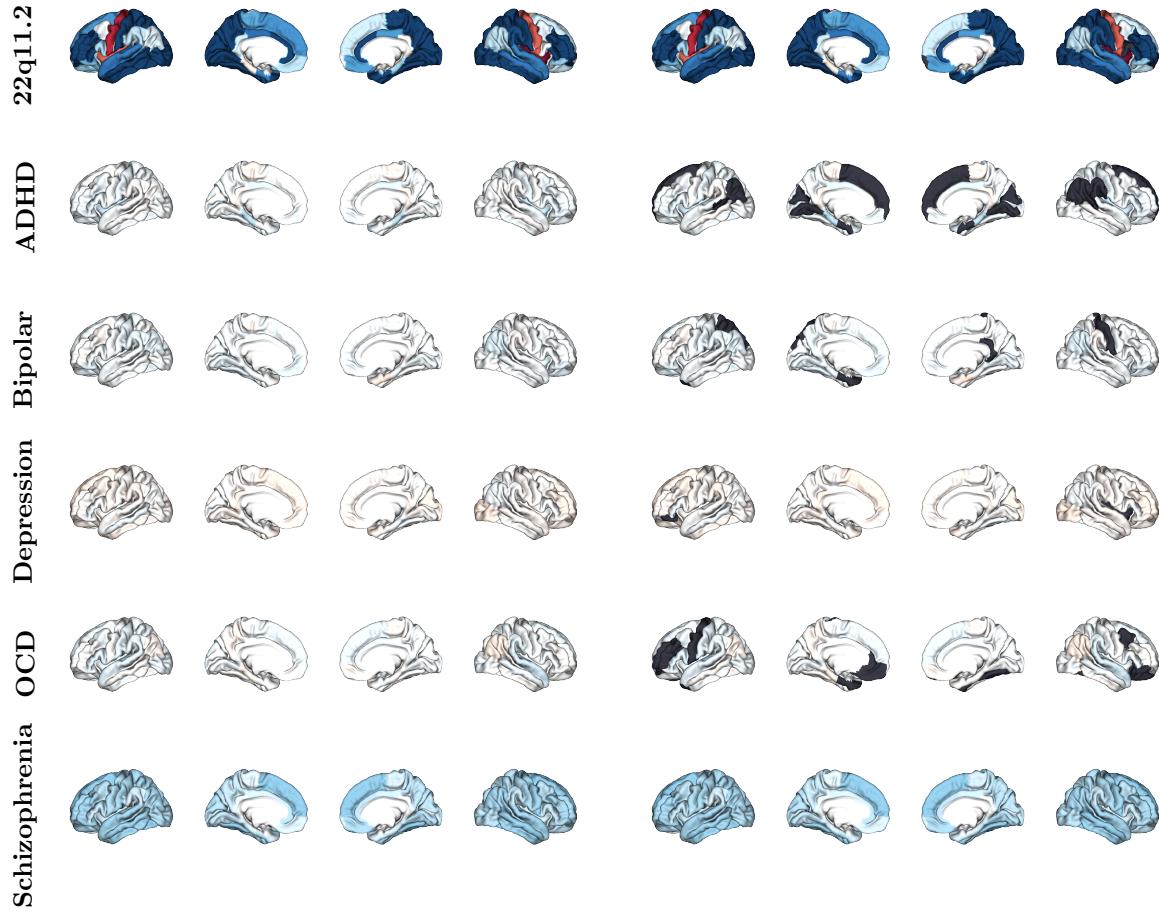


Figure 18: ENIGMA cortical area Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.