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, arising from floating-point rounding and computational effects, is rarely quantified in neuroimaging despite many biomarkers relying on subtle morphometric differences. We instrumented FreeSurfer, a widely used neuroimaging pipeline, with Monte Carlo Arithmetic to emulate numerical differences across computational environments, processing identical MRI scans from Parkinson's disease patients and controls multiple times. We introduce the Numerical-Anatomical Variability Ratio (NAVR), quantifying numerical variation within subjects relative to anatomical differences between subjects. In multiple cortical and subcortical regions, numerical variation reached nearly one-third of the anatomical signal, altering statistical conclusions about group differences and clinical correlations across repeated analyses. Applying NAVR-based thresholding to published ENIGMA consortium brain maps revealed that numerous small regional effects lie below the numerical noise threshold for typical study sizes. Quantifying numerical uncertainty prevents spurious biomarker claims and enhances reproducibility in computational neuroscience.

1 Introduction

Complex neuroimaging pipelines transform raw MRI data into putative biomarkers such as cortical thickness and subcortical volume. Because disease-related effects may be subtle, the reliability of these metrics depends critically on computational stability. However, numerical variability arising from software computations is rarely quantified.

Floating-point arithmetic can cause ostensibly deterministic algorithms to yield subtly different outputs. Whether such numerical noise significantly impacts biological conclusions remains unclear. Preliminary studies suggest substantial effects [4, 8, 7, 1, 15, 13], yet systematic quantification within clinical datasets is lacking.

Parkinson's disease (PD) represents an ideal context to investigate numerical variability, as the identification of reliable structural biomarkers is highly desirable for diagnosis, tracking disease progression, and developing targeted therapies. Structural MRI is particularly attractive for biomarker discovery due to its non-invasive nature, but subtle anatomical changes associated with PD pose significant analytical challenges. Consequently, ensuring that MRI-derived measures are robust to numerical variability is essential to the successful translation of these biomarkers into clinical practice.

We instrumented FreeSurfer 7.3.1 [5] with Monte Carlo arithmetic [9], reprocessing identical MRI scans of PD patients and healthy controls across multiple perturbed numerical states. To systematically quantify the impact, we introduce the Numerical-Anatomical Variability Ratio (NAVR), directly comparing numerical variation within individuals to anatomical differences between individuals.

We find that in several cortical and subcortical regions, NAVR approaches 0.3, indicating that numerical noise can represent nearly one-third of the measured biological variation. Consequently, statistical inferences regarding group differences and clinical correlations can fluctuate significantly. Applying NAVR-based thresholding to published ENIGMA [16] consortium brain maps reveals that numerous previously reported small regional effects likely fall below computational noise thresholds.

Our findings demonstrate that numerical variability is a significant, quantifiable source of uncertainty in neuroimaging studies. The NAVR framework we propose offers researchers a practical approach to quantify and report computational uncertainty, enhancing reproducibility and reliability in clinical neuroscience.

2 Impacts of numerical variability

Floating-point rounding errors, although individually small, can collectively introduce substantial numerical variability. Monte Carlo Arithmetic (MCA) systematically assesses this variability by perturbing floating-point operations with random, zero-mean offsets that preserve mathematical expectations. Repeatedly executing an analysis pipeline with MCA enables us to estimate the variability of results that could arise from routine differences across hardware, software libraries, or computational environments. Unlike ad-hoc comparisons across platforms, MCA specifically isolates numerical effects, free from algorithmic or software-version confounders.

Parkinson's disease provides an ideal scenario for testing numerical stability, as structural MRI-derived measurements such as cortical thickness and subcortical volume typically show subtle group differences ($|d| \approx 0.2-0.4$), yet hold significant promise as potential biomarkers for clinical progression. Furthermore, the availability of large multi-site datasets (e.g., PPMI) and established region-of-interest analyses ensure PD is a thoroughly characterized context for evaluating the impacts of numerical variability. Thus, PD allows us to directly assess whether numerical noise is sufficiently large to obscure clinically meaningful anatomical signals.

We instrumented FreeSurfer 7.3.1 with MCA (virtual precision set to 53 bits for double precision, 24 bits for single precision), and re-analyzed each T1-weighted MRI scan 26 times. The within-subject variability due to numerical differences reached approximately 30% of the observed between-subject anatomical variance in critical cortical and subcortical regions (NAVR $\simeq$ 0.3; Extended Data Fig. ED1). The median NAVR of approximately 0.18 across all cortical areas underscores that numerical uncertainty significantly affects typical neuroimaging analyses.

This numerical instability fundamentally arises from floating-point arithmetic—where subtle rounding errors accumulate over complex calculations. By repeatedly processing the same MRI scans under perturbed computational conditions, we found significant variations directly impacting statistical conclusions. These fluctuations pose a major concern for interpreting subtle anatomical changes in PD, underscoring the need for robust numerical precision in biomarker discovery.

2.1 Numerical variability alters statistical inference

We leveraged Monte Carlo Arithmetic (MCA) to simulate realistic variations that occur naturally between computational environments (e.g., hardware, libraries). Parkinson's disease, characterized by subtle anatomical differences detectable through structural MRI, provides a stringent scenario to evaluate whether numerical instability meaningfully impacts clinical inferences.

We instrumented FreeSurfer 7.3.1 with MCA and reprocessed each MRI scan 26 times. Within-subject numerical variability in cortical thickness and subcortical volume reached [XX] mm and [YY] mm^3, respectively, representing approximately 30% of the total anatomical variability observed between subjects in key regions (NAVR $\simeq$ 0.3; Extended Data Fig. ED1).

Figure [2a] shows that statistical significance in group comparisons between Parkinson's disease patients and controls fluctuated notably across MCA replicates. Some cortical regions alternated between significant and non-significant outcomes ($p < 0.05$), illustrating how numerical instability directly affects clinical interpretation. Similarly, the correlation between regional cortical thickness and disease severity (UPDRS scores) also varied substantially across numerical perturbations (Fig. [2b]), suggesting that apparent biomarker relationships might emerge or vanish purely due to numerical noise.

This variability mirrors multi-team analytical differences seen in previous neuroimaging reproducibility challenges (e.g., NARPS [2]). Importantly, the original IEEE-754 results consistently lie within the range sampled by MCA, confirming that observed variability represents realistic computational scenarios rather than methodological artifacts.

2.2 A framework to quantify the impact of numerical variability

To quantify when numerical variability becomes scientifically concerning, we introduce the Numerical-Anatomical Variability Ratio (NAVR), defined as the ratio of numerical uncertainty to between-subject anatomical variability:

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

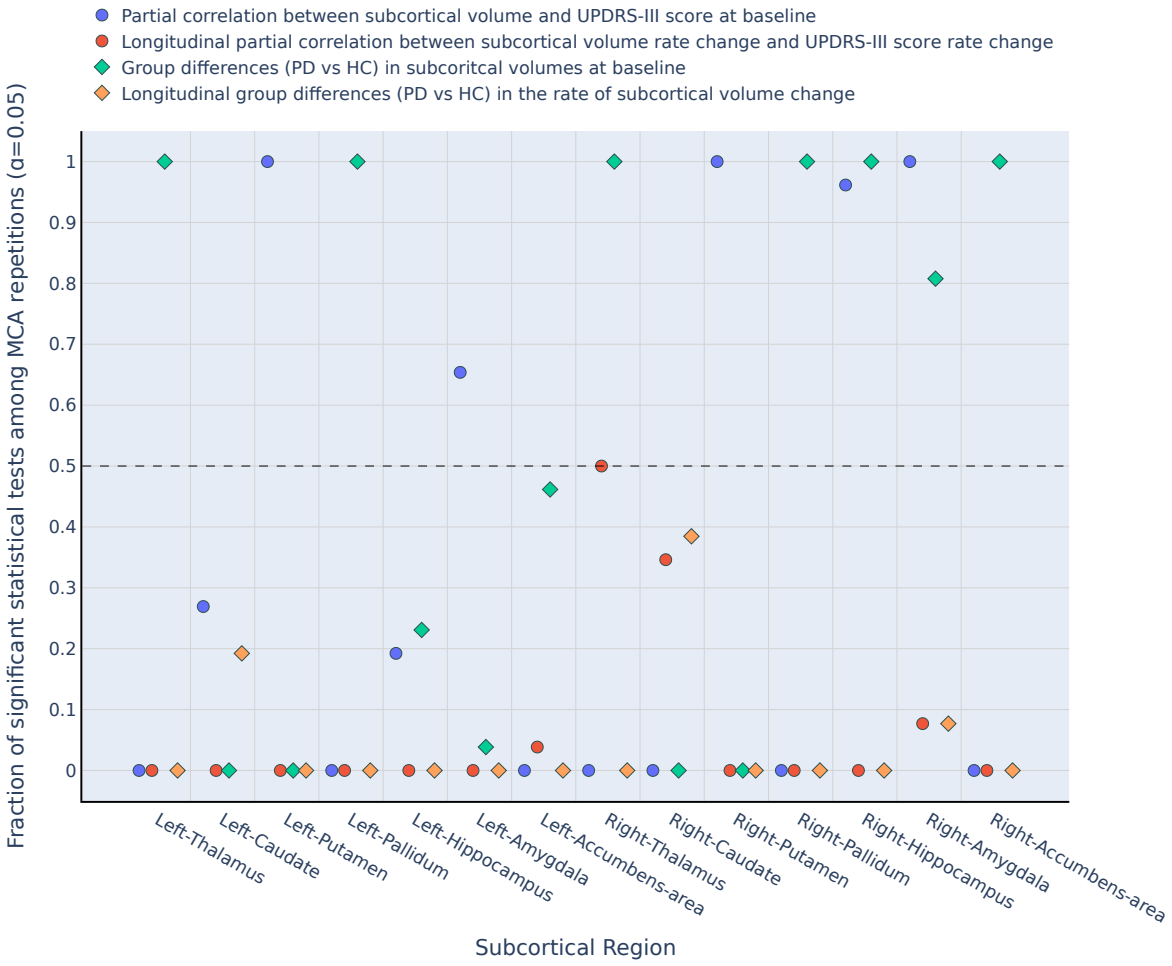


Figure 1: Proportion of significant tests ($p < 0.05$) for subcortical volumes across 26 numerical perturbations. measures.

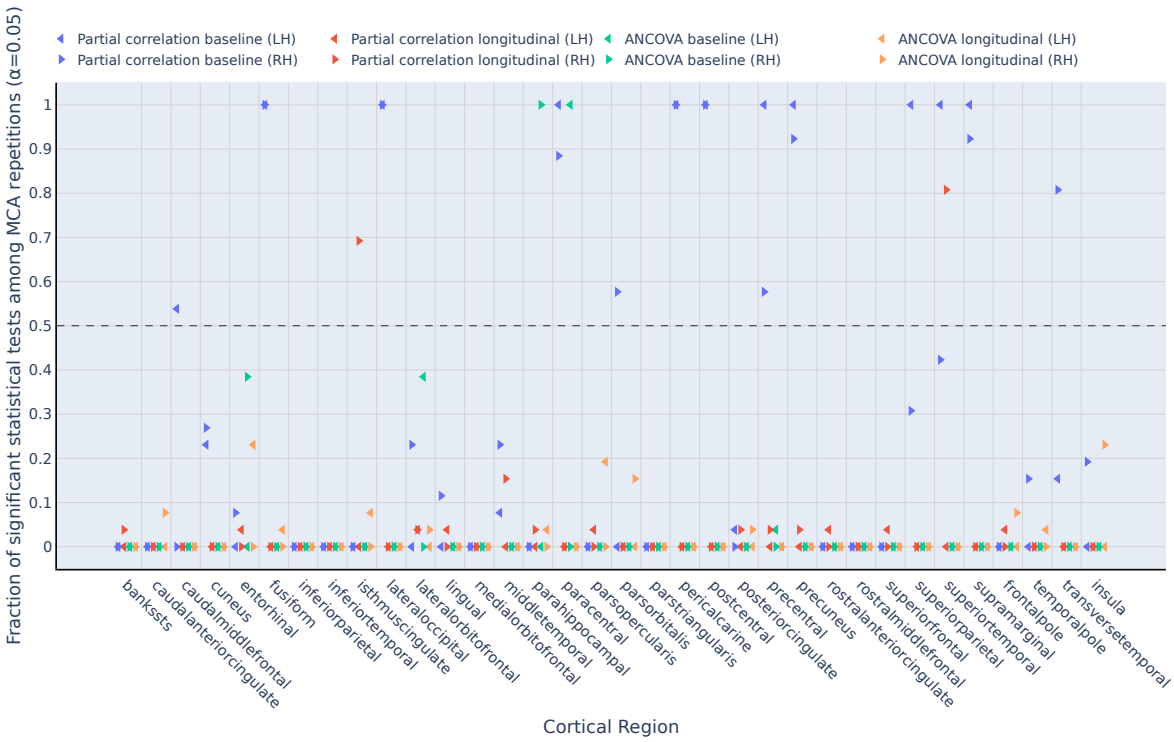


Figure 2: Proportion of significant tests ($p < 0.05$) for cortical thickness across 26 numerical perturbations. measures.

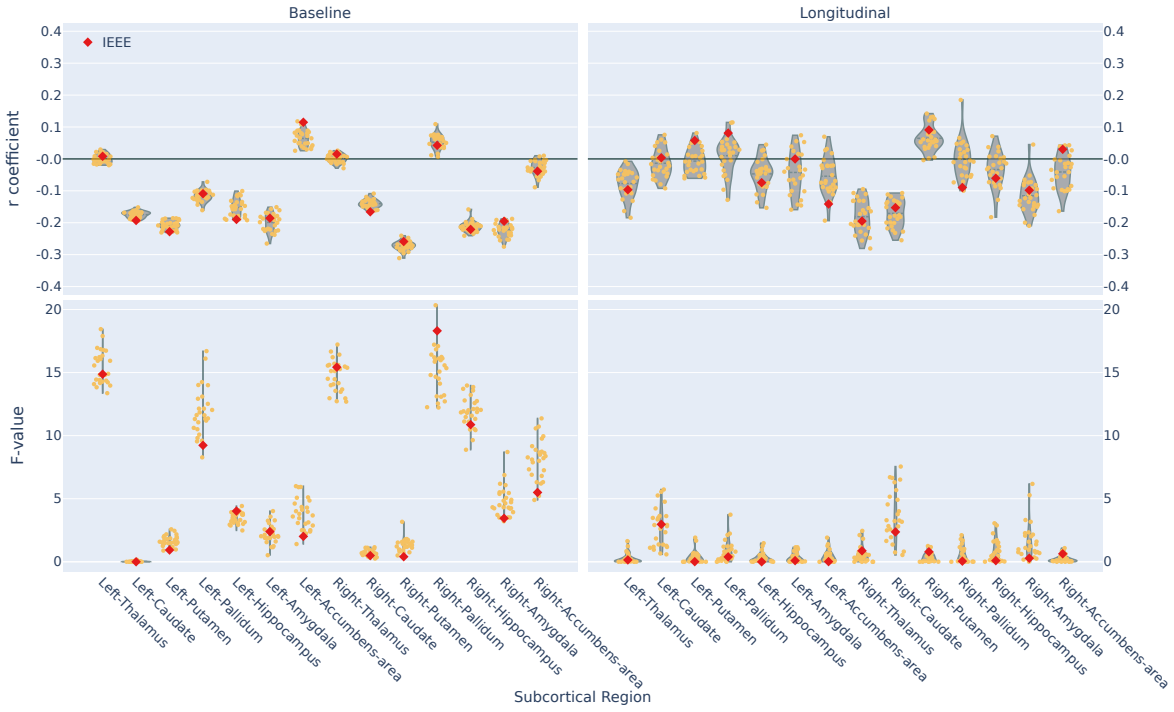
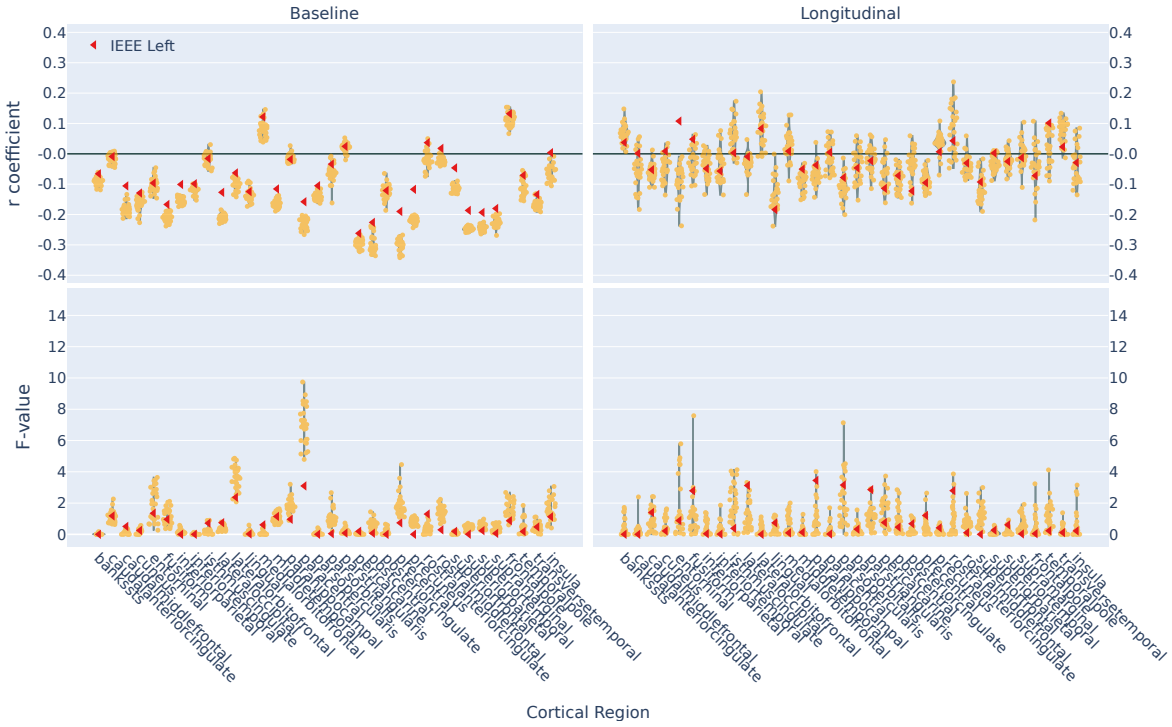
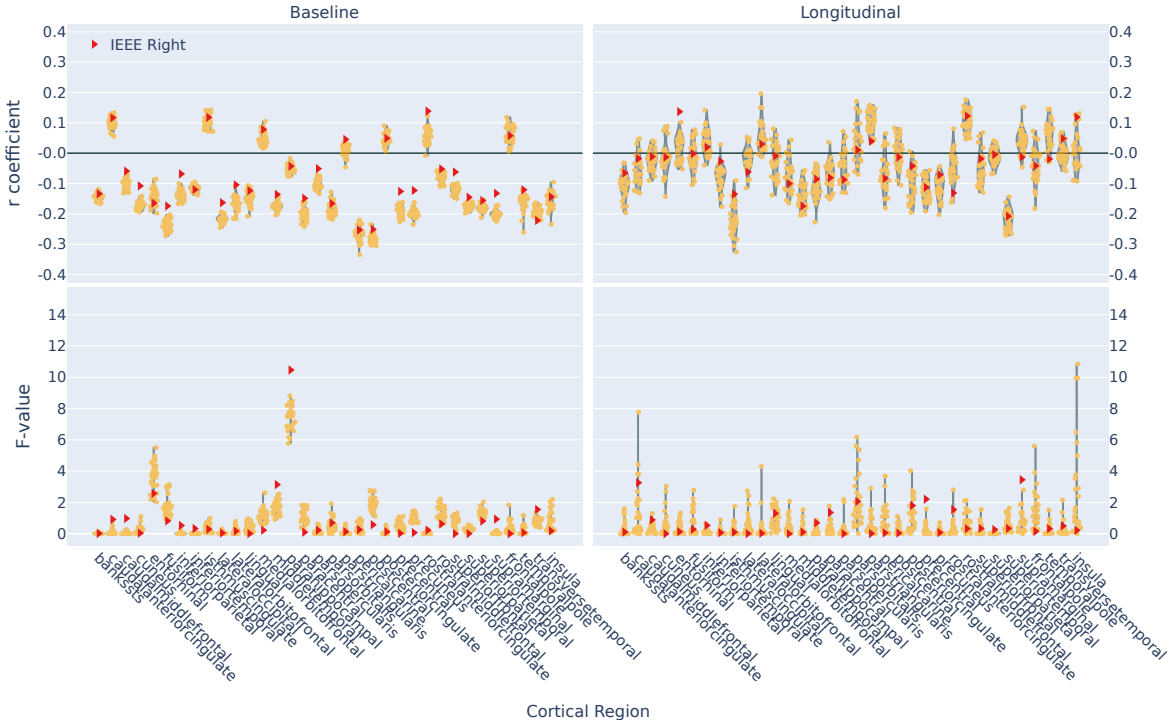


Figure 3: Distribution of partial correlation coefficients (r-values) and F-statistics from ANCOVA across MCA repetitions for subcortical volume measures. Red dots represent the IEEE results. The top row shows r-values, while the bottom row shows F-values. The left column represents baseline analysis,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



(a) Left hemisphere



(b) Right hemisphere

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.

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

Quantifying numerical variability systematically is crucial for reliably interpreting neuroimaging results. NAVR enables researchers to rapidly assess the stability of their measurements relative to biological variation, facilitating the identification of findings potentially obscured by computational noise.

Figures 5b and 5a visualize NAVR across cortical thickness and subcortical volume regions, highlighting areas particularly susceptible to numerical instability. Higher NAVR values indicate greater vulnerability of observed results to computational noise rather than true biological variability.

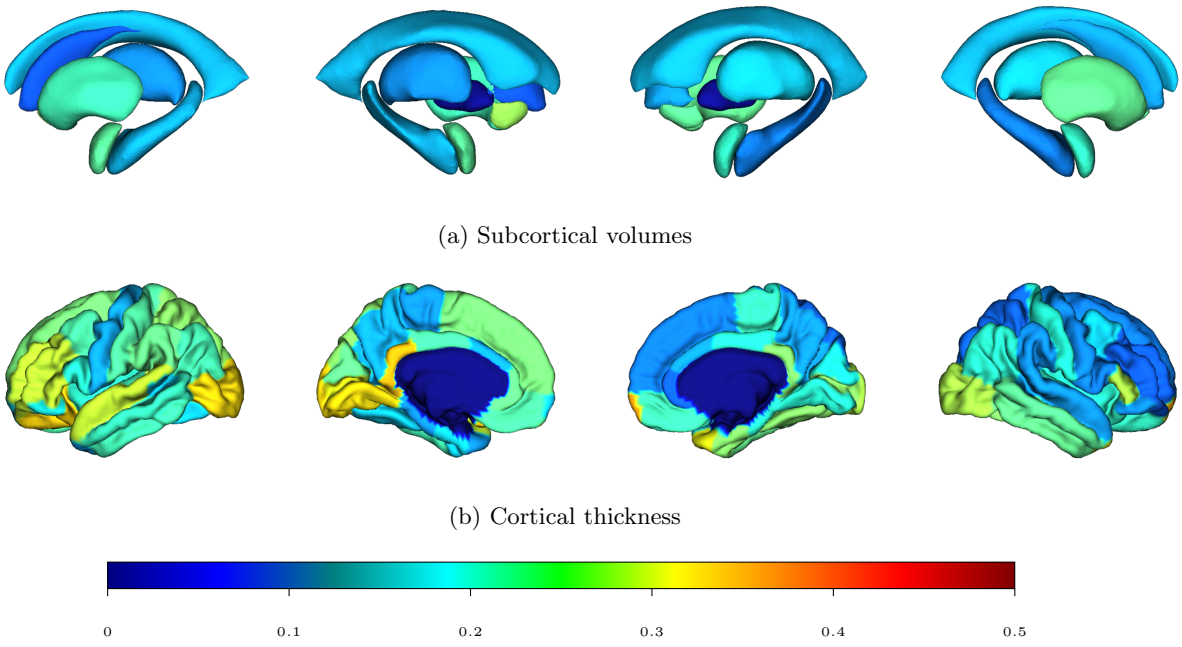
As exhaustive numerical variability assessments can be computationally demanding, we developed an analytical approach for estimating NAVR. This accelerates evaluations of numerical precision and enables rapid re-analysis of existing findings.

We also derived an analytical formula linking NAVR directly to uncertainty in Cohen's d, a widely-used statistical effect size measure:

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

This relationship allows direct estimation of the sample size required for reliable effect size estimation given specific numerical variability. For example, a NAVR of 0.2 implies that a sample size of approximately 1500 participants would be necessary to ensure that numerical variability contributes no more than 0.01 uncertainty to Cohen's d.

To facilitate broad adoption, we provide an online tool[1] to rapidly assess and threshold neuroimaging results based on NAVR values, supporting identification of potentially unreliable findings.



(a) Subcortical volumes

(b) Cortical thickness

Figure 5: 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.

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

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

To assess the broader implications of numerical variability, we applied NAVR to re-evaluate influential findings from the ENIGMA consortium, which has substantially shaped our understanding of psychiatric and neurological disorders. Applying NAVR-based thresholding to ENIGMA's published brain maps identified multiple regions where reported effect sizes fell below the computational noise floor, potentially calling into question their reliability.

Figures 7 and 8 illustrate the impact of applying NAVR thresholds to cortical thickness and subcortical volume maps from ENIGMA. Regions rendered in black indicate areas where reported effect sizes were smaller than numerical variability, suggesting these findings should be interpreted with caution.

This observation highlights potential risks of overestimating small effects and underscores the importance of systematically accounting for numerical uncertainty in neuroimaging research. While ENIGMA's primary findings generally remained robust due to large sample sizes, our analysis indicates that numerous secondary, smaller-scale effects reported in the literature could be compromised by numerical instability.

3 Discussion

Our systematic perturbation of FreeSurfer revealed that numerical variability alone can account for up to 30% of the anatomical variability observed in structural MRI measurements. This level of uncertainty can significantly impact statistical outcomes, leading to the appearance or disappearance of clinically relevant group differences or correlations depending solely on computational conditions. These findings offer a mechanistic explanation for some of the reproducibility challenges reported in clinical neuroimaging.

To move beyond identifying the problem, we introduced the Numerical-Anatomical Variability Ratio (NAVR), a quantitative framework for assessing the relative magnitude of computational noise. By establishing a theoretical link between NAVR and the uncertainty in Cohen's d effect sizes, we provide researchers with a practical tool to assess the robustness of their findings. Our re-analysis of published ENIGMA results illustrates this utility: while large sample sizes confer robustness to core findings, many secondary effects fall below the computational noise floor. This suggests that in smaller exploratory studies, numerical instability may undermine the reliability of reported effects.

Although our primary analysis focused on FreeSurfer 7.3.1 and Parkinson's disease, the underlying numerical issues are general. Floating-point arithmetic is inherently non-associative and sensitive to compiler behavior, hardware architecture, and thread scheduling. As a result, neuroimaging pipelines—though deterministic in design—can produce divergent results across computational environments. Preliminary analyses of FSL YC: cite Niusha and ANTs YC: cite Mathieu indicate that such instability is not unique to FreeSurfer, but likely pervades the field. SPM however seems to be less impacted by numerical variability. Moreover, our PD cohort was relatively homogeneous in age and phenotype, potentially reducing anatomical variance and inflating NAVR values. This highlights the need to apply NAVR across diverse datasets, software packages, and disease contexts to build a comprehensive understanding of computational reliability.

Crucially, this instability is not confined to low-level rounding operations. Image processing workflows involve nonlinear optimization procedures that may converge to different local minima under small perturbations, resulting in substantive changes to derived measures. The situation is analogous to deep learning, where different weight initializations or precision settings can yield distinct model outcomes. In neuroimaging, such instability means that even identical inputs can lead to divergent interpretations—raising serious concerns for both research reproducibility and clinical translation.

NAVR offers a scalable and interpretable metric to quantify this hidden variability. While floating-point rounding is a major source of instability, future work should extend this analysis to other contributors, including algorithmic decisions, preprocessing choices, and data handling practices. A comprehensive understanding of these factors is essential for developing numerically robust software tools.

In conclusion, our results demonstrate that computational uncertainty is as critical as statistical uncertainty in neuroimaging. Incorporating systematic assessments of numerical variability, through tools like NAVR, is necessary to ensure the reproducibility and reliability of neuroimaging-based biomarkers.

4 Methods

4.1 Numerical variability assessment

We employed Monte Carlo Arithmetic (MCA) [9] 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 [13], which extends MCA to mathematical library functions (`exp`, `log`, `sin`, `cos`) through Verificarlo [3], 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.

We processed each visit with FreeSurfer 7.3.1. To sample numerical variability we compiled FreeSurfer with Fuzzy-libm an implementation of Monte Carlo arithmetic (MCA) that injects zero-mean rounding noise into every elementary function call. Virtual precision was set to 53 bits for operations promoted to double and 24 bits for single precision, thereby preserving IEEE-754 expectations but exposing the variance of alternative execution paths. Each subject-visit pair was processed 26 times; failed or quality-control-flagged runs were discarded, and exactly 26 successful runs per pair were retained for analysis.

4.2 Participants

We used structural MRI data from the Parkinson's Progression Markers Initiative (PPMI). Participants included 125 Parkinson's disease patients without mild cognitive impairment (PD-non-MCI) and 106 healthy controls, each providing longitudinal T1-weighted MRI data across two visits. Patients with mild cognitive impairment were excluded to reduce confounding influences.

T1-weighted images were drawn from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org). Inclusion required (i) diagnosis of idiopathic Parkinson's disease (PD-non-MCI) or healthy control (HC); (ii) two usable visits separated by 0.9-2.0 years; and (iii) absence of other neurological disorders. The final dataset comprised 90 healthy controls and 118 PD-non-MCI participants (Extended Data Table 1). The study was approved by the local research ethics boards of all contributing centres, and written informed consent was obtained from every participant.

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 ± 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 scans from PPMI were acquired using standardized protocols (repetition time=2.3 s, echo time=2.98 ms, inversion time=0.9 s, 1 mm isotropic resolution, 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. Images underwent standard preprocessing using FreeSurfer 7.3.1 instrumented with Fuzzy-libm. Each participant's MRI data were processed 26 times under different numerical perturbations to quantify numerical variability. Failed runs were discarded, ensuring exactly 26 successful repetitions per subject.

Longitudinal processing followed the standard FreeSurfer stream [12]: cross-sectional processing of both timepoints, followed by creation of an unbiased within-subject template [10] using robust registration [11]. Downstream analyses used unperturbed FreeSurfer to prevent additional numerical perturbations.

4.4 Numerical Variability Assessment

Numerical variability was quantified using Monte Carlo Arithmetic (MCA), which systematically introduces controlled rounding perturbations to floating-point operations, simulating differences in computational environments. Perturbations were applied using Fuzzy-libm, an MCA implementation that introduces rounding variability into standard mathematical functions (`exp`, `log`, `sin`, `cos`, ...), integrated through the Verificarlo compiler. Virtual precision was set at 53 bits for double precision and 24 bits for single precision, closely simulating real-world computational variability.

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

To quantify computational stability relative to anatomical 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:

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

Anatomical variability captures inter-subject differences:

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

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, $\varepsilon_{ij}^{(\text{anat})} \sim \mathcal{N}(0, \sigma^2_{\text{anat}})$ captures anatomical variation, and $\varepsilon_{ij}^{(\text{num})} \sim \mathcal{N}(0, \sigma^2_{\text{num}})$ represents numerical uncertainty.

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

$$\sigma_d = \frac{2}{\sqrt{N}} \nu_{\text{nav}} \tag{3}$$

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

All MCA instrumentation scripts, FreeSurfer build instructions and analysis notebooks are available at [GitHub URL to be inserted]. Exact commit hashes are archived on Zenodo (DOI [to be added]) to ensure bit-level reproducibility.

7 Acknowledgements

The analyses were conducted on the Virtual Imaging Platform [6], 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

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[6] 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.

[7] 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.

[8] 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.

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

[10] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *Neuroimage*, 57(1):19–21, 2011.

[11] Martin Reuter, H Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4):1181–1196, 2010.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

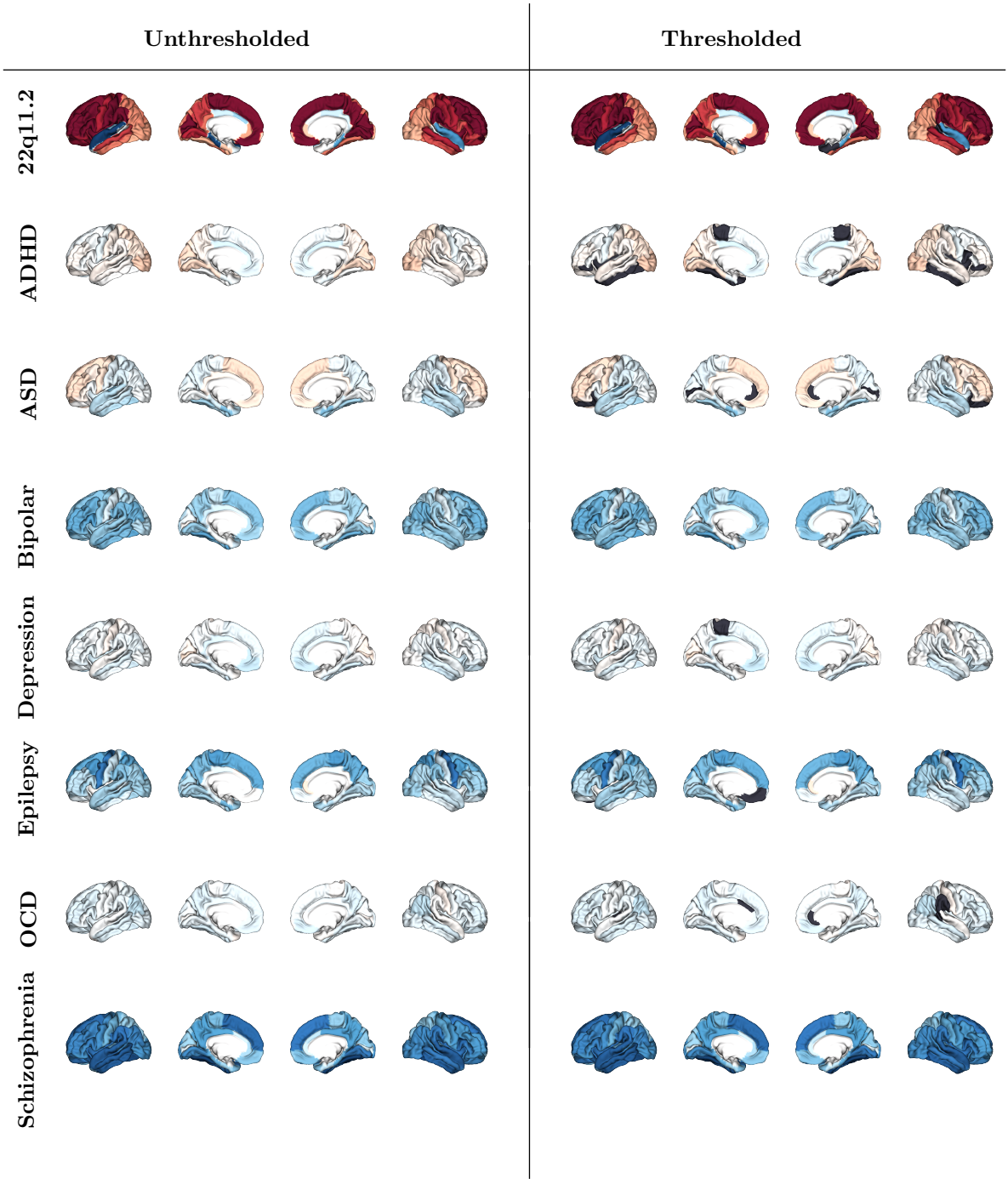


Figure 6: ENIGMA cortical thickness Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions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

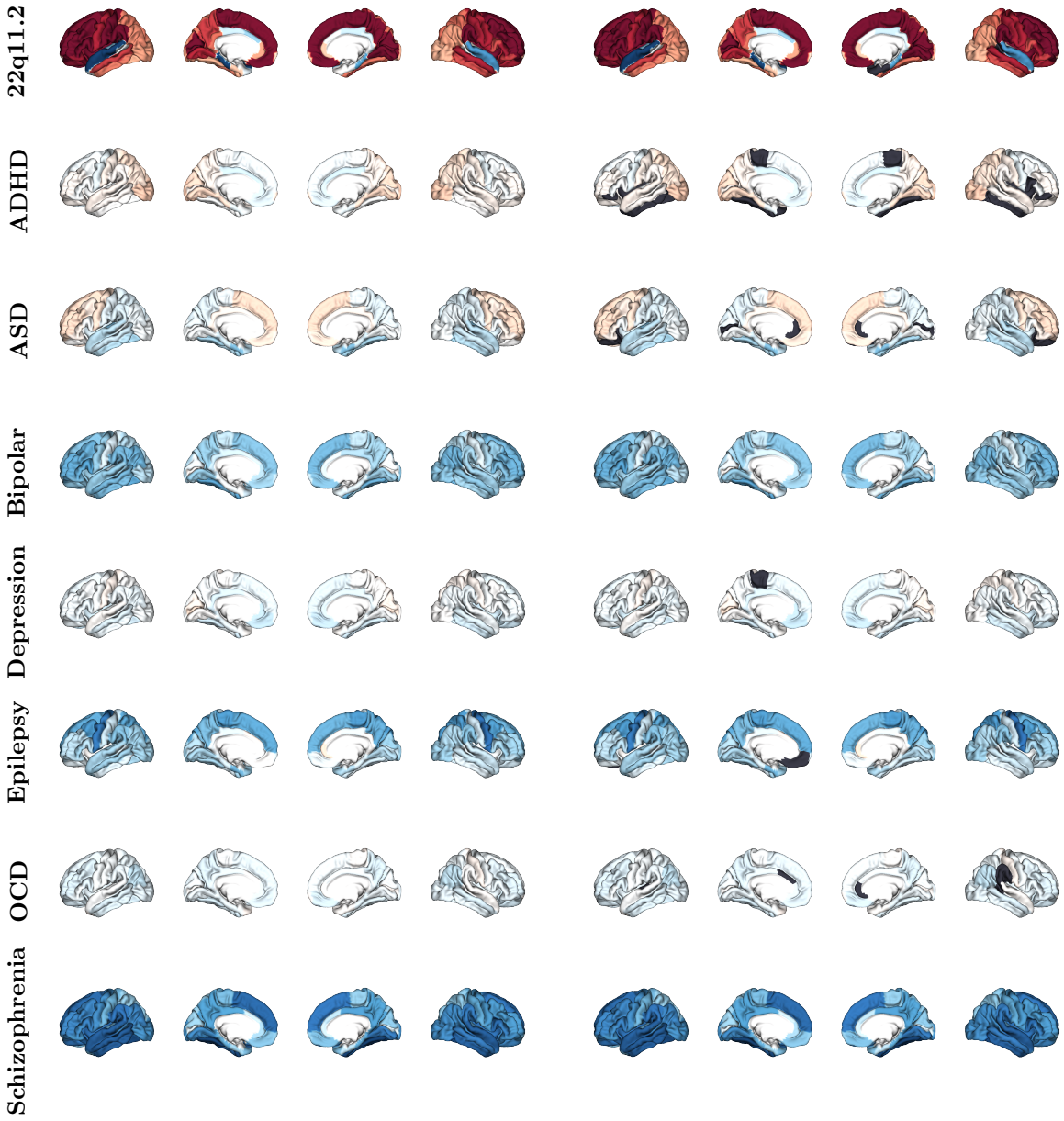


Figure 7: ENIGMA cortical thickness Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions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

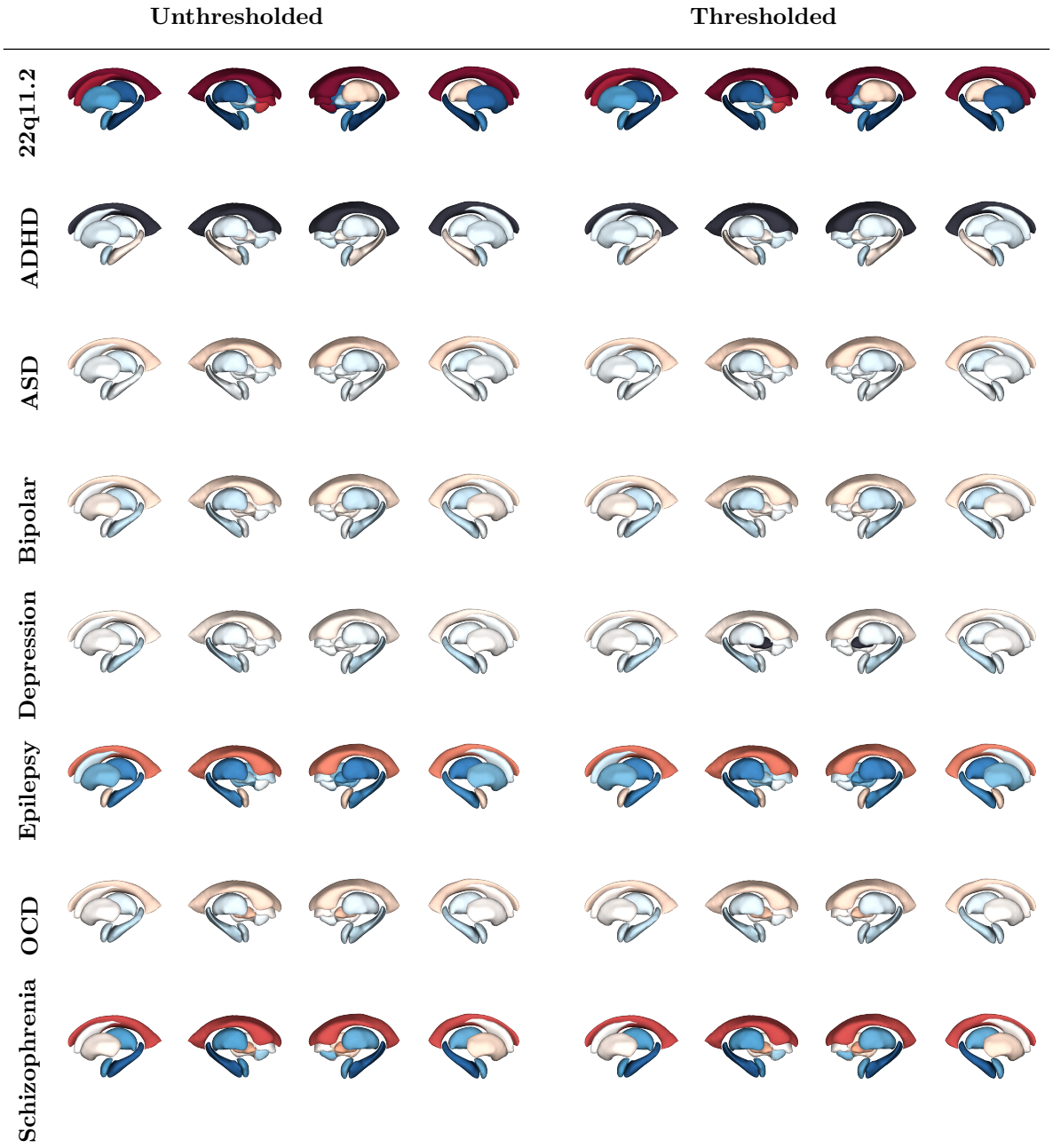


Figure 8: ENIGMA subcortical volume Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.

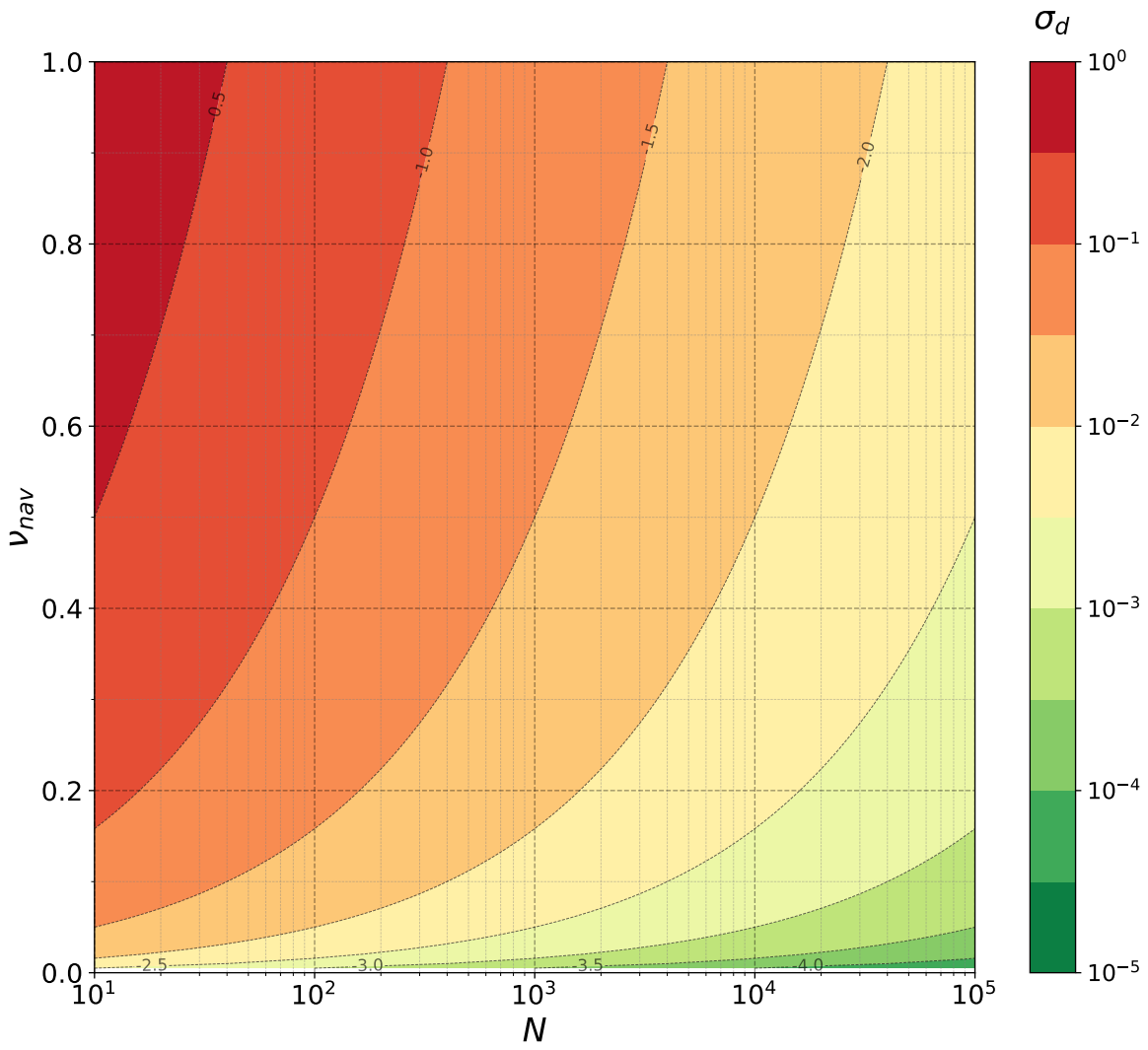


Figure 9: Relationship between ν_{nav} and population sample size N for predicting the uncertainty in Cohen's d effect size estimation. The contour lines represent different ν_{nav} values, showing how numerical variability scales with sample size. With a typical ν_{nav} value of 0.2, to maintain reliable 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` package[2] (version 0.4.0). `significantdigits` implements the Centered Normality Hypothesis approach described in [14]:

$$\hat{s_i} = -\log_2 \left| \frac{\hat{\sigma_i}}{\hat{\mu_i}} \right| - \delta(n, \alpha_s, p_s),$$

where $\hat{\sigma_i}$ and $\hat{\mu_i}$ are the average and standard deviation over the repetitions, and

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

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:

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

.

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 11). Subcortical volumes have a similar precision with 1.33 ± 0.22 significant digits on average (Figure 12). 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 10). 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

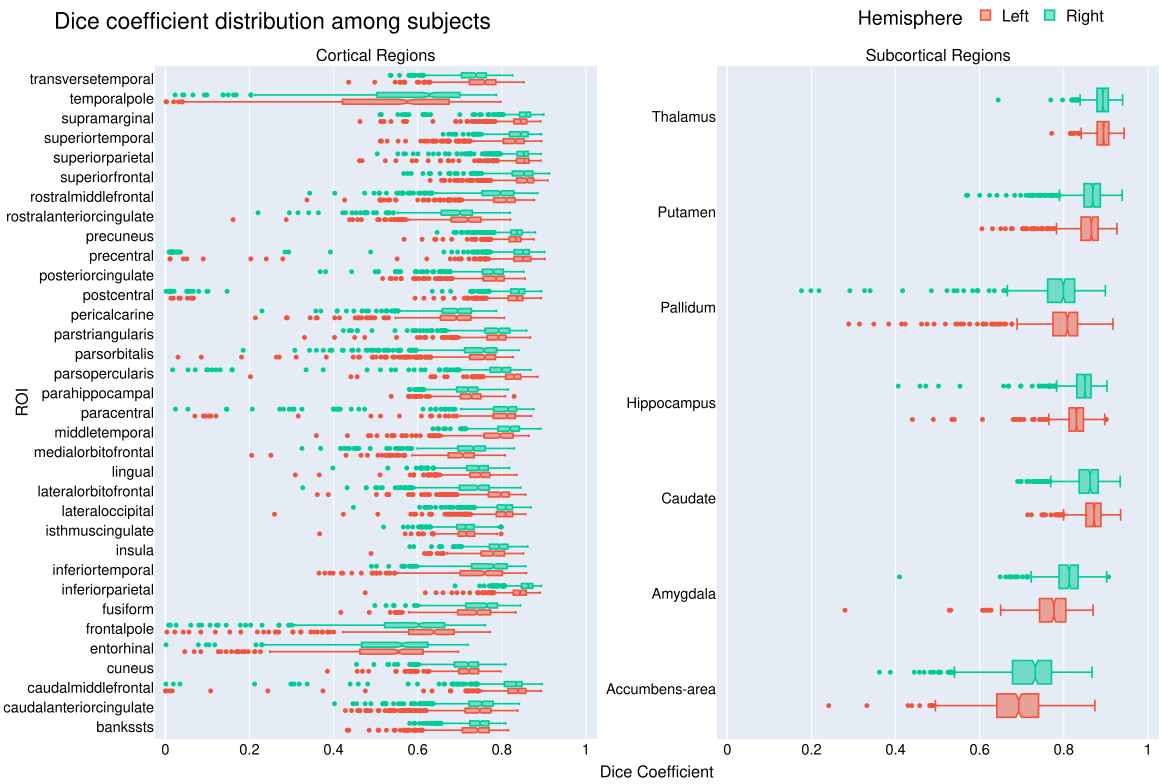


Figure 10: Dice coefficient.

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
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

Continued on next page

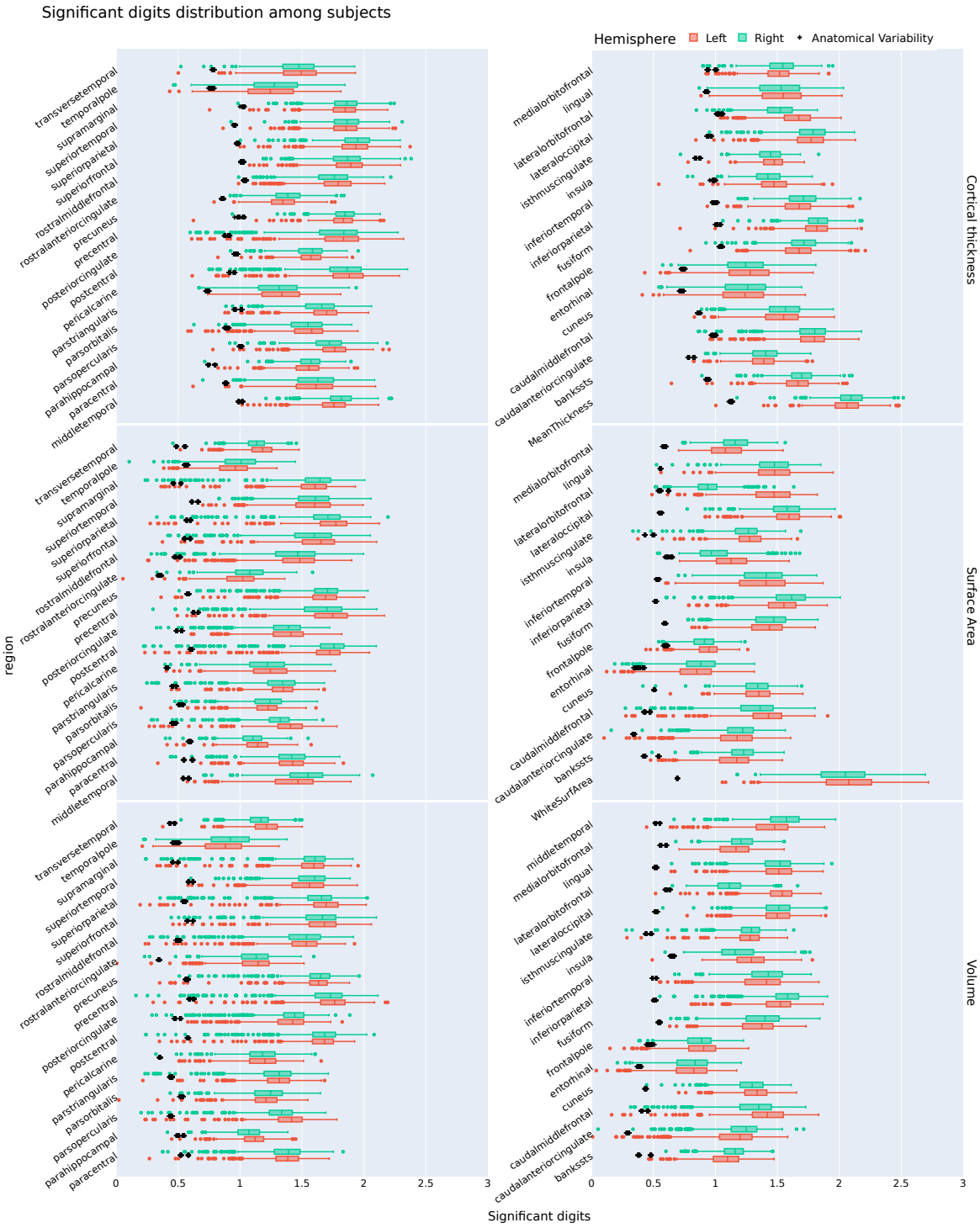


Figure 11: Number of significant digits for each cortical region and metric.

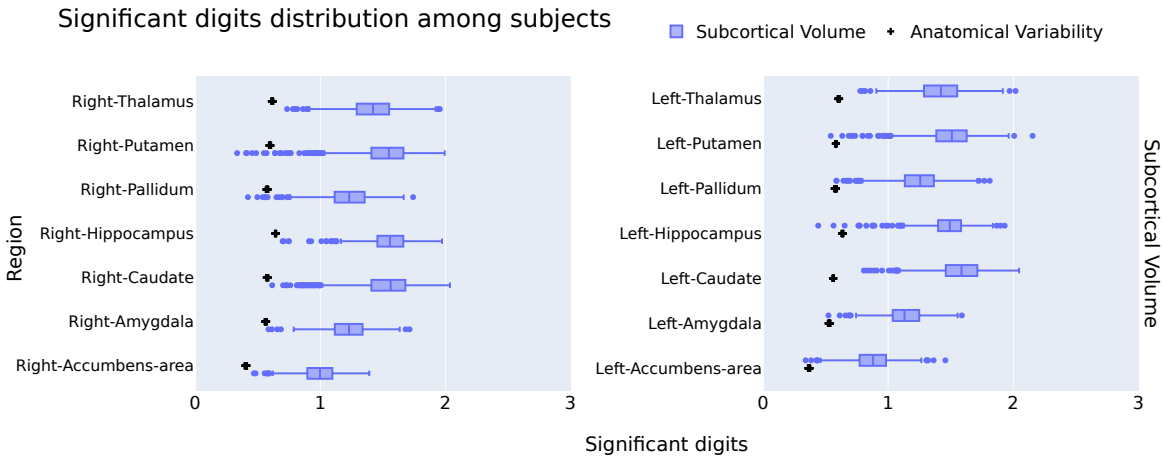


Figure 12: Number of significant digits of subcortical volume for each subcortical region.

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 13 and 14 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 15 and 16 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.

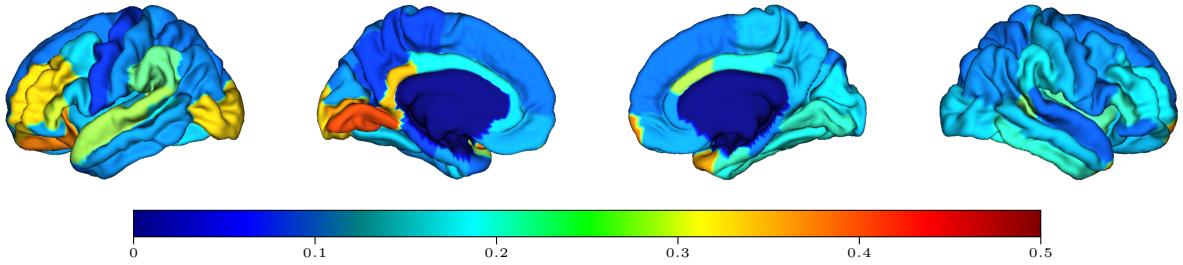


Figure 13: Numerical-Anatomical Variability Ratio (ν_{nav}) for cortical surface area across regions and groups. Higher ν_{nav} values indicate greater computational uncertainty relative to biological variation. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav}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

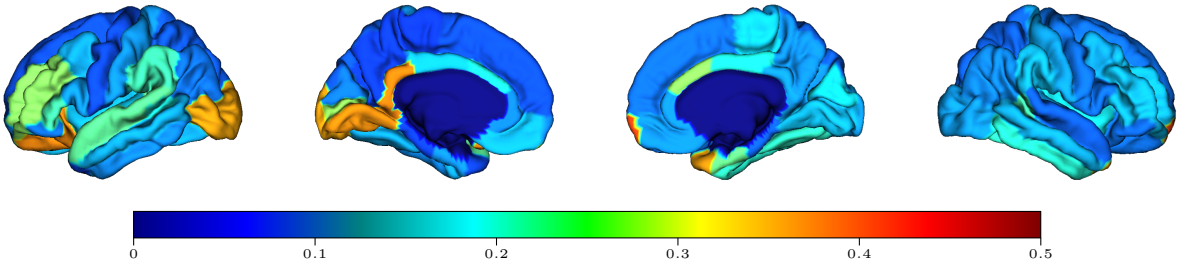


Figure 14: Numerical-Anatomical Variability Ratio (ν_{nav}) for cortical volume across regions and groups. Higher ν_{nav} values indicate greater computational uncertainty relative to biological variation. The color scale indicates the ν_{nav} value, with warmer colors indicating higher ν_{nav} values.

C.2.2 Distribution of statistical tests coefficients

Figures 17 and 18 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.

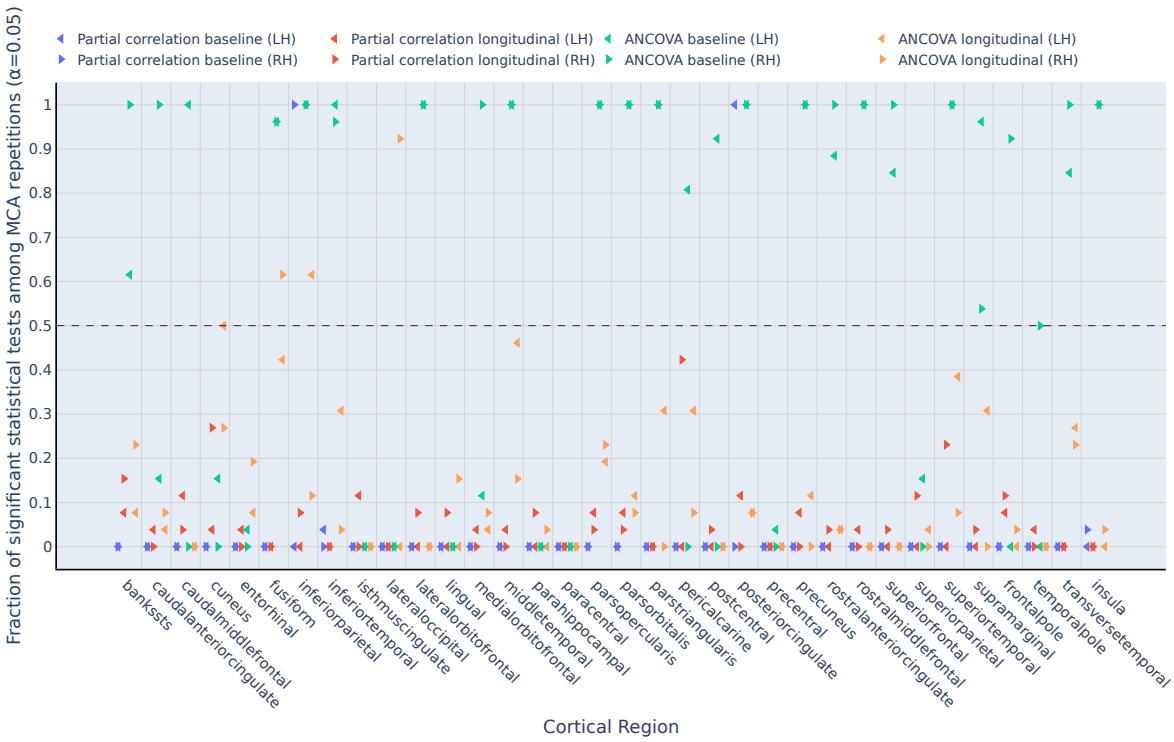


Figure 15: Consistency of statistical tests for cortical area across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency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

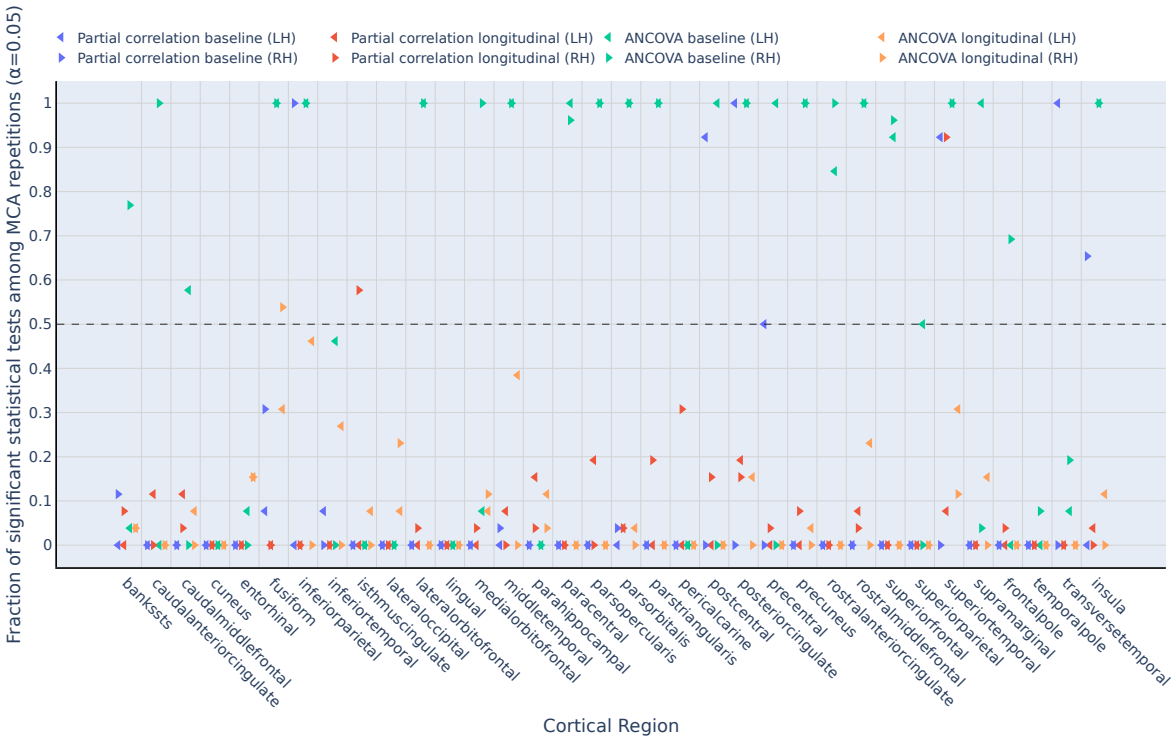
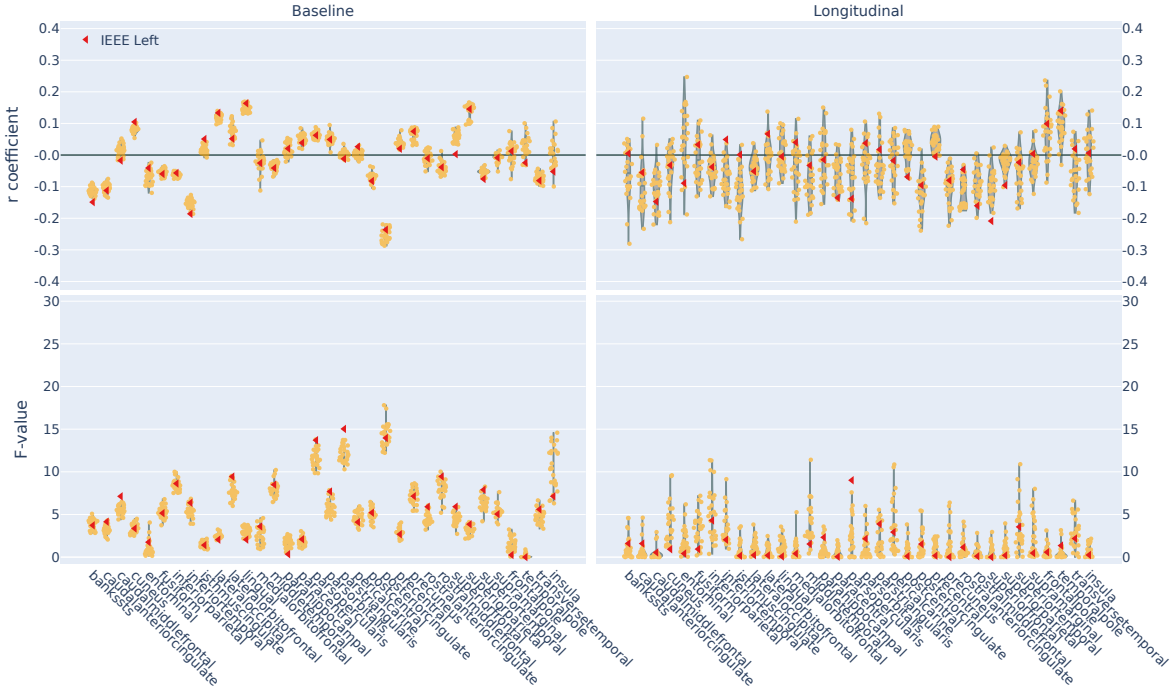
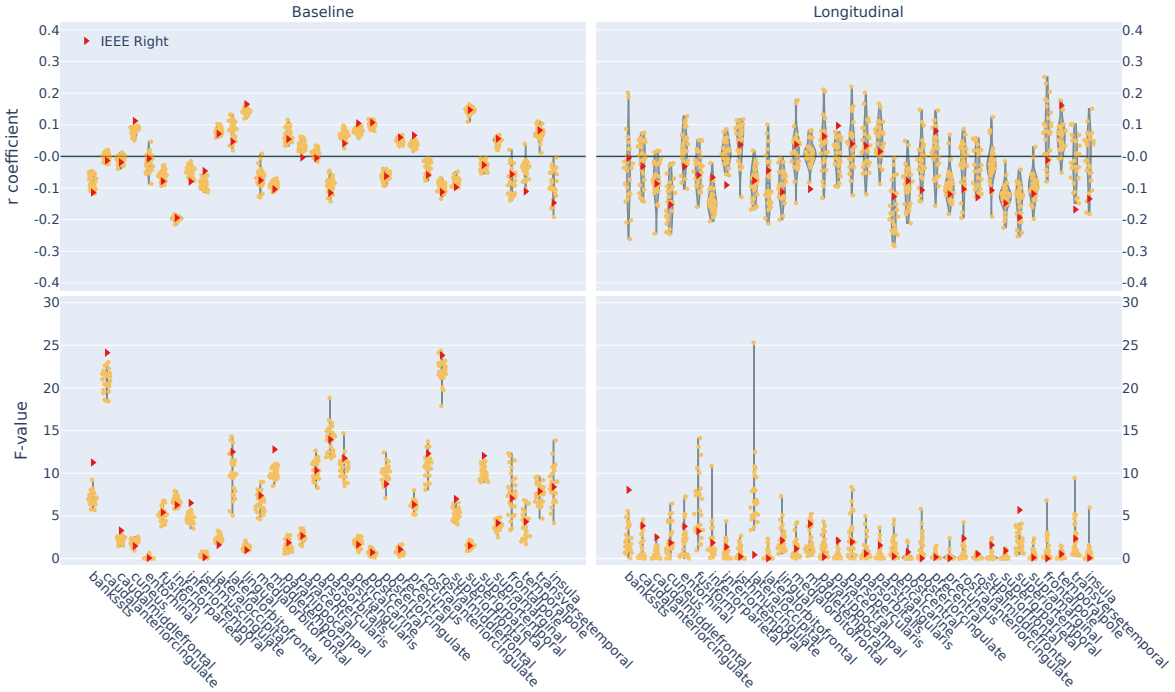


Figure 16: Consistency of statistical tests for cortical volume across all subjects and regions. The plot shows the percentage of subjects for which the statistical test was significant ($\alpha = 0.05$) for each region. The consistency varies across regions, with some regions showing higher consistency than others.

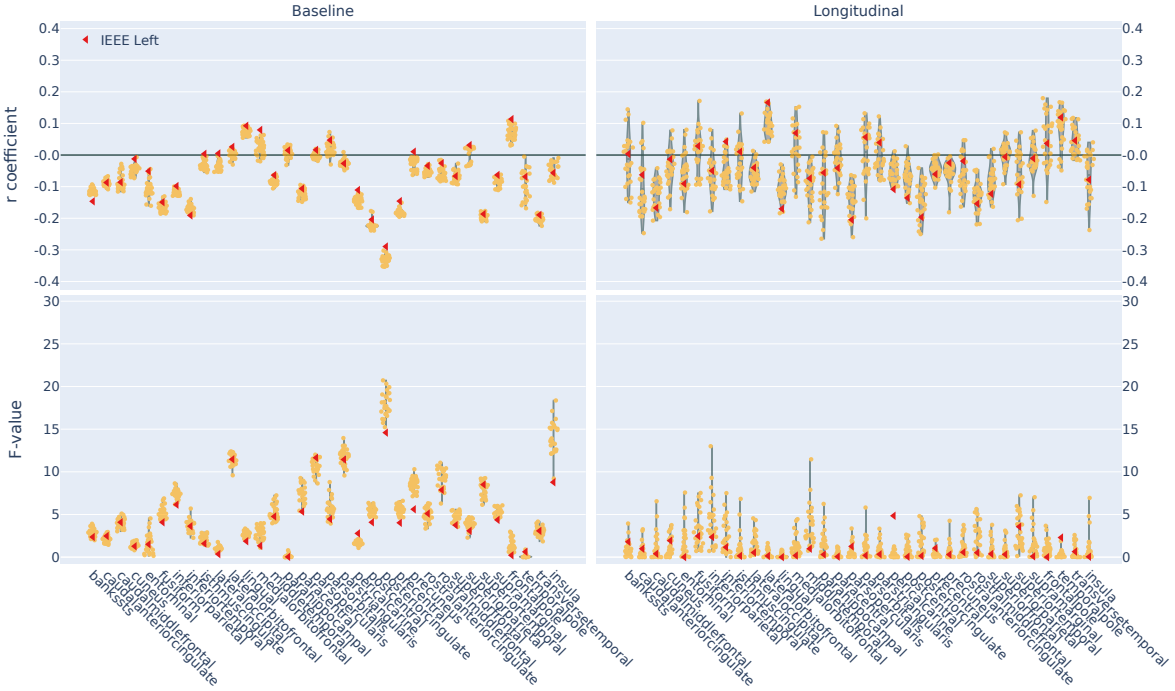


(a) Left hemisphere

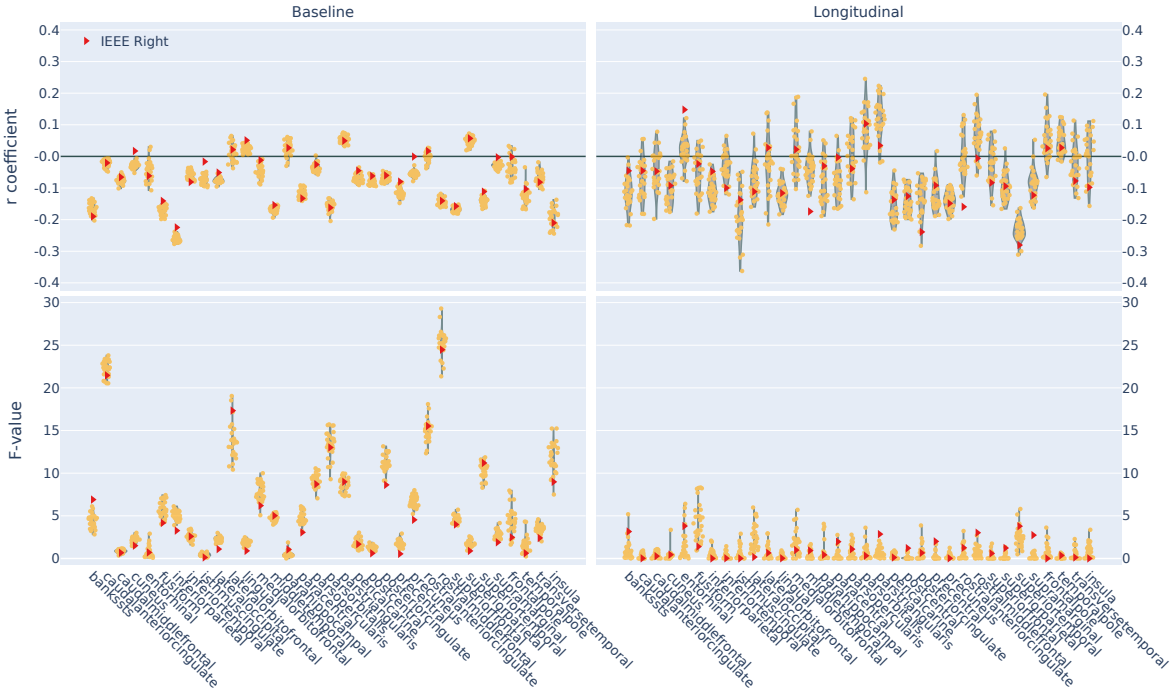


(b) Right hemisphere

Figure 17: 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 18: 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.

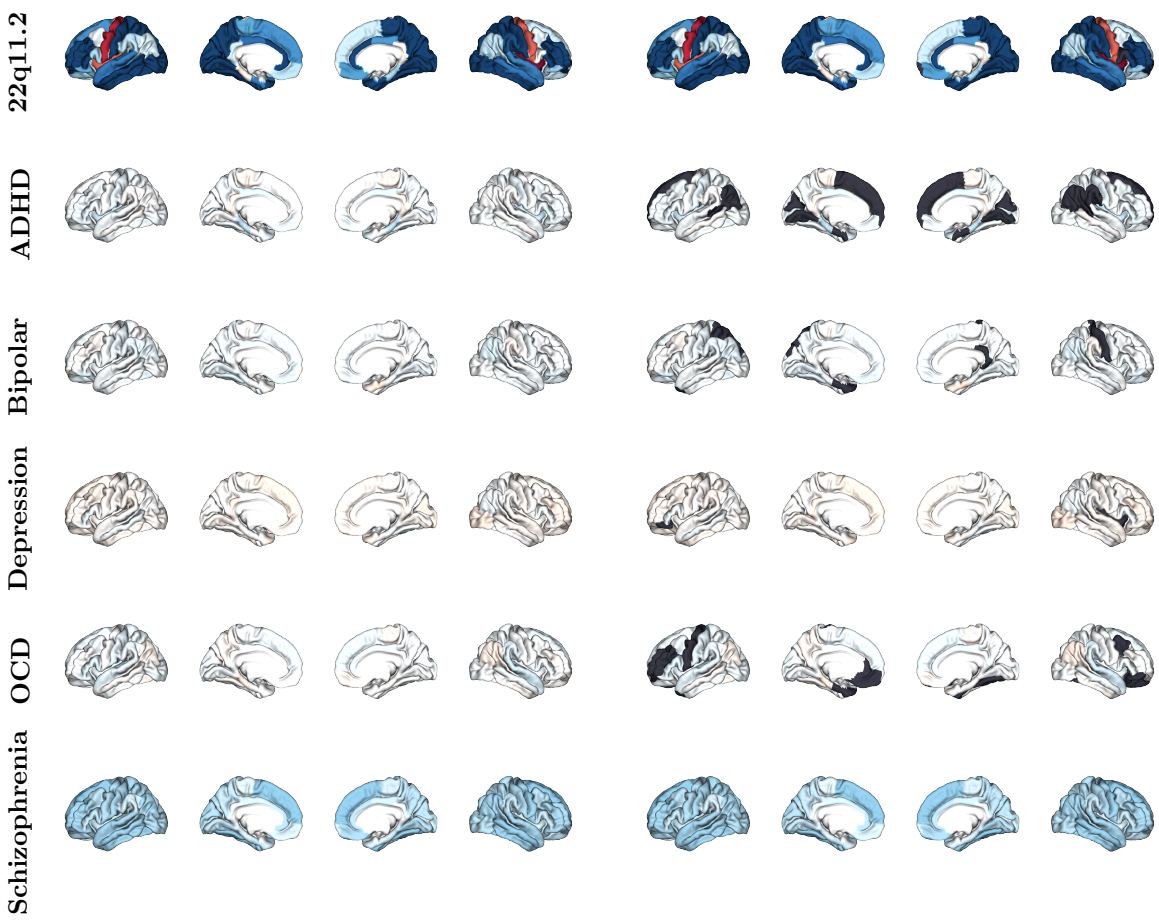


Figure 19: ENIGMA cortical area Cohen's d maps showing unthresholded effect sizes (left) and effect sizes thresholded by the ν_{nav} framework (right) for different disorders. Black regions indicate areas where Cohen's d values fall below the numerical variability threshold, demonstrating regions where reported effect sizes may be unreliable due to computational uncertainty.