MDPI

Article

Brain morphometry and cognitive features in prediction of irritable bowel syndrome

(20241218al) Submission deadline (Diagnostics): 2024-12-30

Application of Artificial Intelligence in Gastrointestinal Disease https://www.mdpi.com/journal/diagnostics/special_issues/CS1956QYIM

Arvid Lundervold [1,2] 0000-0002-0032-4182, Ben René Bjørsvik [2], Julie Billing [3], Birgitte Berentsen [4,5] 0000-0003-3574-7078, Gülen Arslan Lied [4,5] 0000-0002-1827-5008, Elisabeth K Steinsvik [5] 0000-0002-8371-1988, Trygve Hausken [5] 0000-0001-7080-8396, [4], Daniela M. Pfabigan [3] 0000-0002-4043-1695, Astri J. Lundervold [3] 0000-0002-6819-6164*

1 Department of Biomedicine, University of Bergen, Bergen, Norway
2 Medical-AI, Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen, Norway
3 Department of Biological and Medical Psychology, University of Bergen, Norway
4 Department of Clinical Medicine, University of Bergen, Bergen 5021, Norway
5 National Center for Functional Gastrointestinal Disorders, Department of Medicine, Haukeland University Hospital, Bergen 5021, Norway
* Correspondence:Astri.Lundervold@uib.no

Abstract: *Background:* Irritable bowel syndrome (IBS) is a common condition within the spectrum of gut-brain disorders, characterized by abdominal pain, bloating, altered bowel habits, and different patterns of psychological distress. While brain-gut interactions are increasingly recognized in IBS pathophysiology, the relationship between brain morphometry, cognitive function, and clinical presentation remains poorly understood. *Objectives:* To investigate whether multivariate analysis of brain morphometric measures and cognitive test performance can distinguish patients with IBS from healthy controls (HCs), and to evaluate the relative importance of structural and cognitive features in this discrimination. *Methods:* In this cross-sectional study, 49 patients with IBS and 29 HCs underwent structural magnetic resonance imaging (MRI) brain examination and completed the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). Brain morphometry was analyzed using two versions of FreeSurfer software (v6.0.1 and v7.4.1). IBS severity was assessed using the IBS-Severity Scoring System (IBS-SSS). We employed both univariate and multivariate statistical and machine learning approaches, including cross-validation, to analyze morphometric and cognitive measures. *Results:* Univariate and multivariate analyses showed limited discrimination between IBS and HC groups using morphometric measures alone. However, when combining morphometric and cognitive measures in a machine learning framework, the model achieved 93% sensitivity in identifying IBS patients, albeit with 78% specificity. Feature importance analysis highlighted the significance of subcortical structures (particularly hippocampus, caudate, and putamen) and two cognitive domains (recall and verbal skills) in group discrimination. Software version comparison revealed substantial impact on morphometric measurements. *Conclusions:* Contributing with a comprehensive open-source framework for data analysis, our findings suggest that the combination of brain morphometry and cognitive measures provides better discrimination between IBS and HC groups than either measure alone. The identified importance of subcortical structures and specific cognitive domains supports a complex brain-gut interaction in IBS. These results emphasize the need for multimodal approaches in IBS research and careful consideration of methodological factors in brain morphometry studies.

Keywords: Irritable bowel syndrome; structural MRI; brain morphometry, cognition; supervised classification; machine learning

Citation: Lundervold, A., Bjørsvik, B.R.; Billing, J.; ....; Pfabigan, S.M. and Lundervold, A.J. Brain morphometry, gender and cognition in IBS. *Diagnostics* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Introduction

Irritable bowel syndrome (IBS) represents a prevalent and complex gastrointestinal (GI) disorder, affecting approximately 10% of the global population [1]. The syndrome is clinically defined by a characteristic symptom pattern: recurrent abdominal pain associated with defecation, accompanied by alterations in bowel habits [2], and can be divided into clinical phenotypes based on predominant bowel patterns [3] and overall symptom severity [4]. The clinical presentation is heterogeneous, with experiences ranging from mild discomfort to severe symptoms that substantially impair quality of life and daily functioning [4]. Notably, women are disproportionately affected, a difference that appears to arise from a complex interplay of biological factors (including hormonal influences), healthcare-seeking behaviors, and sociocultural determinants [5–8]. Such epidemiological patterns highlight the multifactorial nature of IBS and underscore the importance of considering both biological and psychosocial factors in its study and treatment.

A bidirectional relationship between gastrointestinal symptoms of IBS and psychological functioning is well-documented [9]. While gastrointestinal symptoms can trigger or exacerbate psychological distress, anxiety and depression may in turn amplify the intensity and frequency of abdomal pain [10]. Recent research has expanded this psychobiological framework to include cognitive function, revealing a more nuanced picture of brain-gut interactions in IBS. Although cognitive impairments have been demonstrated at the group level [11,12], these deficits seem to characterize specific subgroups rather than being a universal feature of IBS [9,13]. This heterogeneity in psychological and cognitive presentations aligns with contemporary models of the gut-brain axis [14,15], which conceptualize IBS as a disorder of disrupted neural-enteric communication. In these models, the brain serves as the central integration hub for processing and interpreting the complex array of visceral signals, emotional responses, and cognitive processes that may be involved in IBS.

The relationship between brain structure and cognitive function has evolved from simple localization models to more sophisticated network-based frameworks [16,17]. This network perspective gained particular relevance for understanding IBS through Mayer et al.'s [18] seminal paper in 2015, which proposed that alterations in brain networks could directly influence multiple cognitive domains in IBS patients (see also [19]). Recent empirical support for this systems-level approach comes from Li et al. [20], who identified several associations between symptom severity and regional brain volumes, including positive correlations with subcortical structures (globus pallidus, caudate, and putamen) and negative correlations with cortical regions (anterior cingulate, dorsolateral prefrontal cortex, anterior and midcingulate cortices) and subcortical areas (anterior insula, hippocampus, parahippocampal cortex, thalamus). Of special interest to the present study, they also showed that these brain regions were linked to cognitive performance on tests of language skills and memory function.

Studies of abdominal pain and visceral stimulation have consistently demonstrated involvement of distributed brain networks, encompassing both cortical and subcortical structures [21,22]. Building on this network perspective, Skrobisz et al. [23] conducted a comprehensive morphometric analysis in patients with non-specific digestive disorders, including IBS. Using FreeSurfer software (version 6.0.1), they analyzed 36 brain regions, including subcortical, cortical, and global measures derived from structural magnetic resonance imaging (MRI). Their univariate analyses revealed reduced thalamic volume in IBS patients compared to healthy controls, though volumes remained larger than in patients with inflammatory bowel diseases. While these findings suggest structural brain differences in IBS, univariate approaches may not capture the full complexity of brain-gut interactions. Therefore, our study builds upon Skrobisz et al.'s work in two key ways. First, we examine the robustness of their findings by comparing analyses using both FreeSurfer v6.0.1 and a more recent version, allowing us to differentiate between software-dependent and

true biological effects. Second, we extend beyond univariate analyses by implementing multivariate approaches, including supervised machine learning techniques, to capture complex patterns in brain morphometry that might better characterize IBS. This dual approach - methodological validation and advanced pattern analysis - aims to provide a more comprehensive understanding of the structural brain differences associated with IBS. Finally, responding to Skrobisz et al.'s [23] call for integrating clinical measures, we investigated whether combining cognitive performance data with morphometric features would enhance the accuracy of IBS versus HC classification.

Our study has four key aims:

A To replicate the morphometric differences between IBS patients and HC reported in [23] using the same FreeSurfer software version (FS 6.0.1) and a similar univariate analysis approach as in the original study.
B To evaluate consistency between FreeSurfer versions by comparing morphometric segmentation outcomes from version 6.0.1 (used in [23]) and version 7.4.1 in our dataset ($n = 78$).
C To assess whether morphometric features from FS 7.4.1 (both cross-sectional and longitudinal analyses) can differentiate IBS from HC groups through: (i) Univariate group comparisons, (ii) Multivariate analyses incorporating feature covariance, (iii) Machine learning classification, (iv) Feature importance analysis of successful classifications.
D To determine whether incorporating cognitive performance data enhances the morphometric-based machine learning classification, and if so, identify the most discriminative features between IBS and HC groups.

Materials and Methods

Participants

This study is part of the Bergen Brain-Gut project, a prospective clinical investigation conducted at Haukeland University Hospital, Norway (2020-2022; protocol detailed in Berentsen et al. [24]). We enrolled 78 participants (49 IBS patients and 29 healthy controls [HCs]), all $\geq$18 years old. Recruitment occurred through media advertisements, informational flyers, and direct referrals from the hospital's outpatient clinic. A trained nurse screened all candidates using standardized inclusion and exclusion criteria (Table 1). Eligible participants underwent comprehensive assessment including gastrointestinal measures, psychometric testing, and multiparametric magnetic resonance imaging (MRI).

Sample size determination balanced multiple considerations. Although we did not conduct an a priori power analysis due to limited effect size data for brain morphometric differences in IBS at study inception, our sample size meets or exceeds comparable neuroimaging studies in functional gastrointestinal disorders [23,25–27]. We included only participants with complete key measures and high-quality MRI scans suitable for automated brain segmentation, optimizing data quality while maximizing sample size.

Inclusion criteria	Exclusion criteria
Rome-IV criteria: Recurrent abdominal pain average at least 1 day/week during the last 3 and months, and associated alterations in bowel habits at least 6 months before diagnosis. Other causes are excluded.	Pharmacological treatment affecting GI tract, including medication for anxiety and depression, diabetes, coeliac disease, IBS, Polycystic ovary syndrome, active Helicobacter pylori infection, Parkinson's disease, amyotrophic lateral sclerosis, or Psychiatric disorders.
Normal diet at least 3 weeks before inclusion IBS score equal to or above 175	Treated with antibiotics for the last 3 months Diets such as vegetarian or vegan Use of probiotics or low-FODMAP diet within the last 3 weeks Previous intestinal surgery, except appendectomy Metallic implants, claustrophobia, incompatible with MRI Travel outside Europe last 3 weeks Plan to travel in the near future Pregnancy

Table 1. *Exclusion and inclusion criteria for the IBS patients*. Source: Retrieved from [24].

Measures

Age and sex (not genetically verified) were self-reported by the participants at baseline.

The IBS-Severity Scoring system (IBS-SSS)

The IBS-Severity Scoring system is a questionnaire used to assess the severity and frequency of GI-related IBS symptoms [28]. The questionnaire includes five items related to (i) abdominal *pain intensity*, (ii) abdominal *pain frequency*, (iii) abdominal *distention/bloating*, (iv) dissatisfaction with *bowel habits*, and (v) interference with *quality of life* – over the past 10 days. The maximum score for each question is 100. A sum of scores < 75 is used to define "no or minimal problems", and the scores in the ranges $[75, 175)$, $[175, 300]$, and > 300 as "mild", "moderate", and "severe" IBS symptoms, respectively [28]. In the present study, an IBS-SSS score ≥ 175 was used as the inclusion criteria for the IBS group. Almost all HCs obtained an IBS-SSS score at the lowest level (< 75), with some reporting a score within the mild level ($[75, 175)$).

Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)

RBANS provides a quick and comprehensive assessment of five key cognitive domains, allowing the identification of specific areas of cognitive impairment, and takes about 30 minutes to administer. RBANS is sensitive to mild cognitive impairment, has good reliability and validity, can track changes over time, and is useful for both screening and detailed assessment. The five cognitive domains are (i) *immediate memory* (e.g., story memory and list learning tasks), (ii) *visuospatial/constructional skills* (e.g., copying geometric designs and identifying line orientation), (iii) *language* (e.g., picture naming and semantic fluency tasks), (iv) *attention* (e.g., digit span and coding tasks), and (v) *delayed memory* (e.g., recall of previously learned stories or lists). All participants performed the Norwegian A version of RBANS, administrated by a nurse trained by a clinical neuropsychologist, following the test manual's instructions [29]. The test battery comprises ten subtests, which are combined into five index scores and a total score. These scores are expressed both as raw and as age-corrected scaled scores. The scaled scores have a mean value of 100 and a standard deviation of 15 and are based on performance in a normative group matched to population statistics of 2012 in Norway, Sweden, and Denmark. We used these scaled scores on each of the five RBANS indices for a pairwise correlation analysis between brain morphometric measures and cognitive performance.

MRI Data Acquisition

All neuroimaging data were acquired using a 3 Tesla Siemens Biograph mMR PET/MRI scanner equipped with a standard 12-channel head coil. The comprehensive multiparametric imaging protocol consisted of five sequences: a 3D T1-weighted MPRAGE (TA = 5:35), T2-weighted structural imaging (TA = 5:12), gradient echo (GRE) field mapping (TA = 0:54), resting-state functional MRI using echo-planar imaging (EPI) with integrated motion correction (TA = 9:48), and diffusion-weighted imaging with 30 gradient directions and three b-values (TA = 8:34). The total examination time was approximately 45 minutes.

For the current morphometric analyses, we utilized only the high-resolution T1-weighted images, acquired using a 3D MPRAGE (Magnetization Prepared Rapid Gradient Echo) sequence. The acquisition parameters included a spatial resolution of 1.0 mm isotropic ($1 \times 1 \times 1$ mm^3) across 192 sagittal slices, with repetition time (TR) of 2500 ms, echo time (TE) of 2.26 ms, and inversion time (TI) of 900 ms. The field of view (FOV) was set to 256×256 mm^2 with a corresponding matrix size of 256×256, and parallel imaging was employed using GRAPPA with an acceleration factor of 2.

Figure 1 shows a representative T1-weighted image from our dataset, demonstrating the high tissue contrast necessary for accurate morphometric analysis. The corresponding FreeSurfer-generated segmentation mask, which forms the basis for our morphometric measurements, is illustrated in Figure 2. These images exemplify the quality standards maintained throughout our dataset.

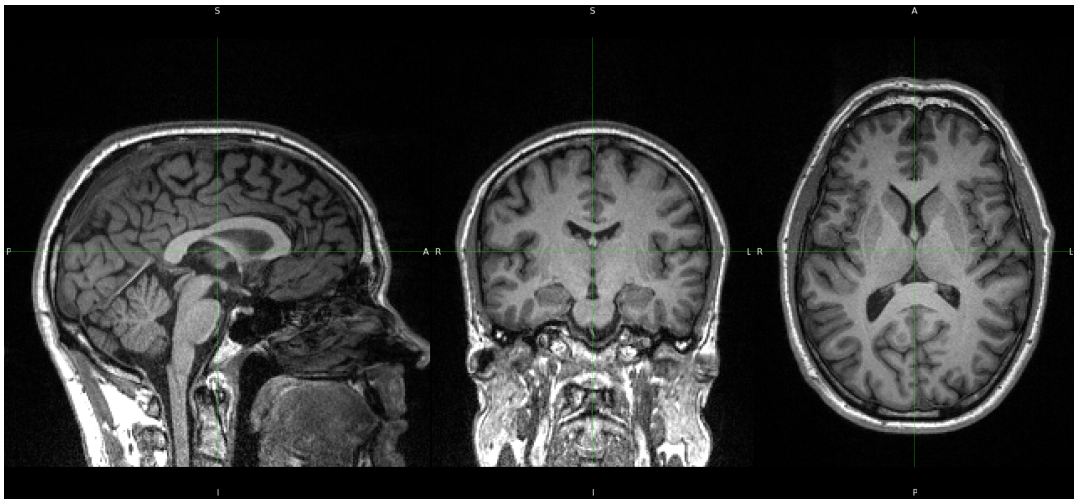


Figure 1. *3D T1-weighted MPRAGE recording from BGA_046*. Panels left to right: Sagittal, Coronal, and Axial section, respectively.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/01-freesurfer-freeview-t1-aseg-bga-046.ipynb

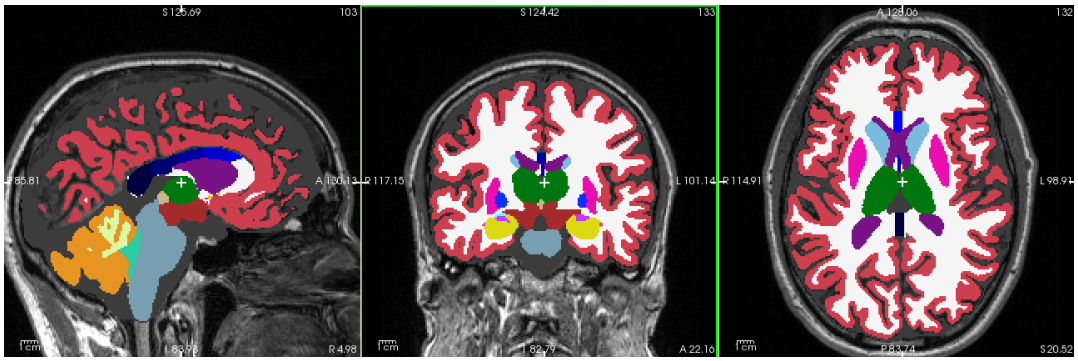


Figure 2. *The color-coded* `aseg` *segmentation mask by FreeSurfer 7.4.1 overlaid on 3D T1-w MPRAGE from BGA_046*. Panels left to right: Sagittal, Coronal, Axial section, respectively. The white cross is located in the medial part of left thalamus. Thalamus: green, Hippocampus: yellow, Caudate: light blue, Putamen: pink, Pallidum: purple, Cortex: red, White-Matter: white. See also the Appendix Fig. A2.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/01-freesurfer-freeview-t1-aseg-bga-046.ipynb

Brain Morphometry Analysis using FreeSurfer

Image processing and morphometric analyses were performed using FreeSurfer (https://freesurfer.net), a widely-validated open-source software suite for analyzing brain MRI data [30]. To address both methodological and biological questions, we conducted parallel analyses using two FreeSurfer versions: version 6.0.1, which was employed in the reference study by Skrobisz et al. [23], and the current version 7.4.1.

The evolution of FreeSurfer's capabilities is particularly relevant to our investigation of brain structure in IBS. Version 7.0 (July 2020) introduced significant improvements in subcortical segmentation accuracy, while version 7.4.1 (June 2023) further enhanced the precision of limbic system structures, notably the hippocampus and amygdala. Additionally, version 7.4.1 provides superior compatibility with multi-modal imaging data and implements refined longitudinal processing algorithms. Since our multimodal MRI examinations were part of a longitudinal IBS intervention study (Berentsen et al. [24]), we also used the longitudinal stream capability of FreeSurfer 7.4.1 to compare baseline longitudinal analysis with a cross-sectional analysis of the first MRI examination.

For both versions, we focused on the automated segmentation of subcortical structures using FreeSurfer's `aseg` pipeline, which identifies and quantifies the volume of distinct brain regions (detailed in Table A.1). This dual-version approach serves two purposes: first, it enables direct comparison with Skrobisz et al.'s [23] findings, and second, it allows us to assess the impact of software evolution on morphometric measurements on a fixed dataset, and differences in cross-sectional and longitudinal stream analysis to discriminate HC and IBS from brain morphometric features. This methodological consideration is crucial, as previous studies have demonstrated that version-dependent variations in automated segmentation can significantly influence morphometric results [31–36]. By analyzing our data with both versions, we can distinguish between genuine biological differences and methodologically-induced variations in brain morphometry.

The enhanced accuracy of version 7.4.1 is particularly relevant for our investigation of IBS, as it provides more reliable quantification of brain regions implicated in visceral sensation, pain processing, emotional regulation, and cognitive function.

We will also like to add that *in vivo* brain segmentation technologies move very fast. Recently (November 2024), the FreeSurfer 8.0.0-beta version enables histological super granularity with identification and volume measurements from more than 300 distinct regions per hemisphere (cf. Fig. A2). The `aseg` mask provides less than 40 brain regions and their volumes within the intracranial space.

Statistical Analysis and Machine Learning Approaches

All analyses were implemented in Python (version 3.10), with complete computational workflows and reproducibility materials available in our public GitHub repository

(https://arvidl.github.com/ibs-brain). Our analytical approach combined traditional statistical methods with advanced machine learning techniques, employing both parametric and non-parametric approaches as appropriate for the data distributions.

For group comparisons, statistical significance was assessed using a threshold of $p < 0.05$, with Bonferroni correction applied to control for multiple comparisons. Effect sizes were quantified using Cliff's Delta [37], a robust non-parametric measure particularly suitable for non-normally distributed data [38]. Following established conventions, we interpreted Cliff's Delta (absolute) values as negligible (0.00-0.14), small (0.15-0.33), medium (0.34-0.47), or large (0.48-1.00).

Relationships between variables were evaluated using Spearman's rank correlation coefficient (ρ), chosen for its robustness to non-normality and ability to capture monotonic relationships [39]. Correlation strengths were classified as weak (0.20-0.39), moderate (0.40-0.59), strong (0.60-0.79), or very strong (0.80-1.00). Values below 0.20 were considered negligible to minimize the risk of over-interpreting weak associations.

To ensure reproducibility and transparency, all analysis scripts, including data preprocessing steps, statistical analyses, and visualization code, are documented in Jupyter notebooks accessible through our GitHub repository. These notebooks provide detailed documentation of parameter choices, statistical assumptions, and analytical decisions.

Our analysis strategy addressed four interconnected research objectives, progressing from replication to more advanced multivariate approaches:

Research Objectives and Analytical Approach

A - Replication Analysis :

Is it possible to replicate the morphometric findings in Skrobisz et al. [23] regarding IBS versus HC discrimination, using the same FreSurfer-derived features and the same FreeSurfer version?

- **(i)** By employing a feature-by-feature (univariate) comparison incorporating effect size?
- **(ii)** By employing a novel consistency score, combining several metrics for replication assessment?

B - Software Version Comparison :

Are there IBS versus HC disparities in morphometric feature values between FreeSurfer 6.0.1 and FreeSurfer 7.4.1 applied to the same set ($n = 78$) of T1-weighted recordings in our Bergen cohort?

What is the difference in the results between FreeSurfer 7.4.1 cross-sectional analysis versus FS 7.4.1 longitudinal stream?

- **(i)** By employing a feature-by-feature comparison?
- **(ii)** Employing a multivariate comparison, incorporating covariance structures in the morphometric features?

C - Morphometric Classification Analysis :

Is it possible to separate IBS individuals from HC based on morphometric features?

- **(i)** By employing a feature-by-feature comparison (FS 7.4.1)?
- **(ii)** Employing a multivariate comparison, incorporating covariance structures in the morphometric features?
- **(iii)** By predicting IBS versus HC from the morphometric features using a machine learning framework (ML)?
- **(iv)** Identifying the importance of morphometric measures in the model with the best prediction?

D - Integrated Morphometric-Cognitive Analysis :
Would adding cognitive performance as a predictor improve the accuracy of separating IBS from HC?

- **(i)** By employing a feature-by-feature comparison?
- **(ii)** Employing a multivariate comparison, incorporating covariance structures in the cognitive features?
- **(iii)** By predicting IBS versus HC from morphometric and cognitive characteristics using a machine learning framework (ML)?
- **(iv)** Identifying the importance of morphometric and cognitive measures included in the model with the best prediction?

This hierarchical analytical framework progresses from basic replication to more advanced multivariate approaches, enabling both methodological validation and novel insights into IBS-related brain structure and function.

Statistical Analysis Framework

Given the complexity of our research questions and the combination of traditional and advanced analytical methods, we implemented a comprehensive statistical framework encompassing both univariate and multivariate approaches. Here we detail our analytical strategy and its methodological justification.

Exploratory and Univariate Analyses

Initial analyses followed established protocols, as in [23], beginning with exploratory data analysis of numerical features and cross-tabulation of categorical variables (Group: HC/IBS; Sex: F/M). For univariate comparisons (Objectives A-D), we employed both parametric (independent t-tests) and non-parametric (Mann-Whitney U) tests, depending on normality assessments. Multiple comparison correction used the Bonferroni method, and *effect sizes* were quantified using Cohen's d (for parametric tests) and Cliff's delta [37], else. Cliff's delta (δ) between two groups X and Y is defined as $\delta = \frac{U}{n_x n_y} - 0.5$, where U is the Mann-Whitney U statistic, n_x is the number of observations in group X, and n_y is the number of observations in group Y. The resulting Cliff's delta (δ) ranges from -1 to +1, where $\delta = +1$ indicates that all values in group X are greater than all values in group Y, $\delta = -1$ indicates that all values in group X are less than all values in group Y, and $\delta = 0$ indicates complete overlap between the two groups.

Permutation Testing

To address small sample sizes and potential non-normal distributions, we employed permutation testing (1,000 iterations) to assess statistical significance. For each test, we computed an observed test statistic (sum of squared differences between group means) and generated a null distribution by randomly reassigning group labels. The empirical p-value was calculated as the proportion of permuted statistics exceeding the observed value. This non-parametric approach provides robust statistical inference while naturally controlling for multiple comparisons.

Multivariate Approaches - assessing multivariate normality

For multivariate analyses (Objectives B-D), we first assessed multivariate normality using two complementary methods: Mardia's test and the more comprehensive Henze-Zirkler's test (see Appendix A.2 for details).

Advanced Distance Metrics

The Mahalanobis distance [40] quantifies the distance between a point P and a distribution D while accounting for data correlations [41]. Unlike Euclidean distance, it

incorporates the covariance structure through the formula $D = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$, where x represents the data point, μ is the mean vector, and Σ^{-1} is the inverse covariance matrix.

Remark: While Cohen's d ($d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$) measures standardized univariate group differences, the Mahalanobis distance extends this concept to multivariate space. In comparing IBS and HC groups, the squared Mahalanobis distance relates proportionally to Hotelling's T^2 statistic, a multivariate analog of the squared t-statistic. Unlike Cohen's d, which has standardized effect size interpretations (small: 0.2, medium: 0.5, large: 0.8), Mahalanobis distance interpretation depends on data dimensionality and covariance structure. To handle outliers and non-normality common in neuroimaging data, we implemented a *robust Mahalanobis distance*. This modification employs winsorization (clipping values at 10th/90th percentiles) and replaces arithmetic means with medians (see Appendix A.3).

Prediction of class belonging using machine learning

In tasks **C(iii)** and **D(iii)** we applied a comprehensive machine learning framework, utilizing morphometric features derived from FreeSurfer (`aseg`) to develop predictive models for two distinct classification tasks. We employed *PyCaret* (https://pycaret.org), an open-source, low-code machine learning library in Python, to develop and evaluate our classification models.

Machine Learning Model Development

Our machine learning approach followed a systematic protocol designed to ensure robust classification while addressing the challenges of limited sample size and potential overfitting. The analysis pipeline consisted of several carefully constructed stages optimized for neuroimaging data classification. Initial data preparation used a stratified sampling approach, partitioning the data set into training (70%) and testing (30%) sets while preserving the distribution of IBS/HC status across both partitions. This stratification was crucial for maintaining representative samples and ensuring valid model evaluation, particularly given our modest sample size and the inherent complexity of neuroimaging data. Model development utilized PyCaret's comprehensive machine learning framework to evaluate multiple classification algorithms, ranging from traditional approaches to advanced ensemble methods. The classifier suite included linear models (Logistic Regression with L1 and L2 regularization), non-linear algorithms (Support Vector Machines [SVM] with various kernels), tree-based methods (Random Forests, Gradient Boosting Machines including XGBoost [42] and LightGBM), and instance-based learners (K-Nearest Neighbors). This diverse algorithm selection allowed exploration of different decision boundaries and pattern of feature interaction.

To ensure robust model assessment and mitigate overfitting risks, we implemented a nested 10-fold cross-validation strategy for model selection. This approach provided unbiased performance estimates while preventing data leakage between model selection and evaluation phases. The final model selection prioritized both predictive performance and model interpretability, considering the clinical relevance of our findings. See Table A1 for an illustration.

Model Performance Assessment

To address the class imbalance between IBS and HC groups, we implemented multiple complementary performance metrics. While classification *accuracy* served as a baseline measure, we employed additional metrics such as: the *F1 score* (harmonic mean of precision and recall) to balance false positive and negative rates; the Receiver Operating Characteristic Area Under Curve (*ROC-AUC*) to assess discrimination ability across classification thresholds; and *Cohen's Kappa* [43] to evaluate classification agreement beyond chance-level performance.

We generated *confusion matrices* to examine error patterns and potential classification biases. For analyses incorporating cognitive function, we used macro-averaged versions of these metrics, ensuring equal weighting across performance levels despite uneven class distributions. Performance assessment followed a dual-track strategy, evaluating models on both cross-validated training data and the held-out test set. This approach enabled us to assess both learning capacity and generalization ability, crucial considerations for clinical applications.

Feature Importance and Model Interpretability Analysis

To understand how morphometric and cognitive features contribute to classification performance, we implemented two complementary approaches to feature importance analysis: permutation importance and SHAP (SHapley Additive exPlanations) values.

The *permutation importance* [44] analysis quantifies feature relevance by measuring model performance degradation when individual features were randomly permuted. Through multiple iterations per feature, we calculated the mean decrease in model performance, providing a model-agnostic measure of feature importance.

The *SHAP analysis*, grounded in cooperative game theory [45], provided both global and local interpretation frameworks. The global analysis aggregated SHAP values across cases to identify consistently important features, while the local analysis examined feature contributions to individual predictions. We visualized these results using SHAP summary plots, which integrated both magnitude and directionality of feature effects.

By combining permutation importance with SHAP analysis, we gained complementary insights into feature relevance: permutation importance revealed features critical to overall model performance, while SHAP analysis illuminated feature interactions and their contributions to specific predictions. This approach helped identify key neurobiological features distinguishing IBS patients from healthy controls, while exploring relationships between brain structure, sex differences, and cognitive function.

Results

Sample Demographics and Clinical Characteristics

The study enrolled 78 participants, comprising 49 patients with IBS and 29 healthy controls. Demographic analysis revealed comparable age distributions between groups (median age: IBS = 34 years, controls = 33 years). Female participants predominated in both cohorts, representing 77.6% (38/49) of the IBS group and 69.0% (20/29) of the control group, reflecting the typical gender distribution observed in IBS populations.

Symptom severity, quantified using the IBS Symptom Severity Scale (IBS-SSS), demonstrated clear differentiation between groups. The IBS cohort exhibited predominantly moderate to severe symptomatology, while healthy controls reported minimal gastrointestinal symptoms, aligning with our inclusion criteria. Six participants (three from each group) had missing IBS-SSS data, which we addressed through multiple imputation stratified by group and gender to maintain statistical robustness. Detailed demographic and clinical characteristics are presented in Table 2.

Table 2. *Demographic and Clinical Characteristics of the Study Sample*

Group	Age	IBS_SSS	Sex	N	Missing
	Median (IQR)	Median (IQR)	F/M (%)		IBS_SSS
HC (N=29)	33.0 (23.0)	21.0 (30.0)	69.0/31.0	29	3
IBS (N=49)	34.0 (14.0)	264.0 (95.0)	77.6/22.4	49	3

Age is reported in years; IBS-SSS scores range from 0-500, with higher scores indicating greater symptom severity. IQR = Interquartile Range; F/M = Female/Male ratio expressed as percentages.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/02-demographics-and-clinical-characteristics.ipynb

Replication analysis of Skrobisz (2022) using the Bergen cohort (with FS 6.0.1)

In our Bergen cohort, we sought to replicate the morphometric findings reported by Skrobisz et al. [23] comparing IBS patients with healthy controls. Table 3 presents our comparative analysis using identical methodological parameters: 35 estimated Total Intracranial Volume (eTIV)-normalized regional brain volumes derived from FreeSurfer 6.0, matching the analytical approach of the original study.

Table 3. *Comparison of eTIV-normalized regional brain volumes between the two cohorts.*

Brain Region	Skrobisz Cohort (FS 6.0)				Bergen Cohort (FS 6.0.1)			
	HC (N=19)		IBS (N=20)		HC (N=29)		IBS (N=49)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Left-Cerebellum-WM	0.00992	0.00113	0.00971	0.00107	0.01050	0.00092	0.01048	0.00092
Left-Cerebellum-Cortex	0.03628	0.00302	0.03553	0.00256	0.03894	0.00344	0.03931	0.00373
Left-Thalamus	0.00511	0.00037	0.00500	0.00024	0.00523	0.00046	0.00514	0.00039
Left-Caudate	0.00239	0.00025	0.00228	0.00021	0.00236	0.00026	0.00236	0.00031
Left-Putamen	0.00336	0.00033	0.00324	0.00028	0.00348	0.00038	0.00344	0.00039
Left-Pallidum	0.00140	0.00012	0.00135	0.00010	0.00140	0.00015	0.00137	0.00011
Left-Hippocampus	0.00270	0.00021	0.00272	0.00020	0.00291	0.00027	0.00290	0.00024
Left-Amygdala	0.00118	0.00013	0.00113	0.00015	0.00122	0.00010	0.00120	0.00010
Left-Accumbens-area	0.00031	0.00005	0.00034	0.00006	0.00043	0.00007	0.00042	0.00006
CSF	0.00061	0.00009	0.00060	0.00012	0.00067	0.00012	0.00070	0.00014
Right-Cerebellum-WM	0.00908	0.00106	0.00922	0.00100	0.00997	0.00089	0.00998	0.00085
Right-Cerebellum-Cortex	0.03652	0.00321	0.03616	0.00264	0.03972	0.00344	0.03998	0.00376
Right-Thalamus	0.00488	0.00030	0.00475	0.00024	0.00512	0.00044	0.00507	0.00036
Right-Caudate	0.00244	0.00024	0.00236	0.00024	0.00244	0.00024	0.00244	0.00030
Right-Putamen	0.00336	0.00030	0.00330	0.00028	0.00351	0.00037	0.00349	0.00035
Right-Pallidum	0.00136	0.00012	0.00133	0.00010	0.00132	0.00013	0.00130	0.00011
Right-Hippocampus	0.00282	0.00022	0.00285	0.00021	0.00301	0.00024	0.00298	0.00023
Right-Amygdala	0.00125	0.00012	0.00120	0.00012	0.00128	0.00009	0.00127	0.00010
Right-Accumbens-area	0.00034	0.00004	0.00036	0.00005	0.00043	0.00005	0.00043	0.00006
WM-hypointensities	0.00047	0.00015	0.00048	0.00013	0.00079	0.00031	0.00069	0.00025
CC_Posterior	0.00065	0.00013	0.00065	0.00010	0.00065	0.00010	0.00070	0.00011
CC_Mid_Posterior	0.00038	0.00007	0.00036	0.00007	0.00037	0.00007	0.00040	0.00007
CC_Central	0.00039	0.00009	0.00043	0.00008	0.00039	0.00009	0.00039	0.00010
CC_Mid_Anterior	0.00041	0.00009	0.00044	0.00013	0.00038	0.00008	0.00041	0.00011
CC_Anterior	0.00062	0.00010	0.00061	0.00008	0.00062	0.00010	0.00065	0.00010
BrainSegVol	0.75340	0.01784	0.74913	0.01647	0.80464	0.02487	0.80558	0.02397
BrainSegVolNotVent	0.74137	0.01880	0.73857	0.01836	0.79224	0.02511	0.79132	0.02490
lhCortexVol	0.15339	0.00620	0.15313	0.00871	0.16670	0.00800	0.16693	0.00951
rhCortexVol	0.15490	0.00690	0.15467	0.00859	0.16614	0.00828	0.16646	0.00939
CortexVol	0.30829	0.01298	0.30780	0.01715	0.33283	0.01611	0.33339	0.01880
lhCerebralWhiteMatterVol	0.15101	0.00748	0.15058	0.00742	0.15990	0.00858	0.15915	0.00876
rhCerebralWhiteMatterVol	0.15103	0.00757	0.15075	0.00727	0.15925	0.00829	0.15827	0.00938
CerebralWhiteMatterVol	0.30205	0.01500	0.30133	0.01461	0.31915	0.01678	0.31742	0.01808
SubCortGrayVol	0.03930	0.00194	0.03855	0.00162	0.04092	0.00258	0.04063	0.00236
TotalGrayVol	0.42105	0.01376	0.41884	0.01868	0.45307	0.02208	0.45396	0.02432

Note: All volumes are normalized to estimated total intracranial volume (eTIV)
HC = Healthy Controls; IBS = Irritable Bowel Syndrome; SD = Standard Deviation; WM = White Matter
Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/03-replication-anaysis-fs6.ipynb

The volumetric comparison of brain structures between IBS patients and healthy controls across both cohorts reveals distinct patterns. While the Bergen cohort demonstrates

systematically larger volumes (6-8% for global measures, reaching up to 35% for specific structures such as the *nucleus accumbens*), the within-cohort comparisons between IBS and healthy control groups show remarkable consistency in global brain eTIV-normalized volumes. Specifically, `BrainSegVol` values remain nearly identical within each cohort (Skrobisz: HC 0.753 ± 0.018, IBS 0.749 ± 0.016; Bergen: HC 0.805 ± 0.025, IBS 0.806 ± 0.024). Cortical measurements demonstrate similar stability, with total cortical volume (`CortexVol`) showing minimal between-group differences in both cohorts. In subcortical structures, we observed subtle variations, notably a slight trend toward volume reduction in IBS patients' subcortical gray matter (`SubCortGrayVol`), though these differences remain within standard deviation bounds. White matter volumes maintain consistency between groups within cohorts, with an interesting pattern of white matter hypointensities emerging in the Bergen cohort. Corpus callosum segments exhibit relatively uniform volumes across all groups. Several methodological factors warrant consideration: the disparate cohort sizes (Skrobisz: HC $n = 19$, IBS $n = 20$; Bergen: HC $n = 29$, IBS $n = 49$), potential variations in FreeSurfer versions (6.0 versus 6.0.1), and differences in operating systems may contribute to the systematic volumetric differences observed between cohorts. While normalization to estimated total intracranial volume (eTIV) facilitates direct comparisons within cohorts by controlling for head size variation, it does not fully account for between-cohort differences.

Figure 3 presents a detailed reproducibility analysis, illustrating the differences in eTIV-normalized brain region volumes between HC and IBS across both cohorts. The plot contrasts effect sizes from the Skrobisz (2022) cohort (x-axis) against the Bergen cohort (y-axis), with the diagonal line representing perfect agreement. We employed Cohen's d values for region-wise effect size calculations, as the availability of only parametric summary statistics from the Skrobisz study precluded non-parametric effect size measures. For each eTIV-normalized brain region volume and cohort, we calculated the pooled standard deviation as:

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

where n_1 and n_2 are the sample sizes, and s_1 and s_2 are the standard deviations of the two groups, IBS and HC, respectively. Cohen's d effect size was then computed as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups. The 95% confidence interval for d was calculated using:

$$CI_{95\%} = d \pm 1.96\sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}$$

where the standard error term accounts for both sampling variance and uncertainty in the effect size estimate.

An *overall reproducibility score* (S) was developed for each brain region to quantify cross-cohort consistency through three complementary metrics: directional consistency (σ), confidence interval overlap (ω), and effect magnitude (ϵ). The score is computed as: $S = \sigma + \omega + \epsilon$, where the binary indicator σ equals 1 if the direction of effect is consistent between cohorts and 0 otherwise, the binary indicator ω equals 1 if the 95% confidence intervals overlap and 0 otherwise, and ϵ represents the minimum absolute effect size observed across cohorts.

This composite metric prioritizes brain regions exhibiting robust cross-cohort replication, with ϵ providing additional weight to stronger effects. Higher scores (S) indicate greater reproducibility of morphometric findings across independent study populations

and analysis pipelines, thereby establishing a quantitative framework for identifying the most reliable neuroanatomical alterations in IBS.

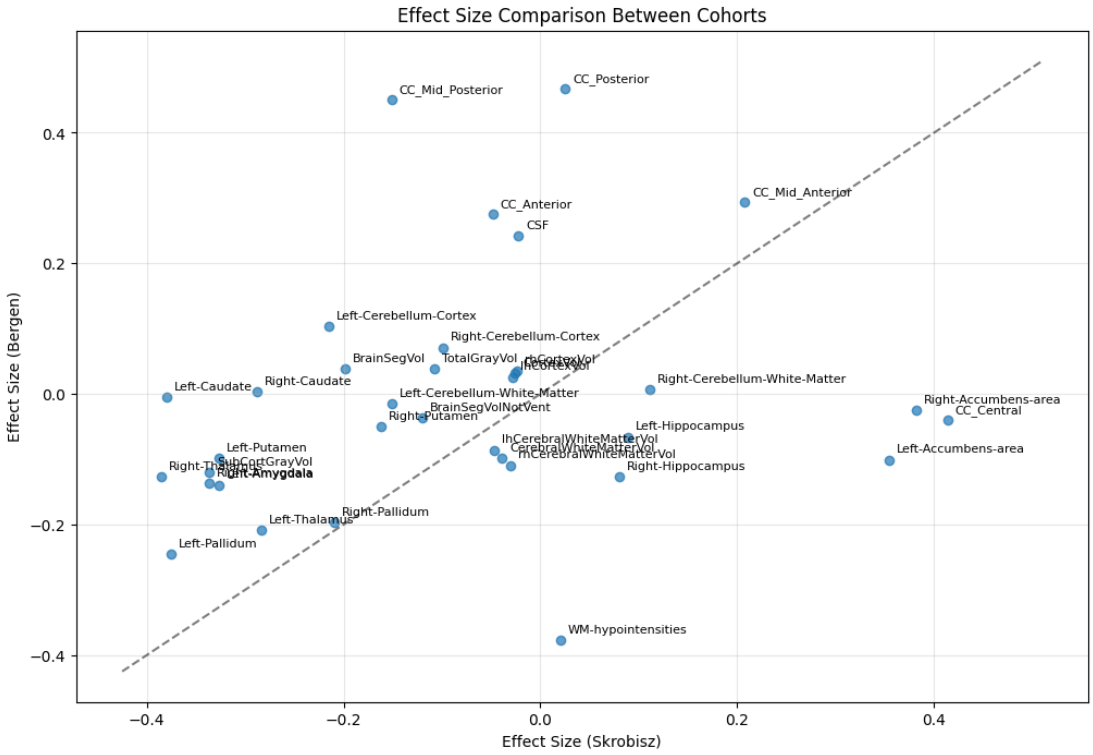


Figure 3. *The large disparity of region-wise effect size of IBS versus HC in comparison between the Skrobisz (2022) cohort and the Bergen cohort*. Scatterplot of calculated Cohen's d effect sizes for each region in both cohorts (see text for details).

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/03-replication-anaysis-fs6.ipynb

The effect size comparison between cohorts revealed moderate correlation ($r = 0.203$, $p = 0.243$). Directional consistency analysis demonstrated that 51.4% of brain regions maintained consistent IBS versus HC differences across cohorts. Notably, all brain regions exhibited overlapping 95% confidence intervals between cohorts, indicating that despite differences in point estimates, the between-cohort variations did not reach statistical significance given measurement uncertainty. Five regions demonstrated particularly strong cross-cohort consistency, achieving the highest overall reproducibility scores (S): mid-anterior corpus callosum (`CC_Mid_Anterior`), `Left-Pallidum`, `Left-Thalamus`, `Right-Pallidum`, and `Left-Amygdala`. These structures showed overall scores ranging from 2.14 to 2.26, suggesting robust replication of IBS-related alterations. Conversely, several regions exhibited marked between-cohort divergence. White matter hypointensities demonstrated particularly discordant effects, while specific corpus callosum segments (`CC_Posterior` and `CC_Mid_Posterior`) showed stronger effects in the Bergen cohort. Cerebellar regions clustered near the origin, indicating consistently modest effects across both cohorts. The overall pattern suggests limited agreement between cohorts in IBS-related brain alterations. While specific structures show robust reproducibility, the widespread dispersion around the diagonal reference line, coupled with moderate correlation, indicates substantial heterogeneity in morphometric findings between these independent samples. This variability may reflect genuine biological heterogeneity in IBS-related brain alterations or methodological differences between studies.

Figure 4 plots a ranking of brain regions on how consistently they show similar patterns between the cohorts.

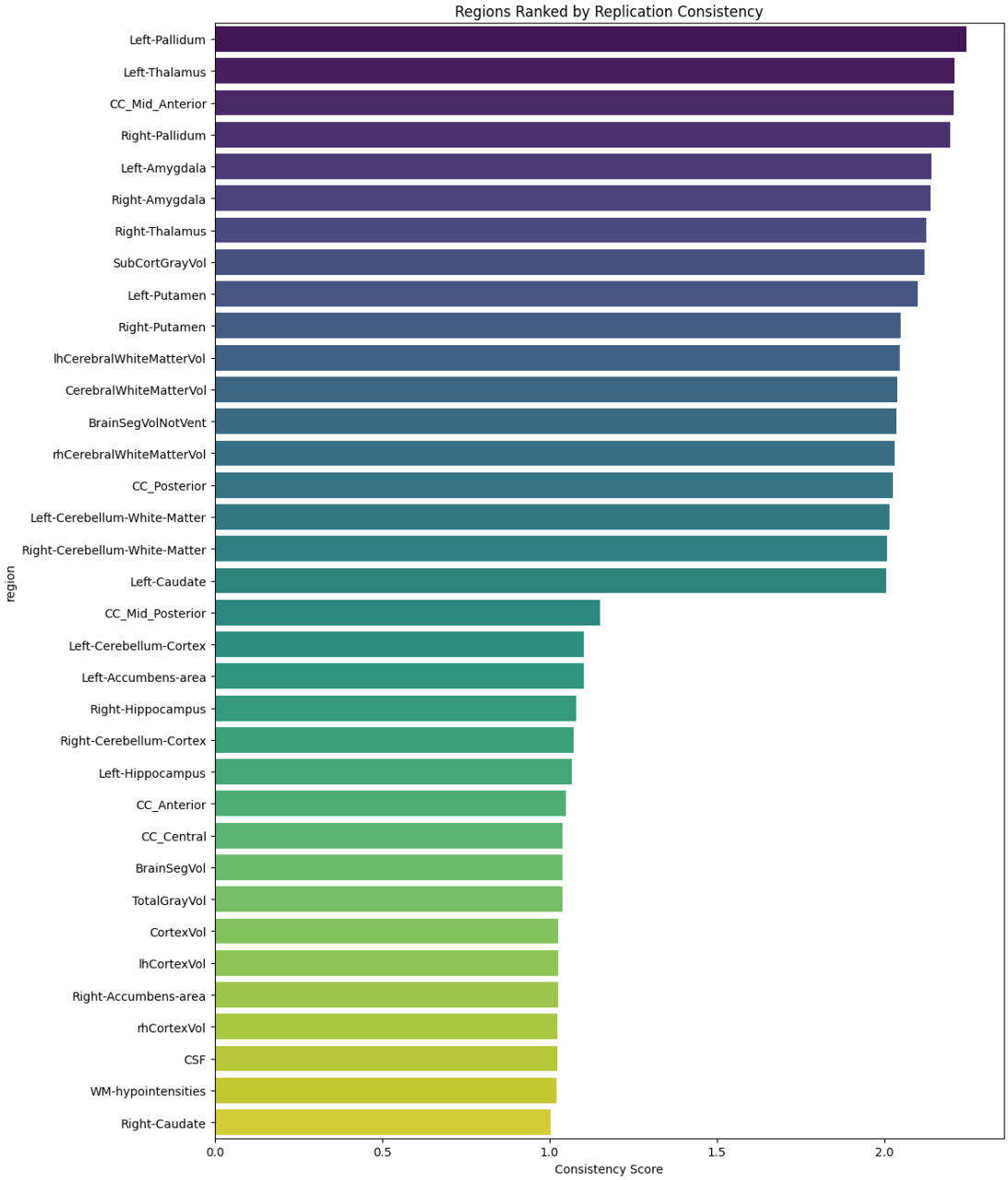


Figure 4. *Brain regions ranked by cross-cohort reproducibility scores*. The composite score (S) integrates directional consistency (σ), confidence interval overlap (ω), and effect magnitude (ϵ) between the Skrobisz and Bergen cohorts. Higher scores indicate greater reproducibility of IBS-related volumetric alterations across independent samples. See text for detailed scoring methodology.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/03-replication-anaysis-fs6.ipynb

The reproducibility analysis revealed varying degrees of cross-cohort consistency in brain structural alterations associated with IBS. Several regions demonstrated robust reproducibility, with the `Left-Pallidum`, `Left-Thalamus`, and `CC_Mid_Anterior` achieving overall scores (S) exceeding 2.0. These high-scoring regions exhibited both directional consistency and complete confidence interval overlap, coupled with substantial effect magnitudes, suggesting reliable IBS-related volumetric alterations across independent samples. Conversely, regions including the `Right-Caudate`, `Right-Cerebellum-Cortex`, `Left-` and `Right-Hippocampus`, `CC_Mid_Posterior`, and `Left-Cerebellum-Cortex` showed lower reproducibility (scores approximately 1.1). While these regions maintained confidence interval overlap, they lacked directional consistency between cohorts, suggesting greater

variability in IBS-related effects. Despite systematic between-cohort differences in eTIV-normalized volumes, certain regions demonstrated consistent relative patterns of alteration. However, our attempt to replicate the specific morphometric differences reported by Skrobisz et al. (2022) yielded limited success. This suggests that structural brain alterations in IBS may be more heterogeneous than previously recognized, potentially reflecting the complex nature of IBS pathophysiology or methodological variations across studies.

To assess the robustness of brain morphometry measurements in IBS research, we conducted a comprehensive analysis of the Bergen cohort data using multiple FreeSurfer processing pipelines. This systematic evaluation examined the stability of morphometric measurements and IBS versus healthy control (HC) group differences across different analytical approaches: FreeSurfer versions (6.0.1 versus 7.4.1) and processing streams within FreeSurfer 7.4.1 (cross-sectional versus longitudinal). Our interventional study design enabled the application of the longitudinal processing stream, providing an additional dimension for assessing measurement reliability. Unlike our previous replication analysis of the Skrobisz (2022) cohort, which relied on summary statistics, this comparison utilized complete morphometric data from all participants, allowing for more detailed assessment of measurement consistency.

Cross-Version Comparison of FreeSurfer Morphometric Measurements

We examined the consistency of volumetric measurements between FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional stream) in quantifying brain structural differences between IBS patients and healthy controls (HC). Table A2 in the Appendix presents group-wise summary statistics (mean and standard deviation) for both IBS patients and healthy controls, derived from the `aseg.stats` files generated by each FreeSurfer version. Figure 5 presents a scatter plot matrix illustrating version-wise comparisons for each brain region. Individual plots display FS 6.0.1 volumes against corresponding FS 7.4.1 measurements, with HC and IBS participants distinguished by blue and red markers, respectively. Reference identity lines facilitate direct assessment of cross-version measurement concordance.

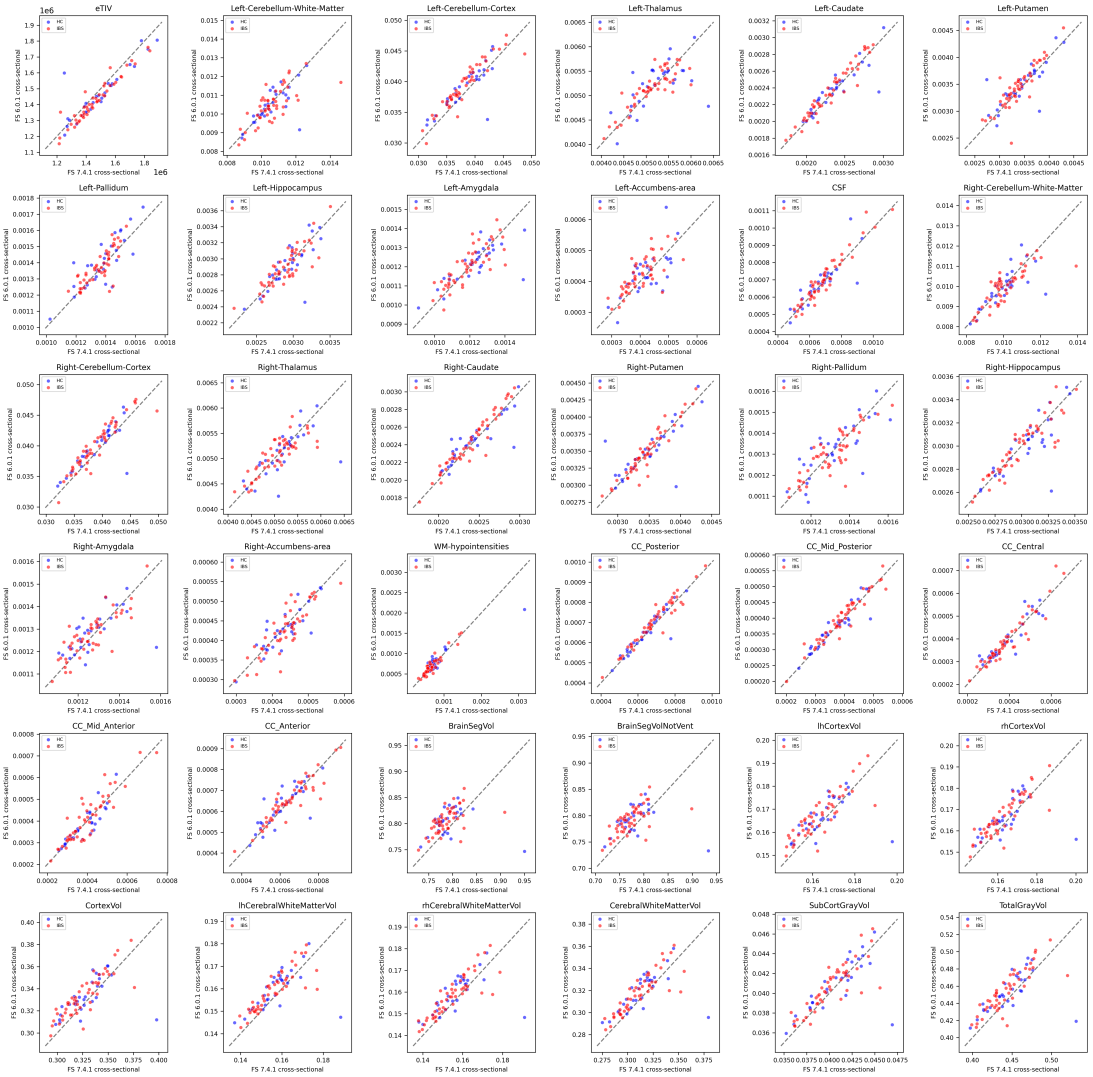


Figure 5. *Comparison of FreeSurfer-derived regional brain volumes across versions 6.0.1 and 7.4.1 (cross-sectional processing).* Scatter plots show eTIV-normalized volumes for each brain region, with version 6.0.1 values on the *y*-axis versus version 7.4.1 on the *x*-axis. The eTIV-volume [mm^3] is shown in the upper left panel. Blue and red markers denote healthy controls and IBS patients, respectively. Identity lines indicate perfect cross-version agreement. See text for detailed analysis.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-comparing-FS-versions-on-same-dataset.ipynb

The scatter plot matrix demonstrates varying degrees of consistency between FreeSurfer versions 6.0.1 and 7.4.1 across different brain regions. Subcortical regions, particularly the thalamus, caudate, putamen, and partly the hippocampus, show strong cross-version agreement with minimal deviation from the identity line. However, systematic differences emerge in several structures: the amygdala and the accumbens demonstrate moderate version-dependent variability, with data points showing systematic deviation from perfect concordance. Corpus callosum segments display region-specific variations in cross-version agreement, with `CC_Anterior` and `CC_Mid_Anterior` showing more pronounced differences compared to other segments. Importantly, the distribution patterns of IBS (red) and healthy control (blue) groups remain fairly consistent across versions, suggesting that while absolute volume estimates may differ between FreeSurfer versions, the relative group differences are largely preserved.

Notably, several regions exhibit strong correlations between versions but with systematic offsets from the identity line, indicating consistent biases between FreeSurfer versions

6.0.1 and 7.4.1. For example: The cortical measurements (`lhCortexVol` and `rhCortexVol`) and `lh-` and `rhCerebralWhiteMatterVol` and `TotalGrayVol` show a clear parallel offset above the identity line, indicating that FreeSurfer 6.0.1 consistently produces higher volume estimates compared to version 7.4.1. This systematic bias appears consistent across the full range of eTIV-normalized volumes and both subject groups. Similar parallel offsets are visible in `Left-` and `Right-Cerebellum-Cortex` and subcortical structures like the `Left-Pallidum` and `Left-Caudate`. Moreover, the eTIV shows systematic higher volumes in version 7.4.1 compared to version 6.0.1 measurements.

Several key structures exhibit individual outliers that warrant attention. In `eTIV`, a single measurement shows substantial deviation, suggesting potential segmentation challenges in this particular case. The `Left-` and `Right-Hippocampus` both show isolated outliers (visible as blue points) significantly deviating from the otherwise tight correlation pattern, indicating potential segmentation inconsistencies between versions for these specific control subjects. The `Left-Thalamus` displays a particularly notable outlier (blue point) that deviates substantially below the main correlation pattern, suggesting a case where version 7.4.1 produced a markedly lower volume estimate compared to version 6.0.1. Similar isolated discrepancies appear in both `Left-` and `Right-Amygdala` measurements, where single data points (again from the control group) deviate notably from the otherwise consistent version correlation. These individual outliers likely represent cases where the segmentation algorithms in the two FreeSurfer versions interpreted the anatomical boundaries differently, possibly due to image quality issues, anatomical variants, or differences in how the versions handle boundary cases. The fact that many of these outliers appear in the control group (blue points) suggests that these discrepancies are not specifically related to IBS pathology but rather to technical aspects of the segmentation process.

These observations underscore the importance of version consistency in morphometry-based classification studies and suggest that meta-analyses or multi-site studies should carefully account for FreeSurfer version effects in their analytical pipelines.

In this context, Figure 6 depicts a scatter plot matrix comparing brain region volumes between two pipelines (cross-sectional and the longitudinal stream) using the *same* FreeSurfer 7.4.1 version, highlighting potential discrepancies.

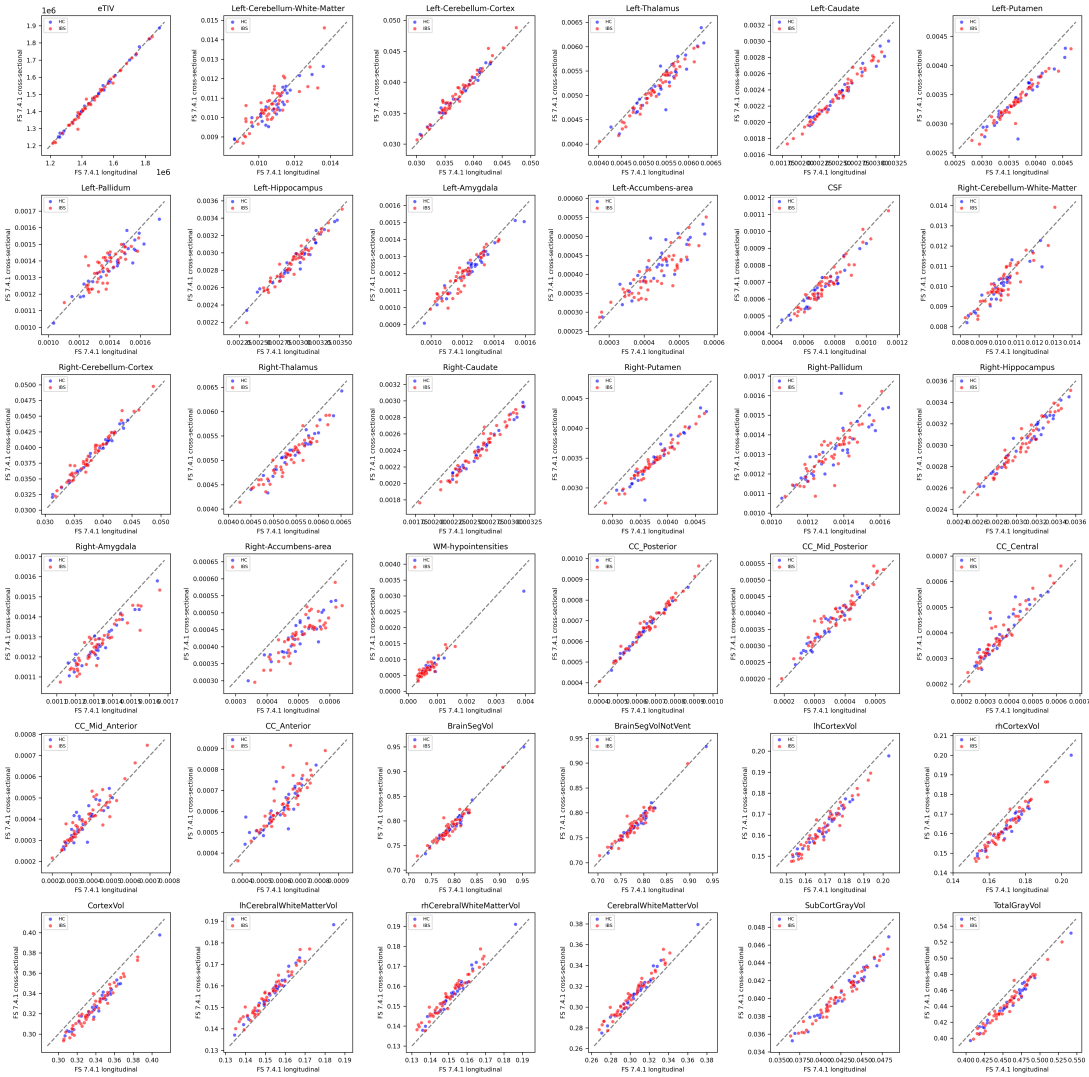


Figure 6. *Comparison of regional brain volumes between FreeSurfer 7.4.1 cross-sectional and longitudinal processing streams.* Scatter plots show eTIV-normalized volumes [eTIV in mm^3] for each brain region, with cross-sectional values on the *y*-axis versus longitudinal values on the *x*-axis. Blue and red markers denote healthy controls and IBS patients, respectively. Identity lines indicate perfect cross-stream agreement. See text for detailed analysis.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-comparing-FS-versions-on-same-dataset.ipynb

The comparison between FreeSurfer 7.4.1's cross-sectional and longitudinal processing streams reveals distinct patterns of agreement and systematic variation across brain regions. Global measurements (`BrainSegVol`, `BrainSegVolNotVent`) demonstrate strong cross-stream consistency, with tight clustering along the identity line. However, substantial systematic differences emerge in several key structures. Most notably, cortical volumes (`lhCortexVol`, `rhCortexVol`) exhibit a clear systematic bias, with longitudinal processing consistently producing higher volume estimates compared to the cross-sectional stream. This pattern contrasts with `Left-` and `Right-Cerebellum-Cortex`, where longitudinal processing yields systematically lower estimates. Subcortical structures display varying degrees of processing stream sensitivity: the putamen and caudate show consistent offsets from the identity line, while pallidum and accumbens measurements demonstrate greater scatter. Corpus callosum segments (`CC_Anterior`, `CC_Mid_Anterior`, `CC_Central`) reveal processing stream-dependent variations that differ from those observed in other structures. Looking at the `eTIV` plot in the top-left panel, it shows remarkably high consistency be-

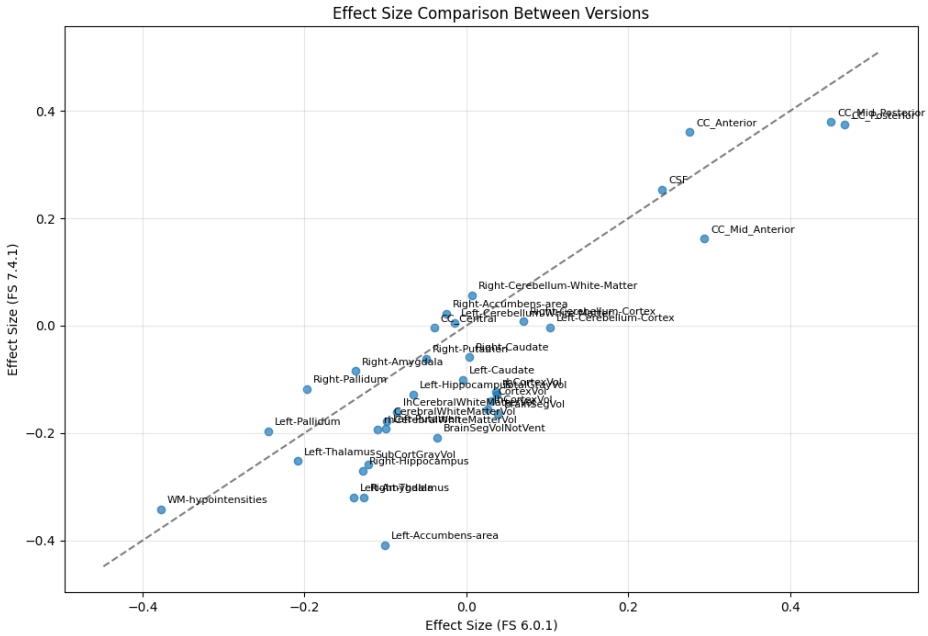
tween cross-sectional and longitudinal processing streams. The data points cluster tightly along the identity line across the full range of values (approximately 1.2-1.8 × 10^6 mm^3), with minimal deviation. This strong agreement in `eTIV` estimations between processing streams is particularly noteworthy because `eTIV` serves as the normalization factor for all other volumetric measurements. The consistency suggests that any observed differences in other brain regions are not attributable to variations in total intracranial volume estimation between processing streams, but rather reflect genuine methodological differences in how the two streams segment specific structures.

Importantly, these systematic biases maintain consistency across both IBS and healthy control groups, as evidenced by the parallel patterns of red and blue markers. This indicates that while absolute volume estimates differ between processing streams, the relative group differences remain largely preserved. These findings underscore the critical importance of maintaining consistent processing stream selection when conducting cross-sectional comparisons or longitudinal analyses in clinical studies.

The summary statistics by the mean and standard deviation for Freesurfer v. 7.4.1 cross-sectional and v. 7.4.1 longitudinal stream, respectively, are shown in the Appendix as Table A3.

Figure 7 illustrates the differential impact of FreeSurfer processing choices on IBS versus healthy control effect sizes across brain regions. Panel (a) compares effect sizes between FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional), while panel (b) contrasts effect sizes derived from FreeSurfer 7.4.1's cross-sectional and longitudinal processing streams, enabling assessment of both version and pipeline-specific influences on group differences.

a)



b)

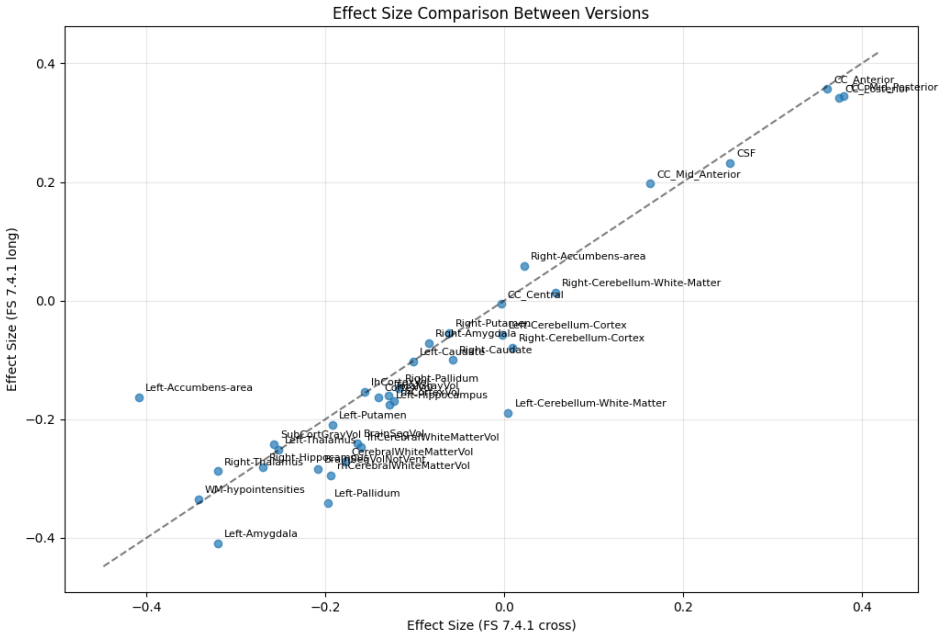


Figure 7. *Effect sizes (Cohen's d) of IBS versus healthy control group differences across brain regions: comparison of FreeSurfer methodological variants.* a) Cross-sectional processing stream comparison between FreeSurfer versions 6.0.1 and 7.4.1. b) Processing stream comparison within FreeSurfer 7.4.1 (cross-sectional versus longitudinal). See text for detailed analysis.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-computing-FS-versions-on-same-dataset.ipynb

The scatter plots reveal distinct patterns in how FreeSurfer methodological choices affect IBS versus healthy control effect sizes across brain regions. Panel (a), comparing

FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional), demonstrates moderate agreement with notable version-specific variations. Key corpus callosum segments (`CC_Anterior`, `CC_Mid_Posterior`) show the strongest positive effect sizes (approximately 0.4) and maintain relative consistency across versions. In contrast, the `Left-Accumbens-area` exhibits the strongest negative effect (approximately -0.4), with its magnitude varying between versions. Panel (b), comparing cross-sectional and longitudinal streams within FreeSurfer 7.4.1, shows that corpus callosum segments maintain their position as regions with the strongest positive effects, while the `Left-Amygdala` and `Left-Accumbens-area` show pronounced negative effects. Most subcortical structures cluster more tightly around the diagonal compared to the version comparison in panel (a). The longitudinal versus cross-sectional comparison demonstrates greater overall consistency than the version comparison, as evidenced by tighter clustering along the diagonal reference line. This suggests that processing stream selection within FreeSurfer 7.4.1 introduces less variability in effect size estimates than version changes. However, specific regions, particularly in the limbic system, show sensitivity to processing stream choice. This systematic comparison highlights that while both FreeSurfer version and processing stream selection affect effect size estimates, version differences generally introduce more variability than processing stream choices within the same version.

Figure 8 quantifies the reproducibility of IBS versus healthy control group differences across brain regions under different FreeSurfer methodological variants. Panel (a) ranks regions by their effect size consistency (S) between FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional), while panel (b) presents regional rankings based on effect size stability between cross-sectional and longitudinal processing streams within FreeSurfer 7.4.1, enabling systematic assessment of both version and pipeline-dependent variations.

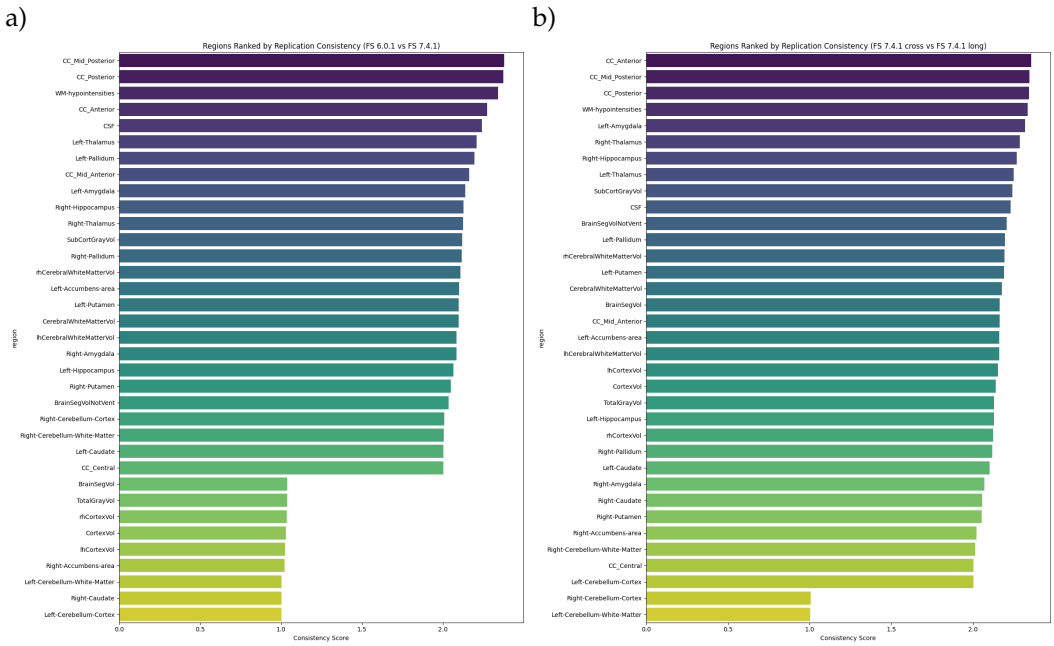


Figure 8. *Brain regions ranked by reproducibility of IBS versus healthy control differences across FreeSurfer methodological variants.* Composite scores (S) combine directional consistency (σ), confidence interval overlap (ω), and effect magnitude (ϵ). Panel (a) compares FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional); panel (b) contrasts cross-sectional and longitudinal processing streams within FreeSurfer 7.4.1. See text for detailed analysis.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-computing-FS-versions-on-same-dataset.ipynb

The regional consistency scores reveal distinct patterns in how FreeSurfer methodological choices affect the reproducibility of IBS versus healthy control differences. Panel (a),

comparing FreeSurfer versions 6.0.1 and 7.4.1 (cross-sectional), shows a gradual distribution of consistency scores ranging from 1.0 to 2.5. Corpus callosum regions (`CC_Mid_Posterior`, `CC_Posterior`) demonstrate the highest consistency, while cerebellar structures show the lowest. Subcortical regions exhibit intermediate consistency, suggesting moderate stability across FreeSurfer versions. Panel (b), comparing cross-sectional and longitudinal streams within FreeSurfer 7.4.1, reveals a more distinct clustering pattern. The `CC_Anterior` and `CC_Mid_Posterior` maintain high consistency, but notably, limbic structures like the `Left-Amygdala` and `Right-Thalamus` show improved consistency compared to their version-wise rankings. This suggests that these regions are more sensitive to FreeSurfer version changes than to processing stream selection. The overall pattern indicates stronger methodological stability when varying processing streams within FreeSurfer 7.4.1 compared to cross-version analyses. Importantly, comparing these methodological variations within the same cohort yields higher consistency scores than the previous cross-cohort comparison (Fig. 4), highlighting the substantial impact of cohort-specific factors on brain morphometric findings in IBS research.

Multivariate analyses: IBS versus HC

The multivariate normality of brain structural data was assessed across three FreeSurfer processing streams using Mardia's test (examining skewness and kurtosis) and the Henze-Zirkler's test. For FS 6.0.1, Mardia's test revealed significant deviations in both skewness ($b_{1,p} = 2.33 \times 10^{14}$, $p < 0.001$) and kurtosis ($b_{2,p} = -8.77$, $p < 0.001$) for the full sample, with similar patterns in the IBS group but different skewness characteristics in the HC group. For FS 7.4.1 cross-sectional, both groups showed significant non-normality, with particularly extreme values in the IBS group (kurtosis statistic = 153.63, $p < 0.001$). The FS 7.4.1 longitudinal analysis also indicated significant departures from multivariate normality across all groups. The Henze-Zirkler's test showed some numerical instability issues, evidenced by extreme values and negative test statistics, suggesting that its results should be interpreted with caution. Overall, these findings consistently indicate significant departures from multivariate normality across all FreeSurfer versions and subject groups, with particularly pronounced effects in the IBS group. This suggests that robust statistical methods should be employed for subsequent analyses of group differences in brain structure.

In this context, the robust Mahalanobis distance analysis was implemented to quantify the multivariate separation between IBS and HC groups across different FreeSurfer processing streams while accounting for potential outliers and non-normality in the neuroimaging data. The computation employs winsorization at the 10th and 90th percentiles to mitigate the impact of extreme values, followed by robust location estimation using medians instead of means. The analysis revealed decreasing Mahalanobis distances across FreeSurfer versions: FS 6.0.1 showed the largest separation ($D = 9.348$, $F = 0.598$, $p = 0.939$), followed by FS 7.4.1 cross-sectional ($D = 6.068$, $F = 0.252$, $p \approx 1.000$) and FS 7.4.1 longitudinal ($D = 5.163$, $F = 0.183$, $p \approx 1.000$). However, none of these distances reached statistical significance (all $p > 0.05$), suggesting that the multivariate brain volume differences between IBS and HC groups are not statistically meaningful across any of the FreeSurfer processing streams. The consistently high p-values and low F-statistics indicate that, despite the apparent numerical differences in Mahalanobis distances, there is insufficient evidence to conclude that the IBS and HC groups differ significantly in their multivariate brain volume profiles. This analysis, incorporating 35 brain regions and accounting for their covariance structure, suggests that the volumetric differences between IBS and HC groups are not robust enough to clearly distinguish between the groups in a multivariate framework.

To further investigate potential group differences beyond the initial Mahalanobis distance analysis, we employed a machine learning framework with cross-validation to assess IBS versus healthy control discriminability and identify the most diagnostically relevant brain structures. This complementary approach enables systematic evaluation of

multivariate patterns while accounting for potential interactions between brain regions.

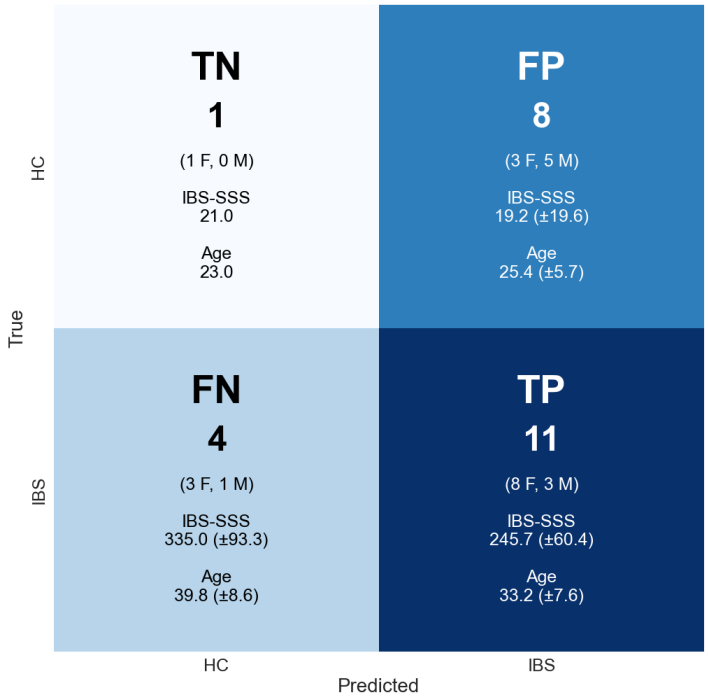
Machine Learning-Based Classification Using Brain Morphometry

We evaluated the discriminative power of brain morphometric features for IBS versus healthy control classification using the PyCaret machine learning library. Multiple classification algorithms were trained and compared (Appendix Fig.A1) using FreeSurfer 7.4.1 longitudinal stream measurements from the Bergen cohort (Table 2). We applied a binary classification framework to distinguish between healthy controls (0) and IBS patients (1) based on brain morphometric features. The dataset comprised 78 participants characterized by 37 numerical features, partitioned into training (n=54) and test (n=24) sets. We employed stratified 10-fold cross-validation to maintain consistent class proportions across folds. Feature preprocessing included mean-based imputation and standardization to zero mean and unit variance, particularly crucial for features with widely differing scales (e.g., raw eTIV values $> 1.2 \cdot 10^6$ versus eTIV-normalized measures < 1). Given the modest dataset size, analyses were performed using CPU computation. All random processes were controlled through a fixed session identifier to ensure reproducibility.

Model performance evaluation across 15 classification algorithms revealed Extreme Gradient Boosting (XGBoost) as the superior approach for IBS versus healthy control discrimination based on brain morphometry (details in Fig A1). XGBoost achieved the highest performance metrics: accuracy (0.72), AUC (0.68), recall (0.72), precision (0.74), and F1 score (0.71). The model's Cohen's Kappa (0.40) and Matthews Correlation Coefficient (0.42) indicate substantial improvement over chance-level classification. K-Nearest Neighbors demonstrated the second-best performance, while Logistic Regression and Support Vector Machines showed moderate discriminative ability. Several algorithms, including AdaBoost and Linear Discriminant Analysis, performed near chance level, as benchmarked against a dummy classifier baseline. XGBoost's superior performance suggests its ability to capture complex, nonlinear relationships in brain morphometric features that distinguish IBS from healthy controls.

The best-performing model (XGBoost) demonstrated mixed classification performance on the hold-out test set, as shown in Figure 9a. The model correctly identified 73% of IBS patients (11/15 cases; 8 female, 3 male; IBS-SSS: 245.7 $\pm$ 60.4; age: 33.2 $\pm$ 7.6). However, specificity was low at 11%, with 8 of 9 healthy controls misclassified as IBS (3 female, 5 male; IBS-SSS: 19.2 $\pm$ 19.6; age: 25.4 $\pm$ 5.7), yielding an overall accuracy of 50% (12/24). This asymmetric performance reveals systematic patterns: correctly classified IBS patients showed higher symptom severity scores (IBS-SSS), female predominance, and higher mean age compared to misclassified controls. The strong bias toward IBS classification suggests that while brain morphometric features contain discriminative information, additional refinement is needed for reliable diagnostic application.

a)



b)

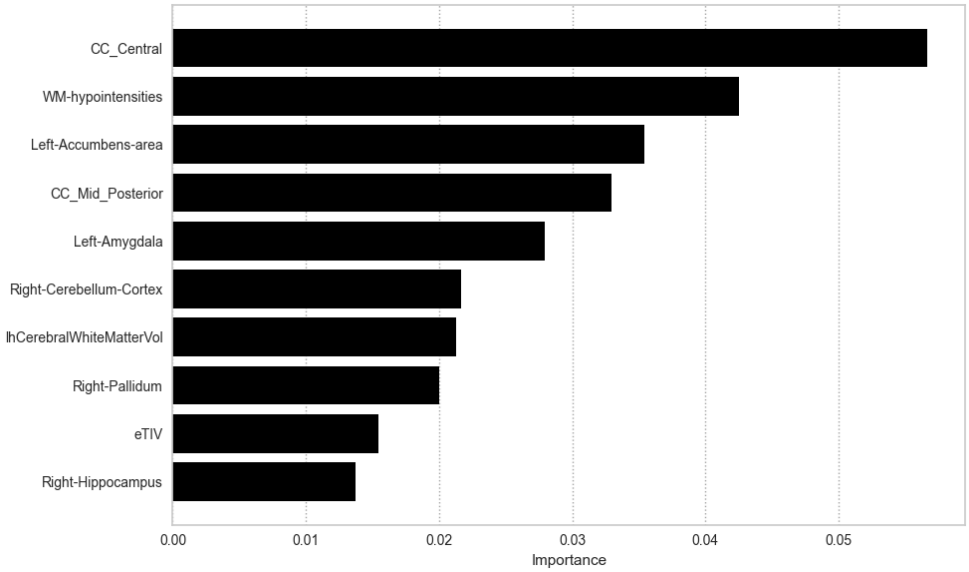


Figure 9. *Machine learning-based discrimination between IBS and healthy controls using brain morphometry.* a) Confusion matrix showing prediction outcomes from XGBoost classification on the test dataset, with quadrants indicating true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). b) Ten most discriminative brain regions identified through permutation importance analysis in the XGBoost model.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/05-predicting-IBS-vs-HC-from-morphometric-measures.ipynb

Permutation importance analysis revealed the relative contribution of brain regions to IBS versus healthy control classification. The central corpus callosum (`CC_Central`)

emerged as the most discriminative feature ($\approx 0.057 \pm 0.038$), followed by white matter hypointensities ($\approx 0.043 \pm 0.029$) and the left nucleus accumbens ($\approx 0.035 \pm 0.040$). A second tier of discriminative regions includes the mid-posterior corpus callosum ($\approx 0.033 \pm 0.029$) and left amygdala ($\approx 0.028 \pm 0.045$), while cerebellar structures showed moderate importance (right cerebellar cortex ($\approx 0.022 \pm 0.026$). Notably, several traditionally studied regions in IBS, including the hippocampus ($\approx 0.014 \pm 0.045$) and total intracranial volume ($\approx 0.015 \pm 0.028$), demonstrated relatively lower discriminative power. This hierarchy suggests that white matter structures, particularly corpus callosum segments, may play a more prominent role in IBS-related brain alterations than previously recognized. However, the permutation importance ranking should be interpreted cautiously given the large standard deviations and the model's modest classification performance (50% accuracy, 73% sensitivity but only 11% specificity). While the ranking identifies features that contribute most to the model's decisions, these contributions come from a model that shows strong bias toward IBS classification and poor discriminative ability for healthy controls.

To gain deeper insight into how individual brain regions influence the model's classification decisions, we employed SHAP (SHapley Additive exPlanations) analysis. Figure 10 visualizes the contribution of each morphometric feature to individual predictions, with SHAP values indicating both the direction and magnitude of each feature's impact. This analysis extends beyond traditional feature importance rankings by revealing how specific volumetric measurements drive classification outcomes on a case-by-case basis. High feature values (red) and low feature values (blue) can contribute differently to the model's decisions, providing a more nuanced understanding of the relationship between brain morphometry and IBS classification than permutation importance alone. The figure reveals complex patterns in how morphometric features influence predictions. For example, high values (red) in the right caudate tend to push predictions toward IBS (positive SHAP values), while low values (blue) in this region tend to predict healthy control. This asymmetric impact of feature values suggests nonlinear relationships between brain structure volumes and IBS classification that may not be captured by simpler univariate analyses.

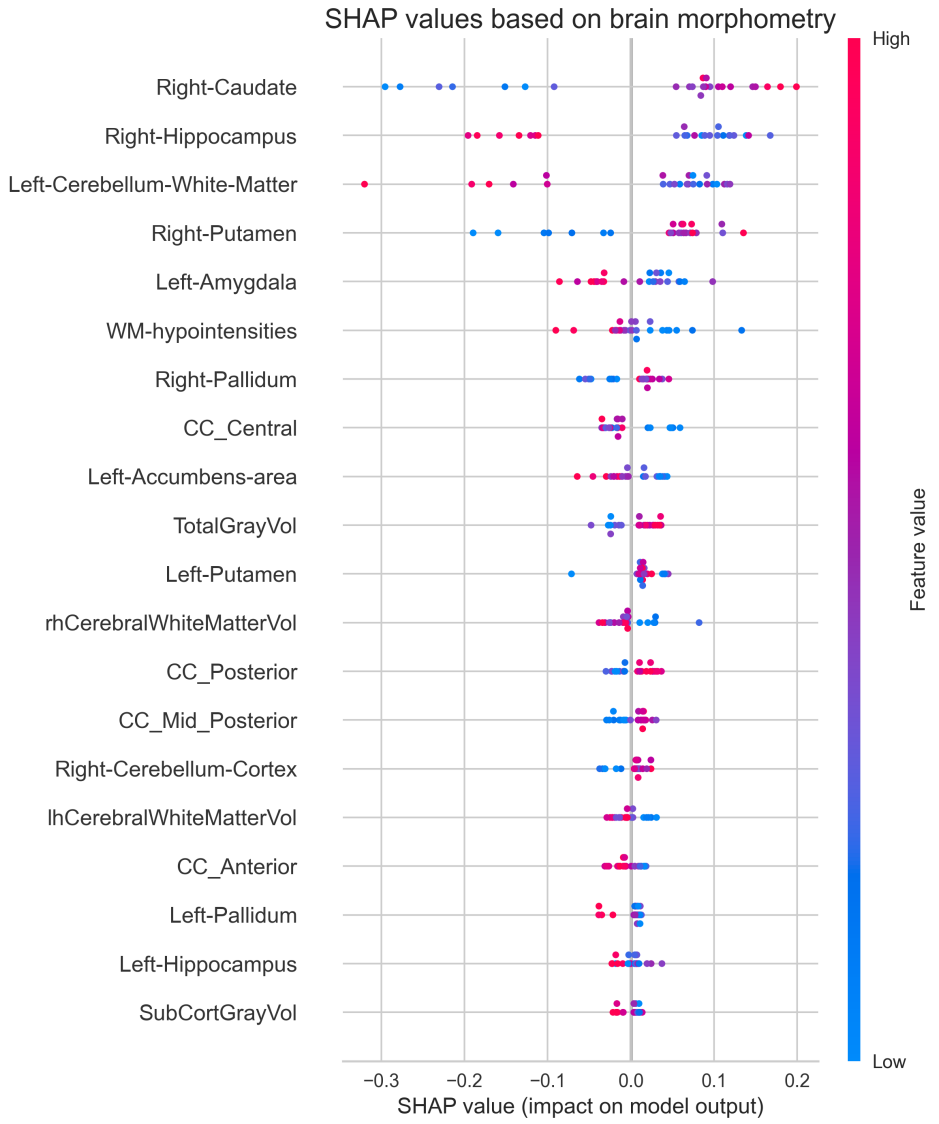


Figure 10. *Feature contribution analysis using SHAP values for brain morphometry-based classification*. SHAP values (x-axis) quantify each region's impact on classification probability, with negative values favoring healthy control classification and positive values favoring IBS. Each point represents a single prediction, with color indicating the relative magnitude of the morphometric measurement (red: high values, blue: low values). The distribution of points reveals how feature values influence model predictions across individual cases.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/05-predicting-IBS-vs-HC-from-morphometric-measures.ipynb

The SHAP analysis reveals more nuanced feature contributions than the permutation importance ranking, while also showing some notable consistencies. `CC_Central` ranks highest in permutation importance and shows meaningful SHAP values, but with complex patterns where both high and low values contribute to classification. Similarly, `CC_Mid_Posterior` shows similar importance in both analyses, with relatively consistent effects. White matter features, particularly `WM-hypointensities`, rank high in both analyses, suggesting robust importance, with SHAP patterns indicating that higher values tend to predict healthy controls. Among subcortical structures, the `Left-Accumbens-area` appears important in both analyses, with SHAP values showing that lower volumes tend to predict IBS. The `Left-Amygdala` shows moderate importance in both analyses, with high values generally predicting healthy controls. Notable differences emerge: the `Right-Caudate` shows strong SHAP value patterns but does not appear in the top permutation importance

features, while the `Right-Hippocampus` ranks lower in permutation importance but shows distinct SHAP patterns. This comparison suggests that while some features (like corpus callosum regions and white matter hypointensities) show consistent importance across methods, the SHAP analysis reveals more complex relationships between feature values and model predictions. This richer characterization of feature contributions might explain some of the model's classification biases, particularly given the observed asymmetric effects where high and low values of the same feature can have different impacts on predictions. However, these feature contribution analyses must (again) be interpreted in the context of the model's modest classification performance (50% accuracy, 73% sensitivity, 11% specificity). The SHAP values and permutation importance rankings identify features that drive the model's decisions, but given the strong bias toward IBS classification, these patterns may reflect systematic misclassification rather than truly discriminative neuroanatomical markers. The complex feature interactions revealed by SHAP analysis might partially explain the model's poor specificity, suggesting that while consistent morphometric patterns exist, they are insufficient for reliable diagnostic classification without additional clinical information.

Univariate Analysis of Cognitive Performance

To assess potential cognitive differences between IBS patients and healthy controls, we analyzed performance across multiple cognitive domains using the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). Table 4 presents group comparisons of the full-scale score and five cognitive domain indices, using non-parametric statistics to account for potential non-normal distributions.

Table 4. *A non-parametric analysis comparing cognitive features in the IBS and HC groups*

Variable	HC	IBS	p-value	Cliff's delta
Fullscale_RBANS	103.0 (93.0-108.0)	91.0 (85.0-100.0)	0.002	0.213
Memory_Index	100.0 (86.0-109.0)	86.0 (81.0-105.0)	0.031	0.147
Visuospatial_Index	97.0 (90.0-107.0)	96.0 (90.0-105.0)	0.763	0.021
Verbal skills Index	105.0 (95.0-113.0)	95.0 (89.0-111.0)	0.087	0.116
Attention_Index	98.0 (89.0-108.0)	97.0 (83.0-101.0)	0.118	0.107
Recall_Index	107.0 (92.0-113.0)	95.0 (85.0-100.0)	0.006	0.186

Values presented as median (interquartile range, IQR). Group differences assessed using Mann-Whitney U tests (uncorrected p-values), where we multiply by 6 to get the corrected values. Effect sizes quantified using Cliff's delta, where positive values indicate higher scores in healthy controls.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/06-morphometry-cognition-exploration.ipynb

Comparison of cognitive performance between IBS patients and healthy controls using RBANS revealed domain-specific differences. After Bonferroni correction ($\alpha = 0.05/6$), two measures showed significant group differences: the Full-scale RBANS score ($p_{corrected} = 0.012$) was lower in IBS patients (median = 91.0, IQR = 85.0-100.0) compared to controls (median = 103.0, IQR = 93.0-108.0), with a small to moderate effect size (Cliff's $\delta = 0.213$). Similarly, the Recall Index demonstrated significantly lower performance ($p_{corrected} = 0.036$) in IBS patients (median = 95.0, IQR = 85.0-100.0) compared to controls (median = 107.0, IQR = 92.0-113.0; $\delta = 0.186$). Other cognitive domains showed no significant differences after correction: Memory Index ($p_{corrected} = 0.186$, $\delta = 0.147$), Visuospatial Index ($p_{corrected} = 1.000$, $\delta = 0.021$), Verbal skills Index ($p_{corrected} = 0.522$, $\delta = 0.116$), and Attention Index ($p_{corrected} = 0.708$, $\delta = 0.107$). These findings suggest selective cognitive differences in IBS, particularly affecting overall cognitive function and recall abilities, while other domains remain relatively preserved. The use of Bonferroni correction provides strong control of Type I error (false positive) rate in these multiple comparisons, though its conservative nature may increase the risk of Type II errors (failing to detect true differences).

Relationship Between Brain Morphometry and Cognitive Performance

To investigate potential links between brain structure and cognitive function, we examined pairwise correlations between regional brain volumes and RBANS cognitive scores. Figure 11 presents a correlation matrix using Spearman's rank correlation, capturing both linear and nonlinear monotonic relationships while remaining robust to outliers and non-normal distributions. This comprehensive analysis includes both morphometric features (regional volumes normalized by eTIV) and cognitive performance measures across multiple domains.

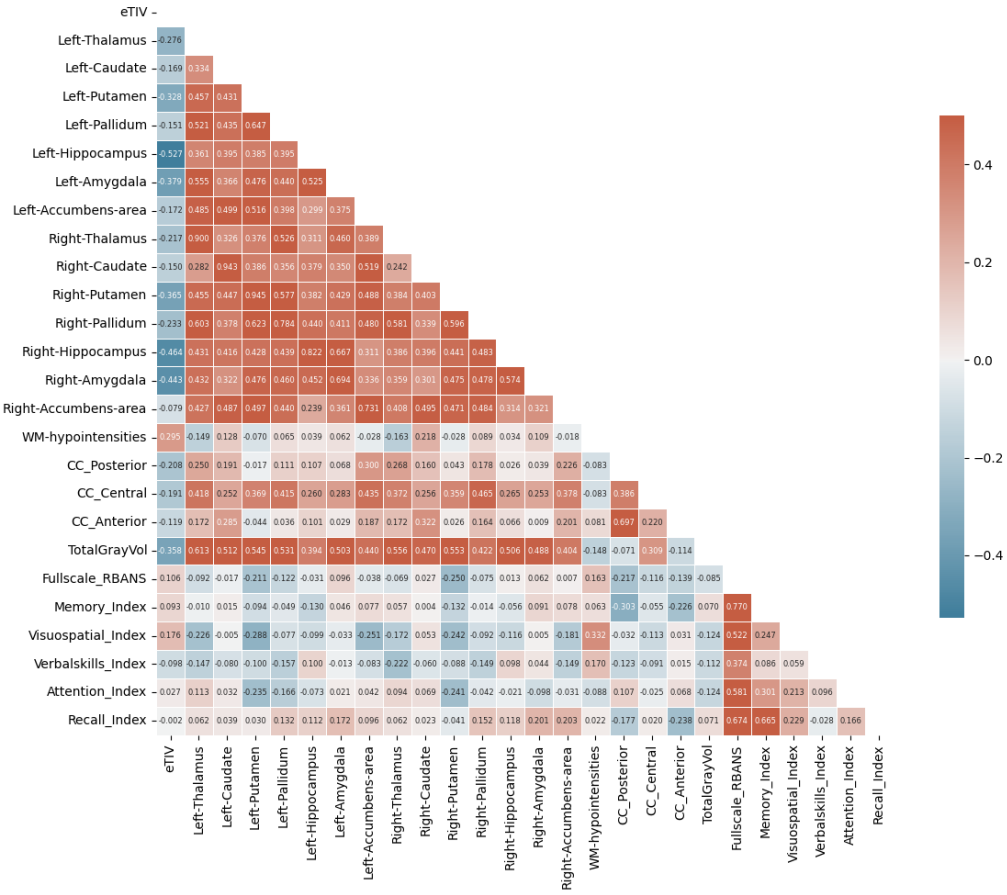


Figure 11. *Brain structure-cognition relationships visualized through Spearman rank correlations*. The correlation matrix shows pairwise associations between regional brain volumes (eTIV-normalized) and RBANS cognitive scores. Red indicates positive correlations, blue indicates negative correlations, with color intensity reflecting correlation strength.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/06-morphometry-cognition-exploration.ipynb

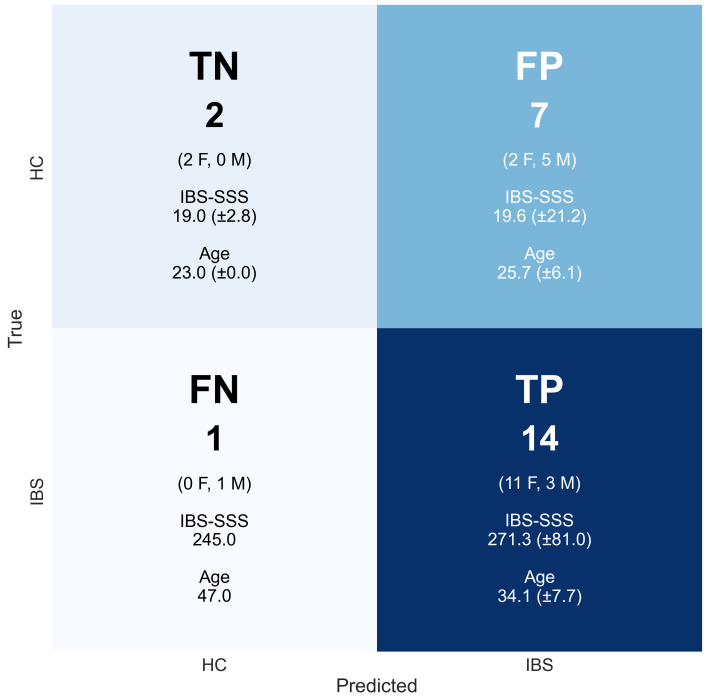
The correlation analysis reveals three distinct patterns of relationships. First, strong bilateral symmetry is evident in subcortical structures, with high correlations between corresponding left and right regions (hippocampus: $\rho \approx 0.8$, amygdala: $\rho \approx 0.7$, putamen: $\rho \approx 0.9$). Second, anatomical relationships appear preserved, with `TotalGrayVol` showing expected moderate correlations with subcortical structures ($\rho \approx 0.4 - 0.6$) and corpus callosum segments displaying varying degrees of inter-relationship ($\rho \approx 0.2 - 0.7$). However, the structure-function relationships, as measured by correlations between brain

morphometry and cognitive performance, are notably weak. The `Fullscale_RBANS` shows minimal correlations with regional volumes ($|\rho| < 0.25$), and even theoretically linked relationships, such as between `Memory_Index` and medial temporal structures, demonstrate weak associations ($|\rho| < 0.15$). An unexpected finding is the moderate correlation between `WM-hypointensities` and `Visuospatial_Index` ($\rho = 0.33$), while other structure-function correlations remain weak ($|\rho| < 0.15$). Within cognitive measures, moderate to strong inter-correlations exist among most RBANS indices, particularly between memory-related measures (`Recall_Index` and `Memory_Index`: $\rho = 0.67$), suggesting preserved cognitive domain relationships despite weak associations with brain structure. This pattern indicates that the relationship between brain morphometry and cognitive function in IBS may be more complex than direct structure-function mappings would suggest.

Multimodal Classification of IBS Using Brain Structure and Cognitive Measures

To evaluate whether combining brain morphometry with cognitive performance improves diagnostic classification, we implemented machine learning models using both feature types. We systematically compared classification performance between models trained on morphometric features alone versus those incorporating both morphometric and cognitive measures. Figure 12a presents the detailed classification outcomes, while Figure 12b shows the relative importance of combined features in the model's decision-making. Table 5 quantifies the impact of feature combination through multiple performance metrics. Results are shown for the XGBoost model (ranked 2nd best, after `knn`).

a)



b)

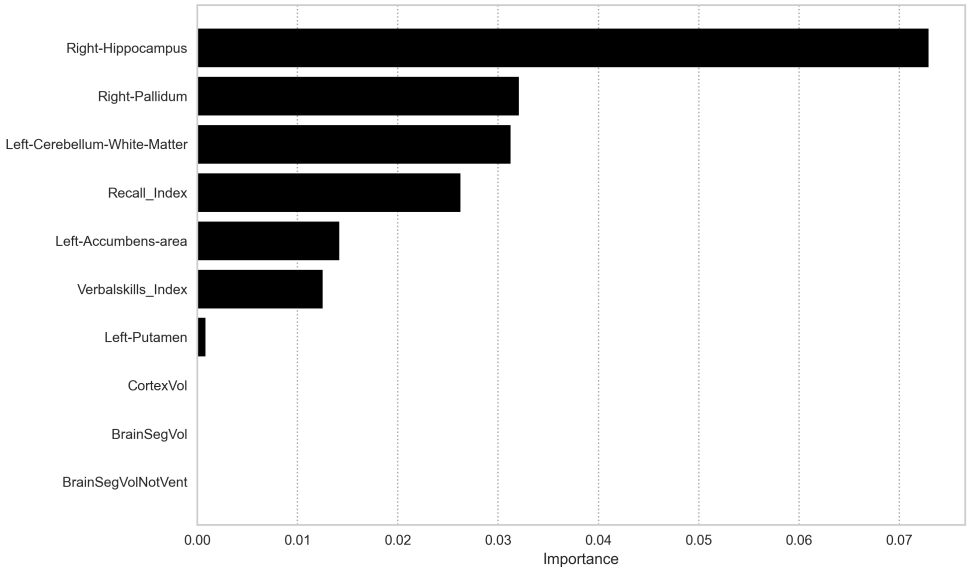


Figure 12. *Multimodal machine learning classification of IBS combining brain morphometry and cognitive measures*. a) Confusion matrix from XGBoost model testing, showing classification outcomes with detailed participant characteristics per quadrant (TN: true negatives, FP: false positives, FN: false negatives, TP: true positives). Gender distribution (F: female, M: male), IBS-SSS scores, and age are reported for each category. b) Feature importance ranking derived from the model, showing relative contribution of brain structural and cognitive measures to classification decisions.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/07-predicting-IBS-vs-HC-from-morphometry-and-cognition.ipynb

The confusion matrix in Figure 12a illustrates the XGBoost model's classification performance using combined brain morphometry and cognitive features. The model demonstrates high sensitivity but poor specificity in IBS detection. Among IBS patients, 14 of 15 were correctly identified (93.3% sensitivity), with these true positives showing characteristic IBS-SSS scores (271.3 ± 81.0) and female predominance (11F/3M). However, specificity was low (22.2%), with only 2 of 9 healthy controls correctly classified. The misclassification patterns reveal notable demographic and clinical features. The false positives (7 controls misclassified as IBS) show a male predominance (5M/2F) and lower age (25.7 ± 6.1 years) compared to true positives, despite normal IBS-SSS scores (19.6 ± 21.2). The single false negative case presents distinct characteristics: male, older (47.0 years), with substantial symptom severity (IBS-SSS: 245.0). These classification outcomes suggest that while the combined morphometric and cognitive features enable sensitive IBS detection, they lack specificity. The gender-specific misclassification patterns and age-related differences in classification accuracy indicate potential demographic influences on the model's performance. These findings highlight both the promise and limitations of multimodal classification approaches in IBS diagnosis.

Feature importance analysis (Fig. 12b) reveals the relative contributions of brain structural and cognitive measures to IBS classification. The right hippocampus emerges as the most discriminative feature (importance $\approx 0.073 \pm 0.036$), followed by the right pallidum and left cerebellar white matter (importance $\approx 0.032 \pm 0.026$, and $\approx 0.031 \pm 0.029$, respectively). Notably, cognitive performance, represented by the Recall Index and Verbal skills Index, ranks among the top discriminative features, suggesting that the integration of cognitive measures enhances classification performance. The ranking highlights a mixed contribution of structural and cognitive features, with subcortical structures (`Right-Hippocampus`, `Right-Pallidum`, `Left-Accumbens-area`) showing particularly strong discriminative power. Global brain measures (`CortexVol`, `BrainSegVol`, `BrainSegVolNotVent`) demonstrate minimal importance, suggesting that regional rather than global alterations better distinguish IBS from healthy controls. This importance ranking should be interpreted in the context of the model's classification performance metrics, where despite improved sensitivity with combined features, specificity remains low. The prominence of memory-related structures and cognitive measures aligns with the observed group differences in RBANS scores, providing a potential neurobiological basis for cognitive alterations in IBS.

Table 5 quantifies the impact of incorporating cognitive measures into the morphometry-based classification through comprehensive performance metrics. The addition of cognitive features to brain morphometry (M $\cup$ C) substantially improved model performance across multiple dimensions: sensitivity increased from 73.3% to 93.3%, accuracy from 50.0% to 66.7%, and the F1 score from 0.647 to 0.778. While specificity remained modest, it showed improvement from 11.1% to 22.2%. The Matthews Correlation Coefficient (MCC) shifted from -0.185 to 0.228, indicating enhanced overall classification performance when combining both feature types.

Table 5. *Classification performance metrics comparing XGBoost models trained on morphometric features alone (M) versus combined morphometric and cognitive features (M $\cup$ C).* Metrics include sensitivity (TPR), specificity (TNR), precision (PPV), accuracy (ACC), and additional measures of classification reliability.

Feature set	TPR	TNR	PPV	NPV	FPR	FNR	FDR	ACC	BACC	F1	MCC
M	0.733	0.111	0.579	0.200	0.889	0.267	0.421	0.500	0.422	0.647	-0.185
M $\cup$ C	0.933	0.222	0.667	0.667	0.778	0.067	0.333	0.667	0.578	0.778	0.228

M: morphometric; C: cognitive features. See list of abbreviations for the rest of column names denoting 11 different metrics.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/07-predicting-IBS-vs-HC-from-morphometry-and-cognition.ipynb

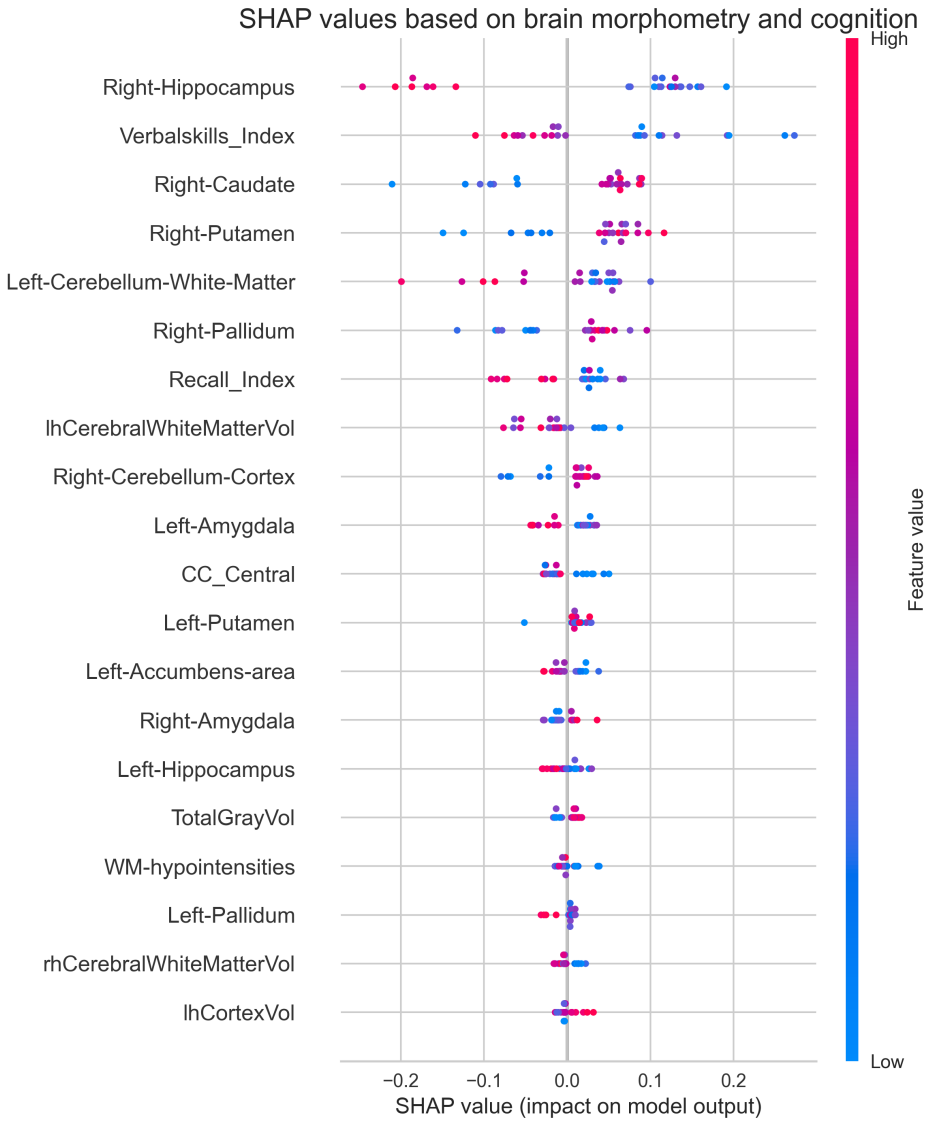


Figure 13. *Feature contribution analysis using SHAP values for multimodal brain morphometry and cognition-based classification*. SHAP values (x-axis) quantify each region's impact on classification probability, with negative values favoring healthy control classification and positive values favoring IBS. Each point represents a single prediction, with color indicating the relative magnitude of the morphometric measurement (red: high values, blue: low values). The distribution of points reveals how feature values influence model predictions across individual cases.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/07-predicting-IBS-vs-HC-from-morphometry-and-cognition.ipynb

SHAP analysis reveals the complex interactions between brain structure, cognitive performance, and IBS classification. The right hippocampus demonstrates the strongest feature impact, with higher volumes (red) generally predicting healthy control status and lower volumes (blue) predicting IBS. The Verbal skills Index emerges as the second most influential feature, showing a distinct pattern where lower scores tend to predict IBS classification. Among subcortical structures, the right caudate and putamen show notable but contrasting patterns. The right caudate exhibits a clustered distribution with clear value-dependent effects, while the right putamen shows more dispersed impact across participants. Left cerebellar white matter demonstrates moderate influence, with its effect direction varying based on volume. The overall pattern suggests a hierarchical organization of discriminative features, where both structural and cognitive measures contribute to classification decisions. Lower-ranked features, including global measures

(`TotalGrayVol`, `lhCortexVol`) and white matter hypointensities, show minimal impact on model predictions, suggesting that regional rather than global alterations better characterize IBS-related brain differences.

Discussion (`in progress`)

Our integrated analysis of brain structure and cognitive function in IBS yields several principal findings with important methodological and clinical implications. First, despite using identical FreeSurfer versions and processing pipelines, we were unable to replicate the morphometric differences between IBS and healthy controls reported by Skrobisz et al. [23]. Specifically, while they found reduced thalamic volume in IBS patients, our analysis showed no significant volumetric differences in this or other brain regions. This non-replication warrants careful consideration - it may reflect true biological heterogeneity in IBS, differences in patient characteristics between cohorts, or highlight critical methodological sensitivities in brain morphometry studies.

The systematic comparison of FreeSurfer versions 6.0.1 and 7.4.1 revealed substantial version-dependent variations in morphometric measurements. Global brain volumes demonstrated systematic offsets (6-8%), with even larger discrepancies (up to 35%) in specific structures like the nucleus accumbens. While relative group differences were largely preserved across versions, absolute measurements showed consistent biases, particularly in subcortical and limbic regions. For example, cortical measurements (`lhCortexVol`, `rhCortexVol`) exhibited parallel offsets above the identity line, indicating that FreeSurfer 6.0.1 consistently produced higher volume estimates compared to version 7.4.1. Similar systematic biases were observed in cerebellar structures and several subcortical regions. These version-dependent variations have critical implications for multi-site studies and meta-analyses. The preservation of relative group differences suggests that within-study comparisons remain valid, but absolute measurements may not be directly comparable across studies using different FreeSurfer versions. This finding underscores the importance of harmonized processing pipelines in neuroimaging research, particularly for studies investigating subtle structural alterations in clinical populations. The observed systematic biases also highlight the need to carefully consider software version effects when conducting replication studies or meta-analyses of brain morphometry findings.

Our machine learning analyses reveal a complex relationship between brain structure, cognitive function, and IBS classification. While morphometric features alone showed limited discriminative power (sensitivity 73.3%, specificity 11.1%), the integration of cognitive measures substantially improved classification accuracy. The combined model achieved notably higher sensitivity (93.3%), correctly identifying 14 of 15 IBS patients, with these true positives showing characteristic IBS-SSS scores (271.3 ± 81.0) and female predominance (11F/3M). However, specificity remained modest (22.2%), with only 2 of 9 healthy controls correctly classified. This asymmetric performance pattern, particularly the high false positive rate among male controls (5M/2F, age 25.7 ± 6.1 years), suggests that while brain structural and cognitive alterations may be characteristic of IBS, they are not necessarily specific to the condition.

Feature importance analysis provides insight into this classification pattern. The right hippocampus emerged as the most discriminative feature (importance $\approx 0.073 \pm 0.036$), followed by subcortical structures (right pallidum, left cerebellar white matter) and cognitive measures (Recall Index, Verbal skills Index). This hierarchy, consistently identified through both permutation importance and SHAP analyses, suggests a fundamental relationship between memory-related neural circuits and IBS pathophysiology. The prominence of hippocampal measurements aligns with emerging evidence for altered brain-gut-behavior interactions in IBS [46,47], particularly regarding the role of memory systems in visceral

symptom processing and learned pain responses.

Gender-specific classification patterns emerged as a particularly noteworthy finding. The model showed high sensitivity for female IBS patients (correctly identifying 11 of 14 females) but demonstrated systematic bias in healthy control classification, with a notable tendency to misclassify male controls as IBS (5 of 7 false positives were male). This asymmetric performance pattern parallels known epidemiological sex differences in IBS [57,58] and suggests that brain structural and cognitive alterations may manifest differently in males and females. Such gender-specific patterns could reflect underlying differences in disease mechanisms, stress responses, or symptom presentation, with important implications for both diagnostic strategies and therapeutic approaches. Future studies may need to consider sex-specific normative ranges for brain structural and cognitive measures in IBS assessment.

The study presents several key methodological contributions to the field of irritable bowel syndrome research. At the forefront is an advanced machine learning approach that integrates brain structural and cognitive measures, moving beyond traditional single-modal assessments. This type of computational methodology represents a significant advancement in the use of neuroimaging techniques in general, offering a more sophisticated analytical framework for understanding the complex neurological underpinnings of complex conditions such as IBS. By developing a machine learning model with high sensitivity, the study opens new avenues for more objective diagnostic strategies, even though the current specificity suggests the need for further refinement. Moreover, the study's cohort and analytical approach contribute methodologically by establishing a robust dataset and openly available code that could be tested and further developed for future investigations. Moreover, the computational neuroimaging methodology developed in this research, supporting data-driven approaches to understanding IBS from a neurological perspective, has broader implications, potentially offering insights that could be applied to other neurological and psychiatric conditions with complex neuroimaging presentations.

Some limitations of our study should be acknowledged. The moderate sample size may limit generalizability, though our cohort is comparable to or larger than many neuroimaging studies in IBS. The cross-sectional design precludes inference about causality or temporal dynamics of observed alterations. Additionally, while our machine learning approach achieved high sensitivity, the limited specificity suggests that brain structural and cognitive measures alone may be insufficient for definitive IBS diagnosis.

Future research should focus on longitudinal studies to understand the temporal stability of observed alterations and their relationship to symptom fluctuations. Integration of additional data modalities, such as functional connectivity and microbiome measures, may further enhance our understanding of IBS pathophysiology and improve diagnostic accuracy. These findings have important implications for both research methodology and clinical practice. The demonstrated impact of software versions on morphometric measurements emphasizes the need for standardized processing pipelines in neuroimaging studies. The improved classification accuracy achieved through multimodal analysis suggests that comprehensive assessment approaches, incorporating both structural and cognitive measures, may be valuable in clinical settings.

References

1. Black, C.J.; Ford, A.C. Global burden of irritable bowel syndrome: trends, predictions and risk factors. *Nature Reviews Gastroenterology & hepatology* **2020**, *17*, 473–486. https://doi.org/10.1038/s41575-020-0286-8.
2. Lovell, R.M.; Ford, A.C. Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. *Clinical Gastroenterology and Hepatology* **2012**, *10*, 712–721. https://doi.org/10.1016/j.cgh.2012.02.029.
3. Bonetto, S.; Fagoonee, S.; Battaglia, E.; Grassini, M.; Saracco, G.M.; Pellicano, R. Recent advances in the treatment of irritable bowel syndrome. *Polish Archives of Internal Medicine* **2021**, *131*, 709–715. https://doi.org/10.20452/pamw.16067.
4. Drossman, D.A.; Tack, J. Rome Foundation clinical diagnostic criteria for disorders of gut-brain interaction. *Gastroenterology* **2022**, *162*, 675–679. https://doi.org/10.1053/j.gastro.2021.11.019.
5. Heitkemper, M.M.; Cain, K.C.; Jarrett, M.E.; Burr, R.L.; Hertig, V.; Bond, E.F. Symptoms across the menstrual cycle in women with irritable bowel syndrome. *Official Journal of the American College of Gastroenterology | ACG* **2003**, *98*, 420–430. https://doi.org/10.1111/j.1572-0241.2003.07233.x.
6. Meleine, M.; Matricon, J. Gender-related differences in irritable bowel syndrome: potential mechanisms of sex hormones. *World Journal of Gastroenterology: WJG* **2014**, *20*, 6725. https://doi.org/10.3748/wjg.v20.i22.6725.
7. Kim, Y.S.; Kim, N. Sex-gender differences in irritable bowel syndrome. *Journal of Neurogastroenterology and Motility* **2018**, *24*, 544.
8. Toner, B.B.; Akman, D. Gender role and irritable bowel syndrome: literature review and hypothesis. *Official journal of the American College of Gastroenterology | ACG* **2000**, *95*, 11–16. https://doi.org/10.1111/j.1572-0241.2000.01698.x.
9. Lundervold, A.J.; Billing, J.E.; Berentsen, B.; Lied, G.A.; Steinsvik, E.K.; Hausken, T.; Lundervold, A. Decoding IBS: a machine learning approach to psychological distress and gut-brain interaction. *BMC Gastroenterology* **2024**, *24*, 267. https://doi.org/10.1186/s12876-024-03355-z.
10. Shiha, M.G.; Aziz, I. Physical and psychological comorbidities associated with irritable bowel syndrome. *Alimentary Pharmacology & Therapeutics* **2021**, *54*, S12–S23. https://doi.org/10.1111/apt.16589.
11. Lam, N.C.Y.; Yeung, H.Y.; Li, W.K.; Lo, H.Y.; Yuen, C.F.; Chang, R.C.C.; Ho, Y.S. Cognitive impairment in irritable bowel syndrome (IBS): a systematic review. *Brain Research* **2019**, *1719*, 274–284. https://doi.org/https://doi.org/10.1016/j.brainres.2019.05.036.
12. Wong, K.M.F.; Mak, A.D.P.; Yuen, S.Y.; Leung, O.N.W.; Ma, D.Y.; Chan, Y.; Cheong, P.K.; Lui, R.; Wong, S.H.; Wu, J.C.Y. Nature and specificity of altered cognitive functioning in IBS. *Neurogastroenterology & Motility* **2019**, *31*, e13696. https://doi.org/10.1111/nmo.13696.
13. Billing, J.; Berentsen, B.; Lundervold, A.; Hillestad, E.M.; Lied, G.A.; Hausken, T.; Lundervold, A.J. Cognitive function in patients with irritable bowel syndrome: impairment is common and only weakly correlated with depression/anxiety and severity of gastrointestinal symptoms. *Scandinavian Journal of Gastroenterology* **2023**, pp. 1–9. https://doi.org/10.1080/00365521.2023.2256916.
14. Mayer, E.A.; Nance, K.; Chen, S. The Gut-Brain Axis. *Annual Review of Medicine* **2022**, *73*, 439–453. https://doi.org/10.1146/annurev-med-042320-014032.
15. Coss-Adame, E.; Rao, S.S. Brain and gut interactions in irritable bowel syndrome: new paradigms and new understandings. *Current Gastroenterology Reports* **2014**, *16*, 1–8. https://doi.org/10.1007/s11894-014-0379-z.
16. Lezak, M.D. *Neuropsychological assessment*; Oxford University Press, USA, 2004.
17. Park, H.J.; Friston, K. Structural and functional brain networks: from connections to cognition. *Science* **2013**, *342*, 1238411. https://doi.org/10.1126/science.1238411.
18. Mayer, E.A.; Labus, J.S.; Tillisch, K.; Cole, S.W.; Baldi, P. Towards a systems view of IBS. *Nature Reviews Gastroenterology & Hepatology* **2015**, *12*, 592–605. https://doi.org/10.1038/nrgastro.2015.121.
19. Mayer, E.A.; Labus, J.; Aziz, Q.; Tracey, I.; Kilpatrick, L.; Elsenbruch, S.; Schweinhardt, P.; Van Oudenhove, L.; Borsook, D. Role of brain imaging in disorders of brain–gut interaction: a Rome Working Team Report. *Gut* **2019**, *68*, 1701–1715. https://doi.org/10.1136/gutjnl-2019-318308.
20. Li, Z.; Ma, Q.; Deng, Y.; Rolls, E.T.; Shen, C.; Li, Y.; Zhang, W.; Xiang, S.; Langley, C.; Sahakian, B.J.; et al. Irritable Bowel Syndrome Is Associated With Brain Health by Neuroimaging,

Behavioral, Biochemical, and Genetic Analyses. *Biological Psychiatry* **2024**, *95*, 1122–1132. https://doi.org/10.1016/j.biopsych.2023.12.024.
21. Labus, J.S.; Wang, C.; Mayer, E.A.; Gupta, A.; Oughourlian, T.; Kilpatrick, L.; Tillisch, K.; Chang, L.; Naliboff, B.; Ellingson, B.M. Sex-specific brain microstructural reorganization in irritable bowel syndrome. *Pain* **2023**, *164*, 292–304. https://doi.org/10.1097/j.pain.0000000000002699.
22. Nan, J.; Yang, W.; Meng, P.; Huang, W.; Zheng, Q.; Xia, Y.; Liu, F. Changes of the postcentral cortex in irritable bowel syndrome patients. *Brain Imaging and Behavior* **2020**, *14*, 1566–1576.
23. Skrobisz, K.; Piotrowicz, G.; Rudnik, A.; Naumczyk, P.; Sabisz, A.; Markiet, K.; Szurowska, E. Evaluation of subcortical structure volumes in patients with non-specific digestive diseases. *Diagnostics* **2022**, *12*, 2199. https://doi.org/10.3390/diagnostics12092199.
24. Berentsen, B.; Nagaraja, B.H.; Teige, E.P.; Lied, G.A.; Lundervold, A.J.; Lundervold, K.; Steinsvik, E.K.; Hillestad, E.R.; Valeur, J.; Brønstad, I.; et al. Study protocol of the Bergen brain-gut-microbiota-axis study: A prospective case-report characterization and dietary intervention study to evaluate the effects of microbiota alterations on cognition and anatomical and functional brain connectivity in patients with irritable bowel syndrome. *Medicine* **2020**, *99*, e21950. https://doi.org/doi:10.1097/MD.0000000000021950.
25. Seminowicz, D.A.; Labus, J.S.; Bueller, J.A.; Tillisch, K.; Naliboff, B.D.; Bushnell, M.C.; Mayer, E.A. Regional gray matter density changes in brains of patients with irritable bowel syndrome. *Gastroenterology* **2010**, *139*, 48–57. https://doi.org/10.1053/j.gastro.2010.03.049.
26. Blankstein, U.; Chen, J.; Diamant, N.E.; Davis, K.D. Altered brain structure in irritable bowel syndrome: potential contributions of pre-existing and disease-driven factors. *Gastroenterology* **2010**, *138*, 1783–1789. https://doi.org/10.1053/j.gastro.2009.12.043.
27. Bhatt, R.R.; Gupta, A.; Labus, J.S.; Zeltzer, L.K.; Tsao, J.C.; Shulman, R.J.; Tillisch, K. Altered brain structure and functional connectivity and its relation to pain perception in girls with irritable bowel syndrome. *Psychosomatic medicine* **2019**, *81*, 146–154. https://doi.org/10.1097/PSY.0000000000000655.
28. Francis, C.Y.; Morris, J.; Whorwell, P.J. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. *Alimentary pharmacology & therapeutics* **1997**, *11*, 395–402. https://doi.org/10.1046/j.1365-2036.1997.142318000.x.
29. Randolph, C. *Repeatable battery for the assessment of neuropsychological status. Norwegian manual*; NL:Pearson, 2013.
30. Fischl, B. FreeSurfer. *Neuroimage* **2012**, *62*, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.
31. Klauschen, F.; Goldman, A.; Barra, V.; Meyer-Lindenberg, A.; Lundervold, A. Evaluation of automated brain MR image segmentation and volumetry methods. *Human Brain MApping* **2009**, *30*, 1310–1327. https://doi.org/https://doi.org/10.1002/hbm.20599.
32. Jovicich, J.; Czanner, S.; Han, X.; Salat, D.; van der Kouwe, A.; Quinn, B.; Pacheco, J.; Albert, M.; Killiany, R.; Blacker, D.; et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage* **2009**, *46*, 177–192. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.02.010.
33. Gronenschild, E.H.; Habets, P.; Jacobs, H.I.; Mengelers, R.; Rozendaal, N.; Van Os, J.; Marcelis, M. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS One* **2012**, *7*, e38234. https://doi.org/https://doi.org/10.1371/journal.pone.0038234.
34. Glatard, T.; Lewis, L.B.; Ferreira da Silva, R.; Adalat, R.; Beck, N.; Lepage, C.; Rioux, P.; Rousseau, M.E.; Sherif, T.; Deelman, E.; et al. Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics* **2015**, *9*, 12. https://doi.org/https://doi.org/10.3389/fninf.2015.00012.
35. Knussmann, G.N.; Anderson, J.S.; Prigge, M.B.; Dean III, D.C.; Lange, N.; Bigler, E.D.; Alexander, A.L.; Lainhart, J.E.; Zielinski, B.A.; King, J.B. Test-retest reliability of FreeSurfer-derived volume, area and cortical thickness from MPRAGE and MP2RAGE brain MRI images. *Neuroimage: Reports* **2022**, *2*, 100086. https://doi.org/10.1016/j.ynirp.2022.100086.
36. Debiasi, G.; Mazzonetto, I.; Bertoldo, A. The effect of processing pipelines, input images and age on automatic cortical morphology estimates. *Computer Methods and Programs in Biomedicine* **2023**, *242*, 107825. https://doi.org/https://doi.org/10.1016/j.cmpb.2023.107825.
37. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin* **1993**, *114*, 494. https://doi.org/10.1037/0033-2909.114.3.494.

38. Meissel, K.; Yao, E.S. Using Cliff's delta as a non-parametric effect size measure: an accessible web app and R tutorial. *Practical Assessment, Research, and Evaluation* **2024**, *29*. https://doi.org/0.7275/pare.1977.
39. Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* **1904**, *15*, 72–101. https://doi.org/10.2307/1412159.
40. Mahalanobis, P.C. On the generalized distance in statistics **1936**. *2*, 49–55.
41. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The Mahalanobis distance. *Chemometrics and intelligent laboratory systems* **2000**, *50*, 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.
42. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
43. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **1960**, *20*, 37–46. https://doi.org/10.1177/001316446002000.
44. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
45. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2020**, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
46. Casamitjana, A.; Mancini, M.; Robinson, E.; Peter, L.; Annunziata, R.; Althonayan, J.; Crampsie, S.; Blackburn, E.; Billot, B.; Atzeni, A.; et al. A next-generation, histological atlas of the human brain and its application to automated brain MRI segmentation. *bioRxiv* **2024**. https://doi.org/10.1101/2024.02.05.579016.

Author Contributions: "Conceptualization of the present study, A.L., A.J.L., D.M.P., J.B., B.R.B.; methodology, A.L., A.J.L., B.R.B., D.M.P.; formal analysis, A.L.; data collection G.A.L., E.S. T. H., B.B., and A.J.L.; writing original draft preparation A.J.L.; review and editing: all authors, project administration, B.B.; funding acquisition, T.H., A.L. All authors have read and agreed to the published version of the manuscript. During the preparation of this work, the authors used Claude 3.5 Sonnet (Anthropic) to improve the readability of certain parts of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content included in this publication.

Funding: This research was funded by Research Council of Norway (grant ID FRIMED-BIO276010) and Helse Vest's Research Funding (grant ID HV912243) and by the Trond Mohn Research Foundation, grant number BFS2018TMT0, and from The Research Council of Norway, project number 294594.

Institutional Review Board Statement: The B-BGM project was approved by the Southeast Regional Ethical Committees (REC) for medical and health research ethics in Norway (REK2015-01621). All participants provided written consent to participate, and the project was conducted following the ethical requirements from the Declaration of Helsinki. The project is registered at www.clinicaltrials.gov (#NCT04296552).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The complete analysis workflow is publicly available at https://arvidl.github.com/ibs-brain, comprising reproducible Jupyter notebooks containing all analysis code and visualizations, cleaned datasets in CSV format, Conda environment configuration for exact replication, and source code for generating all tables and figures presented in the Results section. The computational analyses were developed with assistance from the Claude 3.5 Sonnet large language model integrated within the Cursor (Anysphere) AI code editor and development environment.

Acknowledgments: We sincerely thank all patients and healthy volunteers for their participation in the Bergen Brain-Gut Microbiota (B-BGM) project. We also thank all the present and previous members of the B-BGM project.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in the manuscript:

AUC	Area Under Curve
CM	Confusion matrix
Cohen's d	effect size
Cliff's delta	non-parametric effect size
DGBI	Disorders of the gut-brain interaction
FS	FreeSurfer
GI	Gastrointestinal
GitHub	Web-based platform for code sharing with version control
HC	Healthy Control
IBS	Irritable bowel syndrome
IBS-SSS	IBS Severity Scoring System
IQR	Inter Quartile Range
ML	Machine learning
MRI	Magnetic Resonance Imaging
RBANS	Repeatable Battery for the Assessment of Neuropsychological Status
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SD	Standard deviation
SHAP	SHapley Additive exPlanations
TA	Time of Acquisition
XGBoost	eXtreme Gradient Boosting

ML-performance	Definition
TPR	TP/(TP+FN) (true positive rate, sensitivity, recall)
TNR	TN/(TN+FP) (true negative rate, specificity)
PPV	TP/(TP+FP) (positive predictive value, precision)
NPV	TN/(TN+FN) (negative predictive value)
FPR	FP/(FP+TN) (false positive rate)
FNR	FN/(TP+FN) (false negative rate)
FDR	FP/(TP+FP) (false discovery rate)
ACC	(TP+TN)/(TP+FP+FN+TN) (accuracy)
BACC	(Sensitivity + Specificity) / 2 (balanced accuracy)
F1	1/((1/PPV) + (1/TPR)) (F1-score, harmonic mean of precision and recall)
MCC	((TP*TN)-(FP*FN))/sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)) (Matthews corr.coeff)

Appendix A Supplementary definitions, tables, and figures

Appendix A.1 FreeSurfer segmented brain regions obtained from `aseg`

Table A1. FreeSurfer segmented brain regions (`aseg`) with short descriptions of functional anatomy

Region	Description
eTIV	Estimated Total Intracranial Volume
Left-Cerebellum-White-Matter	White matter in the left cerebellum
Left-Cerebellum-Cortex	Gray matter (cortex) in the left cerebellum
Left-Thalamus	Left thalamus. *Thalamus* is a relay center for sensory and motor signals. In IBS, thalamic activity will contribute to pain perception and visceral hypersensitivity
Left-Caudate	Left caudate nucleus. *Nucleus caudatus* is involved in motor and motility control and learning
Left-Putamen	Left putamen. *Putamen* is part of the basal ganglia involved in motor control and may contribute to habitual responses to gastrointestinal discomfort
Left-Pallidum	Left globus pallidus. *Globus pallidus* is involved in regulating voluntary movement and gut motility patterns
Left-Hippocampus	Left hippocampus. *Hippocampus* is crucial for memory formation and spatial navigation, and in IBS, involved in contextual fear learning related to gastrointestinal symptoms
Left-Amygdala	Left amygdala. *Amygdala* is involved in processing emotions, fear, and anxiety
Left-Accumbens-area	Left nucleus accumbens. *Nucleus accumbens* is involved in reward and motivation, stress responsivity, and pain modulation
CSF	Cerebrospinal Fluid
Right-Cerebellum-White-Matter	White matter in the right cerebellum
Right-Cerebellum-Cortex	Gray matter (cortex) in the right cerebellum
Right-Thalamus	Right thalamus
Right-Caudate	Right caudate nucleus
Right-Putamen	Right putamen
Right-Pallidum	Right globus pallidus
Right-Hippocampus	Right hippocampus
Right-Amygdala	Right amygdala
Right-Accumbens-area	Right nucleus accumbens
WM-hypointensities	White matter hypointensities (dark on T1-w sequences), can be associated with small vessel disease, demyelination, inflammation, fluid accumulation
CC_Posterior	Posterior part of the corpus callosum
CC_Mid_Posterior	Mid-posterior part of the corpus callosum
CC_Central	Central part of the corpus callosum
CC_Mid_Anterior	Mid-anterior part of the corpus callosum
CC_Anterior	Anterior part of the corpus callosum
BrainSegVol	Total volume of brain segmentation
BrainSegVolNotVent	Brain segmentation volume without ventricles
lhCortexVol	Volume of the left hemisphere cortex
rhCortexVol	Volume of the right hemisphere cortex
CortexVol	Total cortical volume (left + right)
lhCerebralWhiteMatterVol	Volume of left hemisphere cerebral white matter
rhCerebralWhiteMatterVol	Volume of right hemisphere cerebral white matter
CerebralWhiteMatterVol	Total cerebral white matter volume (left + right)
SubCortGrayVol	Volume of subcortical gray matter
TotalGrayVol	Total gray matter volume

Appendix A.2 Multinomality testing: Mardia's test and Henze-Zirkler test

Mardia's test extends the univariate concepts of skewness and kurtosis to multivariate distributions. For a p-dimensional random vector X, multivariate normality implies specific properties of its third and fourth moments. The test examines these moments through multivariate measures of skewness and kurtosis. Given a sample of n observations, $X_1, \ldots, X_n$, the sample measures are computed using Mahalanobis distances. The multivariate skewness is defined as $b_{1,p} = \frac{1}{n^2}\sum_{i,j=1}^{n}[(X_i - \bar{X})^T S^{-1}(X_j - \bar{X})]^3$, where $\bar{X}$ is the sample mean vector and S is the sample covariance matrix. The multivariate kurtosis is defined as $b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \bar{X})^T S^{-1}(X_i - \bar{X})]^2$. Under the null hypothesis of multivariate normality, $nb_{1,p}/6$ follows asymptotically a chi-square distribution with $\frac{p(p+1)(p+2)}{6}$ degrees of freedom, and $(b_{2,p} - p(p+2))/\sqrt{8p(p+2)/n}$ follows approximately a standard normal distribution.

The Henze-Zirkler test is based on a non-negative functional distance between two distribution functions, specifically between the empirical characteristic function of the standardized data and the characteristic function of the standard normal distribution. The test statistic is defined as $HZ_n = n\left(1+2\beta^2\right)^{p/2}\left[D_n - \left(1+\beta^2\right)^{-p/2}\right]$, where $\beta = \frac{1}{\sqrt{2}}$ is the smoothing parameter and $D_n = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\beta^2}{2}d_i^2\right)$, with d_i^2 being the squared Mahalanobis distances $d_i^2 = (X_i - \bar{X})^T S^{-1}(X_i - \bar{X})$, with $\bar{X}$ being the sample mean vector and S the sample covariance matrix. The test is invariant under affine transformations and has good power against a broad range of alternatives. Under the null hypothesis of multivariate normality, the Henze-Zirkler test statistic HZ_n follows approximately a lognormal distribution with parameters μ and σ^2 that depend on the sample size n and dimension p as follows: $\mu = -\frac{1}{2}\log(1+2\beta^2) - \frac{p}{2}\log(1+\beta^2) + \log\left(1+\frac{p\beta^4}{2(1+2\beta^2)}\right)$ $\sigma^2 = 2\left[-\log(1-\frac{2\beta^4}{(1+2\beta^2)^2}) + \frac{p\beta^4}{(1+2\beta^2)(1+\beta^2)}\right]$ where $\beta = \frac{1}{\sqrt{2}}$ is the smoothing parameter. This means that under H_0: $\log(HZ_n) \sim N(\mu + \frac{\log(n)}{2}, \frac{\sigma^2}{n})$. The test rejects the null hypothesis of multivariate normality for large values of the test statistic.

While both tests assess multivariate normality, they capture different aspects of departure from normality. Mardia's test specifically examines the third and fourth moments of the distribution, making it particularly sensitive to asymmetry and tail behavior. The Henze-Zirkler test, based on characteristic functions, can detect various types of departures from normality, including those that might not be captured by moment-based methods. Using both tests provides a more comprehensive assessment of multivariate normality, though careful attention must be paid to numerical stability, particularly in high-dimensional settings or with small sample sizes.

Appendix A.3 Robust Mahalanobis distance between IBS and HC

Our computation of a robust Mahalanobis distance method begins with winsorization of the data to reduce the impact of outliers. For each feature x_i, values are trimmed at the 10th and 90th percentiles such that $x_{win} = x_{(0.1)}$ if $x < x_{(0.1)}$, x if $x_{(0.1)} \leq x \leq x_{(0.9)}$, and $x_{(0.9)}$ if $x > x_{(0.9)}$, where $x_{(\alpha)}$ represents the α-th quantile. Following winsorization, robust location estimation is performed using the median instead of the mean: $\hat{\mu}robust = \text{median}(Xwin)$. The pooled covariance matrix is then computed using the winsorized data as $\hat{\Sigma}pooled = \frac{(nHC-1)\hat{\Sigma}HC+(nIBS-1)\hat{\Sigma}IBS}{nHC+n_{IBS}-2}$. The robust Mahalanobis distance is calculated as $D_{robust} = \sqrt{(\hat{\mu}IBS - \hat{\mu}HC)^T\hat{\Sigma}pooled^{-1}(\hat{\mu}IBS - \hat{\mu}HC)}$. To assess the statistical significance of this distance, Hotelling's T^2 statistic is transformed to an F-statistic: $F = \frac{nHCn_{IBS}}{(n_{HC}+n_{IBS})(n_{HC}+n_{IBS}-2)p}D^2_{robust}$. Under the null hypothesis of no group difference, this follows an F-distribution with degrees of freedom p and $n_{HC} + n_{IBS} - p - 1$, where p is the number of features. The p-value is computed as p-value $= 1 - F_{p,n_{HC}+n_{IBS}-p-1}(F)$. This robust approach provides a more reliable measure of group separation when the

data contains outliers or deviates from multivariate normality, as is often the case with neuroimaging data. The use of robust estimators (median and winsorized covariance) makes the distance measure less sensitive to extreme values while maintaining the ability to detect genuine multivariate differences between groups.

Appendix A.4 Comparing Freesurfer 6.0.1 and FreeSurferr 7.4.1 cross-sectional

Table A2 gives the summary statistics, mean and standard deviation from HC and IBS patients in the Bergen cohort on each of the 35 included brain regions (also reported by Skrobisz et al. [23]) derived from the `aseg.stats` files using cross-sectional Freesurfer 6.0.1 and Freesurfer 7.4.1, respectively. For eTIV [mm^3] computed with each of the two versions, we found the mean (SD) as follows:

FS6-cross - HC: 1468820 (155501); IBS: 1426237 (136413), and
FS7-cross - HC: 1494273 (171472); IBS: 1462311 (144145), respectively.

Table A2. Comparison of Brain Region Volumes in IBS Patients and Healthy Controls.

Brain Region	Bergen Cohort FS 6.0.1				Bergen cohort FS 7.4.1			
	HC (N=29)		IBS (N=49)		HC (N=29)		IBS (N=49)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Left-Cerebellum-WM	0.010496	0.000915	0.010483	0.000924	0.010603	0.000932	0.010607	0.001027
Left-Cerebellum-Cortex	0.038939	0.003435	0.039314	0.003733	0.038066	0.003526	0.038056	0.003684
Left-Thalamus	0.005232	0.000464	0.005144	0.000393	0.005236	0.000522	0.005114	0.000458
Left-Caudate	0.002356	0.000259	0.002355	0.000306	0.002346	0.000283	0.002317	0.000294
Left-Putamen	0.003479	0.000385	0.003441	0.000386	0.003438	0.000396	0.003370	0.000329
Left-Pallidum	0.001405	0.000154	0.001374	0.000107	0.001380	0.000136	0.001358	0.000095
Left-Hippocampus	0.002913	0.000272	0.002896	0.000242	0.002926	0.000251	0.002895	0.000243
Left-Amygdala	0.001218	0.000097	0.001203	0.000105	0.001228	0.000133	0.001190	0.000111
Left-Accumbens-area	0.000427	0.000069	0.000421	0.000057	0.000424	0.000061	0.000400	0.000057
CSF	0.000670	0.000120	0.000702	0.000141	0.000658	0.000114	0.000689	0.000130
Right-Cerebellum-WM	0.009973	0.000891	0.009979	0.000851	0.010052	0.000934	0.010108	0.001015
Right-Cerebellum-Cortex	0.039719	0.003445	0.039978	0.003760	0.038881	0.003534	0.038912	0.003673
Right-Thalamus	0.005120	0.000438	0.005071	0.000358	0.005190	0.000455	0.005053	0.000413
Right-Caudate	0.002438	0.000240	0.002439	0.000301	0.002418	0.000286	0.002402	0.000285
Right-Putamen	0.003506	0.000366	0.003489	0.000351	0.003487	0.000402	0.003466	0.000322
Right-Pallidum	0.001323	0.000126	0.001301	0.000107	0.001321	0.000137	0.001306	0.000118
Right-Hippocampus	0.003013	0.000240	0.002983	0.000229	0.003049	0.000230	0.002986	0.000235
Right-Amygdala	0.001284	0.000087	0.001271	0.000098	0.001269	0.000107	0.001260	0.000106
Right-Accumbens-area	0.000428	0.000053	0.000427	0.000061	0.000434	0.000054	0.000435	0.000057
WM-hypointensities	0.000791	0.000306	0.000688	0.000253	0.000787	0.000481	0.000667	0.000244
CC_Posterior	0.000652	0.000096	0.000702	0.000113	0.000645	0.000096	0.000685	0.000113
CC_Mid_Posterior	0.000369	0.000067	0.000401	0.000071	0.000366	0.000069	0.000394	0.000073
CC_Central	0.000395	0.000089	0.000391	0.000105	0.000390	0.000091	0.000390	0.000101
CC_Mid_Anterior	0.000379	0.000081	0.000409	0.000113	0.000384	0.000078	0.000400	0.000105
CC_Anterior	0.000623	0.000096	0.000650	0.000101	0.000608	0.000098	0.000646	0.000112
BrainSegVol	0.804644	0.024872	0.805581	0.023967	0.792112	0.037690	0.786845	0.028349
BrainSegVolNotVent	0.792235	0.025106	0.791323	0.024898	0.779857	0.037538	0.772948	0.030305
lhCortexVol	0.166698	0.008003	0.166929	0.009510	0.164771	0.010207	0.163181	0.010196
rhCortexVol	0.166137	0.008276	0.166462	0.009388	0.164149	0.010295	0.162912	0.009834
CortexVol	0.332835	0.016110	0.333391	0.018798	0.328920	0.020369	0.326092	0.019888
lhCerebralWhiteMatterVol	0.159895	0.008578	0.159148	0.008757	0.157377	0.010114	0.155820	0.009472
rhCerebralWhiteMatterVol	0.159252	0.008291	0.158267	0.009384	0.156808	0.010522	0.154840	0.009950
CerebralWhiteMatterVol	0.319147	0.016780	0.317415	0.018079	0.314184	0.020552	0.310659	0.019351
SubCortGrayVol	0.040924	0.002583	0.040629	0.002364	0.040864	0.002871	0.040194	0.002433
TotalGrayVol	0.453068	0.022076	0.453961	0.024324	0.446625	0.027101	0.443252	0.025556
eTIV [mm^3]	1468820.2	155501.4	1426237.4	136412.8	1494273.2	171472.3	1462310.8	144145.1

Note: All volumes are normalized to estimated total intracranial volume (eTIV)
HC = Healthy Controls; IBS = Irritable Bowel Syndrome; SD = Standard Deviation; WM = White Matter
Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-computing-FS-versions-on-same-dataset.ipynb

Appendix A.5 Comparing FreeSurfer 7.4.1 cross-sectional and longitudinal stream

Table A3 gives the summary statistics, mean and standard deviation from HC and IBS patients in the Bergen cohort on each of the 35 included brain regions derived from the `aseg.stats` files using Freesurfer 7.4.1 cross-sectional analysis and Freesurfer 7.4.1 longitudinal stream, respectively. For eTIV [mm^3] computed with each of the two versions, we found the mean (SD) as follows:

FS7-cross - HC: 1494273 (171472; IBS: 1462311 (144145), and
FS7-long - HC: 1492944 (171478); IBS: 1464197 (143328).

Table A3. Comparison of Brain Region Volumes in Bergen cohort, FS 7.4.1 cross-sectional vs. FS 7.4.1 longitudinal stream

	FS 7.4.1 cross-sectional				FS 7.4.1 longitudinal stream			
	HC (N=29)		IBS (N=49)		HC (N=29)		IBS (N=49)	
Brain Region	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Left-Cerebellum-White-Matter	0.010603	0.000932	0.010607	0.001027	0.010888	0.001076	0.010688	0.001044
Left-Cerebellum-Cortex	0.038066	0.003526	0.038056	0.003684	0.037438	0.003633	0.037232	0.003575
Left-Thalamus	0.005236	0.000522	0.005114	0.000458	0.005385	0.000517	0.005262	0.000473
Left-Caudate	0.002346	0.000283	0.002317	0.000294	0.002506	0.000313	0.002474	0.000315
Left-Putamen	0.003438	0.000396	0.003370	0.000329	0.003702	0.000418	0.003621	0.000369
Left-Pallidum	0.001380	0.000136	0.001358	0.000095	0.001415	0.000156	0.001373	0.000098
Left-Hippocampus	0.002926	0.000251	0.002895	0.000243	0.002970	0.000273	0.002925	0.000244
Left-Amygdala	0.001228	0.000133	0.001190	0.000111	0.001253	0.000138	0.001203	0.000111
Left-Accumbens-area	0.000424	0.000061	0.000400	0.000057	0.000440	0.000072	0.000429	0.000070
CSF	0.000658	0.000114	0.000689	0.000130	0.000712	0.000127	0.000742	0.000130
Right-Cerebellum-White-Matter	0.010052	0.000934	0.010108	0.001015	0.010218	0.001008	0.010231	0.000955
Right-Cerebellum-Cortex	0.038881	0.003534	0.038912	0.003673	0.038471	0.003722	0.038176	0.003654
Right-Thalamus	0.005190	0.000455	0.005053	0.000413	0.005475	0.000474	0.005341	0.000456
Right-Caudate	0.002418	0.000286	0.002402	0.000285	0.002608	0.000310	0.002577	0.000303
Right-Putamen	0.003487	0.000402	0.003466	0.000322	0.003788	0.000418	0.003766	0.000375
Right-Pallidum	0.001321	0.000137	0.001306	0.000118	0.001350	0.000156	0.001330	0.000120
Right-Hippocampus	0.003049	0.000230	0.002986	0.000235	0.003102	0.000245	0.003034	0.000241
Right-Amygdala	0.001269	0.000107	0.001260	0.000106	0.001332	0.000115	0.001323	0.000115
Right-Accumbens-area	0.000434	0.000054	0.000435	0.000057	0.000503	0.000065	0.000507	0.000063
WM-hypointensities	0.000787	0.000481	0.000667	0.000244	0.000757	0.000644	0.000607	0.000274
CC_Posterior	0.000645	0.000096	0.000685	0.000113	0.000632	0.000097	0.000669	0.000112
CC_Mid_Posterior	0.000366	0.000069	0.000394	0.000073	0.000350	0.000066	0.000375	0.000075
CC_Central	0.000390	0.000091	0.000390	0.000101	0.000364	0.000082	0.000363	0.000091
CC_Mid_Anterior	0.000384	0.000078	0.000400	0.000105	0.000361	0.000071	0.000379	0.000101
CC_Anterior	0.000608	0.000098	0.000646	0.000112	0.000587	0.000096	0.000620	0.000093
BrainSegVol	0.792112	0.037690	0.786845	0.028349	0.798892	0.038555	0.790867	0.029736
BrainSegVolNotVent	0.779857	0.037538	0.772948	0.030305	0.785834	0.038695	0.776074	0.031471
lhCortexVol	0.164771	0.010207	0.163181	0.010196	0.170575	0.010496	0.168994	0.010110
rhCortexVol	0.164149	0.010295	0.162912	0.009834	0.170608	0.010675	0.168900	0.009646
CortexVol	0.328920	0.020369	0.326092	0.019888	0.341183	0.021084	0.337894	0.019676
lhCerebralWhiteMatterVol	0.157377	0.010114	0.155820	0.009472	0.153875	0.009612	0.151551	0.009309
rhCerebralWhiteMatterVol	0.156808	0.010522	0.154840	0.009950	0.152915	0.009898	0.150025	0.009752
CerebralWhiteMatterVol	0.314184	0.020552	0.310659	0.019351	0.306790	0.019437	0.301576	0.018991
SubCortGrayVol	0.040864	0.002871	0.040194	0.002433	0.042919	0.003186	0.042213	0.002739
TotalGrayVol	0.446625	0.027101	0.443252	0.025556	0.459994	0.028590	0.455681	0.025742

Note: All volumes are normalized to estimated total intracranial volume (eTIV)
HC = Healthy Controls; IBS = Irritable Bowel Syndrome; SD = Standard Deviation; WM = White Matter
Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/04-computing-FS-versions-on-same-dataset.ipynb

Appendix A.6 Training 15 binary classifiers and their assessment

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.7200	0.6833	0.7200	0.7383	0.7124	0.4031	0.4204	0.0190
knn	K Neighbors Classifier	0.6867	0.6292	0.6867	0.7022	0.6693	0.2982	0.3265	0.1790
lr	Logistic Regression	0.6267	0.5500	0.6267	0.3938	0.4833	0.0000	0.0000	2.7970
svm	SVM - Linear Kernel	0.6267	0.5333	0.6267	0.3938	0.4833	0.0000	0.0000	0.0100
gbc	Gradient Boosting Classifier	0.6267	0.4667	0.6267	0.6544	0.5920	0.1978	0.2357	0.0260
dummy	Dummy Classifier	0.6267	0.5000	0.6267	0.3938	0.4833	0.0000	0.0000	0.0080
dt	Decision Tree Classifier	0.6133	0.6083	0.6133	0.6656	0.6096	0.2039	0.2374	0.1710
nb	Naive Bayes	0.6067	0.5583	0.6067	0.4801	0.5177	0.0450	0.0578	0.1650
ridge	Ridge Classifier	0.6067	0.2917	0.6067	0.3878	0.4726	-0.0364	-0.0408	0.0150
lightgbm	Light Gradient Boosting Machine	0.5867	0.5625	0.5867	0.5003	0.5239	0.0276	0.0270	0.0500
et	Extra Trees Classifier	0.5300	0.5812	0.5300	0.4621	0.4821	-0.0735	-0.0713	0.0300
rf	Random Forest Classifier	0.5267	0.5021	0.5267	0.5167	0.4987	-0.0492	-0.0293	0.0370
qda	Quadratic Discriminant Analysis	0.5133	0.5042	0.5133	0.4528	0.4698	-0.1175	-0.1328	0.0090
lda	Linear Discriminant Analysis	0.5000	0.3792	0.5000	0.5650	0.4773	-0.0224	0.0141	0.0100
ada	Ada Boost Classifier	0.4867	0.3708	0.4867	0.4466	0.4536	-0.0624	-0.0992	0.0210

```
CPU times: user 5.24 s, sys: 596 ms, total: 5.84 s
Wall time: 38.1 s
```

Figure A1. *Binary classification models trained using PyCaret.* Based on 36 morphometric features derived from the longitudinal stream of Freesurfer 7.4.1 applied to T1-weighted examinations in the Bergen cohort described in Tab. 2. For each of the 15 models, seven performance metrics on the O (HC) versus 1 (IBS) prediction were obtained as the means after stratified 10-fold cross-validation, i.e., for each iteration (out of 10), nine folds are combined to form the training set (90% of data), the remaining fold becomes the validation set (10% of data). The models are ranked according to accuracy (see text for more details).

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/05-predicting-IBS-vs-HC-from-morphometric-measures.ipynb

Appendix A.7 High resolution histological atlas segmentation of T1-weighted MPRAGE recording

As a proof of concept, Figure A2 displays a high resolution segmentation of the T1-weighted recording from subject BGA_046 in the Bergen cohort. This is based on the NextBrain project (https://github-pages.ucl.ac.uk/NextBrain) described in [46]. The NextBrain project provides a sophisticated brain segmentation module that utilizes a probabilistic atlas to identify 333 distinct regions of interest (ROIs) per hemisphere in *in vivo* brain scans. The segmentation process employs a Bayesian algorithm, making it adaptable to various MRI pulse sequences including T1-weighted, T2-weighted, and FLAIR. The software offers two implementation modes: a comprehensive Bayesian version and a faster alternative. The full version (used in this example), while more computationally intensive, provides detailed segmentation. The faster version utilizes a neural network for pre-computing atlas deformation, significantly reducing processing time to under an hour on standard hardware. Both versions generate outputs including bias-field corrected scans, SynthSeg segmentation, MNI registration, hemisphere-specific segmentations, and volumetric measurements in CSV format. The system employs a sophisticated Gaussian mixture model for tissue classification, with customizable parameters for bias field correction and tissue

grouping.

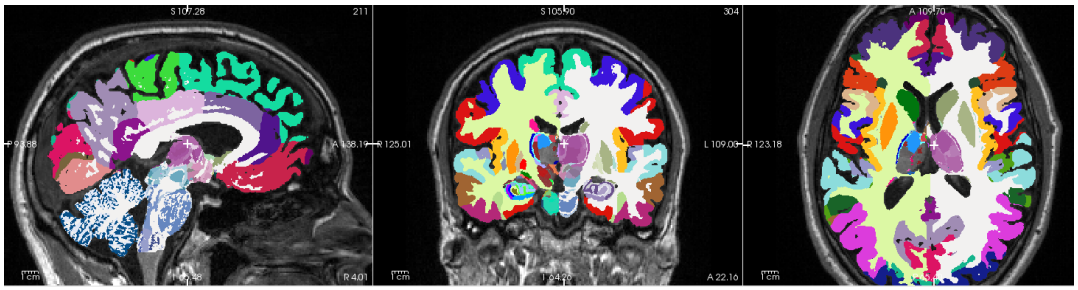


Figure A2. *High resolution Histological atlas segmentation, available in FreeSurfer 8.0.0-beta, of 3D T1-weighted MPRAGE recording from BGA_046.* Panels left to right: Sagittal, Coronal, and Axial section, respectively. The white cross-bar in the middle of the brain is located in the *paracentral nucleus* of the *left thalamus* at RAS coordinates 4.03, 22.15, 21.90. Cfr. the much coarser granularity of ASEG segmentation in Fig. 2, with the same positioning of the white cross-bar.

Generated by: https://github.com/arvidl/ibs-brain/blob/main/notebooks/01-freesurfer-freeview-t1-aseg-bga-046.ipynb