

Article

Brain morphometry and cognitive features in prediction of irritable bowel syndrome

(20241210a1) Submission deadline (*Diagnostics*): 2024-12-30Application of Artificial Intelligence in Gastrointestinal Disease https://www.mdpi.com/journal/diagnostics/special_issues/CS1956QYIM

Astri J. Lundervold ¹ 0000-0002-6819-6164*, Ben René Bjørsvik ², Julie Billing ¹, Birgitte Berentsen ^{5,6} 0000-0003-3574-7078, Gülen Arslan Lied ^{5,6} 0000-0002-1827-5008, Elisabeth K Steinsvik ⁵ 0000-0002-8371-1988, Trygve Hausken ⁵ 0000-0001-7080-8396, ^{4,†}, Daniela M. Pfabigan ¹ 0000-0002-4043-1695, Arvid Lundervold ^{2,3,†} 0000-0002-0032-4182

¹ Department of Biological and Medical Psychology, University of Bergen, Norway² Department of Biomedicine, University of Bergen, Bergen, Norway³ Medical-AI, Mohn Medical Imaging and Visualization Center, Department of Radiology, Haukeland University Hospital, Bergen, Norway⁴ Department of Clinical Medicine, University of Bergen, Bergen 5021, Norway⁵ National Center for Functional Gastrointestinal Disorders, Department of Medicine, Haukeland University Hospital, Bergen 5021, Norway⁶ Department of Biomedicine, University of Bergen, Bergen, Norway⁵ Medical-AI, Mohn Medical Imaging and Visualization Center, Department of Radiology, Haukeland University Hospital, Bergen, 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.

Citation: Lundervold, A.J.; Bjørsvik, B.R.; Billing, J.; ...; Pfabigan, S.M. and Lundervold, A. Brain morphometry, gender and cognition in IBS.

Diagnostics **2024**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: © 2024 by the authors. Submitted to *Diagnostics* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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

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 abdominal 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. Recent empirical support for this systems-level approach comes from Li et al. [19], 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 mid-cingulate 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 [20,21]. Building on this network perspective, Skrobisz et al. [22] 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 [22] call for integrating clinical measures, we investigated whether combining cognitive performance data with morphometric features would enhance the accuracy of IBS versus HC classification.

We will address these aims in the following way:

- A We aim to replicate the morphometric differences between IBS patients and HC reported in [22] by using the same FreeSurfer software version (FS 6.0.1) and a similar univariate analysis approach as in the original study.
- B We will compare outcomes of the morphometric segmentation of the T1-w recordings in the current dataset ($n = 78$) for the FreeSurfer software version used in [22] (FS 6.0.1) with a more recent version (FS 7.4.1) by calculating pairwise correlations.
- C We will investigate if morphometric features (derived from the FS 7.4.1 brain segmentation, both cross-sectional and longitudinal-stream-based analyses) can be used to distinguish between IBS individuals and HC. To this end, we will employ four different strategies: (i) we will conduct univariate comparisons between IBS individuals and HC; (ii) we will conduct multivariate comparisons between the two groups (by incorporating covariance structures of the morphometric features); (iii) we will apply a machine learning framework (ML) to predict group membership based on morphometric features; and (iv) we will identify feature importance of respective morphometric features should the classification into IBS individuals and HC be successful.
- D We will investigate if the prediction of group membership (IBS vs. HC) is improved when cognitive performance is added as a predictor to the morphometric features ML model. If prediction is improved, we will identify those features that have the strongest influence on distinguishing IBS individuals from HC.

Materials and Methods

Participants

This investigation is part of the Bergen Brain-Gut project, a comprehensive clinical study conducted at Haukeland University Hospital, Norway, between 2020 and 2022 (detailed protocol in Berentsen et al. [23]). Our sample comprised 49 patients with IBS and 29 healthy controls (HCs), all aged 18 years or older. Participant recruitment employed multiple strategies, including media advertisements, informational flyers, and direct referrals from the hospital's outpatient clinic. All potential participants underwent systematic screening by a trained nurse using standardized inclusion and exclusion criteria (Table 1), followed by a comprehensive assessment battery including gastrointestinal measures, psychometric testing, and multiparametric magnetic resonance imaging (MRI).

The final sample size was determined by several factors. While no formal a priori power analysis was conducted as effect sizes for brain morphometric differences in IBS were not well established at the study's inception, our sample size aligns with or exceeds those of similar neuroimaging studies in functional gastrointestinal disorders [22]. For the current analyses, we included only participants with complete data on all key measures and artifact-free MRI scans suitable for automated brain segmentation, ensuring data quality while maintaining the largest possible sample size. This approach balanced statistical power requirements with practical constraints and methodological rigor.

Inclusion criteria	Exclusion criteria
Rome-IV criteria: Recurrent abdominal pain average at least 1 day/week during the last 3 months, and associated alterations in bowel habits at least 6 months before diagnosis. Other causes are excluded. Normal diet at least 3 weeks before inclusion IBS score equal to or above 175	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. 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 [23].*Measures*

Age and sex (not genetically verified) were self-reported by the participants at baseline. 133 134

The IBS-Severity Scoring system (IBS-SSS) 135

The IBS-Severity Scoring system is a questionnaire used to assess the severity and frequency of GI-related IBS symptoms [24]. 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 [24]. 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) 146

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 [25]. 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 ($1 \times 1 \times 1 \text{ mm}^3$) 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 $256 \times 256 \text{ mm}^2$ 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.

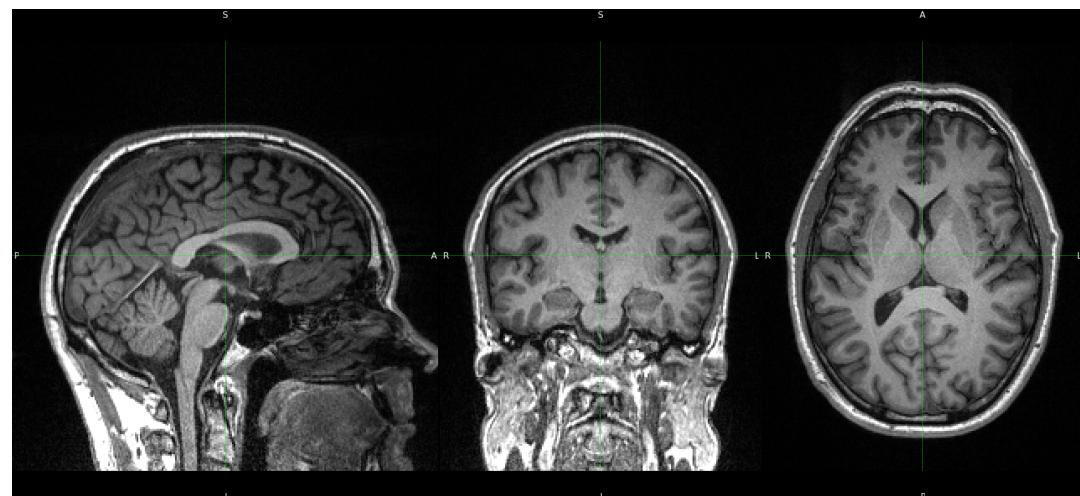


Figure 1. 3D T1-weighted MPRAGE recording from BGA_046. Panels left to right: Sagittal, Coronal, and Axial section, respectively.

Generated by: <https://github.com/arvid1/ibs-brain/blob/main/notebooks/01-freesurfer-freeview-t1-aseg-bga-046.ipynb>

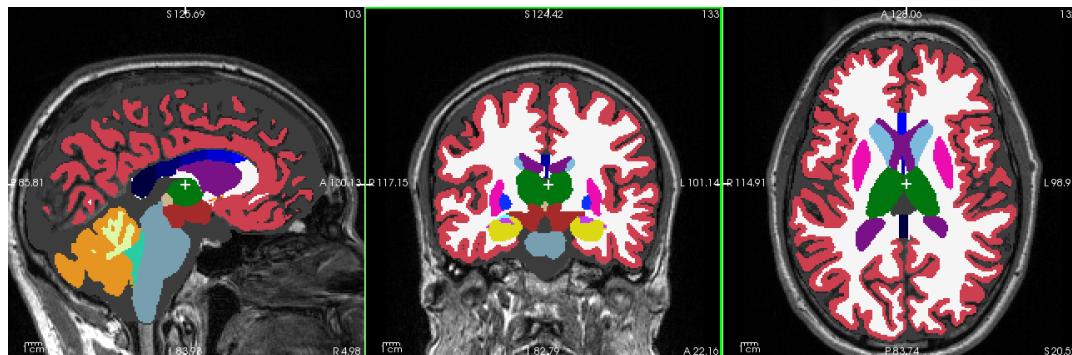


Figure 2. The color-coded ASEG segmentation mask by FreeSurfer 7.4.1 overlaid on 3D T1-w MPRAGE from BGA_046. Panels left to right: Sagittal, Coronal, Axial section, respectively. The white cross is located in the medial part of Left-Thalamus. Thalamus: green, Hippocampus: yellow, Caudate: light blue, 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 [26]. 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. [22], 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. [23]), 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 [22] 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 [27–32].

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. However, by analyzing our data with both versions, we can distinguish between genuine biological differences and methodologically-induced variations in brain morphometry.

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, a robust non-parametric measure particularly suitable for non-normally distributed data. Following established conventions, we interpreted Cliff's Delta 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. 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. [22] regarding IBS versus HC discrimination, using the same FreeSurfer-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 vs. 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 about 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 vs. HC from the morphometric features using a machine learning framework (ML)?
- (iv) Identifying feature importance of the 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? 273
- (ii) Employing a multivariate comparison, incorporating covariance structures in the cognitive features? 274
- (iii) By predicting IBS vs. HC from the morphometric and cognitive features using a machine learning framework (ML)? 276
- (iv) identifying feature importance of the morphometric and cognitive measures included in the model with the best prediction? 278

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 [22], 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, 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. For interpretation, absolute values of delta less than 0.147 indicate a negligible effect, values between 0.147 and 0.33 indicate a small effect, values between 0.33 and 0.474 indicate a medium effect, and values greater than or equal to 0.474 indicate a large effect.

Permutation Testing

To address the challenges of small sample sizes and potential non-normal distributions, we implemented permutation testing with 1,000 iterations. This approach calculates an observed test statistic (sum of squared differences between group means) and generates a null distribution by randomly reassigning group labels and recalculating the statistic. The resulting empirical p-value represents the proportion of permuted statistics exceeding the observed value. This method offers several advantages: robustness to non-normality, natural handling of multiple comparisons, and suitability for small samples in multivariate analyses.

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 is a fundamental measure in multivariate statistics that quantifies the distance between a point P and a distribution D , while accounting for the correlations within the dataset. Unlike the Euclidean distance, the Mahalanobis distance incorporates the covariance structure of the data. The classical Mahalanobis distance is defined as $D = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$, where x represents the data point, μ is the mean vector, Σ is the covariance matrix, and Σ^{-1} denotes its inverse.

Remark: The classical Mahalanobis distance D and Cohen's d represent complementary approaches to measuring group differences, with distinct mathematical foundations and interpretations. While Cohen's d measures the standardized difference between two group means in univariate space ($d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$), the Mahalanobis distance extends this concept to multivariate space by incorporating the covariance structure of the data. The Mahalanobis distance can be viewed as a multivariate generalization of the number of standard deviations a point lies from a distribution's mean, accounting for the correlations between variables through Σ^{-1} . When comparing two groups, as in our case with IBS and HC, the squared Mahalanobis distance is proportional to the Hotelling's T^2 statistic, which itself can be considered a multivariate analog of the squared t-statistic used in univariate analysis. While Cohen's d has well-established guidelines for effect size interpretation (small: 0.2, medium: 0.5, large: 0.8), the interpretation of Mahalanobis distances depends on the dimensionality of the data and the covariance structure, making direct comparison with Cohen's d challenging. However, both metrics serve important roles in quantifying group differences, with the Mahalanobis distance being particularly valuable when dealing with correlated multivariate data, as is common in neuroimaging studies.

The *robust Mahalanobis distance* computation implements a resistant version of the classical Mahalanobis distance to measure the multivariate separation between two groups (IBS and HC) while accounting for outliers and non-normality. Our implementation enhances the classical formula with robust statistical methods to handle outliers and non-normal distributions. The first robust modification employs winsorization, which clips extreme values at the 10th and 90th percentiles while preserving the data structure. We also replace the arithmetic mean with the median. 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 (see Appendix A.3 for details).

Prediction of class belonging using machine learning

In tasks C(iii) - 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 employed a stratified sampling approach, partitioning the dataset into training (70%) and testing (30%) sets while preserving the distribution of key variables (IBS/HC status, sex, and cognitive function levels) 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 ad-

vanced 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 and LightGBM), and instance-based learners (K-Nearest Neighbors). This diverse algorithm selection enabled exploration of different decision boundaries and feature interaction patterns.

To ensure robust model assessment and mitigate overfitting risks, we implemented a nested cross-validation strategy. The outer loop employed 10-fold cross-validation for model selection, while the inner loop optimized hyperparameters through random search with internal cross-validation. This approach provided unbiased performance estimates while preventing data leakage between model selection and evaluation phases. Hyperparameter optimization focused on algorithm-specific parameters crucial for neuroimaging data: regularization strengths for linear models, kernel parameters for SVMs, tree structure parameters for ensemble methods, and neighborhood configurations for instance-based learners. The final model selection prioritized both predictive performance and model interpretability, considering the clinical relevance of our findings.

Model Performance Assessment

Model evaluation employed a multi-faceted approach to ensure comprehensive performance assessment, particularly important given our class imbalance between IBS and HC groups. While overall classification *accuracy* provided an initial performance indicator, we used the implementation of additional metrics to capture nuanced aspects of model behavior and clinical relevance. These are the *F1 score*, computed as the harmonic mean of precision and recall, offered a balanced assessment of model performance by considering both false positives and false negatives. We supplemented this with the Receiver Operating Characteristic Area Under Curve (ROC-AUC), which quantifies discrimination ability across different classification thresholds and is particularly robust to class imbalance. We also included the Kappa measure (*Cohen's Kappa*), used to assess the agreement between the predicted classifications and the true classifications, while accounting for the possibility of agreement occurring by chance. It provides a more robust evaluation of a model's performance compared to simple accuracy, especially in imbalanced datasets. For detailed error analysis, we generated confusion matrices showing the distribution of correct and incorrect predictions across classes, enabling identification of specific classification patterns and potential biases. For the cognitive function analysis, which involved multiple classes, we employed macro-averaged versions of all metrics. This approach, calculating metrics independently for each class before averaging, ensures equal weighting of all classes regardless of their relative frequencies in the dataset. This consideration was particularly crucial given the uneven distribution of cognitive performance levels in our sample.

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. To quantify uncertainty in our performance estimates, we calculated 95% confidence intervals for all metrics using bootstrapping methods. These intervals provide important context for interpreting model reliability and potential clinical utility, particularly given our moderate sample size.

The comprehensive evaluation strategy, combining multiple performance metrics with detailed error analysis and uncertainty quantification, ensures a thorough understanding of model behavior and reliability. This approach is particularly valuable for potential clinical applications, where understanding both model capabilities and limitations is crucial for responsible implementation.

Feature Importance and Model Interpretability Analysis

Understanding the relative contribution of different morphometric features to classification performance is crucial for both methodological validation and biological insight. We implemented two complementary approaches to feature importance analysis, combining model-agnostic methods with sophisticated game-theoretic interpretability techniques.

The first approach employed permutation importance analysis, which quantifies feature relevance by measuring the degradation in model performance when individual features are randomly permuted. For each morphometric feature, we performed multiple permutation iterations, calculating the mean decrease in model performance. This approach provides an intuitive measure of feature importance while remaining independent of the underlying model architecture, enabling consistent interpretation across different classification algorithms.

Our second approach utilized SHAP (SHapley Additive exPlanations) values, grounded in cooperative game theory [33]. This method offers a unified framework for model interpretation at both global and local levels. Global analysis aggregates SHAP values across all cases to identify consistently important features, while local analysis examines feature contributions to individual predictions. This dual-scale interpretation provides crucial insights into how different brain regions contribute to classification decisions, both at the population level and for individual cases.

We applied both analytical approaches to our highest-performing models across all classification tasks. The results were synthesized through comprehensive visualization techniques, including SHAP summary plots depicting feature value distributions and their associated impact on model predictions. These visualizations integrate both the magnitude and directionality of feature effects, providing insight into how specific morphometric characteristics influence classification decisions.

The integration of permutation importance and SHAP analysis provides complementary perspectives on feature relevance. While permutation importance identifies features whose disruption most impacts model performance, SHAP analysis reveals the complex interactions between features and their contributions to specific predictions. This comprehensive approach to model interpretation enhances our understanding of the neurobiological features distinguishing IBS patients from healthy controls, while also illuminating potential relationships between brain structure, sex/gender differences, and cognitive function. The results not only validate our modeling approach but also suggest specific brain regions and networks that may play key roles in IBS pathophysiology.

Results

Sample Demographics and Clinical Characteristics

Our study included 78 participants (49 IBS patients and 29 healthy controls), with demographic and clinical characteristics summarized in Table 2. The age distributions were comparable between groups, with median ages of 34 years in the IBS group and 33 in the healthy control group. Both groups showed a predominance of female participants, with slightly higher representation in the IBS group (38 of 49) compared to the control group (20 of 29).

IBS symptom severity, assessed using the IBS-SSS, demonstrated clear group separation. The IBS group reported substantially higher symptom severity compared to healthy controls, consistent with our inclusion criteria. Missing IBS-SSS data for three participants in each group (total n = 6) were handled through multiple group and gender-based imputation. The observed severity scores in the IBS group indicate predominantly moderate to severe symptomatology, while healthy controls showed minimal gastrointestinal symptoms, as expected.

Table 2. Demographic and Clinical Characteristics of the Study Sample

Group	Age Median (IQR)	IBS_SSS Median (IQR)	Sex F/M (%)	N	Missing 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)

Table 3 addresses the attempt to replicate, in our Bergen cohort, the morphometric findings in Skrobisz et al. [22] regarding IBS vs. HC, using the same metrics (i.e., 35 eTIV-normalized volumes) from FreeSurfer-derived brain regions and the same FreeSurfer version (FS 6.0) as reported in Skrobisz et al. (2022).

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-analysis-fs6.ipynb>

The volumetric analysis comparing brain structures between IBS patients and healthy controls across the two cohorts reveals several noteworthy patterns. The Bergen cohort shows consistently larger volumes (6-8% for global measures, up to 35% for specific struc-

479

480

481

tures like the *accumbens*). However, global brain eTIV-normalized volumes show remarkable consistency between IBS and HC groups within each cohort, with BrainSegVol values being nearly identical (Skrobisz: HC 0.753 ± 0.018 , IBS 0.749 ± 0.016 ; Bergen: HC 0.805 ± 0.025 , IBS 0.806 ± 0.024). Cortical volumes demonstrate similar stability, with total cortical volume (CortexVol) showing minimal between-group differences in both cohorts. Subcortical structures exhibit subtle variations, with a minor trend toward volume reduction in IBS patients' subcortical gray matter (SubCortGrayVol), though these differences remain within standard deviation ranges. The analysis of white matter reveals consistent volumes between groups within cohorts, while white matter hypointensities show intriguing variation in the Bergen cohort. Corpus callosum segments maintain relatively uniform volumes across all groups. Methodologically, it is important to note the different cohort sizes (Skrobisz: HC=19, IBS=20; Bergen: HC=29, IBS=49) and the possible subtle differences in FreeSurfer versions (6.0 vs 6.0.1) and operating systems, which might contribute to the systematic differences observed between cohorts. The normalization to estimated total intracranial volume (eTIV) enables direct comparisons while controlling for head size variation within cohorts but not between cohorts.

A more penetrating reproducibility analysis is illustrated in Figure 3 showing the differences in HC versus IBS of eTIV-normalized brain regions volumes between the Skrobisz (2022) and the Bergen cohorts. The plot displays effect sizes from the Skrobisz cohort on the *x*-axis against the Bergen cohort on the *y*-axis, with a diagonal reference line indicating perfect agreement. We are here using Cohen's d values for region-wise effect size which are a standardized measure of the difference between means. More specifically, for each eTIV-normalized brain region volume and cohort we are calculating 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 score (S) for each brain region was computed as a composite measure combining three key components: binary sign consistency (σ), confidence interval overlap (ω), and effect magnitude (ϵ). Specifically, $S = \sigma + \omega + \epsilon$, where σ equals 1 if the direction of effect is consistent between cohorts and 0 otherwise, ω equals 1 if the 95% confidence intervals overlap and 0 otherwise, and ϵ represents the minimum absolute effect size observed across cohorts. This scoring system yields higher values for brain regions demonstrating consistent effects across cohorts, with the magnitude component (ϵ) providing additional weight to regions showing stronger effects. A higher overall score thus indicates greater reliability and robustness of the observed effects across different study populations and analysis pipelines, offering a quantitative basis for identifying the most consistently replicated findings in brain morphometry.

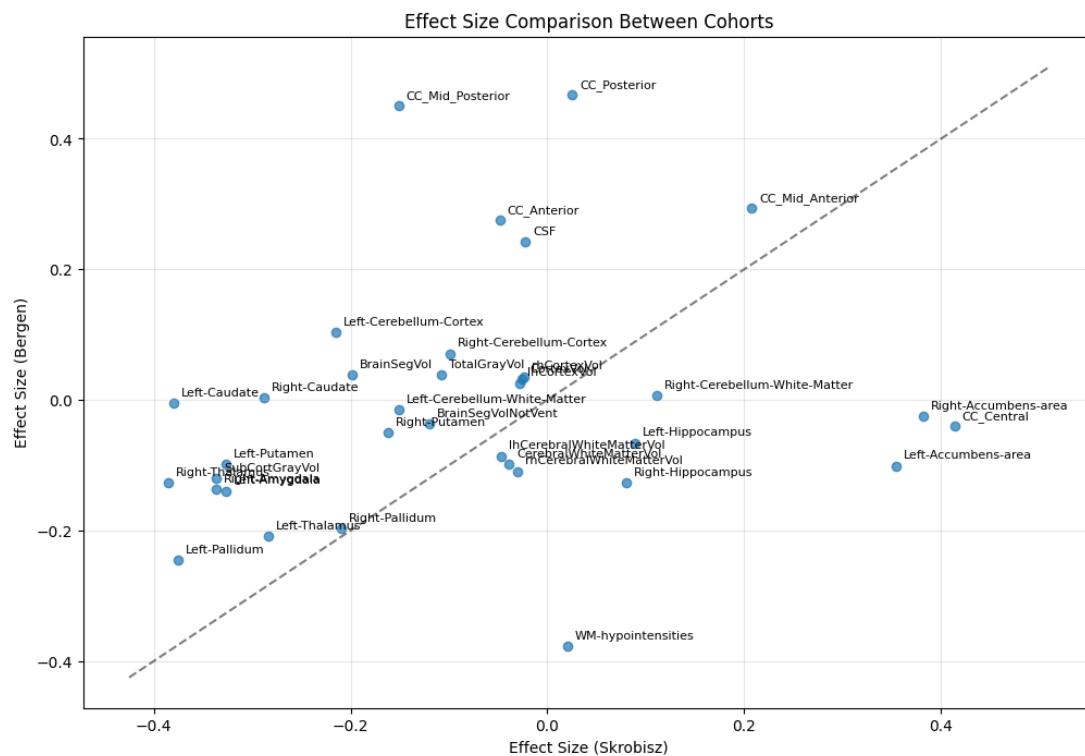


Figure 3. The large disparity of region-wise effect size of IBS vs. HC in comparison between the Skrobisz (2022) cohort and the Bergen cohort. Scatterplot of calculated Cohen's d effect sizes for each region in both cohorts (see text for details).

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

The scatter plot, presenting a comparison of effect sizes between the two cohorts, reveals a moderate effect size correlation of 0.203 ($p=0.243$) between the cohorts. We found a sign consistency of 51.4%, meaning that in approximately half of the brain regions (51.4%), the direction of the difference between IBS and HC groups was the same in both cohorts. We also found a complete (100%) CI overlap rate, i.e., for all brain regions the 95% confidence intervals of the effect sizes from the two cohorts overlapped (data not shown). The relatively low sign consistency suggests poor replication of the directional effects between cohorts. The complete overlap suggests that while the point estimates of effects might differ between cohorts, we cannot conclude that these differences are statistically significant, given the uncertainty in our measurements.

Several brain regions show notable consistency between cohorts, particularly the CC_Mid_Anterior, Left-Pallidum, Left-Thalamus, Right-Pallidum, and Left-Amygdala, which are identified as the most consistent regions with overall scores (S) ranging from 2.14 to 2.26. Some regions demonstrate substantial divergence from the diagonal, suggesting differential effects between cohorts. For instance, WM-hypointensities show particularly discordant effects between cohorts, as do several corpus callosum regions (CC_Posterior and CC_Mid_Posterior) which exhibit stronger effects in the Bergen cohort. The cerebellar regions cluster near the center, indicating relatively modest effects in both cohorts. To conclude our analysis, the pattern of data points suggests modest overall agreement between cohorts, with considerable variability in effect sizes across brain regions. The widespread distribution of points around the diagonal line, combined with the moderate correlation coefficient, indicates that while there is some consistency in findings between cohorts, there are also substantial differences in how IBS-related brain alterations manifest in these two independent samples.

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

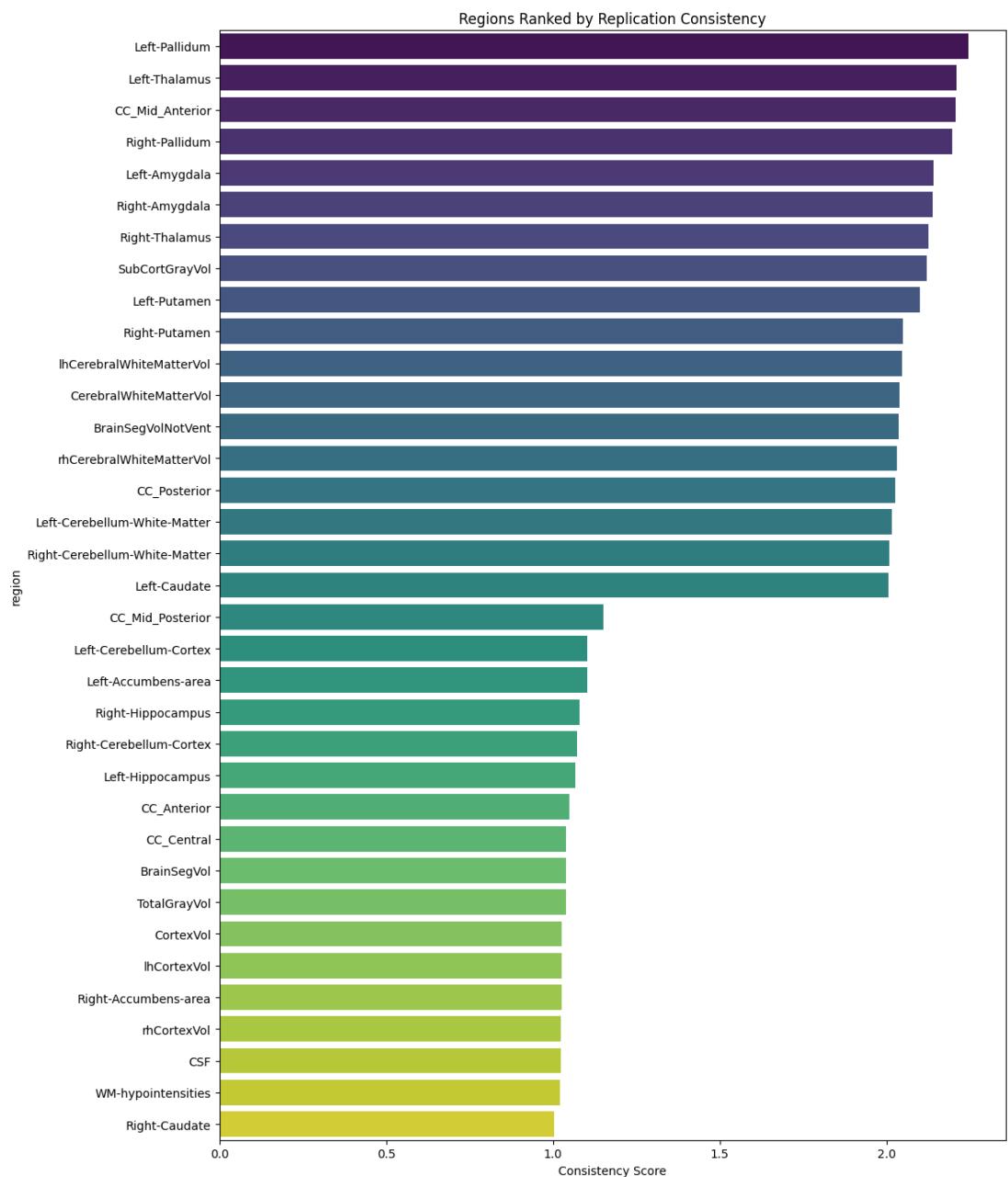


Figure 4. Barplot of regions ranked by their replication consistency . The consistency score used for ranking combines several factors: sign match, which assesses whether the direction of difference is the same in both cohorts; the CI overlap, which checks whether the confidence intervals for the effects overlap between cohorts; and effect magnitude, which evaluates how similar the size of the effect is between cohorts (see text for details).

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

The replication consistency scores and their rankings reveal that several brain regions exhibit high levels of replication consistency across cohorts. The top-ranked regions, such as the Left-Pallidum, Left-Thalamus, and CC_Mid_Anterior, all have an overall score above 2.0, indicating strong consistency. These regions show both a consistent direction of effect (sign match) and complete confidence interval overlap, along with relatively high effect magnitudes. This suggests that these regions are reliably replicated

across studies, making them robust findings. In contrast, regions like the Right-Caudate, Right-Cerebellum-Cortex, Left- and Right-Hippocampus, CC_Mid_Posterior and Left-Cerebellum-Cortex, which have overall scores around 1.1, lack sign consistency despite having complete confidence interval overlap. This indicates that while the magnitude of effects is similar, the direction of effects is not consistent across cohorts, suggesting variability in these findings. Overall, the analysis highlights that while some regions demonstrate strong replication consistency, others show variability, particularly in the direction of effects, which may reflect differences in cohort characteristics or measurement variability. To conclude this consistency analysis, the rankings suggest that while eTIV-normalized volumes differ systematically between cohorts, some regions show more consistent relative patterns than others. This information is valuable for understanding which brain measurements might be more reliable across different studies and analysis pipelines. An interim conclusion is that the attempt to replicate the specific HC vs IBS morphometric differences reported in Skrobisz et al. (2022) was largely unsuccessful, suggesting that reported brain structural differences in IBS might be less robust or more heterogeneous than previously thought.

To evaluate the reproducibility and reliability of brain morphometry measurements in IBS research, we conducted a comparative analysis using multiple FreeSurfer processing pipelines on a *fixed set* of neuroimaging data, using the Bergen cohort. Specifically, we examined how morphometric measurements and IBS versus healthy control (HC) group differences were affected by different FreeSurfer versions (6.0.1 versus 7.4.1) and different processing streams within FreeSurfer 7.4.1 (cross-sectional versus longitudinal). The inclusion of longitudinal stream analysis was enabled by our interventional study design, allowing us to assess the consistency of morphometric findings across different analytical approaches. Moreover, in these comparisons, we had access to all measurements from all subjects and not only summary statistics as in the replication analysis of the Skrobisz (2022) cohort.

Univariate comparisons of morphometric measures using different FreeSurfer versions

In this analysis, we aimed to compare brain region volumes between two FreeSurfer versions, FS 6.0.1 and FS 7.4.1 cross-sectional, to assess their consistency in measuring volumetric differences between IBS patients and healthy controls (HC). The scatter plot matrix in Figure 5 visualizes the relationship between corresponding brain regions across the two versions. Each subplot represents a specific brain region, with volumes from FS 6.0.1 plotted on the y-axis and those from FS 7.4.1 on the x-axis. Data points are color-coded by group, with blue representing HC and red representing IBS. An identity line is included in each plot to facilitate visual assessment of agreement between the two versions.

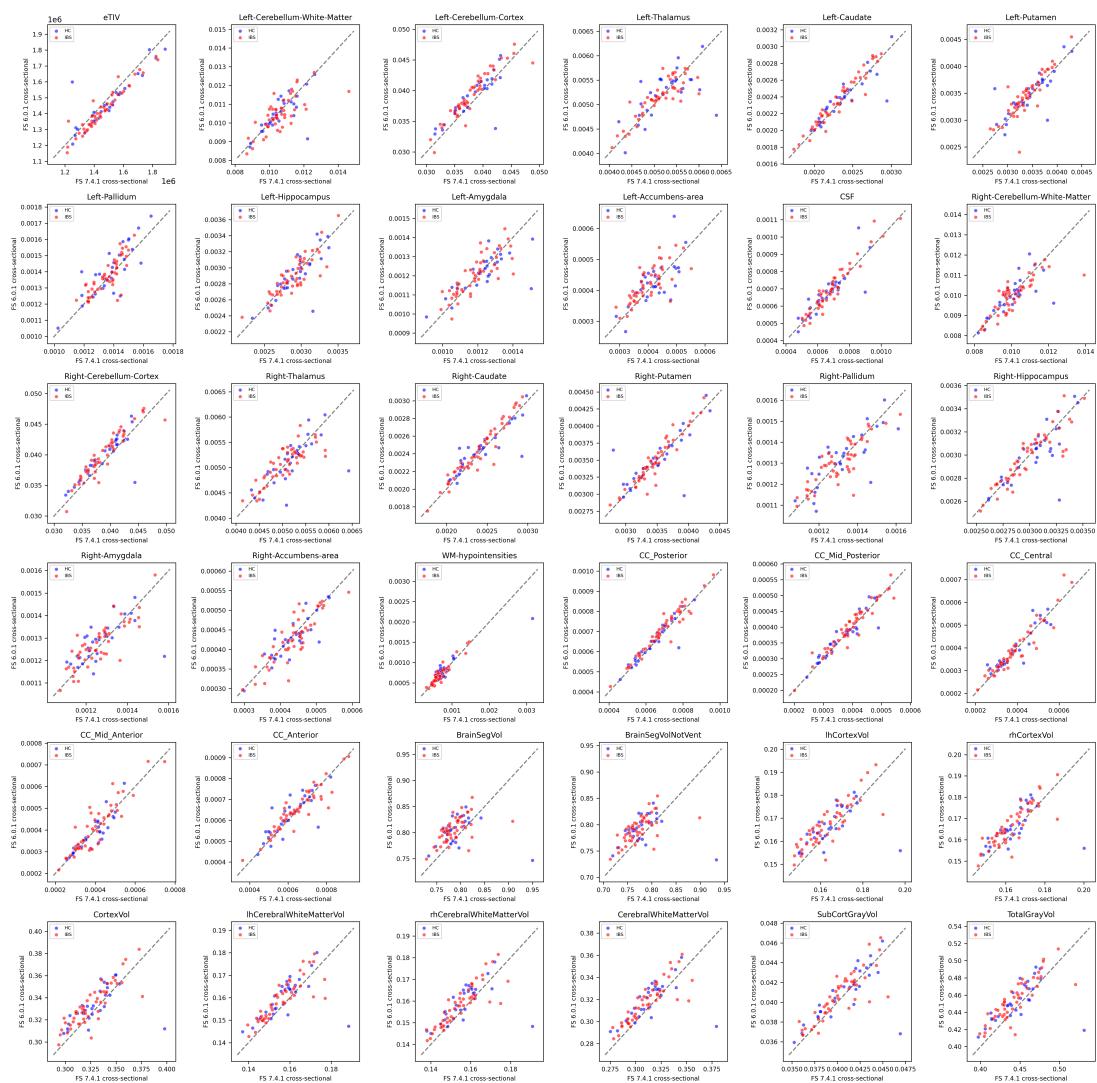


Figure 5. Scatter plot matrix comparing brain region volumes between FreeSurfer versions, FS 6.0.1 and FS 7.4.1 cross-sectional. Each subplot represents a specific brain region, with FS 6.0.1 cross-sectional eTIV-normalized volumes [eTIV in mm³] on the y-axis and FS 7.4.1 cross-sectional on the x-axis. Data points are color-coded by group, with blue for HC and red for IBS. An identity line is included to facilitate visual assessment of agreement between the two FS versions (see text for details).

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

The scatter plot matrix reveals a generally strong agreement between FS 6.0.1 and FS 7.4.1 cross-sectional measurements, as indicated by the clustering of data points around the identity line in most brain regions. This suggests that both versions provide consistent volumetric estimates. However, some variability is observed, particularly in regions such as the corpus callosum and subcortical structures, where deviations from the identity line are more pronounced. These discrepancies may reflect differences in segmentation algorithms or version-specific sensitivity to regional signal intensity variations. Overall, the figure supports the reliability of both FreeSurfer versions in capturing brain volume differences, though minor inconsistencies warrant further investigation regarding morphometry-based group prediction.

The summary statistics by the mean and standard deviation in HC and IBS patients on each of the 35 included brain regions (also reported by Skrobisz et al. [22]) derived from the aseg.stats files using cross-sectional Freesurfer 6.0.1 and Freesurfer 7.4.1, respectively,

is shown in the Appendix as Table A2.

Figure 6 depicts a scatter plot matrix comparing brain region volumes between two FreeSurfer pipelines: FS 7.4.1 cross-sectional and FS 7.4.1 longitudinal. This visualization allows for a comprehensive comparison of the cross-sectional versus the longitudinal stream measures brain region volumes, highlighting potential discrepancies.

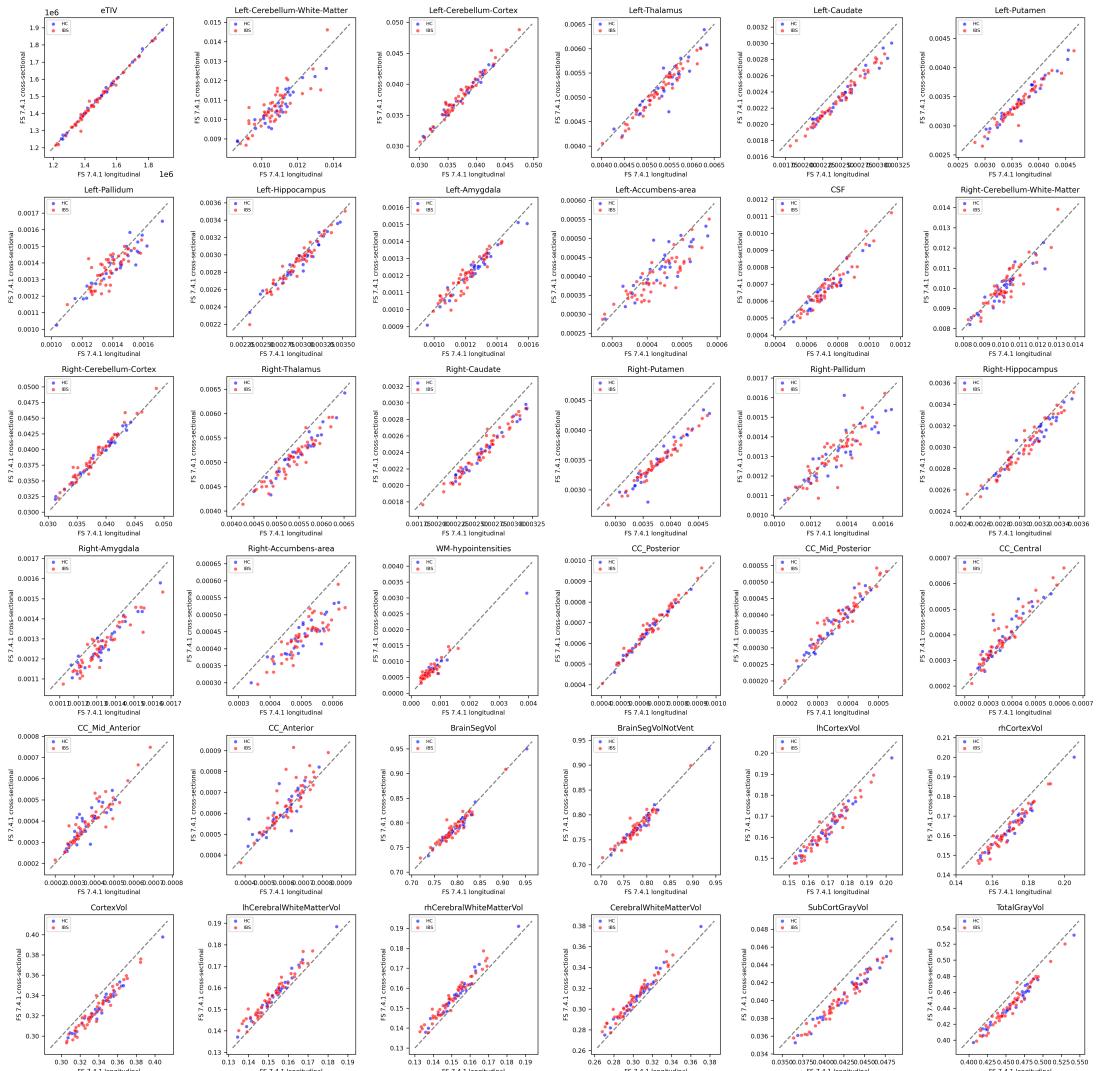


Figure 6. Scatter plot matrix comparing brain region volumes between FS 7.4.1 cross-sectional and FS 7.4.1 longitudinal stream. Each subplot represents a specific brain region, with FS 7.4.1 cross-sectional eTIV-normalized volumes [eTIV in mm³] on the y-axis and FS 7.4.1 longitudinal on the x-axis. Data points are color-coded by group, with blue for HC and red for IBS. An identity line is included to facilitate visual assessment of agreement between the two processing streams (see text for details).

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

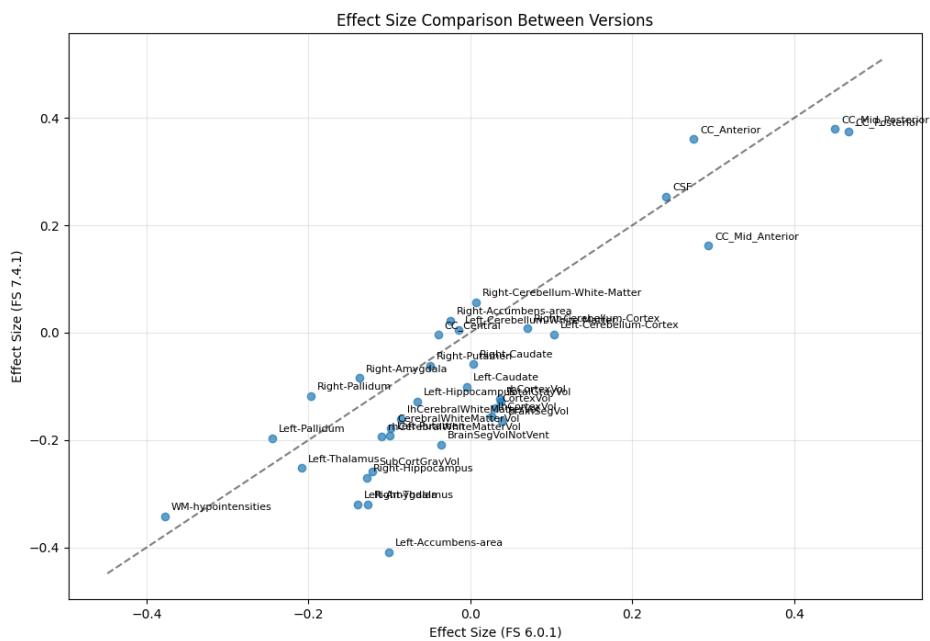
The scatter plot matrix reveals a strong overall agreement between FS 7.4.1 cross-sectional and FS 7.4.1 longitudinal measurements, as indicated by the clustering of data points around the identity line in most brain regions. However, some variability is observed, particularly in regions such as the corpus callosum and subcortical structures, where deviations from the identity line are more pronounced. The scatter plot matrix also reveals notable systematic biases between cross-sectional and longitudinal processing streams in several key regions. Particularly striking are the systematic differences in CortexVol, where

the longitudinal stream consistently produces lower volume estimates compared to the cross-sectional approach. Similar systematic biases are observed in subcortical structures, with Right-Putamen and Caudate showing consistent offsets from the identity line. CSF volumes demonstrate an opposite trend, with longitudinal processing typically yielding higher estimates than cross-sectional processing. These systematic differences suggest that the choice of processing stream can significantly impact volume estimates, even within the same FreeSurfer version (7.4.1). Interestingly, these biases appear consistent across both IBS and HC groups, maintaining the relative differences between groups regardless of the processing stream. This systematic variation between cross-sectional and longitudinal processing was not observed in the previous comparison between FS 6.0.1 and FS 7.4.1 cross-sectional analyses, indicating that these differences are specifically related to the processing stream rather than version updates. These findings highlight the importance of maintaining consistent processing approaches when comparing across studies or conducting longitudinal analyses.

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 shows the region-wise effect size of IBS vs. HC dependent on FreeSurfer version (a) and FS 7.4.1 cross-sectional versus longitudinal stream (b).

a)



b)

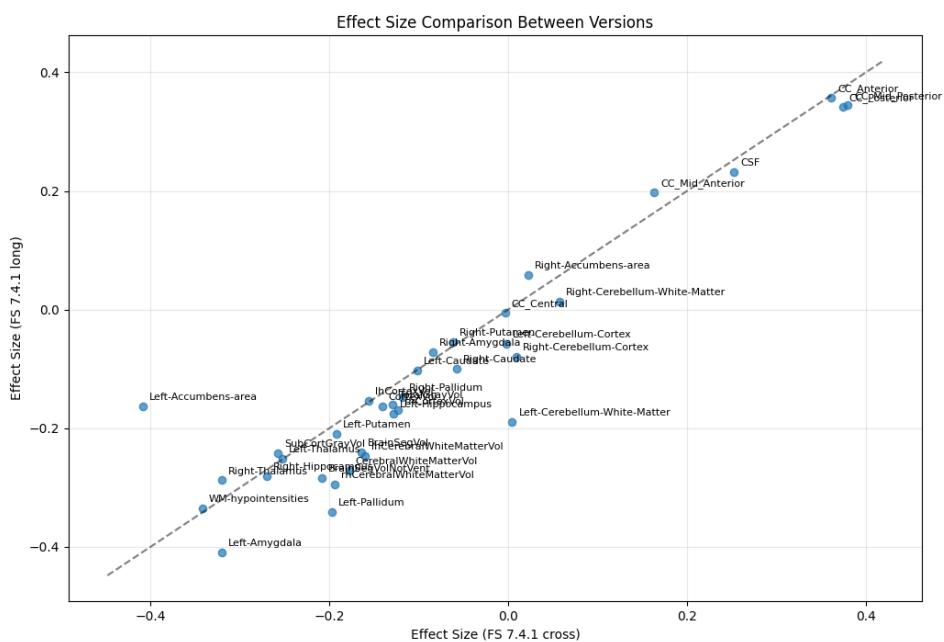


Figure 7. Region-wise effect size of IBS verus HC in comparison of FreeSurfer versions and analysis pipelines. a) FS 6.0.1 cross-sectional versus F 7.4.1 cross-sectional. b) F 7.4.1 cross-sectional analysis versus FS 7.4.1 longitudinal stream (see text for details).

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

Both panels (a) and (b) in Figure 7 display a scatter of brain regions with effect sizes plotted on x and y axes, with a diagonal reference line indicating perfect agreement between versions/methods. In subfigure (a), most data points cluster around the diagonal

line, suggesting general consistency between FreeSurfer versions 6.0.1 and 7.4.1 for cross-sectional analyses. Notable regions showing larger effects include the CC_Anterior and CC_Mid_Posterior regions in the positive direction, and Left-Accumbens-area in the negative direction. Subfigure (b) demonstrates the comparison between cross-sectional and longitudinal approaches within FreeSurfer 7.4.1. The distribution pattern is similar to (a), with strong correlation between methods. The CC_Anterior and CC_Mid_Posterior regions again show the largest positive effect sizes, while regions like Left-Amygdala and Left-Accumbens-area display the strongest negative effects. The tight clustering around the diagonal line suggests that both cross-sectional and longitudinal analyses in FreeSurfer 7.4.1 produce comparable effect size estimates for most brain regions.

In terms of data distribution patterns, panel (b) shows slightly better alignment of points along the diagonal reference line compared to subfigure (a), particularly in the central region of the plot. This suggests that the methodological comparison within the same FreeSurfer version (b) yields more consistent results than the version comparison (a). The effect size ranges also differ slightly. In panel (a), the effect sizes span approximately from -0.4 to 0.5 on both axes, while in panel (b), the x-axis ranges from about -0.4 to 0.4, showing a slightly wider distribution of effects in the analysis of FreeSurfer 6.4.1. Some brain regions show different relative positions between the two plots. For instance, the CC_Mid_Anterior region appears to show more deviation from the diagonal line in panel (a) compared to panel (b), suggesting that this region's measurements are more sensitive to FreeSurfer version changes than to the choice of cross-sectional versus longitudinal analysis method. The spatial distribution of specific anatomical regions also varies between plots, with some regions showing more consistent clustering in panel (b), particularly in the central portion of the plot, indicating that the cross-sectional and longitudinal analyses within FreeSurfer 7.4.1 produce more coherent results for these regions.

Figure 8 presents barplots showing regions ranked by their IBS vs HC separation consistency, dependent on FreeSurfer version (a) and FS 7.4.1 cross-sectional versus longitudinal stream (b) with distinct comparisons between panels.

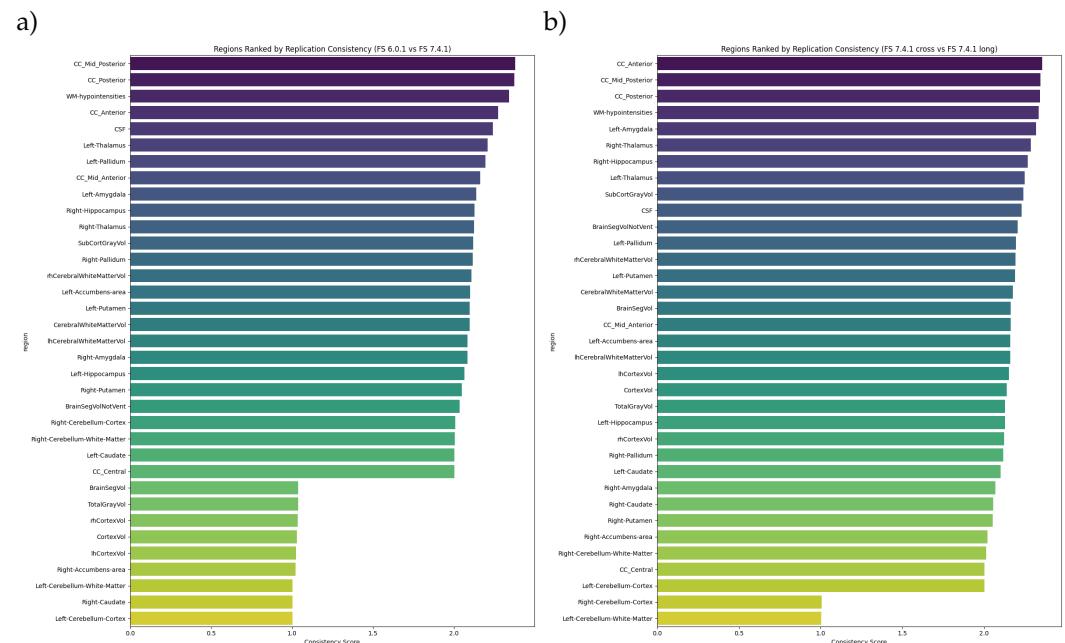


Figure 8. Barplot of regions ranked by their IBS versus HC separation consistency across FS versions.
a) FS 6.0.1 cross-sectional versus F 7.4.1 cross-sectional. b) F 7.4.1 cross-sectional analysis versus FS 7.4.1 longitudinal stream (see text for details).

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

While both panels display consistency scores ranging from 0 to approximately 2.5, they reveal several key differences: Panel (a) compares FreeSurfer versions (6.0.1 vs. 7.4.1 cross-sectional), showing greater variability in consistency scores across regions. The distribution of the highest scores appears more gradual, with scores ranging from approximately 1.0 to 2.5. The highest consistency is observed in CC_Mid_Posterior and CC_Posterior regions, while Left-Cerebellum-Cortex shows the lowest consistency. Panel (b), comparing analysis methods within FreeSurfer 7.4.1 (cross-sectional vs. longitudinal), demonstrates a more pronounced stepwise pattern in consistency scores. The distribution appears more clustered, with a sharper distinction between a larger number of high-consistency and a few low-consistency regions. CC_Anterior and CC_Mid_Posterior regions maintain high consistency, but the ordering of other regions differs notably from panel (a). The relative rankings of brain regions show substantial differences between panels. For instance, Left-Amygdala ranks higher in consistency in panel (b) compared to its position in panel (a), suggesting that this region's measurements are more consistent between analysis methods than between FreeSurfer versions. The most striking difference is in the pattern of consistency scores: panel (b) displays a more uneven, distinct grouping of regions into high and moderate consistency clusters. This suggests that the choice of analysis method (cross-sectional vs longitudinal) produces a pattern of a larger number of higher regional consistency compared to the version comparison.

Finally, we note that the regional consistency scoring of IBS versus HC separation was overall higher when using the same cohort (same set of neuroimaging data) with different FreeSurfer software versions and processing pipelines (Fig. 8) compared to using different cohorts and the same software version (Fig. 4). Through these approaches, we were able to assess both the methodological variation and the cohort variation in morphometric studies of the IBS brain.

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 challenge this initial negative finding obtained using a robust Mahalanobis distance analysis, we implemented a machine learning approach with cross-validation to assess discriminability between IBS and HC in our cohort and reveal the most important brain structures.

Prediction of IBS versus HC from morphometric measures

A large collection of machine learning models was trained (see Appendix Fig A1), using the low-code machine learning PyCaret library in Python. We used data from the Bergen cohort described in Tab. 2, where their morphometric measures (features) were obtained with the longitudinal stream of FreeSurfer 7.4.1.

The cross-validation procedure was set up using PyCaret for a binary classification task with the target variable 'Outcome' (0: HC, 1: IBS). The dataset consists of 78 samples and 37 features, all of which are numeric. The data was split into a training set of 54 samples and a test set of 24 samples. The setup uses a stratified K-Fold cross-validation with 10 folds, ensuring that each fold maintains the same proportion of the target classes (HC and IBS) as the original dataset. Preprocessing was enabled, with numeric features imputed using the mean and categorical features imputed using the mode, not relevant for our complete dataset. Note that PyCaret performs feature normalization by default during the setup process. It scales numeric features to have zero mean and unit variance, which is essential for many machine learning algorithms to perform optimally. This normalization helps ensure that all features contribute equally to the model's performance, cfr. eTIV characteristic values ($> 1.2 \cdot 10^6$) versus the much smaller eTIV-normalized values (< 1). The session was identified by a session ID in order to be reproducible. As the dataset is rather small, in a machine learning context, the setup did not utilize GPU acceleration. This configuration ensures robust model evaluation by reducing overfitting and providing a reliable estimate of model performance across different subsets of the data.

The PyCaret provided performance metrics of the 15 classification models evaluated on the dataset (details in Fig A1). The Extreme Gradient Boosting (xgboost) model outperformed the others, achieving the highest accuracy (0.7200), AUC (0.6833), recall (0.7200), precision (0.7383), and F1 score (0.7124). It also shows the best Cohen's Kappa (0.4031) and MCC (0.4204) values, indicating strong predictive power and agreement beyond chance. The K Neighbors Classifier (knn) followed, with slightly lower performance metrics. Models like Logistic Regression (lr) and SVM showed moderate performance, while others, such as Ada Boost (ada) and Linear Discriminant Analysis (l1da), performed poorly, with accuracy close to random chance. The Dummy Classifier serves as a baseline, highlighting the effectiveness of more sophisticated models. Overall, xgboost was the most effective model for this dataset, suggesting it captures the underlying patterns well.

The confusion matrix in Fig. 9a was obtained with the best model (xgboost) on the hold out data set used for testing. It shows that the majority of the IBS patients in the test set (11/15) were correctly classified, i.e., 73% sensitivity, where $TP = 11$ (8F, 3M; IBS-SSS: 245.7 ± 60.4 ; age: 33.2 ± 7.6). On the other side, all but one HC was misclassified as IBS, i.e., 11% specificity, where $FP = 8$ (3F, 5M; IBS-SSS: 19.2 ± 19.6 ; age: 25.4 ± 5.7). The accuracy was 50% (12/24). Notable patterns include higher IBS-SSS scores in true IBS cases, female predominance among IBS patients, and older age in IBS cases compared to controls. The

model's performance is characterized by good sensitivity but poor specificity, suggesting a bias toward IBS classification.

The bar plot in Figure 9b displays the top 10 features ranked by permutation importance using XGBoost classification on the test dataset. CC_Central exhibits the highest importance (≈ 0.048), followed by WM-hypointensities (≈ 0.042) and Left-Accumbens-area (≈ 0.035). The mid-range importance features include CC_Mid_Posterior (≈ 0.033), Left-Amygdala (≈ 0.029), and Right-Cerebellum-Cortex (≈ 0.023). The lower-ranked features comprise lhCerebralWhiteMatterVol and Right-Pallidum (both ≈ 0.02), while eTIV (≈ 0.015) and Right-Hippocampus (≈ 0.013) show the lowest importance values. Notably, the importance scores suggest that central corpus callosum and white matter hypointensities are the most influential neuroanatomical features in the model's decision-making process, while hippocampal measures contribute relatively less to the classification outcomes.

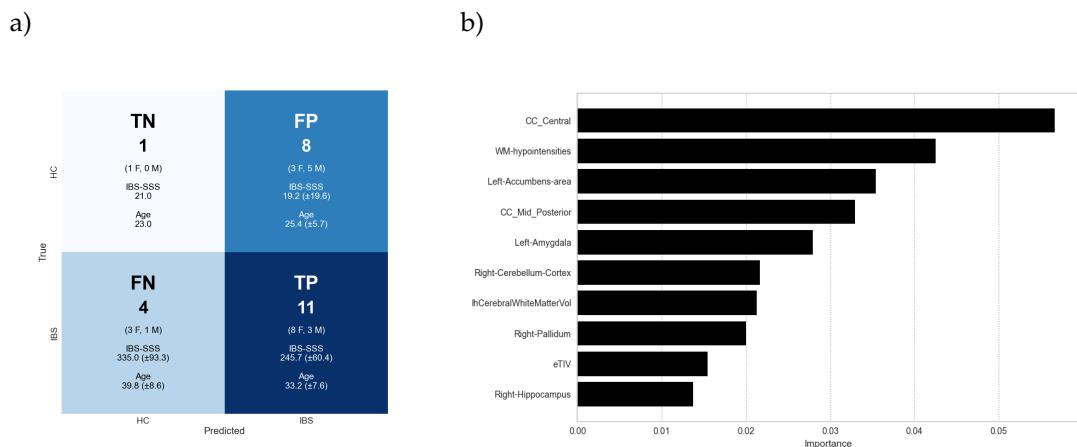


Figure 9. (a) Confusion matrix with annotated quadrants, obtained with xgboost predicting IBS versus HC from brain morphometry, only, in the test dataset. TN = true negative, FP = false positive, FN = false negative, TP = true positive. (b) Top ten morphometric features in permutation importance for predicting IBS versus HC in the test dataset, using xgboost.

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

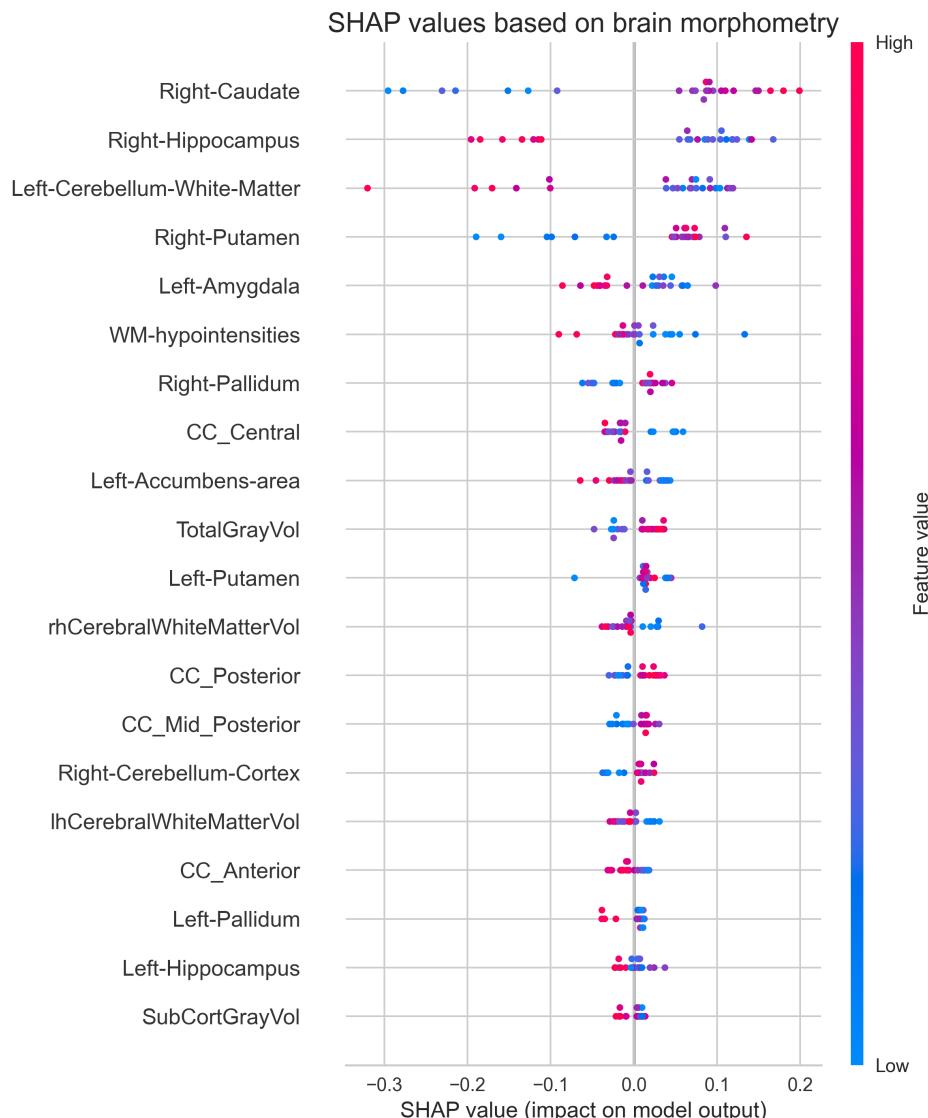


Figure 10. SHAP values with morphometry as predictors. The SHAP values (x-axis) indicate the impact each feature (y-axis) has on the model's output, which is the probability of classifying someone as IBS patient or HC. Values to the right of 0 indicate a positive contribution (towards one class, likely HCs in this case). Values to the left indicate a negative contribution (likely towards IBS). Color gradient (Feature Value): This represents the actual feature value: Red/pink indicates a high feature value. Blue/purple indicates a low feature value.

Figure 10 shows SHAP values, which measure how different brain regions impact a predictive model's output. The right-hemispheric structures (caudate, hippocampus) appear most important for whatever outcome the model is predicting, suggesting these areas might deserve particular attention in patient assessment. There is also a pattern suggesting the basal ganglia network as a whole as important for the model's predictions, with right-hemispheric structures (especially caudate) showing stronger effects than left-hemispheric ones. Other subcortical structures like the amygdala and hippocampus also show notable effects, supporting a broader subcortical involvement in predicting IBS versus HC.

Univariate analysis of the cognitive features

Analysis of RBANS scores between Healthy Controls (HC) and IBS patients, using Bonferroni-corrected Mann-Whitney U tests ($\alpha = 0.05$), reveals significant differences in cognitive performance across specific domains. Note that Table 4 shows the uncorrected p-values,

where we multiply by 5 to get the corrected values. The Fullscale RBANS demonstrates a significant difference ($p_{corrected} = 0.012$) between HC (median = 103.0, IQR = 93.0-108.0) and IBS patients (median = 91.0, IQR = 85.0-100.0), with a small to moderate effect size (Cliff's $\delta = 0.213$). Similarly, the Recall Index shows a significant difference ($p_{corrected} = 0.036$) between HC (median = 107.0, IQR = 92.0-113.0) and IBS patients (median = 95.0, IQR = 85.0-100.0), also with a small to moderate effect size (Cliff's $\delta = 0.186$). The Memory Index, while showing a small effect size (Cliff's $\delta = 0.147$), does not reach statistical significance after correction ($p_{corrected} = 0.186$). The remaining cognitive domains show minimal differences between groups: 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 that IBS patients demonstrate significantly lower overall cognitive function and recall abilities compared to healthy controls, while other cognitive domains such as visuospatial, verbal, and attention skills remain relatively preserved. The use of Bonferroni correction strengthens the robustness of these findings by controlling for multiple comparisons, though it may increase the risk of Type II errors.

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

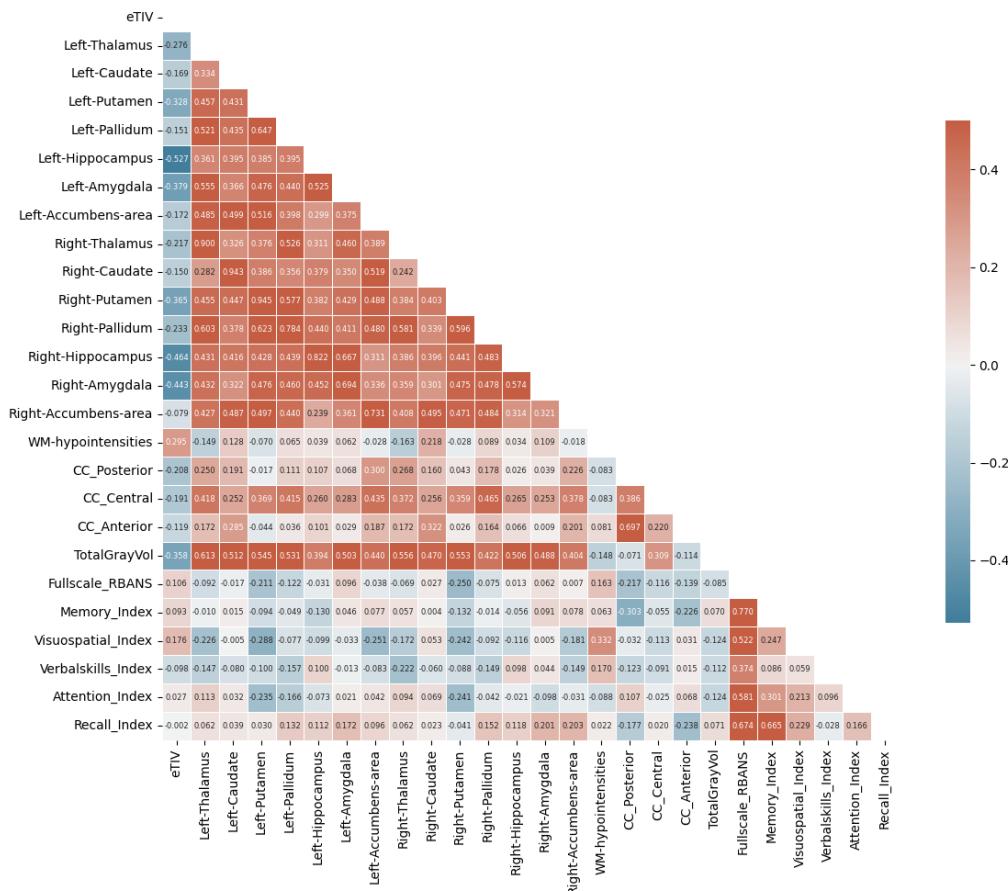
Variable	HC	IBS	p-value	Cliff's delta
Fullscale_RBANS	103.0 (93.0-108.0)	91.0 (85.0-100.0)	0.002	0.213
Memory_Index	100.0 (86.0-109.0)	86.0 (81.0-105.0)	0.031	0.147
Visuospatial_Index	97.0 (90.0-107.0)	96.0 (90.0-105.0)	0.763	0.021
Verbal skills Index	105.0 (95.0-113.0)	95.0 (89.0-111.0)	0.087	0.116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

Median (IQR) of full-scale RBANS scores, and the index scores of the five cognitive domains. p-values derived from the Mann-Whitney U test are uncorrected. Cliff's delta is used to estimate effect sizes.

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

Correlations between morphometric and cognitive features

Figure 11 shows a heatmap to illustrate the pairwise Spearman correlations between the included morphometric volume measures, the RBANS indexes, and the Fullscale RBANS score. The correlation matrix reveals several key patterns: Bilateral subcortical structures show moderate to strong correlations with their contralateral counterparts, demonstrating structural symmetry (e.g., left-right hippocampus: $\rho \approx 0.8$, bilateral amygdala: $\rho \approx 0.7$, bilateral putamen: $\rho \approx 0.9$). TotalGrayVol demonstrates moderate positive correlations with subcortical structures ($\rho \approx 0.4 - 0.6$), reflecting its composite nature. However, the correlations between morphometric measures and cognitive performance indices are notably weak. The Fullscale_RBANS shows minimal correlations with brain structures ($|\rho| < 0.25$), and Memory_Index exhibits weak correlations even with medial temporal structures ($|\rho| < 0.15$). The corpus callosum segments (anterior, central, posterior) show a large span of intercorrelations ($\rho \approx 0.2 - 0.7$) but weak associations with cognitive measures. White matter hypointensities (WM-hypointensities) demonstrate particularly weak correlations across all measures ($|\rho| < 0.15$), except for Visiospatial_Index ($\rho = 0.33$). The correlations between the full scale score and each of the five indexes are moderate to strong for all except for the Verbalskills_Index ($\rho = 0.37$). Among cognitive indices, moderate intercorrelations exist, with the strongest relationship observed between Recall_Index and Memory_Index ($\rho = 0.67$). These findings suggest that while brain structures maintain expected anatomical relationships, their associations with cognitive performance measures are more subtle and complex than might be expected from a simple structure-function relationship model.



Prediction of IBS versus HC from morphometric and cognitive measures

845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860

The confusion matrix presented in Figure 12a summarizes the performance of an XGBoost model (ranked 2nd best, after knn) predicting irritable bowel syndrome (IBS) versus healthy controls (HC) based on brain morphometry and cognitive features within the test dataset. Each cell provides detailed demographic and clinical information about the participants classified into the corresponding category.

True negatives (TN), representing correctly identified HC participants, include 2 individuals (2 females, 0 males) with an average IBS Severity Scoring System (IBS-SSS) score of 19.0 ± 2.8 and a mean age of 23.0 ± 0.0 years. False positives (FP), where HC participants were misclassified as having IBS, consist of 7 individuals (2 females, 5 males) with an IBS-SSS score of 19.6 ± 21.2 and a mean age of 25.7 ± 6.1 years. The false negative (FN) category includes one IBS participant (1 male) misclassified as HC, characterized by a high IBS-SSS score of 245.0 and an age of 47.0 years. True positives (TP), indicating IBS participants correctly identified by the model, encompass 14 individuals (11 females, 3 males) with an average IBS-SSS score of 271.3 ± 81.0 and a mean age of 34.1 ± 7.7 years.

This matrix highlights the model's effectiveness, with relatively strong performance in identifying IBS participants ($TP = 14$) compared to HC participants ($TN = 2$). However, the model exhibits a notable rate of false positives ($FP = 7$), indicating misclassification of HC participants as IBS, which could be related to overlapping features between the two groups. The lone false negative suggests a potential challenge in identifying older IBS participants with distinct clinical characteristics, as evidenced by the high IBS-SSS score and advanced age compared to the TP group. This result underscores the complexity of differentiating IBS from HC based on these features and may indicate areas for further optimization of the model, or access to larger cohorts.

To better assess the difference in `xgboost` performance using the morphometric features (M), only and using the full dataset, including cognition (C), we have produced Table 5.

Table 5. Performance Metrics from Confusion Matrices using `xgboost` with different feature sets.

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 features. C: cognitive features. TPR: sensitivity; TNR: specificity; PPV: precision; ACC: accuracy. 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>

From Tab. 5 we see that using M alone, the model showed moderate sensitivity (0.733) but struggled with specificity (0.111), resulting in poor overall classification quality ($MCC = -0.185$). In contrast, combining M and cognitive features ($M \cup C$) substantially improved sensitivity (0.933), specificity (0.222), and overall classification performance ($MCC = 0.228$), demonstrating that integrating cognitive features enhances the model's ability to differentiate IBS from healthy controls.

Figure 12b illustrates how different brain measurements and cognitive abilities influence our model's predictions. Each row shows a different measurement, with the most influential features at the top. The dots represent individual participants, with red dots indicating higher values and blue dots indicating lower values for each measurement. The position of each dot shows whether that particular value pushes the model's prediction higher (towards the right) or lower (towards the left). The Right Hippocampus, shown at the top, has the strongest overall impact: when some people have high values (red dots), this strongly pushes predictions in one direction, while low values (blue dots) push predictions in the other direction. Verbal skills also show important patterns, with a wide spread across different participants, suggesting that both high and low verbal skills can be important indicators, depending on the individual case. Among subcortical regions, Right Caudate and Right Putamen demonstrate particularly notable influences. These patterns reveal meaningful differences between our two groups in both brain structure and cognitive abilities. Particularly important are the variations we see in verbal abilities and in memory-related brain regions like the hippocampus, which could inform how we approach patient care and communication strategies.

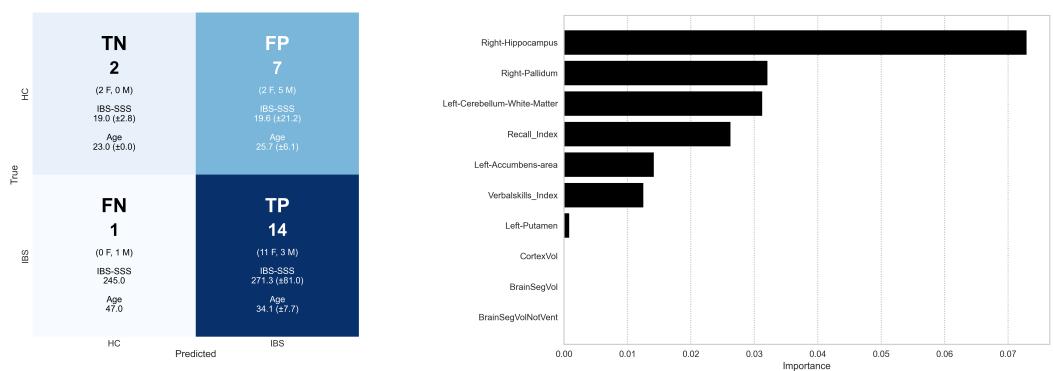


Figure 12. (a) Confusion matrix with annotated quadrants, obtained with xgboost predicting IBS versus HC, jointly from brain morphometry and cognition, in the test dataset. TN = true negative, FP = false positive, FN = false negative, TP = true positive. (b) Top ten morphometric features in permutation importance for predicting IBS versus HC in the test dataset, using xgboost.

Generated by: https://github.com/arvid1/ibs-brain/blob/main/notebooks/07_predicting-IBS-vs-HC-from-morphometry-and-cognition.ipynb

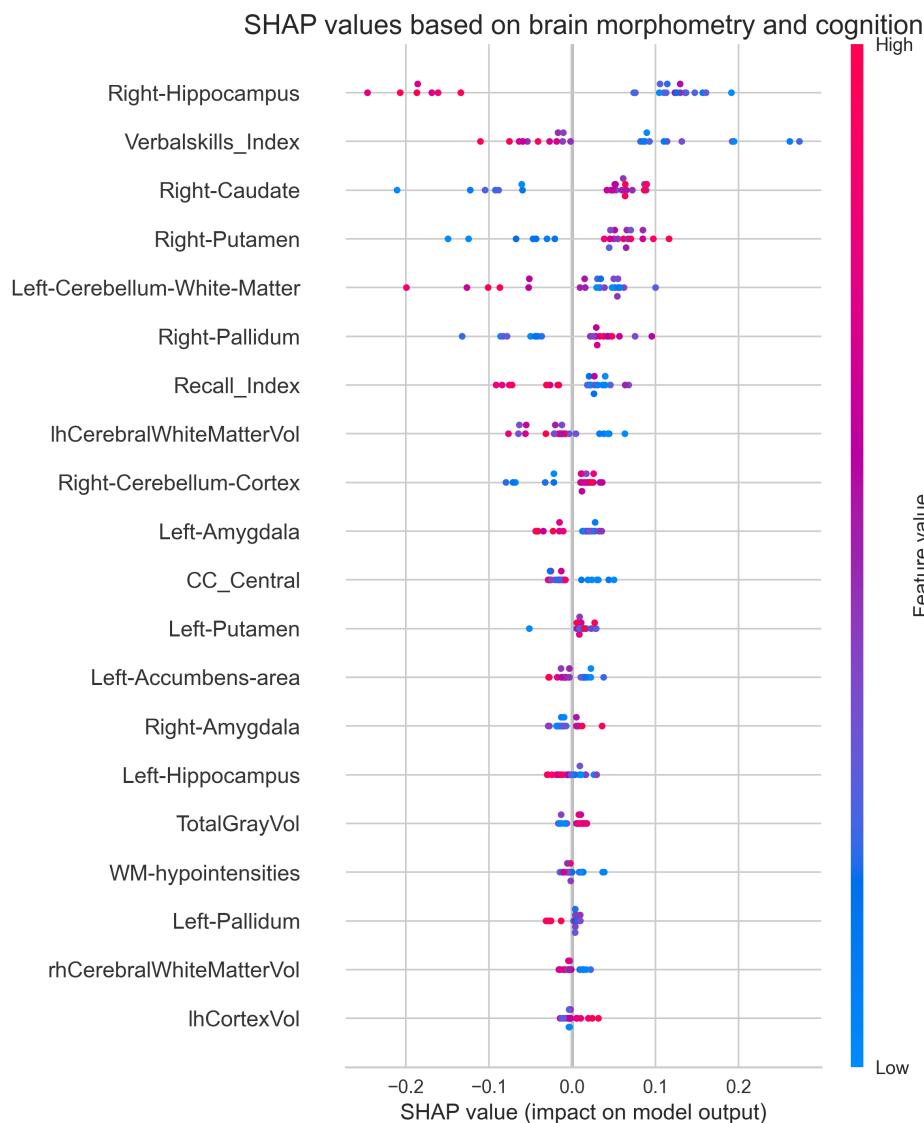


Figure 13. SHAP values with morphometry and cognition as predictors.

Note: X-axis (SHAP values): The SHAP values indicate the impact each feature has on the model's output, which is the probability of classifying someone as IBS patient or HC.

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

Discussion

Our study yields two key methodological insights and one substantial findings regarding brain structure and function in IBS. First, we were unable to replicate the morphometric differences between the IBS and healthy control group reported by Skrobisz et al. [22], regardless of whether we used FreeSurfer version 6.0.1 or 7.4.1. Second, we observed substantial discrepancies in morphometric measurements between these software versions, highlighting the critical importance of considering methodological factors in neuroimaging research. Application of advanced multivariate and machine learning techniques to investigate brain-behavior relationships in IBS is another primary contribution. While morphometric features alone proved insufficient for reliable group discrimination, the integration of cognitive performance measures with brain morphometry substantially improved classification accuracy. Specifically, our analyses revealed that two cognitive domain indices, combined with volumetric measures of subcortical structures—particularly the hippocampus and basal ganglia—provided robust discrimination between IBS patients and

healthy controls. The consistency of these findings across different approaches to feature importance analyses strengthens their validity and suggests a fundamental relationship between brain structure, cognitive function, and IBS symptomatology. This observation aligns with views of IBS as a disorder involving complex interactions between central nervous system function and gastrointestinal symptoms, rather than purely peripheral manifestations.

Brain Structures involved in discriminating between IBS and HC

Our results showed that subcortical structures, particularly within the basal ganglia, played a key role in distinguishing IBS patients from healthy controls. While traditionally associated with motor control, the basal ganglia also critically influence reward processing, habit formation, and pain modulation - functions directly relevant to IBS symptomatology, and the impact on patients' experience of gastrointestinal symptoms. These findings align with recent results from a UK Biobank study [19], which also highlighted the importance of hippocampal and basal ganglia structures, including the Pallidum and Caudate, in IBS. Beyond the basal ganglia, several other subcortical structures relevant to IBS symptomatology emerged as discriminators. The nucleus accumbens, fundamental to reward processing and motivation, may mediate the emotional and motivational aspects of chronic pain in IBS. Dysfunction in this structure could explain the intensified emotional distress and pain sensitivity commonly reported by IBS patients [9]. Similarly, the amygdala appears significant, particularly given its connection to pain-modulation and emotion-processing networks, including the prefrontal cortex and insula. This aligns with previous research [34] demonstrating enhanced amygdala-insula connectivity in IBS patients. Although our results differ from Skrobisz et al.'s [22] findings regarding thalamic involvement, other studies have supported its role in IBS. Diffusion tensor imaging has revealed altered thalamic organization in IBS patients, with reduced fractional anisotropy and increased mean diffusivity [35]. These alterations suggest compromised structural integrity of thalamic circuits, potentially affecting pain processing and sensory integration. The involvement of corpus callosum should also be mentioned, as interhemispheric integration is crucial for visceral sensation processing, pain modulation [36] as well as in mental disorders [37]. Taken together, our findings support that integrated neural signatures are involved in predicting IBS [38].

Integration of Cognitive Performance and Brain Structure in IBS

The enhanced diagnostic accuracy by including cognitive measures strongly support that IBS should be understood as a disorder of the gut-brain interaction [14,39]. The brain's integral role in cognitive, emotional, and autonomic regulation suggests that these manifestations are fundamentally interconnected rather than merely coincidental. The prominent role of hippocampal volume was a principal finding. The fundamental role of Hippocampus in cognitive processing is well known [40], and was supported by the Recall index being identified as another feature with strong importance. The role of Verbal skills was more surprising. Although research has established connections between memory systems and language processing, particularly in semantic memory organization [41], a negligible correlation between the two indices suggests that IBS affects multiple cognitive domains through independent mechanisms.

Our findings may also have implications for other somatic and psychiatric disorders, like Alzheimer's disease, Parkinson's disease, and major depression. The gut-brain axis are involved in all these diseases, which also are characterized by cognitive impairment. Recent research has identified potential pathways linking gut microbiota alterations to neurological function, particularly through inflammatory responses and tryptophan metabolism [42,43]. The emergence of the microbiota-gut-brain axis as a key framework [44] offers new perspectives on how peripheral inflammation might influence both brain structure

and cognitive function in IBS. This integrated view suggests that cognitive assessment, combined with brain morphometry, might provide valuable insights not only for IBS but for a broader spectrum of gut-brain disorders.

Brain-Gut Axis: Implications for Understanding and Treating IBS

Our findings should have important implications for clinical practice and treatment development. The observed relationship between brain structure, cognitive function, and IBS symptomatology suggests that effective interventions should target multiple domains simultaneously. Such a multifaceted approach recognizes IBS as a complex disorder requiring coordinated intervention across multiple domains.

Future research directions should expand upon these findings through multimodal investigation. Integration of functional neuroimaging, gut microbiome analysis, and broader clinical assessment [19] could provide a more comprehensive understanding of IBS pathophysiology. Particularly crucial will be longitudinal studies to determine the temporal relationship between brain changes and symptom development. Such studies would allow us to track the evolution of cognitive and structural alterations over time, identify early markers of disease progression, and evaluate the impact of various therapeutic interventions. This temporal perspective is essential for understanding whether observed brain changes represent cause or consequence of IBS symptoms.

This comprehensive approach to understanding IBS aligns with the emerging paradigm of precision medicine. By considering the full spectrum of biological, cognitive, and behavioral manifestations, we may better identify patient subgroups and develop more personalized treatment strategies. The integration of brain structure, cognitive function, and clinical symptoms represents a promising framework for advancing both our understanding and treatment of this complex disorder. Ultimately, this integrated perspective may lead to more effective, personalized interventions that address the full range of IBS manifestations.

Strengths and Limitations: Critical Evaluation and Future Directions

Although contributing through its multimodal analytical approach, several limitations warrant discussion. The primary limitation concerns sample size, which particularly constrained our ability to conduct robust sex/gender-based analyses. This limitation is especially noteworthy given the evidence for substantial sex/gender differences in IBS presentation, progression, and treatment response [45]. The importance of sex/gender considerations in IBS research has become increasingly apparent. Clinical presentations show clear sex-based patterns, with IBS-C predominating in women and IBS-D in men [46]. These differences reflect complex interactions between biological and environmental factors. Sex hormones, particularly estrogen and progesterone, influence both gastrointestinal function and pain processing in the central nervous system [47]. Recent research has revealed sex-based differences extending to gut microbiota composition [48] and sensory processing. Notably, Labus et al. [20] demonstrated enhanced sensory sensitivity in women with IBS, potentially related to sex-specific morphometric variations in brain structure.

An inability to fully account for IBS symptom severity in our analyses was another limitation. Recent work by Li et al. [19] has demonstrated that symptom severity correlates significantly with both cognitive performance and brain volumetric measures, particularly in regions associated with emotional processing and cognitive control. This finding suggests that future studies should incorporate detailed severity measures to better understand the relationship between symptom intensity and brain-behavior patterns.

The present study's methodological strengths lie in its comprehensive multivariate approach to analyzing brain-behavior relationships. This approach better captures the complex interactions between multiple brain regions and cognitive measures, providing a more nuanced understanding than traditional univariate analyses. However, we acknowledge

important limitations. While our sample exceeds that of many comparable studies, multivariate analyses and machine learning approaches typically benefit from larger datasets. To address this limitation, we implemented robust cross-validation procedures, including 10-fold validation and hold-out test sets, thereby reducing the risk of overfitting and enhancing the generalizability of our findings. Moreover, the results point to several important directions for future research. First, larger-scale studies are needed to validate and extend our multivariate findings. Such studies should maintain rigorous methodological standards while increasing statistical power. Second, standardization of neuroimaging analysis protocols, including careful documentation of software versions and processing parameters, is crucial for reproducibility. Third, the field would benefit from systematic investigation of how different analysis approaches might influence morphometric findings in IBS research. Overall, future studies should consider implementing standardized protocols for both imaging and cognitive assessment, facilitating meta-analyses and enabling more direct comparisons across studies. This standardization, combined with transparent reporting of methodological details, would strengthen the field's ability to build cumulative knowledge about brain-gut interactions in IBS. Longitudinal studies represent a particularly important future direction. Such studies could address crucial questions about the temporal dynamics of brain-gut interactions in IBS, including whether observed structural and cognitive changes precede or follow symptom development. Longitudinal data would also enable better prediction of disease trajectories and treatment responses, potentially informing personalized interventions such as dietary modifications (e.g., Low FODMAP diet) or targeted cognitive interventions. The combination of longitudinal design with multimodal assessment (including brain structure, cognitive function, and clinical symptoms) could provide unprecedented insights into the development and progression of IBS.

Conclusions and Future Directions

The present study advances our understanding of brain-gut interactions in IBS through several key contributions. First, our comprehensive multivariate analyses reveal the inherent complexity of IBS pathophysiology, demonstrating that single-modality approaches may be insufficient for characterizing this multifaceted disorder. While we did not replicate previously reported volumetric differences in thalamic structure, our machine learning analyses uncovered more subtle and complex patterns of brain-behavior relationships. Particularly noteworthy was the finding that morphometric features gain discriminative power when integrated with cognitive measures, especially in subcortical regions including the hippocampus and basal ganglia. These results strongly support a systems-level conceptualization of IBS, where the condition emerges from complex interactions between neural structure, cognitive function, and gastrointestinal symptoms. This perspective suggests that effective characterization and treatment of IBS requires consideration of multiple biological and cognitive markers rather than focusing on isolated symptoms or structures. The successful integration of structural and functional measures in our analyses points toward more sophisticated approaches for both diagnosis and treatment planning. Moving forward, several research priorities emerge from our findings. Large-scale validation studies are needed to confirm the reliability and generalizability of our brain-cognition relationships across diverse patient populations. Such studies should incorporate standardized protocols for both imaging and cognitive assessment to facilitate cross-study comparisons. Longitudinal investigations are particularly crucial for understanding how these markers evolve over time and relate to treatment response. Additionally, future research should explore how individual differences in brain structure and cognitive function might predict treatment outcomes, potentially enabling more personalized therapeutic approaches. Ultimately, our findings suggest that advancing IBS treatment may require a fundamental shift toward integrated, multimodal assessment approaches that capture both structural and functional aspects of brain-gut interactions. This more comprehensive understanding of IBS pathophysiology could lead to more effective, personalized interven-

tions that address the full spectrum of patient symptoms and experiences.

1067

1068

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>.
1069
1070
1071
1072
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>.
1073
1074
1075
1076
1077
1078
1079
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>.
1075
1076
1077
1078
1079
1080
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>.
1080
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>.
1081
1082
1083
1084
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>.
1085
1086
1087
1088
1089
7. Kim, Y.S.; Kim, N. Sex-gender differences in irritable bowel syndrome. *Journal of Neurogastroenterology and Motility* **2018**, *24*, 544.
1089
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>.
1090
1091
1092
1093
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>.
1094
1095
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>.
1096
1097
1098
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>.
1099
1100
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>.
1101
1102
1103
1104
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>.
1105
1106
1107
1108
1109
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>.
1110
1111
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>.
1112
1113
1114
16. Lezak, M.D. *Neuropsychological assessment*; Oxford University Press, USA, 2004.
1115
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>.
1116
1117
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>.
1118
1119
1120
19. 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>.
1121
1122
1123
1124
20. 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>.
1125
1126
1127

21. 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. [1128](https://doi.org/10.3390/brainimaging2020140901)
22. 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. [1129](https://doi.org/10.3390/diagnostics12092199)
23. 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. [1130](https://doi.org/10.1097/MD.00000000000021950)
24. 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. [1131](https://doi.org/10.1046/j.1365-2036.1997.142318000.x)
25. Randolph, C. *Repeatable battery for the assessment of neuropsychological status. Norwegian manual*; NL:Pearson, 2013. [1132](#)
26. Fischl, B. FreeSurfer. *Neuroimage* **2012**, *62*, 774–781. [1133](https://doi.org/10.1016/j.neuroimage.2012.01.021)
27. 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. [1134](https://doi.org/https://doi.org/10.1002/hbm.20599)
28. 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. [1135](https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.02.010)
29. 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. [1136](https://doi.org/https://doi.org/10.1371/journal.pone.0038234)
30. 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. [1137](https://doi.org/https://doi.org/10.3389/fninf.2015.00012)
31. 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 MPAGE and MP2RAGE brain MRI images. *Neuroimage: Reports* **2022**, *2*, 100086. [1138](https://doi.org/10.1016/j.ynrirp.2022.100086)
32. 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. [1139](https://doi.org/https://doi.org/10.1016/j.cmpb.2023.107825)
33. 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. [1140](https://doi.org/10.1038/s42256-019-0138-9)
34. Qi, R.; Liu, C.; Ke, J.; Xu, Q.; Ye, Y.; Jia, L.; Wang, F.; Zhang, L.; Lu, G. Abnormal amygdala resting-state functional connectivity in irritable bowel syndrome. *American Journal of Neuroradiology* **2016**, *37*, 1139–1145. [1141](#)
35. Ellingson, B.M.; Mayer, E.; Harris, R.J.; Ashe-McNally, C.; Naliboff, B.D.; Labus, J.S.; Tillisch, K. Diffusion tensor imaging detects microstructural reorganization in the brain associated with chronic irritable bowel syndrome. *PAIN®* **2013**, *154*, 1528–1541. [1142](https://doi.org/10.1016/j.pain.2013.04.010)
36. Ito, A.; Yang, S.; Shinto, E.; Shinto, A.; Toyofuku, A.; Kurata, J. Interhemispheric and Corticothalamic White-Matter Dysfunction Underlies Affective Morbidity and Impaired Pain Modulation in Chronic Pain. *Anesthesia & Analgesia* **2022**, pp. 10–1213. [1143](https://doi.org/10.1213/ANE.00000000006992)
37. Piras, F.; Vecchio, D.; Kurth, F.; Piras, F.; Banaj, N.; Ciullo, V.; Luders, E.; Spalletta, G. Corpus callosum morphology in major mental disorders: a magnetic resonance imaging study. *Brain Communications* **2021**, *3*, fcab100. [1144](https://doi.org/10.1093/braincomms/fcab100)

38. Bhatt, R.R.; Gupta, A.; Labus, J.S.; Liu, C.; Vora, P.P.; Stains, J.; Naliboff, B.D.; Mayer, E.A. A neuropsychosocial signature predicts longitudinal symptom changes in women with irritable bowel syndrome. *Molecular Psychiatry* **2022**, *27*, 1774–1791. <https://doi.org/10.1038/s41380-021-01375-9>. 1186
1187
1188
1189
39. Zhao, M.; Hao, Z.; Li, M.; Xi, H.; Hu, S.; Wen, J.; Gao, Y.; Antwi, C.O.; Jia, X.; Yu, Y.; et al. Functional changes of default mode network and structural alterations of gray matter in patients with irritable bowel syndrome: a meta-analysis of whole-brain studies. *Frontiers in Neuroscience* **2023**, *17*, 1236069. <https://doi.org/10.3389/fnins.2023.1236069>. 1190
1191
1192
1193
40. Lisman, J.; Buzsáki, G.; Eichenbaum, H.; Nadel, L.; Ranganath, C.; Redish, A.D. Viewpoints: how the hippocampus contributes to memory, navigation and cognition. *Nature Neuroscience* **2017**, *20*, 1434–1447. 1194
1195
1196
41. Burgess, N.; Maguire, E.A.; O’Keefe, J. The human hippocampus and spatial and episodic memory. *Neuron* **2002**, *35*, 625–641. [https://doi.org/10.1016/s0896-6273\(02\)00830-9](https://doi.org/10.1016/s0896-6273(02)00830-9). 1197
1198
42. Sibelli, A.; Chalder, T.; Everitt, H.; Workman, P.; Windgassen, S.; Moss-Morris, R. A systematic review with meta-analysis of the role of anxiety and depression in irritable bowel syndrome onset. *Psychological Medicine* **2016**, *46*, 3065–3080. <https://doi.org/10.1017/S0033291716001987>. 1199
1200
1201
43. Carloni, S.; Rescigno, M. The gut-brain vascular axis in neuroinflammation. In Proceedings of the Seminars in Immunology. Elsevier, 2023, Vol. 69, p. 101802. <https://doi.org/10.1016/j.smim.2023.101802>. 1202
1203
1204
44. Ishioh, M.; Nozu, T.; Okumura, T. Brain Neuropeptides, Neuroinflammation, and Irritable Bowel Syndrome. *Digestion* **2024**, *105*, 34–39. <https://doi.org/10.1159/000533275>. 1205
1206
45. van Kessel, L.; Teunissen, D.; Lagro-Janssen, T. Sex-gender differences in the effectiveness of treatment of irritable bowel syndrome: A systematic review. *International Journal of General Medicine*, pp. 867–884. <https://doi.org/10.2147/IJGM.S291964>. 1207
1208
1209
46. Lee, O.Y.; Mayer, E.A.; Schmulson, M.; Chang, L.; Naliboff, B. Gender-related differences in IBS symptoms. *Official journal of the American College of Gastroenterology | ACG* **2001**, *96*, 2184–2193. 1210
1211
47. Chang, L.; Heitkemper, M.M. Gender differences in irritable bowel syndrome. *Gastroenterology* **2002**, *123*, 1686–1701. <https://doi.org/10.1053/gast.2002.36603>. 1212
1213
48. Vemuri, R.; Sylvia, K.E.; Klein, S.L.; Forster, S.C.; Plebanski, M.; Eri, R.; Flanagan, K.L. The microgenderome revealed: sex differences in bidirectional interactions between the microbiota, hormones, immunity and disease susceptibility. In Proceedings of the Seminars in Immunopathology. Springer, 2019, Vol. 41, pp. 265–275. 1214
1215
1216
1217
49. 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>. 1218
1219
1220
1221

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.”. 1222
1223
1224
1225
1226

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. 1227
1228
1229

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). 1230
1231
1232
1233
1234

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. 1235
1236

Data Availability Statement: The implementation of the complete workflow, the setup of the corresponding conda environment, the cleaned input dataset in .csv format, and code for all tables and figures in the Results section are available as *Jupyter notebooks* at <https://arvidl.github.com/ibs-brain>. 1237
1238
1239
1240

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	1242
CM	Confusion matrix	1243
Cohen's d	effect size	1244
Cliff's delta	effect size	
DGBI	Disorders of the gut-brain interaction	
FS	Freesurfer	
GI	Gastrointestinal	
GitHub	Meeting platform for collaboration	
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	Operating Characteristic	
SHAP	SHapley Additive exPlanations	1248
SD	Standard deviation	
SHAP	SHapley Additive exPlanations	
XGBoost	eXtreme Gradient Boosting	
Performance	measures	
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

1249

Appendix A.1 FreeSurfer segmented brain regions obtained from aseg

1250

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. <i>Thalamus</i> 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. <i>Nucleus caudatus</i> is involved in motor and motility control and learning
Left-Putamen	Left putamen. <i>Putamen</i> is part of the basal ganglia involved in motor control and may contribute to habitual responses to gastrointestinal discomfort
Left-Pallidum	Left globus pallidus. <i>Globus pallidus</i> is involved in regulating voluntary movement and gut motility patterns
Left-Hippocampus	Left hippocampus. <i>Hippocampus</i> is crucial for memory formation and spatial navigation, and in IBS, involved in contextual fear learning related to gastrointestinal symptoms
Left-Amygdala	Left amygdala. <i>Amygdala</i> is involved in processing emotions, fear, and anxiety
Left-Accumbens-area	Left nucleus accumbens. <i>Nucleus accumbens</i> 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 Multinormality 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, \dots, 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 $HZ_n = n(1 + 2\beta^2)^{p/2} [D_n - (1 + \beta^2)^{-p/2}]$, where $\beta = \frac{1}{\sqrt{2}}$ is the smoothing parameter and $D_n = \frac{1}{n} \sum_{i=1}^n \exp(-\frac{\beta^2}{2} d_i^2)$, 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: $\mu = -\frac{1}{2} \log(1 + 2\beta^2) - \frac{p}{2} \log(1 + \beta^2) + \log\left(1 + \frac{p\beta^4}{2(1+2\beta^2)}\right)$ $\sigma^2 = 2\left[-\log\left(1 - \frac{2\beta^4}{(1+2\beta^2)^2}\right) + \frac{p\beta^4}{(1+2\beta^2)(1+\beta^2)}\right]$ 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: $\hat{\mu}_{robust} = \text{median}(X_{win})$. The pooled covariance matrix is then computed using the winsorized data as $\hat{\Sigma}_{pooled} = \frac{(n_{HC}-1)\hat{\Sigma}_{HC}+(n_{IBS}-1)\hat{\Sigma}_{IBS}}{n_{HC}+n_{IBS}-2}$. 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: $F = \frac{n_{HC}n_{IBS}}{(n_{HC}+n_{IBS})(n_{HC}+n_{IBS}-2)p} D_{robust}^2$. 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 $p\text{-value} = 1 - F_{p,n_{HC}+n_{IBS}-p-1}(F)$. 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.

1300
1301
1302
1303
1304

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

1305
1306
1307
1308
1309
1310
1311
1312

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. [22]) 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

1313
1314
1315
1316
1317
1318
1319
1320

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

Brain Region	FS 7.4.1 cross-sectional				FS 7.4.1 longitudinal stream			
	HC (N=29)		IBS (N=49)		HC (N=29)		IBS (N=49)	
	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	0.0002402	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

1321

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 0 (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

1322

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 [49]. 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

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

grouping.

1338

1339

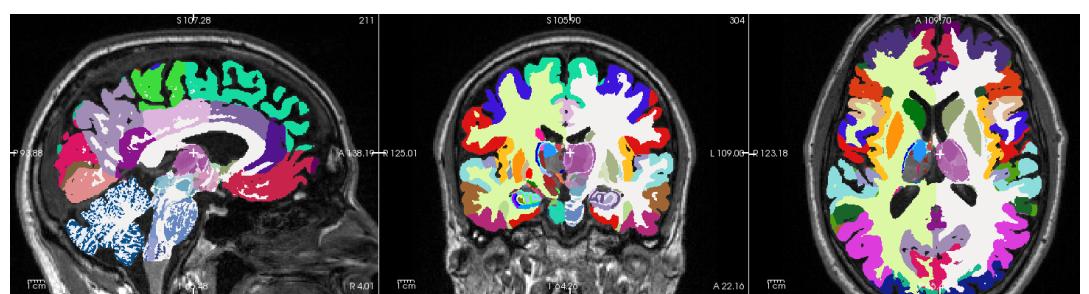


Figure A2. High resolution Histological atlas segmentation, available in FreeSurfer 8.0.0-beta, of 3D T1-weighted MPRAGE recording from BGA_046. Panels left to right: Sagittal, Coronal, and Axial section, respectively. The white cross-bar in the middle of the brain is located in the *paracentral nucleus* 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>