Choice of processing pipelines for T1-weighted brain MRI impacts association and prediction analyses.

Elise Delzant, Olivier Colliot, Baptiste Couvy-Duchesne

Table des matières

Title:

Choice of processing pipelines for T1-weighted brain MRI impacts association and prediction analyses

Authors

Elise Delzant[1], Olivier Colliot[1], Baptiste Couvy-Duchesne[1,2]

Affiliation

[1]Sorbonne Université, Institut du Cerveau – Paris Brain Institute, CNRS, Inria, Inserm, AP-HP , Hôpital de la Pitié-Salpêtrière, Paris, France

[2]Institute for Molecular Bioscience, the University of Queensland, St Lucia, QLD, Australia

Acknowledgements

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program (reference ANR-19-P3IA-0001, project PRAIRIE 3IA Institute and reference ANR-10-IAIHU-06, project Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6, from the European Union's Horizon Europe Framework Programme (grant number 101136607, project CLARA) and by Inria in the context of the Inria-University of Queensland international team.The research is also supported by a CJ Martin fellowship (NHMRC 1161356).

Abstract

The vast amount of data from the UK Biobank, offers an unprecedented opportunity to improve robustness and reproducibility in neuroimaging. In particular, little is known about the impact of MRI processing pipelines on neuroimaging results (robustness to processing).

Using 39,655 T1w brain MRI images from the UK Biobank, we compared five commonly used high-dimensional grey-matter representations of the brain, based on three major neuroimaging software: FSL (volume-based representation), CAT12/SPM (both surface and volume-based representation) and FreeSurfer (cortical and subcortical surface-based representation). We evaluated the impact of the choice of the pipeline on morphometricity estimates (defined as the percentage of trait variance captured by all brain features), sensitivity to confounders, false positive rates, detection of robust associations and prediction analysis, across 29 traits (e.g Maternal smoking, Diabetes, Alzheimer's disease).

We observed that most processing pipelines resulted in some non-normal data distribution in vertex/voxel measurements, particularly FSLANAT and CAT12 Surface. In addition, all processing were highly sensitive to imaging confounders (e.g., head motion, signal to noise ratio and brain position in the scanner), which should be taken into account whenever possible. We found that FSL and FreeSurfer maximized morphometricity estimates across all traits considered. However, a fraction of the signal they captured was unique, highlighting that we should expect inconsistency in the brain regions detected using different processing, which we confirmed in a vertex/voxel-wise analysis. Overall, volume-based representations of the grey matter detected more significant clusters with a higher replication rate compared with the surface-based processing. Finally, all three volume-based processing also yielded more performant linear prediction. Clusters comprising a single voxel/vertex exhibited lower replication rate, indicating they should be taken with a grain of salt. Overall, FSLVBM emerged as a top- performing all-rounder, as it maximized morphometricity, replicability and predictive accuracy.

We extensively compared 5 commonly used representations of the grey matter, which included 29 traits and several types of analyses (association, prediction). We quantified the strengths and limitations of the different pipelines, which can help researchers make more informed choices to select the MRI processing best suited to their research question. Our findings can serve as benchmark for other processing that may capture more information and enhance robustness of findings. Our results also suggest that multi-processing analysis (e.g.

ensemble learning) could maximize brain-based prediction by leveraging the unique signal contained in each process.

Introduction

Brain MRI (Magnetic Resonance Image) can provide direct insights into the brain structures, regions or functions that are associated with behaviour but also with disorders of the brain. Additionally, the emergence of large datasets becoming available allows us to perform analysis at an unprecedented scale. These datasets could help tackle the reproducibility crisis in the neuroimaging field, as it has been partly attributed to the small sample sizes[1]. For example, the UK Biobank has released MRI images from over 40,000+ volunteers, together with detailed information about chronic disorders, lifestyle and behaviour. Beyond new knowledge about the brain, the UK Biobank can help us progress towards more robust methodology and in particular understand how analytic choices may influence results.

T1w brain MRI images provide a detailed mapping of the grey matter structure and can be processed using different imaging software. There are two main representations of the grey-matter: the first one is volume-based (also called Voxel-Based-Morphometry pipeline, -VBM[2]), which quantifies the grey-matter density for each voxel (3d pixel). In comparison, surface-based approaches project a mesh over the grey matter to examine[3] cortical thickness, surface area or volume at the vertex level. Despite their widespread utilization and long-term use, there are no guidelines to choose from the imaging pipelines, leading to a lack of robustness in the results. Robustness to processing methods can be defined as the ability to identify consistent findings across variations in methods (for a unique dataset)[4]. Robustness complements reproducibility (ability to identify consistent findings with the same method and data) and replicability (ability to identify consistent findings across datasets, using the same method). Previous work has highlighted the variability induced by different surface-based processing pipelines[5] on structural MRI: indeed, by examining sex-differences and age-related changes, they showed that processing identified considerably different regions, as well as low similarity between processed grey matter brain maps. Similarly, when comparing VBM pipelines, a previous study showed that all pipelines

resulted in slightly different brain measurements for the same individual and that the choice of image processing impacted age prediction and sex classification[6]. It is therefore important to determine and quantify how the choice of processing software affects neuroimaging results to ultimately improve robustness to processing methods. Notably, the challenge of robustness is not unique to T1w processing but is also a common concern in other neuroimaging modalities[4], such as functional neuroimaging[7].

The multiverse approach has been proposed to overcome the robustness issue[7]: a single analyst or collaborative teams perform analyses with the same data but different pipelines, and the results are compared or merged. However, applying this method to large datasets (e.g. the UK Biobank) is computationally demanding and would entail so much image processing that the multiverse cannot be extensively explored. In addition, combining analyses performed on volumes and surfaces is challenging, as there is no simple way to overlay the association maps. Instead of reconciling all the branches of the multiverse, another option may be to compare the branches to identify the best one(s) or prune the least efficient ones. Recently, Furtjes et al. used morphometricity to compare different atlas-based representations of grey-matter structure in the UK Biobank. They showed that more detailed atlases captured more information (larger morphometricity)[8]. Similarly, morphometricity has been used to compare the amount of information captured by different cortical mesh and varying levels of smoothing, in surface based processing [9].

If morphometricity can help benchmark multiverse branches to identify the MRI processing that retains the most information, other criteria are to be considered, depending on the study objectives. For example, Brain-Wide Association Study (BWAS) aims to assess the relationship between a particular trait and each brain measurement, and its power to detect associations depends on the morphometricity as well as other factors such as sample size[1], multiple testing correction methods[10] or trait complexity (i.e. how many brain regions contribute to the morphometricity). Similarly, brain-based prediction is also influenced by the training sample size and the number of brain features, which warrants for more detailed evaluation.

We sought to benchmark some of the main branches of the multiverse of T1w MRI processing, to evaluate the robustness of findings across processing. We conducted analyses utilizing five standard and commonly used high-dimensional representations (voxel-based or vertices-based) of the grey matter. These T1w processings correspond to the 3 most common neuroimaging software for structural MRI, with default options and off-the-shelf implementations. We compared the grey-matter representations in terms of morphometricity, but we also

examined their sensitivity to imaging and sample confounders, their false positive rate, as well as their ability to detect replicable associations in brain-wide association studies or to develop brain-based predictors. We considered 29 different traits of interest to evaluate how results generalize across traits and diseases, and we used 39,655 T1w images from the UK-Biobank. Our results should help researchers make informed decisions about MRI processing options to study grey matter structures, which might depend on their trait of interest or study objectives. Our results may also guide the exploration of the multiverse by revealing the most promising branches.

Material and Methods

UK Biobank data

We considered all participants from the UK Biobank who had undergone a brain MRI exam at the time of data extraction. Of note, the UK Biobank is a large-scale database that comprises more than 500,000 volunteers across the United Kingdom and it is not a representative sample of the United Kingdom population, with an over-representation of healthy and educated individuals [11].

We excluded from our analysis participants labelled as 'unusable' [12] by the UK Biobank (due to low-quality MRI images) and the one who opted out of the study. We selected individuals with a usable T1w structural brain image (Data Field 20252) and available T1 surface files (Data Field 20263), which served as input of the processing (**Figure 1**). We excluded participants whose T2-FLAIR was not used (because deemed unusable) in the FreeSurfer processing conducted by the UK Biobank.

Our final sample comprised 39,826 participants, which we split into a discovery and a replication sample based on their assessment center. The discovery sample consisted of participants from the first center (Cheadle, greater Manchester): 23,288 adults, aged 63.1 on average (SD=7.5) with 52% women. The replication sample consisted of 16,538 adults, aged 62.2 on average (SD=7.7) with 54% women collected across the three other imaging centers (Reading, Newcastle and Bristol).

Informed consent was obtained from all UK Biobank participants. Procedures are controlled by a dedicated Ethics and Guidance Council (http://www.ukbiobank.ac.uk/ethics). IRB approval was also obtained from the

North West Multi-Centre Research Ethics Committee. This research has been conducted using the UK Biobank Resource under Application Number 53185.

Phenotypes of interest in the UK Biobank

Traits of interest

We focused on 29 variables of interest, that have a known or hypothesized link with grey-matter structure.

We considered maternal smoking around birth (Data Field ID 1787), which demonstrated large morphometricity from surface based processing[9]. We also considered several traits relating to fertility and sexual behaviours, which have been widely studied in biology and socials science, and are intimately linked to feelings, behavior and well-being: for example, Grinde et al. discussed the idea that sexual activity is not solely related to procreation, but is also correlated to human feelings and well-being[13]. We considered age first had sexual intercourse (Data Field 2139) age at first live birth (Data Field 2754) and number of children (number children fathered reported by men, Data Field ID 2405; number of live births reported by women, Data Field 2734), to investigate how sexual behaviors modulates the grey-matter (behavioral approach, dependence symptom, impulsivity). For example, previous studies have demonstrated that compulsive sexual behavior disorder is associated with reduced gray matter volume[14]. Similar findings have been observed in women with stimulant addiction who are in a period of abstinence[15].

We included socio-economic variables such as age when completed full time education (Data Field 845) and the English Indices of Multiple Deprivation (Data Field 26410) of the area of residence.

We also considered ongoing substance use, such as number of cigarettes smoked based on current (Data Field 1239) and past tobacco smoking (Data Field 1249), frequency of drinking alcohol (Data Field 20414 - excluding former drinkers[8] -Data Field 20404), and cannabis initiation (Data Field 20453).

Several self-reported clinical traits are available in the UK Biobank. We included Bipolar Disorder (type I or II) and Major Depression (Data Field 20126); recent restlessness (over the last two weeks, how often they have been so restless that it is hard to sit still, Data Field 20516); sleeplessness or insomnia (in the past 4 weeks before the MRI, Data Field 1200); lifetime history of Diabetes (Data Field 2443), Tinnitus (Data Field 4803), stroke (constructed from the date of the earliest reported stroke, either self-reported or hospital-reported, Data Field

42006), high blood pressure (constructed from age when high blood pressure was diagnosed (Data Field 2966), Parkinson's disease (either self-reported or hospital-reported, Data Field 42032) and Alzheimer's disease report (hospital-reported, Data Field 42020). For the four precedent phenotypes, all diagnoses were included (both prior to and following MRI). But, for Alzheimer's disease, all participants were diagnosed after the imaging visit (2.6 years after on average). While, for Parkinson's disease, most participants already had the disease before the MRI (2.1 years before on average).

We studied the "g" factor of general cognitive ability[8] that we constructed using the lavaan package in R, based on different measures of cognitive ability: Verbal Numeric Reasoning (Data Field 20016 & 20191), Trail Making (Data Field 6350 & 20157), Matrix Pattern Completion (Data Field 6373), Tower Rearranging (Data Field 21004), Symbol Digit Substitution (Data Field 23324 & 20159), Pairs Matching (Data Field 399) and Reaction Time (Data Field 20023). The variance explained by this new "g" factor is 35% of the variance contained in individual cognitive measures, which is consistent with previous reports[8].

We constructed a multisite Chronic Pain score from 0 to 7, as the sum of body sites at which chronic pain (for at least 3 months) was reported (Data Field 6159, 3404, 3414, 3571, 3741, 3773, 3799, 4067). We excluded pains that lasted less than 3 months, and non-specific "all-over the body" pain as in a previous publication[16].

Putative confounders

Lastly, we considered a set of known or hypothesized confounding covariates of brain imaging studies[12]. Quantifying their association with the different grey-matter representations can pinpoint image processing that is more sensitive to confounders but can also suggest which covariates to use in brain imaging analyses.

This set resulted in the UK Biobank assessment center (Data Field 54), head positioning in the MRI scanner (X,Y and Z positions -Data Field 25756 & 25757 & 25758), inverted signal-to-noise ratio (Data Field 25734), mean rs-fMRI head motion averaged across space and time points (Data Field 25741) discrepancy between T1 brain image and standard-space brain template (linearly aligned) (Data Field 25731), the volume of the brain (gray + white matter) (Data Field 25009) and intensity scaling for T1 (Data Field 25925). We constructed the time since the first MRI (of each centre) using the date of attending the assessment centre for the neuroimaging visit (Data Field 53).

We additionally considered as confounders age (Data Field ID 34), sex (Data Field ID 31), body mass index (Data Field ID (21001) and hip and waist circumferences (Data Field ID 49 & 50) as they exhibited large morphometricity in a previous study and could impact morphometricity estimates of other phenotypes. Of note, the body-size phenotypes could also be traits of interest in other studies.

Brain MRI Processing

We processed all our images with five different pipelines (**Figure 1**). Two used surface-based representations of the grey-matter structure (vertices) while the other three used volume-based representations (voxels).

FSLVBM and FSLANAT

We processed T1w images (from data Field 20252) with FSL[17] in two ways that appeared to give slightly different brain measurements for the same individual in a previous paper[6]. The first one (FSLVBM) performs the full FSL pipeline from the raw T1w image (*T1_brain.nii.gz*, **Figure 1**), while the second one (FSLANAT) takes as input the brain segmentation already performed by the UK Biobank (*T1_brain_pve_1.nii.gz*, **Figure 1),** to reduce processing time. The main difference between the two processing lies in the order of the processing steps. In short, FSLANAT registers the brain into a common space (MNI) before segmenting the grey matter. On the other hand, FSLVBM segments the grey matter in the native space before registering the segmented images to MNI. In both processing, images are non-linearly registered to a study-specific template. Since it was not computationally possible to create a template with all the images, we compared the variability induced by using templates derived from different sample populations. We generated templates using samples of varying sizes (90,300,600,900,990,1998,3000), all sex-balanced and evenly distributed across the three main centres. Subsequently, we registered a test set of 300 participants (also representative in terms of sex and site) on these templates. We then calculated, for each participant, the correlation between the registered image and that registered to the template made with the largest sample (3000). We observed that within-subject correlation was high and stopped increasing ($r>0.995$) when using templates derived from more than 600 participants (**SFigure 1**). Thus, we registered all images to the template made with 600 scans. The resulting images of both FSL processings have dimensions 91x109x91 voxels, each of size 2mm.

CAT12 Volume-based and Surface-based

We applied another processing pipeline to the raw T1w defaced images (*T1_brain.nii.gz*, from Data Field 20252) created by the ENIGMA consortium, which relies on the Computational Anatomy toolbox CAT12 for SPM[18]. We used the standalone version of the ENIGMA CAT12 toolbox that does not require a MATLAB license. The toolbox can output voxel-based (volume) and vertex-based (surface) representations of the grey-matter structure. The volume-based method (referred to as CAT12 Volume) uses the output from the standard SPM 'unified segmentation' that is initialized using Tissue Probability Maps and then proceeds with skull-stripping. Registration is done using using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL) but with a predefined DARTEL and Geodesic Shooting templates in MNI space. The ENIGMA-CAT12 software uses a pre-generated template derived from 555 healthy control subjects of the IXI database. Our final output is the modulated-warped-GM-partial volume segmentation (*mwp1.nii.gz*). Each image has a size of 85x103x85 voxels, each of size 2mm.

The surface-based approach (referred to as CAT12 Surface) extracts cortical thickness using a projection-based method[19], performed on the outputs of the volume-based processing described above. After the initial surface reconstruction, topological defects are repaired using spherical harmonics[20], which is followed by a surface refinement. The individual surfaces are spatially registered to the FreeSurfer "FsAverage" template using a spherical mapping. Finally, the local thickness values are transferred onto the FreeSurfer "FsAverage" template (164k mesh), which we extracted without any spatial smoothing. We obtained 163,842 vertices that measure local cortical thickness for each hemisphere.

FreeSurfer

The UK Biobank team processed the T1w and T2w (FLAIR) images using FreeSurfer 6.0[21], and we downloaded the resulting processed images, available in bulk data 20263 (**Figure 1**). In short, the processing utilized T1w and T2w (FLAIR) images together which improves pial surface reconstruction. FreeSurfer "recon-all" routine implements all required steps to process the images, including a projection onto the "FsAverage" cortical mesh. From the downloaded bulk data, we ran the recon-all -qcache option to obtain the surface files. In addition, we applied the ENIGMA-shape protocol[22][23] to the output of the FreeeSurfer processing, which computes vertex-wise measurements (radial thickness and log-jacobian, analogous to a surface area) for seven subcortical structures. We obtained 163,842 vertices for each cortical hemisphere and modality (thickness and surface area) and 13,560 vertices for the subcortical volumes for each type of measurement (radial thickness and

surface area). In the following analyses, we considered 'FreeSurfer Cortical Thickness', which only retains cortical thickness measurements to facilitate comparison to CAT12 Surface outputs. 'FreeSurfer All Modalities' includes all cortical and subcortical measurements (thickness and surface area).

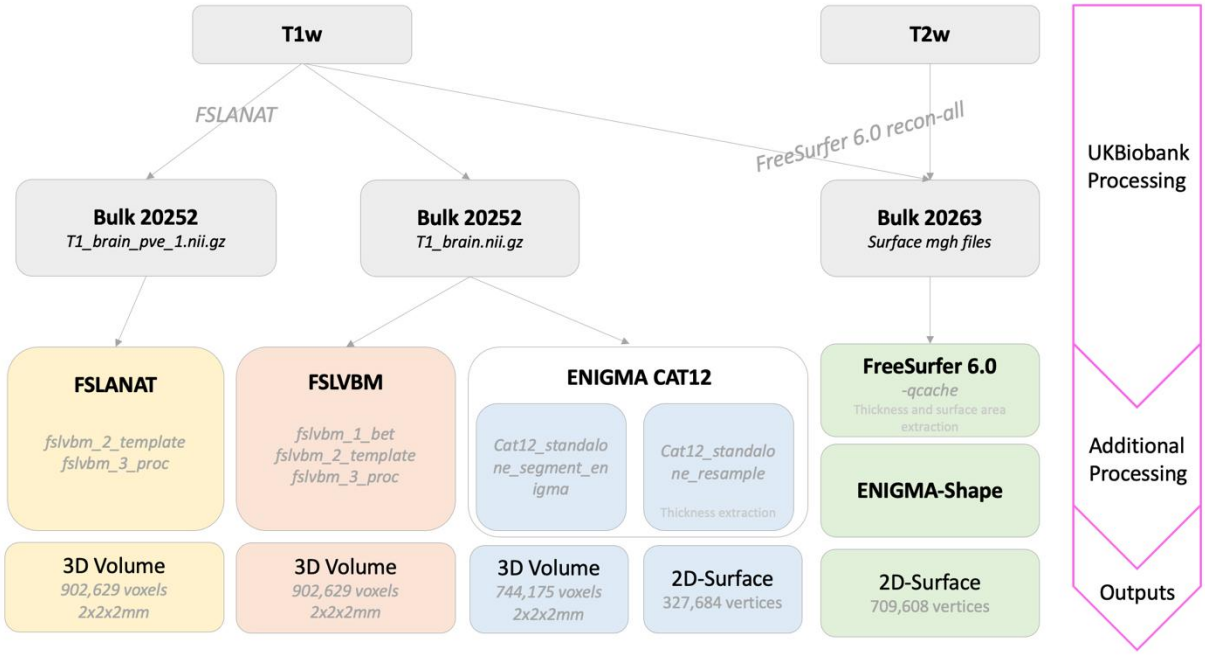


Figure 1 : Overview of the five processing pipelines. We considered three voxel-based morphometry pipelines (FSLANAT, FSLVBM, CAT12 Volume) and two surface-based morphometry pipelines (CAT12 Surface and FreeSurfer). We downloaded bulk files (individual zip files used to access large and/or complex items such as imaging data) provided by the UK Biobank and performed the additional processing steps to extract the vertex or voxel wise data.

Voxels and vertices Quality Control

For all three volume-based processing (FSLANAT, FSLVBM, CAT12-Volume), we excluded the voxels with mean lower than 0.1 and variance lower than 0.01, to retain the grey-matter voxels that are non-null across most participants. This resulted in 181,544 voxels for the FSLVBM pipeline, 184,637 voxels for the FSLANAT method and 192,483 voxels for the CAT12 volume-based processing (**SFigure 2**).

For all three surface-based processing (CAT12-Surface, FreeSurfer Cortical Thickness and FreeSurfer All Modalities), we excluded the vertices that contained 0 for all participants. This resulted in 299,881 vertices for CAT12 Surface and FreeSurfer Cortical Thickness, and 654,002 vertices for FreeSurfer All Modalities.

We exported all brain measurement tables into .bod format (a binary format) to optimize the memory requirement and computational time of analyses with OSCA[24] software.

SumR^2, kurtosis and skewness of measurements

We sought to quantify the amount of correlation between the brain measurements of each processing, which reflects the inherent smoothness of the data, and the strength of the connectome between correlated brain regions. We regressed out the covariates from the vertices/voxels values, as they might increase the correlation between brain measurements. Thus, we computed (for each processing), for each i-th vertex/voxel the sum of all $R_{i,j}^2$ (square of Pearson's correlation between i-th vertex/voxel and j-th one), with j varying from 1 to p (p being the total number of vertices/voxels given processing). The sum of R^2 ($sumR^2_i$) for the i-th voxel/vertex quantifies the amount of correlation with all other brain measurements:

$$SumR_i^2 = \sum_{j=1}^{p} R_{i,j\ adjusted}^2$$

Where $R^2_{adjusted} = R^2 - \frac{1-R^2}{N\_2}$ is the unbiased estimator of R^2, since the standard estimator of the Pearson correlation has upward bias of approximately 1/N (with N being the sample size). To reduce computation, we calculated $sumR^2$ using a subset of 1,000 UK Biobank individuals, representative in terms of site and sex.

In addition, we calculated the kurtosis and skewness of each vertex/voxel measurement to evaluate departures from normality in the distributions. Of note, kurtosis is a measure of the "tailedness" of the distribution, and skewness of the asymmetry.

Parcellation of vertex/voxels using cortical, subcortical and cerebellar atlases

We used complementary atlases, to annotate voxels and vertices across the different grey-matter regions (cortical, cerebellar and subcortical). For the cortical voxels/vertices, we used the Jülich-Brain atlas v3.0.3[25] ,

which is provided for both volume-based and surface-based processing. Indeed, this atlas is available in FreeSurfer FsAverage space and in Colin 27 (volume-based) space. The Jülich-Brain atlas comprises 157 different regions of interest.

For the subcortical nuclei in volume-based processing, we used the HarvardOxford atlas[26] provided by FSL and comprising 21 regions of interest (e.g., hippocampal subfields) aligned to the MNI-152 NLIN template. For surface-based processing, the vertices are associated with one of 7 subcortical nuclei (e.g., hippocampus) as part of processing.

For the cerebellum (only mapped in volume-based processing), we used the Dierdrichsen atlas[27] , also provided by FSL, aligned to the MNI-152 NLIN template and resulting in 28 anatomical structural regions.

All three Harvard-Oxford (subcortical atlas), Dierdrichsen (cerebellar atlas) and FSL processings are registered in the MNI-152 NLIN coordinate system[28]. However, CAT12 Volume (SPM) uses a slightly different space (MNI 152 linear)[29]. Therefore, we projected the Harvard-Oxford and Dierdrichsen MNI coordinates into the SPM space to obtain a subcortical and a cerebellar atlas aligned to the SPM voxels. Similarly, the Julich MNI space is also different (Colin 27, and we projected the atlas to FSLs and SPM coordinates. To achieve that, we used some co-registration functions provided by ANTs[30] (*antsRegistration, antsApplyTransforms*) and MrTrix3[31] (*mrtransform*) software.

Morphometricity

We estimated, for each processing, the percentage of trait variance captured by all brain features (vertices or voxel measurements), which has been coined “morphometricity”[32]. To estimate this morphometricity, we fitted the following linear mixed model[9]:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{b} + \mathbf{e} \quad (\mathbf{1})$$

With $Y_{N,1}$ a vector containing N observations of our trait of interest

$X_{N,c}$ a matrix of c covariates

$\beta_{c,1}$ a vector of fixed effects

b a vector of brain random effects with $b \sim N(0,B\sigma^2_b)$ and σ_b^2 the total trait variance captured by all voxel or vertex-wise measurements

e a vector of error terms with $e \sim N(0, I\sigma^2_e)$ and σ_e^2 the residual variance accounted for by the term error.

I the identity matrix

B is a matrix of variance-covariance between individuals calculated from all standardized brain measurements which we will refer to as a Brain Relatedness Matrix (BRM)[9]. These BRM are calculated with OSCA[24]. We utilized information contained in BRM to perform QC and resulted in 172 participants excluded due to extreme/outlying BRM-values (**SFigure 3**).

Finally, morphometricity is expressed as an R^2, which quantifies the proportion of variance explained by the brain measurements:

$$R^2 = \frac{\sigma_b^2}{\sigma_e^2 + \sigma_b^2}$$

The linear mixed model is implemented in OSCA*[24]*, (option –reml) a C++ software that contains efficient functions for data management and estimation of the model parameters using Restricted Maximum Likelihood (REML).

We repeated the analyses after Rank Inverse Normalisation of the brain measurements to investigate whether morphometricity estimates were impacted by non-normal distributions of brain measurement.

We extended the LMM above (1) to test whether the morphometricity of different processing explains the same trait variance or whether each captures a different or complementary proportion of trait variance. We fitted a model (**Eq. 2**) with two random effects $b_{Processing\ 1}$ and $b_{Processing\ 2}$, each corresponding to a processing. We compared this model to a reduced (nested) model containing only a single random effect (**Eq. 1**), using a likelihood ratio test, which follows a chi-square distribution."

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{b_{Processing\ 1}} + \boldsymbol{b_{Processing\ 2}} + \mathbf{e} \quad (\mathbf{Eq.\ 2})$$

With $b_i \sim N(0, Bi\sigma^2_{bi})$, $i \in (Processing\ 1, Processing\ 2)$ and all other parameters left unchanged.

As surface-based processing does not provide cerebellum measurements, we tested whether the complementary proportion of trait variance comes from the cerebellum. We performed sensitivity analyses that focused on FSLVBM (without cerebellum measurements) and FreeSurfer.

Lastly, we applied LMM with multiple random effects to decompose the morphometricity into the (conditional) contributions of the cortical, subcortical and cerebellar (when available) measurements. The LMM model then becomes:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{b_{Cortical}} + \boldsymbol{b_{Subcortical}} + \boldsymbol{b_{Cerebellar}} + \mathbf{e} \quad (\mathbf{Eq.\ 3})$$

This model extends the LMM hypothesis of a single normal distribution of effects to incorporate three distributions, each corresponding to different part of the grey-matter. For each part, we reported the proportion of variance captured, which collectively should sum to the overall morphometricity.

Brain-Wide Association Study

Family Wise Error Rate

We evaluated, for each processing, the false positive rate (Family Wise Error Rate; FWER) when using Bonferroni correction. Bonferroni's correction is a straightforward method that controls for FWER by setting a corrected significance threshold at alpha/nTests (nTests representing the number of brain measurements here). We expect Bonferroni's correction to be overly conservative (FWER<5%) as it assumes independence among the tests, which is not the case here as brain measurements are correlated. We expect FWER to differ between processing, influenced by the amount of correlation between brain measurements and the non-normal

distribution of vertex/voxel-wise measurements. On the other hand, we sought to confirm that non-normal distributions of brain measurements can influence the false positive rate.

To estimate the FWER, we simulated 1000 normally distributed random traits (i.e., not associated with brain measurements). Then, we tested the association between each brain measurement and these traits, controlling for standard covariates that have been recommended for neuroimaging analyses of the UK Biobank: age, sex, total brain volume, grey-matter density, head motion during resting stage fMRI, time since first scan, scanner brain position, as well as body-size covariates (BMI, Waist and Hip circumference). For each processing, we calculated the FWER as the proportion of simulated traits yielding at least one significant association (hence false positive voxel/vertex) after Bonferroni correction. We used qqplots to contrast the distributions of p-values obtained for each processing to that of a null distribution. To evaluate the effect of non-normal brain measurements on the FWER, we repeated the analyses after Rank Inverse Normalisation of the brain measurements.

We iteratively tested the association (bi) between the simulated trait y and the ith voxel/vertex-wise measurement (Xi) using a Generalised Linear Model (GLM: y = bi Xi + covariates + e), –linear option in OSCA[24].

Optimal significance threshold

To perform a fair comparison of the processing at the same level of false positives, we estimated the optimal significance threshold for each processing, which corresponds to FWER=5%. To estimate this threshold, we extracted the minimal p-value per simulated trait (1000 p-values in total) and set our new threshold to the 5th percentile. Indeed, this approach identifies the largest p-value significance threshold that ensures a rate of false positive (FWER) of 5%.

BWAS on UK Biobank traits of interest

We investigated which grey-matter measurements are associated with our UK Biobank traits of interest across the different processings by performing brain-wide association studies. We controlled for all covariates

(demographics, body size and neuroimaging covariates) in the analyses. We used several criteria to compare the results obtained using the different processing.

We first compared the association effect sizes by reporting the mean absolute z-score $\frac{|\beta|}{Standard\ Error}$) across all voxels/vertices for each trait and processing. We used the absolute z-score to focus on the magnitude of associations regardless of sign and the scale of the vertex/voxel-wise measurement.

Then, for each trait and each processing, we reported the number of significant vertices/voxels (using the optimal significance threshold that ensures comparable FWER=5% for all processing). We also reported the number of significant clusters as well as their sizes. To identify clusters, we performed an iterative 3D cluster search using the *vcgKDTree* function implemented in the R-library Rvcg. At each step, the algorithm considered the 10 nearest neighbours from the set of significant vertices/voxels and included them in the cluster if they were also significant. The algorithm stops when all significant vertices have been attributed to a cluster.

Lastly, we evaluated the number of clusters within each region of interest (ROI) and the number of ROIs containing at least one cluster for each processing. We evaluated robustness as the number of common brain regions (i.e., ROI that contain a significant cluster) identified by several processing.

Replicability of findings

We evaluated the replicability of the results by performing similar analyses in the replication sample. Thus, we compared morphometricity estimates from the discovery and replication sample. In addition, we conducted a Brain-Wide Association Study to report the proportion of replicating voxels and clusters. We used a significance threshold of 0.05/NSignifV, with NSignifV the number of significant associations in the discovery sample (across all traits and processing), which corresponds to the total number of vertex/voxel wise associations we took to the replication sample.

Prediction

Prediction from significant vertices/voxels

We evaluated how much the significant vertices/voxels can together account for the traits of interest by evaluating their predictive power in the replication sample. Indeed, the significant associations can tag redundant signals, which does not necessarily translate into increased prediction[33]. Typically, vertices/voxels from the same cluster often capture the same information, but this can also be the case between distant vertices/voxels that are correlated. We selected the most significant vertex/voxel in each cluster and constructed a linear predictor using association weights b_i, estimated in the BWAS (**Eq. 4**). We used OSCA (–score option) to calculate a score for each individual of the replication sample. We evaluated the prediction accuracy (R^2) of these scores in the replication sample using a linear model that controlled for all covariates and tested its significance with a likelihood-ratio-test (nested models, between a full model with predictor and covariates vs. a reduced model with only covariates).

We finally compared the results obtained with those using all significant vertices/voxels. For that, we compute the prediction accuracy using the same model but including all significant voxels instead of only the top ones. This comparison allows us to conclude about the redundancy of the signal captured by all brain measurements.

Prediction from the whole grey matter

Machine learning approaches aim to build performant brain-based predictors that are not limited to statistically significant regions. We built Best Linear Unbiased (BLU) Predictors to evaluate which processing is best suited to machine learning and prediction analyses. BLU Predictors have been widely used in animal and human genetics and have previously been applied to vertex-wise data of the UK Biobank, where they showed similar to superior performances compared to LASSO predictors[9]. BLU Predictors are efficiently implemented in OSCA, and easily scale up to large samples. We trained BLU Predictors in the UK Biobank discovery sample and evaluated them in the replication sample, controlling for all covariates. As previously, we reported prediction accuracy as an R^2 (hence comparable to the morphometricity estimates). We further reported the fraction of morphometricity that BLU Predictors can predict.

Results

Vertices and voxels distribution and correlations

We investigated the distribution of vertices and voxels-wise measurements from each brain MRI processing. We found that some processing exhibited positive kurtosis (>3, **Table 1**). In particular, FreeSurfer-based processing exhibited larger median kurtosis (4.1 for FreeSurfer Thickness and 4.9 for FreeSurfer All Modalities), suggesting that many vertices/voxels have heavier tails than those expected in a normal distribution, leading to increased occurrences of extreme values. Skewness levels suggest that the vertex/voxels distributions were largely symmetrical, except for FreeSurfer processing where skewness indicated the presence of right tails in vertex-wise distributions. CAT12 Surface stood out as the only processing with negative (albeit moderate) median skewness (**Table 1**), implying that, unlike in other processing, its vertices measurements have longer left tails.

Processing	**Nb of measurements**	**Median Kurtosis**	**Median Skewness**	**Median SumR^2**
FSLVBM	181,544	3.0	0.36	42
FSLANAT	184,637	3.1	0.35	48
CAT12 Volume	192,483	3.3	0.24	240
CAT12 Surface	299,881	3.5	-0.076	1471
FreeSurfer Thickness	299,881	4.1	0.79	137
FreeSurfer All modalities	654,002	4.9	1.1	105

Table 1 : Data Description for each processing. First column summarizes number of voxels/vertices brain measurements from the 6 brain MRI processing pipelines. Median kurtosis (resp. skewness) quantifies the non-normality of the grey-matter measurements. Both were computed with the 'moments' package in R. The median sumR2 quantifies the amount of correlation in the structural connectome.

We calculated the SumR^2 to quantify the overall amount of correlation in each structural connectome. CAT12 processing methods, and particularly CAT12 Surface, exhibited large median sumR2 (**Table 1 : Data**

Description for each processing. First column summarizes number of voxels/vertices brain measurements from the 6 brain MRI processing pipelines. Median kurtosis (resp. skewness) quantifies the non-normality of the grey-matter measurements. Both were computed with the 'moments' package in R. The median sumR2 quantifies the amount of correlation in the structural connectome.

), suggesting that their vertices/voxels have a higher degree of correlation. We investigated whether this large median sumR^2 came from the non-normal distribution by re-estimating it after adjustment for covariates and rank-inverse normalization: we found that, it was robust to non-normal distributions (before and after adjustment r=0.999), suggesting that this amount of correlation is inherently linked to the processing method itself.

Next, we investigated if highly correlated regions (with high sumR2) were the same across processing. It was the case for the three volume-based processing, which exhibited similar mean sumR^2 per ROI, with a correlation of r=0.97 between FSLVBM and FSLANAT and R2=0.70 between FSLANAT or FSLVBM and CAT12 Volume. Both cortical thickness processing exhibited a correlation of 0.57. However, the ROI with high sumR2 differed between volume and surface-based processing. For example, the correlation of mean sumR2 across ROI between FreeSurfer and FSLVBM was 0.47. This suggests that the pattern of correlation between brain measurements is highly dependent on the representation used (volume or surface). Detailed mean sumR2 per ROI can be found in **STable 1**.

Overall, some processing exhibited non-normal distributions, and we observed varying levels of correlation among voxels/vertices that can both impact further associations analysis.

Morphometricity of putative confounders

We estimated the morphometricity of several traits known (e.g. head motion) or hypothesized (e.g. time since first scan) to have an effect on brain images, which would quantify how much the possible confounders may contaminate each image processing[34]. We found that all the considered confounders (either imaging or body size) exhibited a large morphometricity (R2 in 20-80%) (**Figure 2**), which implies they are associated with one or several brain regions. Interestingly, cortical thickness measurements (from FS or CAT12) appeared the least associated with possible confounders (median morphometricity across traits 32% -FreeSurfer- and 21% -CAT12,

Figure 4). In comparison, FSL based processing (FSLANAT or FSLVBM) showed the largest associations with the possible confounders (medians morphometricity 63-64%) (**Figure 2**).

Next, we used LMM with multiple random effects to jointly estimate the variance accounted for by cortical, subcortical and cerebellar measurement (when available). For FreeSurfer all modalities, we further separated cortical thickness and surface area. Of note FreeSurfer Thickness and CAT12 Surface only included measurement of cortical thickness and were not included in these additional analyses. We found that across all processings, the different brain parts (cortical, subcortical and cerebellar) were all associated with confounders (**SFigure 4),** which suggest they can cause widespread false positives.

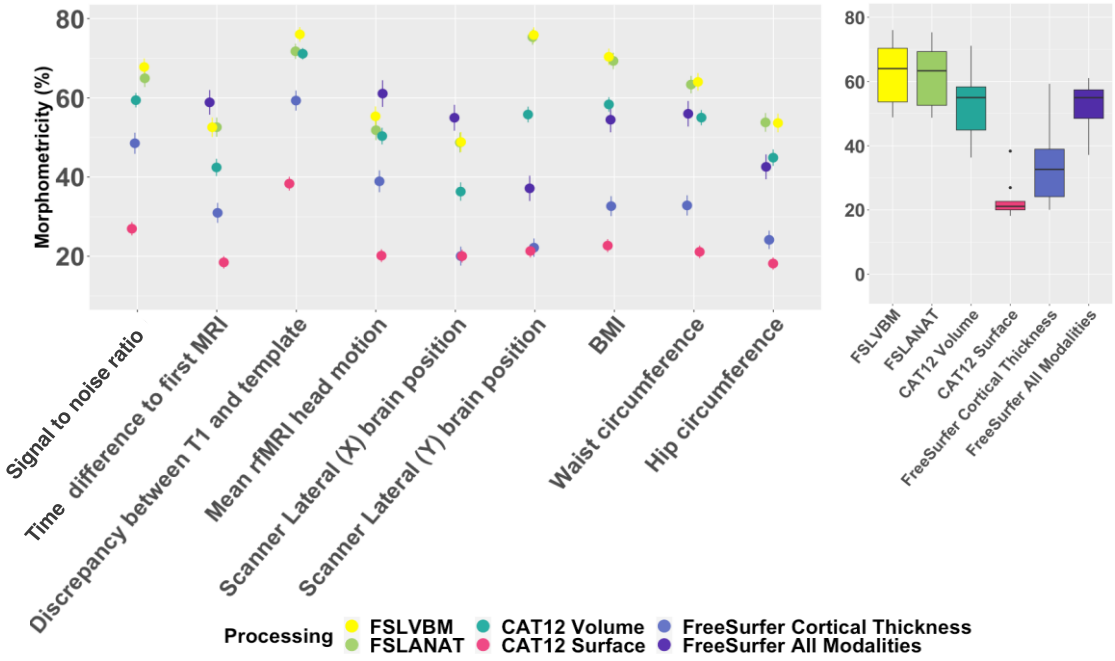


Figure 2 : Morphometricity of possible confounders, imaging and body size. The left panel depicts the morphometricity of each possible confounder (after controlling for age and sex). The different colours correspond to the different processings, and the bars represent the 95% confidence intervals. The boxplot (right panel) summarizes the distribution of morphometricity estimates across all possible confounders

Morphometricity of traits of interest

When controlling for all confounders (age, sex, and those studied in previous section), both FSL processing accounted for the largest proportion of variance (median = 5.8%) followed by FreeSurfer All modalities (median = 4.2%) and CAT12 Volume (median=3.0%). CAT12 Surface yielded the smallest proportion of variance explained (median = 1.5%). Moreover, we detected significant morphometricity (after Bonferroni correction i.e. $p<0.05/(18*6)$) for most traits of interest (except for restlessness, depression score and general cognition score, depicted in grey on **Figure 3).** We observed that morphometricity varied, depending on the trait. Maternal smoking around birth exhibited the largest morphometricity ([13% ; 36%]) followed by diabetes ([5% ; 15%]), high blood pressure ([5% ; 14%]) and alcohol frequency([4% ;14 %]).

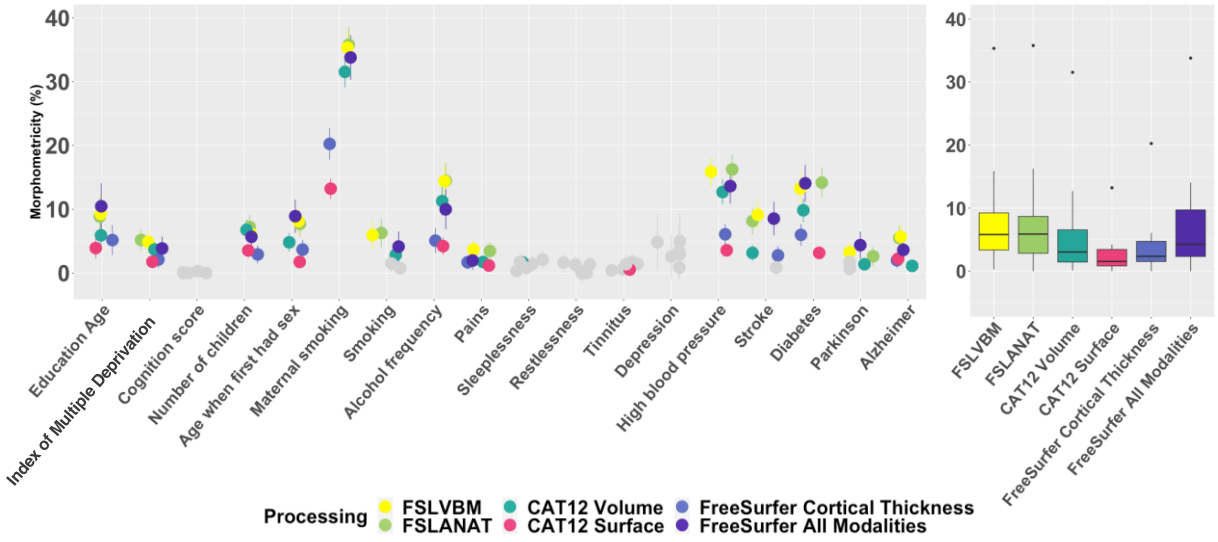


Figure 3 : Morphometricity of traits of interest, controlling for all covariates.

The left panel depicts the morphometricity of each trait of interest, when controlling for all covariates. The different colours correspond to the different processing, and the bars represent the 95% confidence intervals. We represented in grey, the morphometricity estimates that were not significatively different from 0 (after multiple testing correction $pval<0.05(18*6)$)). The boxplot (right panel) summarizes the distribution of morphometricity estimates across all traits of interest for each processing.

When modelling brain parts and measurement types in specific random effects, we found (**SFigure 5**) that, overall, all broad regions contribute to the detected morphometricity. For FSLVBM, the variance in traits of interest was mostly explained by the cortical and subcortical measurements, even though cerebellar measurement also captured some significant information, for maternal smoking for instance. For all traits with a significant morphometricity, we confirmed that cortical and subcortical measurements contributed to the association. We also found that the cerebellum was significantly associated with Maternal smoking and Diabetes. Results remained consistent after we rank-normalized voxels/vertices for each processing, which suggests morphometricity estimates are robust to the non-normal distributions present in the data. (**SFigure 6**).

Consistent morphometricity estimates in the replication sample

We observed mostly comparable estimates of morphometricity in the replication sample (**SFigure 7**), suggesting that morphometricity estimates are replicable. Of note, morphometricity was smaller for the ”time difference to first MRI” for all processing. This is likely due to the fact that the scanners are more recent in the replication sample (average scanner age of 3 years in the discovery vs. 0.9 years in the replication dataset), leading to less variability in the time since the first MRI. In addition, we observed larger estimates of morphometricity from FreeSurfer processing in the replication analysis (**SFigure 7**), particularly for traits such as Mean motion during rfMRI, SNR, Stroke, Parkinson or Maternal smoking. This could be partly explained by the smaller numbers in the replication sample (i.e. larger standard errors), although we cannot rule out differences in scanner (e.g., software) or acquisition that could impact the output from FreeSurfer.

Morphometricity shared between processings vs. unique to each processing

We investigated whether the processings captured a different and/or complementary proportion of the trait variance. We only focused on traits exhibiting significant morphometricity (i.e. we excluded Cognition, Depression and Restlessness). We found that fitting grey-matter measurements from two processings resulted in an overall increase in the proportion of explained variance (**Figure 4**), which suggests that each grey-matter representation captures a unique proportion of trait variance that is missed by other processing.

For example, when combining the two top processing in terms of morphometricity (FSLVBM and FreeSurfer All Modalities), the trait variance accounted for grew to 12.5%, on average, across all traits (vs. 10.1% for

FSLVBM and CAT12 Volume or 8.7% for FS All Modalities and CAT12 Surface). This represents an increase in variance explained, with a ratio of 1.49 and 1.35 (compared to Freesurfer and FSLVBM, respectively). This suggests that both processes capture a unique fraction of the trait variance (35% to 49% of the signal) in addition to the variance they both capture (51% to 65% of the morphometricity).

As expected, CAT12 surface captured less morphometricity than its competitors (average 2%,**Figure 4)**, so the gain of adding an additional set of brain measurements was maximal (range 2.59-4.43). We also noted that FSLANAT and FSLVBM detected mainly the same morphometricity with only about 10% of unique signal (**Figure 4**).

We confirmed that the improvement was significant for each trait (**SFigure 8**). For example, for `Maternal smoking ` FreeSurfer All modalities explained 33.8% of the variance, and FSLVBM 35.3%, but together they accounted for 46%. As FreeSurfer do not measure the cerebellum region, we wondered whether FSLVBM's unique signal is due to the cerebellum. Therefore, we performed a similar analysis including only FSLVBM (without cerebellar measurements) and FreeSurfer All Modalities. Overall, it did not result in any significant difference compared to the volume-based processing unique signal. Therefore, the cerebellum does not fully

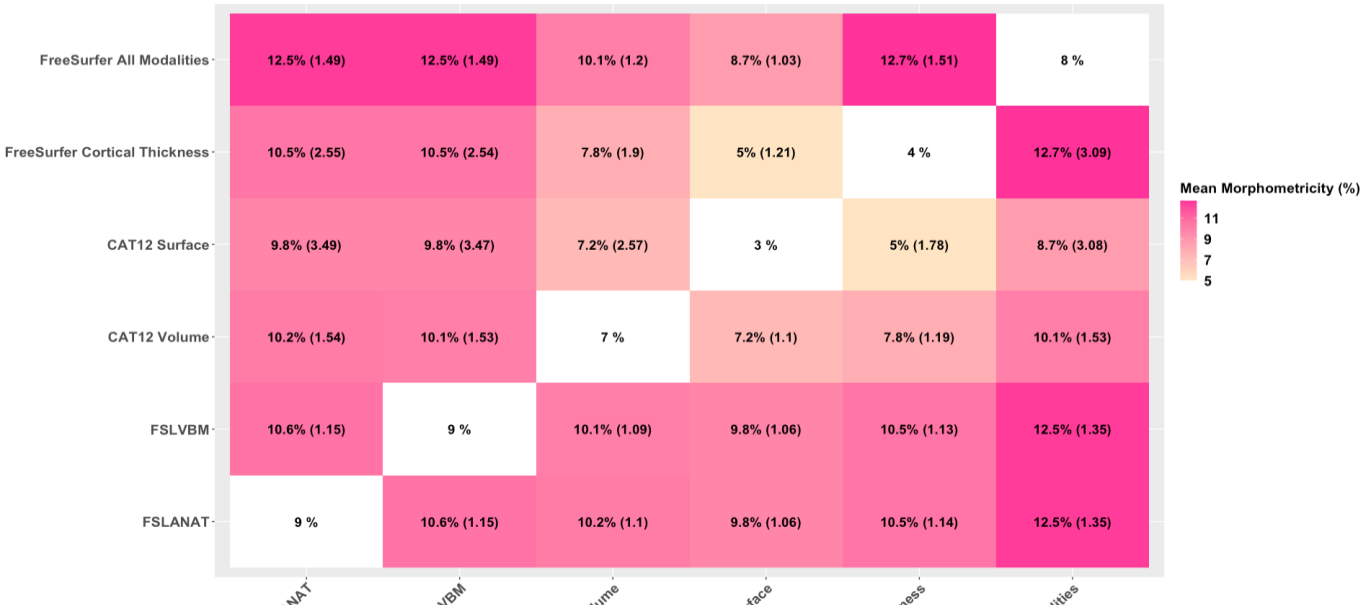
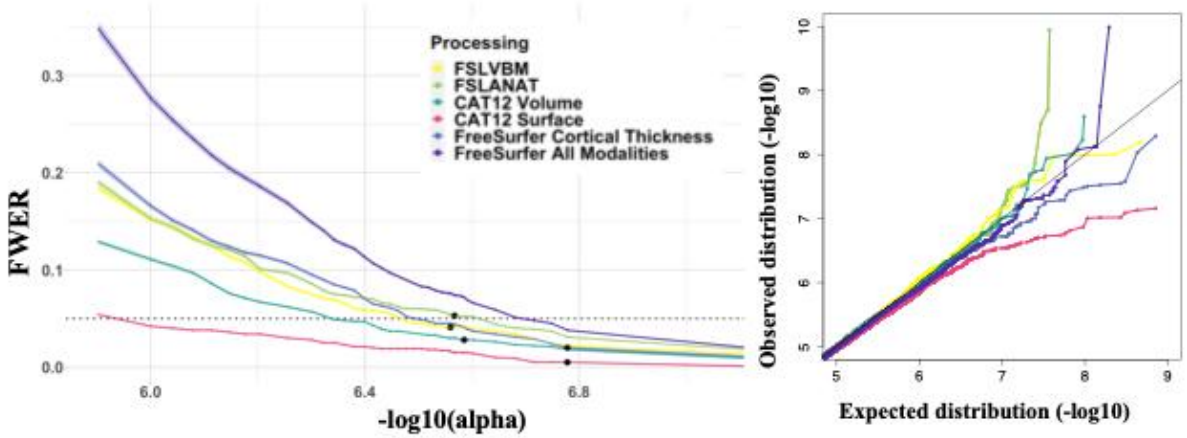


Figure 4 : Heatmap of morphometricity increase when fitting two grey-matter representation in the model. The heatmap shows the average percentage of variance explained (morphometricity) across all traits of interest when combining two processing methods. The rate of increase is shown in parentheses, and diagonal values represents the mean morphometricity of each processing method alone. Rows indicate the reference processing and columns the added processing. For example, the second row of the first column corresponds to the result of FSLANAT added to FreeSurfer Cortical Thickness resulting in an average morphometricity of 10.5%, 2.55 times higher than FreeSurfer Cortical Thickness alone (average morphometricity 4%, second row, 5th column). Symmetrically (6th row, 5th column), adding FreeSurfer Cortical Thickness to FSLANAT only slightly increases morphomemetricity (x1.14,) as FSLANAT explains 9% of variance by itself (6th row, first column).

explain these discrepancies.

False positive rate of vertex/voxel wise association (BWAS)



Processing	FWER (Bonferroni)	Bonferroni Threshold (e-07)	Optimal Threshold (FWER=5%) (e-07)
FSLVBM	4.1%	2.8	3.4
FSLANAT	5.3%	2.7	2.4
CAT12 Volume	2.8%	2.6	4.7
CAT12 Surface	0.5%	1.7	12
FreeSurfer Cortical Thickness	2%	1.7	3.1
FreeSurfer All Modalities	2%	0.76	2.0

Figure 5 : Family Wise Error Rate of BWAS, Q-Qplot of minimal p-values and optimal significant thresholds. The top left panel depicts the Family-Wise Error Rate (FWER) at different significance thresholds. The x-axis shows the significance thresholds in the log scale, to improve readability. The horizontal dashed line represents an FWER of 0.05. The black dots indicate the Bonferroni's significance threshold of each grey-matter processing. The shaded bands correspond to the binomial proportion confidence intervals at 95%, which are narrow <0.02 due to the high number (1000) of simulated traits. The top right panel depicts the QQ-Plot of the minimal p-values for each vertex/voxel across the 1000 simulated traits. Minimal p-values are represented in log10 scale. The bottom table reports the FWER with Bonferroni's correction (i.e. the black dots on the top left plot) as well as the optimal significant threshold when ensuring FWER=5%. For comparison, we also reported the Bonferroni's significance threshold (middle column).

We previously showed that processing exhibited non-normal distributions of the grey-matter, implying that data have higher probabilities of extreme values (high or low) than would be expected in a normal distribution, which may impact association testing.

We evaluated how stringent was Bonferroni's correction, for each processing, by estimating the Family wise error rate (FWER) <5%, under the null hypothesis. We expected to find FWER lower than 5% for all processing, especially those that exhibited a large amount of correlation between voxels/vertices (median sumR2**,Table 1 : Data Description for each processing.** First column summarizes number of voxels/vertices brain measurements from the 6 brain MRI processing pipelines. Median kurtosis (resp. skewness) quantifies the non-normality of the grey-matter measurements. Both were computed with the 'moments' package in R. The median sumR2 quantifies the amount of correlation in the structural connectome.

), as it induces non-independence of the test statistics. Our findings (**Figure 5, top-left panel**) indicated that Bonferroni's correction effectively ensured a FWER<5% except for FSLANAT (FWER=5.3%, **Figure 5**), suggesting that using this processing may result in a false positive rate above 5%). For all other processing, Bonferroni's correction was overly conservative (FWER largely below 5%, **Figure 5**, **bottom table**): FreeSurfer All Modalities exhibited a FWER of 2%, FSLVBM of 4% and CAT12 Volume of 3%. Bonferroni's correction was particularly stringent for CAT12 Surface, as indicated by an FWER of 0.5%, which is consistent with the fact that it is the processing with the larger amount of correlation between vertices (**Table 1 : Data Description for each processing.** First column summarizes number of voxels/vertices brain measurements from the 6 brain MRI processing pipelines. Median kurtosis (resp. skewness) quantifies the non-normality of the grey-matter measurements. Both were computed with the 'moments' package in R. The median sumR2 quantifies the amount of correlation in the structural connectome.

).

Optimal significance threshold

To compare processing to a set FWER level, we derived optimal significance thresholds (intersection of the coloured lines and dashed lines/**bottom table, Figure 5**) that correspond to FWER=5%, therefore less stringent than Bonferroni's correction, for most processing. CAT12 Surface had a new threshold of 1.2e-06 compared to 3.1e-07 for FS Thickness and 2.0e-07 for FS All modalities. For volume-based processing, the optimal threshold for FSLVBM was equal to 3.4e-07, compared to 2.4e-07 for FSLANAT and 4.7e-07 for CAT12 Volume.

On the other hand, to evaluate the impact of distribution on the false positive rate, we looked at the distribution of the minimal p-value per voxel/vertex across the 1000 random traits to focus on the top associations that are mostly likely to reach significance for each processing. The QQ-plot of FSLANAT (green, **Right panel, Figure 5**) showed an inflation for the top percentiles of the distribution meaning that a couple of voxels exhibited more significant p-values distribution than expected by chance (**SFigure 10,** left column) and are more likely to exhibit a significant association. After applying a rank-inverse normal transform to the brain measurements (after covariates adjustments), the Q-Q plot of FSLANAT (**SFigure 10,** right column, green) aligned more closely with the expected distribution, confirming the inflation was due to non-normal distributions.

Additionally, we observed a slight QQ-plot deflation for FS Thickness, FS All Modalities and a more important one for CAT12 Surface. The deflation of pvalues of FreeSurfer Cortical Thickness was reduced after applying rank-inverse normal transform to the brain measurements. On the other hand, the deflation of CAT12 Surface was attenuated but remained, which suggests CAT12 Surface may suffer from a low type II error. We therefore can expect CAT12 Surface to be conservative and detect fewer significant associations.

BWAS of traits of interest

We tested the association between our voxels/vertices and the UK Biobank traits of interest, controlling for all covariates. We did not single out results for the FreeSurfer Cortical Thickness processing, as all its vertices are included in FreeSurfer All Modalities.

Effect sizes of vertex/voxel-wise associations

The statistical power of BWAS depends in part on the effect sizes between traits and vertex/voxel-wise measurements. Thus, we investigated if different processing methods yielded differences in effect sizes, measured by the median absolute z-score across all voxels/vertices.

CAT12 Surface exhibited the largest effect sizes with a median z-score of 0.95 on average and a maximum of 1.5, (**Figure 6, top panel**), followed by CAT12 Volume (median=0.85, max=1.3), compared to FSL processing (FSLANAT median=0.79, max=0.98, similar results for FSLVBM) and FreeSurfer (median=0.73, max=0.97).

Traits with larger morphometricity (Maternal smoking, Diabetes, High blood pressure) also displayed larger average z-scores, implying larger associations in the brain, thus greater statistical power to identify significant vertices/voxels than for the other traits.

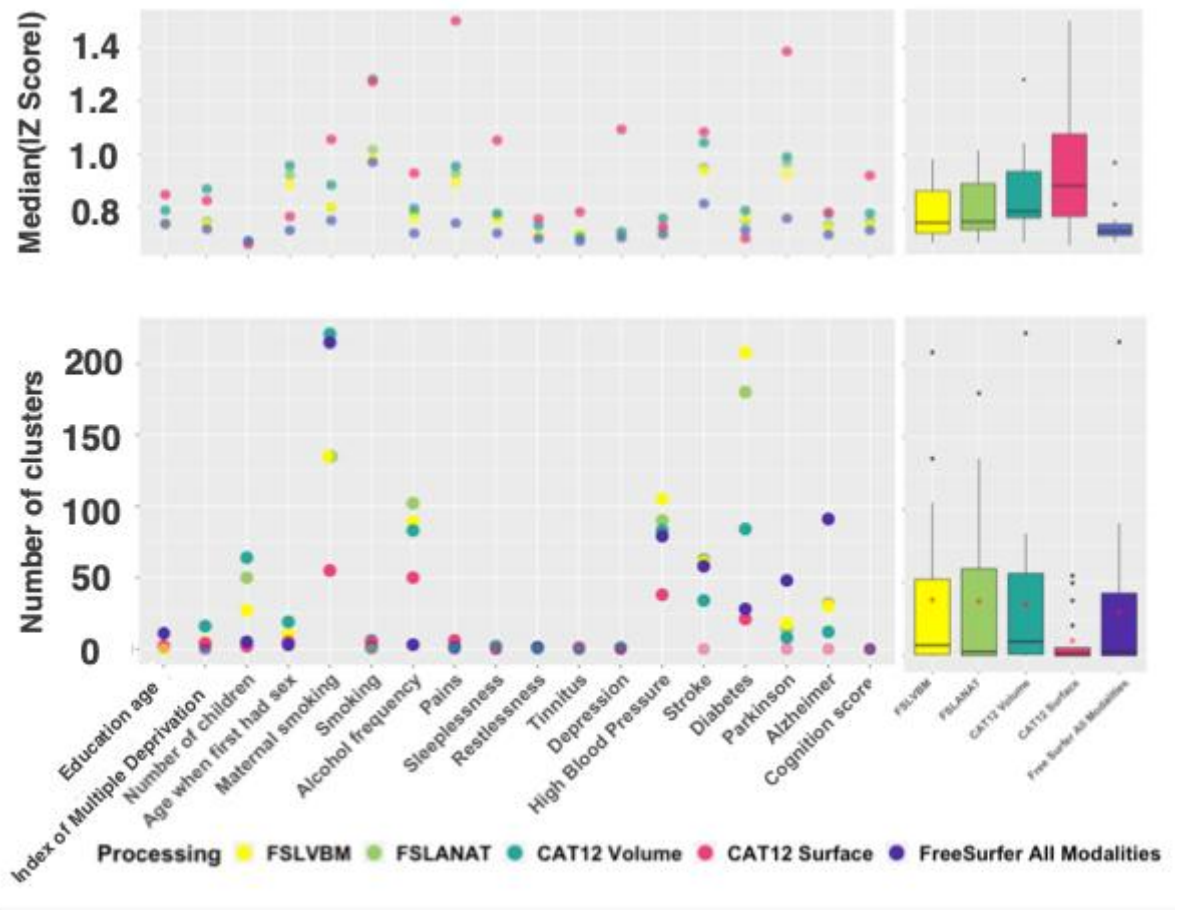


Figure 6 : Median absolute Z-score across the brain and Number of clusters after correcting with optimal significance threshold across all traits and processings. The top-left panel depicts the median absolute z-score for all traits of interest. The top-right boxplot displays the overall distribution of these scores. The bottom-left panel depicts the number of significant clusters for all traits of interest. The bottom-right boxplot displays the overall distribution of the number of clusters across all traits of interest for each processing. Red stars represent mean values.

Number of significant clusters after multiple testing correction

We used the optimal significance threshold (corresponding to FWER = 5%) in the following analysis, to allow for a fair comparison of the processing methods.

We found that (**Figure 6, bottom-panel**Erreur ! Source du renvoi introuvable.) CAT12 Volume and FSLVBM yielded the highest number of clusters across all traits (median=10 for CAT12 Volume and 7.5 for FSLVBM). In comparison, we identified a median of 3 clusters for both FS All Modalities and FSLANAT and 2 for CAT12 Surface. At a trait level, the number of significant clusters varied widely between processing. For example, we identified >200 clusters associated with `Maternal smoking` using FreeSurfer (N=221) and CAT12 Volume (N=215), 135 clusters using FSL and 55 using CAT12 Surface) (**Figure 6, left panel**). There was a correlation (r=0.89) between the number of associated clusters and the morphometricity, indicating that morphometricity is a good predictor of discoverability, although other factors also contribute (e.g., effect sizes and number of tests).

To explore these clusters in more detail, we examined their size and distribution. As expected, processing with large correlation among their voxels/vertices had larger clusters and a smaller proportion of clusters made of a single voxel/vertex: 9% of the CAT12 Surface clusters contained a single vertex (and median size of clusters = 11) , vs. 45% of the CAT12 Volume clusters (median size of clusters = 2) and 60% for all three other processings (with a median size of clusters equals to 1 for all three of them).

ROIs robustly associated across different processings

We also assessed the robustness of associations across different processing, meaning their ability to detect clusters in the same ROIs.

We found that, for all three volume-based processing, the significant clusters implicated more ROIs (median=14 for FSLANAT and 10 for FSLVBM and CAT12 Volume, compared to 7 for FreeSurfer and 4 for CAT12 Surface) (**SFigure 13**). Especially, for Maternal smoking (67 ROIs for CAT12 Volume, 55 for FSLVBM and 45 for FSLANAT) and Diabetes (46 and 47 for FSL processing, and 37 for CAT12 Volume), which is is consistent with the large number of significant clusters associated with this trait.

We focused on all traits that identified significant clusters across all processing methods (**SFigure 14**). We observed the strongest agreement between both FSL processing (although never perfect), and slightly lower agreement between FSL and CAT12 volume processing. The overlap of significant ROI was even lower between FSLVBM and Freesurfer. CAT12 Surface had a poor agreement with all other processing. This robustness pattern was similar across all traits.

For example, for `Maternal smoking` (**SFigure 14**), 32 ROI were robustly implicated by different volume-based processing (for a total of 45-67 ROI detected, hence rates of 47-71%). In comparison, only 12 of the ROIs were detected by FSLVBM and FreeSurfer (out of 27 detected by FreeSurfer), and only 5 were common to FSLVBM and CAT12surface (out of 26 detected by CAT12 Surface).

Brain based Prediction

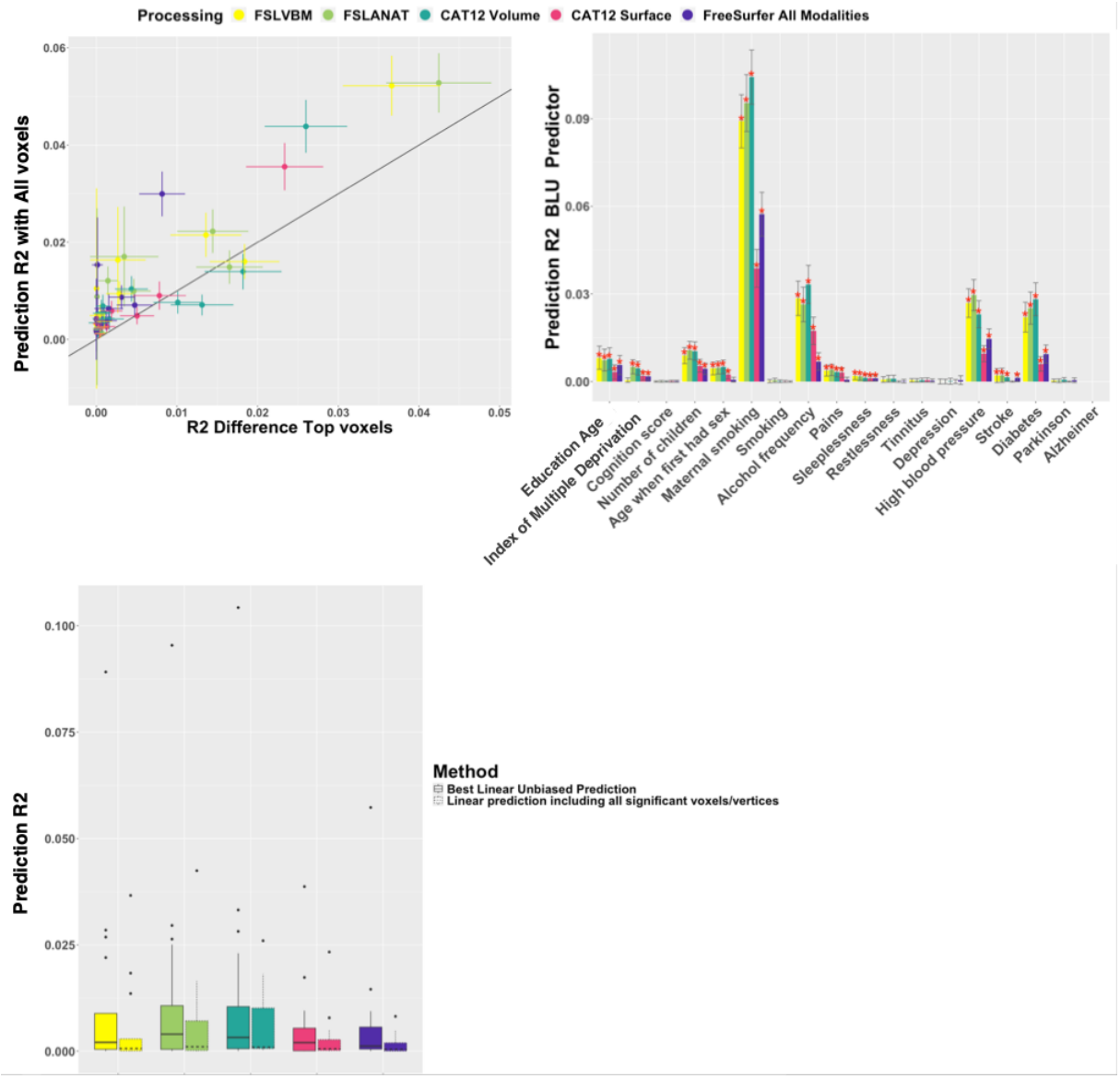


Figure 7 : Linear prediction and best linear unbiased prediction. The top left panel compares the prediction accuracy including only the top-voxels/vertices per cluster (x-axis) versus all significant voxels/vertices (y-axis). The vertical and horizontal bars show the 95% confidence intervals in the two samples. The top right panel depicts the prediction accuracy R2 using the Best Linear Unbiased Prediction method. Red stars indicate significant log-likelihood ratio test after Bonferroni correction ($p < 0.05/6*29$) and vertical bars represent the confidence intervals. The bottom panel recapitulates all three volume-based processing and top phenotypes predicted, the morphometricity estimates (in grey), and the BLU prediction as well as the fraction of morphometricity predicted in brackets.

We compared prediction accuracy achieved in the UK Biobank replication sample from significant voxels/vertices (after optimal significance threshold correction, **Figure 7, left panel**). We contrasted the prediction achieved from the top vertex/voxel per cluster to that coming from all significant vertices/voxels. We observed (**Figure 7, left panel**) that, for several traits and processings, the prediction was greater when including all significant voxels/vertices, which suggests that some clusters contain additional signal that is not fully captured by the most significant voxel/vertex. We expect more significant clusters to translate into greater prediction accuracy, unless several clusters tag the same information. Of note, `Alzheimer's disease` could not be predicted as there were no participants with records of Alzheimer's disease in the replication sample.

The top voxels from volume-based processing yielded greater predictions than the vertices from surface-based processing. In particular, for traits such as `Maternal smoking`, `Alcohol frequency`, `High blood pressure` (**SFigure 15, left panel**), where prediction R2 ranged between [0.018-0.029] for volume-based methods versus R2 in [0.0056-0.014] for both surface-based ones (non-overlapping confidence intervals). Of note, using Bonferroni correction (instead of the optimal one) had little incidence on the prediction achieved from significant vertices/voxels (**SFigure 15, right panel**).

Next, we evaluated the prediction accuracy of Best Linear Unbiased predictors, which capture signals across the whole brain and are not limited to vertices/voxels reaching significance. As expected, BLUP yielded improved prediction compared to only using significant vertices/voxels (**Figure 7, bottom panel**). As previously, the three volume-based processing exhibited better predictions than surface-based processing (R2=0.01 on average versus R2=0.005 for both surface-based processing). The prediction was the largest (and significant across all processing) for `Education age`, `Number of children`, `Maternal smoking`, `Alcohol frequency`,`Sleeplessness`, `High blood pressure` and `Diabetes` (**Figure 7, top-right panel**). . For these traits, we reported the fraction of morphometricity predicted using each processing (**SFigure 16**). Overall, FreeSurfer resulted in a lower fraction of morphometricity predicted (median=6.8%) compared to 17-19% for all 4 other methods. For instance, these 4 processing methods predicted 25-30% of `Maternal smoking` morphometricity, compared to 17% for FreeSurfer. Interestingly, we found that the proportion of predicted morphometricity varied

	Number of significant voxels	Number of replicating	Fraction

from one trait to the next, which suggests that some traits are harder to predict. For example, 27% of the morphometricity of `Maternal smoking` could be predicted (median across processing), compared to 18% for both High blood pressure and Diabetes, and even lower for the other traits (**SFigure 16**).

Finally, as we checked the prediction achieved by clusters of size 1 (**SFigure 17**). For some processing and traits (Number of children, Age when first had sex, Maternal smoking, Alcohol frequency, High blood pressure, Stroke and Diabetes), the prediction reached significance, suggesting that clusters of size one can sometimes tag true associations. However, the predictions remained low (for example R2=0.008 for Maternal smoking [FSLVBM processing], in comparison with R2=0.09 from all significant voxels,). Of note, the clusters of size one identified by FreeSurfer did not significantly predict any traits (R2<0.0006).

Replication rate of significant vertices/voxels and clusters

		voxels	
FSLANAT	3,945	1,436	**0.36**
FSLVBM	4,277	1,489	**0.35**
CAT12 Volume	13,674	5,118	**0.38**
CAT12 Surface	7,077	1,167	**0.17**
FreeSurfer All Modalities	10,936	6,742	**0.62**
	Number of significant clusters	**Number of replicating cluster**	**Fraction**
FSLANAT	676	117	**0.17**
FSLVBM	692	124	**0.18**
CAT12 Volume	635	120	**0.19**
CAT12 Surface	190	41	**0.22**
FreeSurfer All Modalities	544	55	**0.10**

Table 2 : Replicating voxels and clusters using the optimal significance threshold across all traits, for all processing. To control for multiple testing in the replication sample, we used Bonferroni-corrected significance thresholds of 39,909 for the vertex-wise level and 2,737 for the cluster-wise inference.

We investigated whether significant voxels detected in our main analysis replicated an independent UK Biobank dataset (replication sample). To control for multiple testing in the replication sample, we used Bonferroni corrected significance thresholds of 39,909 for the vertex-wise level and cluster-wise inference, which accounts for the total number of discoveries, across traits and processings.

Volume-based processing all showed a good replication rate (35-38% of voxels replicated, and 17-19% clusters). CAT12 Surface displayed a lower replication rate at the vertex level (17%), but comparable at the cluster level (22%). This was due to many vertices from large clusters not replicating, although the core of the cluster reached significance. Moreover, FreeSurfer exhibited a higher replication rate at the vertex level (62%) but a low rate at the cluster level (10%). In fact, the vertex-level replication rate was driven by `High blood pressure` trait: indeed 60% of FreeSurfer significant vertices were associated with `High blood pressure` (versus 26-27% for FSLANAT and FSLVBM processing, 12% for CAT12 Volume and 4% for CAT12 Surface). When excluding this trait, the vertex-wise replication rate was reduced to 12% for FreeSurfer, while the cluster-wise replication rate remained at 8%. Of note, the replication rate of FSLANAT and FSLVBM processing was also driven by associations with High blood pressure` (16-17% of replication rate on all other traits). Similarly, CAT12 Volume replication rate was driven by associations with `Diabetes` (representing 34% of the significant voxels), without which the replication rate reduced to 22%.

The low cluster-level replication rate of FreeSurfer was mostly driven by the clusters of size one. Strikingly, FreeSurfer and FSL-based processing had similar proportions of cluster of size one (60-63%), but the replication rate of these clusters was 8.5% and 7.2% for both FSLANAT and FSLVBM, while it was only 2.1% for FreeSurfer. Of note, the clusters of size one identified with CAT12 Volume or CAT12 Surface also showed a 7% replication rate. When excluding clusters of size one, the cluster-wise replication rate was 33-35% for FSLANAT and FSLVBM, 29% for CAT12 Volume and 23% for CAT12 Surface and FreeSurfer.

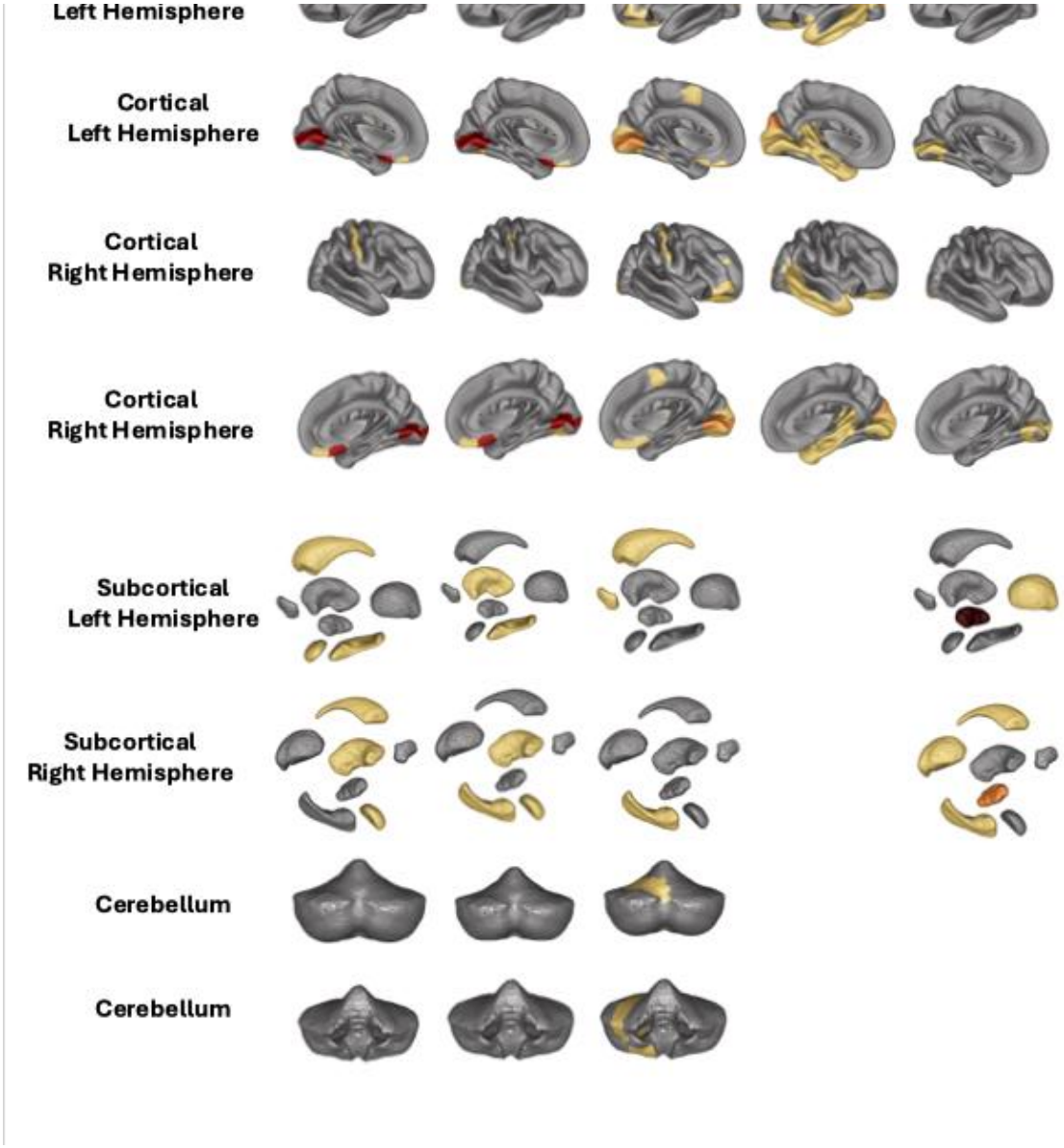
Location and robustness of replicated clusters in the brain

We checked whether the clusters were in same ROIs across all processing (**Figure 8**) focusing on the ones that replicate i.e., where the strength of evidence is stronger. For ‘maternal smoking around birth’, 42 ROI replicated, across all processing. All processing implicated the cortical region ‘Area_h0c1_ V1, _17, _Calcarine Sulcus’. In addition, the three volume-based processing identified clusters in the ROIs Area_Fo1_ (OrbitoFrontal Cortex),

Area_Fo2_(OrbitoFrontal Cortex), CGL_(Metathalamus), STN_(Subthalamus) and Frontal-to-Temporal-II_(GapMap). Overall, 45% of ROIs were robustly identified by at least 2 processing.

For Diabetes, Left Thalamus as well as Right Crus_I, right_VI regions were robustly associated across all processing. Out of 37 ROIs identified in total, 49% ROIs were identified by at least two processing. `High blood pressure` did not exhibit ROIs identified by all processing (**SFigure 19**), but 41% of the regions (out of 34 ROIs) were implicated by at least two processing. Finally, `Alcohol frequency` (**SFigure 20**) had the lowest rate of clusters identified by at least two processing (4/29).

We were not able to define replicating clusters for Alzheimer ‘s disease (there were no patients with Alzheimer’s disease recorded in the replication sample). However, in the discovery cohort, all processing exhibited an association in the right hippocampus and the right amygdala.



Discussion

Using 39,655 individuals from the UK Biobank, we compared 5 major processing of T1w brain MRI (**Figure 1**).

Figure 8 : Location replicating cluster per region of interest, for Maternal smoking.

We assessed whether the choice of processing influenced the total association (morphometricity) between grey matter and 29 traits of interest, the ability to detect associations in BWAS or to build performant prediction scores.

All processing contained some non-normal brain measurements

A first observation was that all processing contained some non-normal brain measurements (**Figure 1**), although it was more pronounced in FreeSurfer (higher kurtosis and positive skewness on average). These non-normal distributions may influence some analyses that are sensitive to outliers or tails of distributions.

All processing and brain parts are sensitive to imaging and body size confounders

Our results confirm that all considered image processings are sensitive to putative imaging confounders (**Figure 2**) (e.g., SNR, head position, head motion and age of the magnet) and to body size measurements (e.g., BMI), as indicated by large morphometricity estimates (20-80%). As for CAT12 Surface, its lower association with confounders (**Figure 2**) does not mean that it is less sensitive, in light of the lowest morphometricity it detects overall. Moreover, we showed that putative confounders are associated across the brain and the different types of measurements (**Figure 2**) (cortical, subcortical, cerebellum, cortical thickness and surface area). This suggests that imaging and body size measurements can create false positive associations in any brain part or grey matter measurement. Therefore, we recommend systematically controlling for these confounders in grey matter analyses, for example, by including them as covariates (in association studies or when evaluating predictors) and/or when selecting matched cases and controls.

FSLVBM, FSLANAT and FreeSurfer maximise the morphometricity, but each processing captures a unique signal

Across 19 traits of interest (and controlling for all putative confounders), we found that the choice of processing impacted morphometricity estimates. For example, CAT12 Surface exhibited the lowest morphometricity across traits (**Figure 3**). On the other hand, FSLVBM and FreeSurfer (all measurements) maximized the morphometricity (with estimates ~ 3x larger than those of CAT12 Surface (range of morphometricity = [0%-13%];**Figure 3,4**) in both the discovery and replication dataset. However, when combined, two processings resulted in a marked increase in morphometricity estimates, suggesting that each method captures a unique signal (**Figure 4**) that is absent in the other processing. For example, for Maternal smoking around birth, 75% of the signal FSLVBM and FreSurfer All Modalities captured was unique, while 25% was shared. We confirmed that the cerebellum alone (not measured in FreeSurfer), could not explain the unique signal captured by

FSLVBM. Our results shed new light on some of the lack of robustness to processing methods in neuroimaging results, as it indicates that some associations are only detectable with a specific processing. This has implications for several analyses. For example, multiverse analyses that seek to combine results across processings should expect processing-specific results, and they may miss relevant associations by focusing on consensus results. In addition, our results suggest that ensemble learning from multiple processings may maximize prediction accuracy by leveraging each processing-specific signal.

Correlated and non-normal brain measurements impact the false positive rate in BWAS

Using simulations, we observed (**Figure 5**) that processings were well calibrated to ensure FWER<5% after Bonferroni correction, except for FSLANAT, which demonstrated a small inflation of false positive (FWER=5.3%) due to non-normal distributions in grey-matter measurements. Our results confirmed that Bonferroni is overly stringent for the other processings (FWER in [0.5%;4.1%]), as it assumes that all tests are independent, even over correlated brain measurement. In particular, BWAS using CAT12 Surface was the most stringent (FWER=0.5% when applying Bonferroni's correction) which is consistent with the very high level of correlation between brain measurements (**Table 1**). However, the low FWER was also attributable to non-normal distributions in voxel-wise measurements from CAT12 Surface that caused a deflation of test statistics (**Figure 5, Figure 9**).

Our findings highlight the necessity of sensitivity analyses in BWAS: either by transforming distributions of brain measurements (e.g., rank inverse normal transformation) and/or visually inspecting associations at a vertex/voxel level to ensure the association is not caused by outliers or tails. We also found that normalizing the distribution of brain measurements (e.g., using RINT) can have a beneficial effect on the false positive rate after Bonferroni correction (it removes inflation of false positives in FSLANAT, and made BWAS analyses on CAT12 Volume less stringent -**SFigure 3**-.). Overall, our results highlight the need for multiple testing correction more efficient than Bonferroni's, as low FWER (<5%) leads to reduced statistical power. One may use the optimal significance threshold we derived from simulations, which calibrate the tests to FWER=5%. Other options include permutations[35] (albeit computationally expensive in large samples) and Random Field Theory[10] (although many variations exist, with no unified implementation for volume and surface-based processing).

The choice of processing influences the size, number and replication rate of brain regions detected in BWAS

We compared results using optimal significance thresholds for each processing, which ensured a comparable FWER of 5%. We found that the number of significant vertices/voxels varied widely between processing (between 3,945 [FSLANAT] and 13,674 [CAT 12 Volume]; **Table 2**). However, this difference was partly due to differences in cluster sizes, which tend to be larger in the presence of greater correlation across brain measurements (**Table 1**). For example, 7,077 voxels were detected with CAT12 Surface for only 190 clusters, which is consistent with the large amount of correlation between CAT12 Surface measurements (**Table 1**). The other processing each led to the detection of 544 (FreeSurfer) to 692 (FSLVBM) clusters across all traits (**Figure 6, Table 2**). Of note, the large number of significant vertex-wise measurements in FreeSurfer was driven by a handful of extremely large clusters (one of 2,547 vertices and two of 900 vertices associated with 'High blood pressure').

The vertex/voxel-wise effect sizes also varied between processings (**Figure 6**), which can contribute to differences in discoverability. Effect sizes were lower for FreeSurfer measurements, which may explain the smaller number of detected clusters compared to FSLVBM. On the other hand, they were maximal for CAT12 Surface, which can somewhat compensate its lower morphometricity, and the deflation of test statistics we have previously observed.

We sought to replicate associations in an independent UK Biobank sample, acquired at different sites. We found a good rate of replication with 17%-22% of the clusters replicating (**Table 2**), although this was lower for FreeSurfer (only 10%). When removing clusters containing a single vertex/voxel, 33-35% of clusters replicated using FSLANAT and FSLVBM, 29% using CAT12 Volume and 23% using CAT12 Surface and Freesurfer.

Over all traits considered, using FSLVBM led to identify more significant clusters, which displayed the largest replication rate. However, it is important to point out that other processing may maximize the number of identified regions depending on the trait. For example, FreeSurfer led to the identification of more brain regions associated with Parkinson's and Alzheimer's disease and CAT12 Volume maximized the number of discoveries with Maternal smoking around birth. This variability among processings could be further explored through

predictive modeling. We would expect that a greater number of associated regions may result in a better prediction accuracy.

Volume based processings yield better accuracy in brain-based prediction

First, we observed that detecting a greater number of significant clusters in BWAS, did not necessarily translate in greater prediction accuracy (in an independent UK Biobank sample). For example, maternal smoking was best predicted from grey-matter regions detected using FSLVBM and FSLANAT (**SFigure 15**), despite FreeSurfer and CAT12 surface yielding >50 extra clusters (**Figure 6**). We could not conclude about the prediction of Alzheimer's and Parkinson's disease (where FreeSurfer detected the most clusters), due to the small number of cases. Across the different traits, associations from volume-based processing gave the best prediction accuracy (**SFigure 15**), which is consistent with their higher replication rate. Of note, clusters of size one contributed little to the prediction from significant regions (**SFigure 17**).

We also found that the top (most significant) voxels/vertices do not always capture the full association from their clusters (**Figure 7**), either because the selection from p-value is not optimal, or because several signals can cohabit in a cluster. This is an important observation when trying to build parcimonious and interpretable predictors that rely on a minimal set of brain regions.

Next, we used BLU predictors that improved prediction accuracy by leveraging signals beyond significant brain regions (**Figure 7**). We showed that several traits could be predicted from grey-matter structure (e.g., Education age, index of multiple deprivations, number of children, age when first had sex, maternal smoking around birth, alcohol frequency, pains, sleeplessness, high blood pressure, stroke and diabetes; **Figure 7**). As previously, volume-based representations of the grey matter maximized prediction accuracy for a handful of traits (e.g., maternal smoking, alcohol frequency, high blood pressure and diabetes).

Overall, traits with the largest morphometricity tended to be better predicted. Yet, the proportion of morphometricity that BLUP scores could recover varied between traits, suggesting that some traits are harder to predict than others. For example, BLUP scores could account (on average) for 28% of the morphometricity of maternal smoking but only 20% of that of alcohol use frequency and 10% of the morphometricity of sleeplessness (**SFigure 16**). This suggests that the number of associated brain regions that contribute to the morphometricity, varies from one trait to the next, being larger for sleeplessness than for maternal smoking. For instance, some previous work displayed the dependence between the prediction R2 and p/N (p being the number of features and N the sample size), for different values of morphometricity and showed that the lower the

morphometricity the larger sample size is needed. In our study, it could be represented with both traits Education Age and Number of children, as they exhibited similar morphometricity estimates (on average 6%) but different percentage of morphometricity predicted (twice larger for `Number of children`), suggesting that Education age is associated with a more diffuse pattern of grey-matter regions.

Clusters comprising a single vertex/voxel should be treated with caution

We found that across all processing methods, clusters of size one replicated less than larger clusters, with FreeSurfer showing particularly low replication rates (2.1% compared to 7-9% for other methods). Moreover, building linear predictors using only size one cluster yielded very few significant results (**Figure 17**) and low prediction accuracy, indicating that while some clusters of this size may represent true signals, the majority may be false positives. Therefore, clusters of size one should be treated with caution, especially in BWAS that rely on FreeSurfer processing. Our results highlight the need for more systematic replication and external validation (e.g., via prediction), to validate neuroimaging findings.

Associated brain regions are not always robust to processing

We evaluated robustness to processing by comparing whether the clusters identified by the different processing belonged to the same ROIs. We observed that the three volume-based processing detected signals located mostly in the same areas (**SFigure 14**), especially between FSLVBM and FSLANAT, even if the overlap was never perfect. Overall, the overlap of significant ROIs was even lower between FSLVBM and Freesurfer, and CAT12 Surface had a poor agreement with all other processing (**SFigure 14**). Our results demonstrate that the choice of preprocessing pipeline contributes to variability in the results, which can explain the lack of robustness of some of the published results. However, non-robust results (i.e., that are only detected with a specific processing) remain of interest, especially since we showed that each processing captures a unique signal. It underscores the need to carefully interpret both robust and non-robust results, as different processing methods can capture unique signal. Moreover, understanding where results are least robust could help understand where the unique signal comes from, and could be used to develop better processing.

FSLVBM is a performant all-rounder

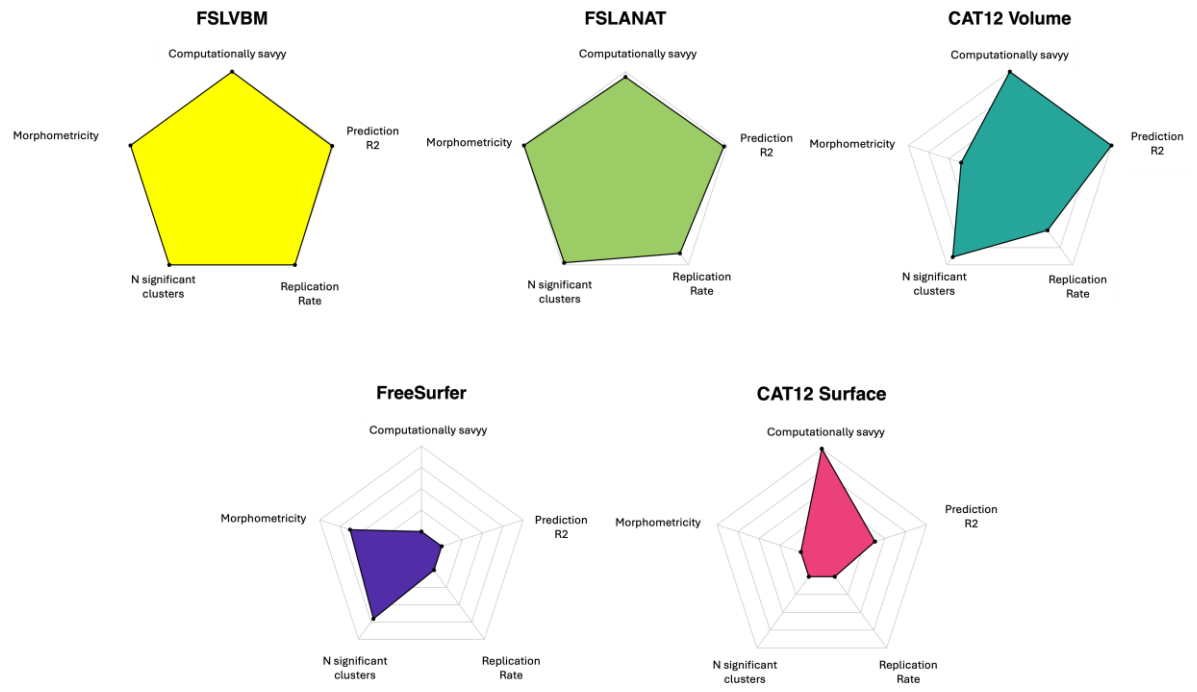


Figure 9 : Comparison of all 5 processing methods. This spider plots depicts the performance of all processing according to 5 metrics. Each color represents a processing method. The average computational cost ranges from 2 hours (FSLVBM, CAT12) to 10 hours (FreeSurfer). The prediction accuracy (R2) is computed across 12 traits shown in **Figure 7** top-right panel, for which prediction was significantly greater than chance. Average accuracy ranges from 3% (FreeSurfer) to 6% (FSLVBM). Morphometricity estimates were computed across all traits of interest (**Figure 3**) and ranges from 1.5% (CAT12 Surface) to 6% 5(both FSL processings). The number (N) of significant clusters across all traits ranges from 190 (CAT12 Surface) to 692 (FSLVBM), and rate of replicating clusters ranges from 23% (CAT12 Surface and FreeSurfer) to 35% (FSLVBM).

Overall, FSLVBM appeared to be the best single processing (green,**Figure 9**), as it maximized morphometricity estimates, prediction, discoverability and replication rate. Although FSLANAT (yellow, **Figure 9**) exhibited mostly similar performance (slightly lower replication rate), we found that it contained voxels with non-normal distributions, which might create false positive associations. In addition, FSLVBM relies on a well-established methodology and only requires about ~2h of processing time per MRI (vs. ~10 for FreeSurfer). Its only downside is that grey-matter density measured at each voxel is not easily interpretable, in contrast to cortical thickness and surface area that are extracted by CAT12 or FreeSurfer. Thus, we would recommend FSLVBM for future analyses of the UK Biobank, especially when several traits are investigated. For researchers interested in specific diseases/traits, our specific analyses (association, prediction) and other processing may be more suited on a case-by-case basis. We have summarized the pros and cons of all considered processing in **Table 3**.

Novel associations between grey-matter structure and traits of the UK Biobank

Beyond methodological considerations, our results have also highlighted several novel links between grey-matter structure and UK Biobank traits. For example, we are the first to report the morphometricity of several traits (e.g. Number of children, Age when first had sex), and to demonstrate that this overall association is partly robust to different grey-matter representation. Examining these traits are interesting to bridge social science and neuroscience: indeed, they may reflect underlying mechanisms involved in social behavior and help to understand how brain biology influences reproductive choices. We confirmed the implication of the cerebellum in several traits (Maternal Smoking **Figure 8**, Diabetes **SFigure 18**) which suggests its function is not limited to movement and motor functions. Specifically, for Diabetes, Right_Crus_I and Right_VI showed significant clusters across three volume-based processing. These results align with some previous work depicting changes in the cerebellar circuit in patients with type II Diabetes[36].

In addition, we identified hundreds of significant clusters associated with our traits of interest, which sheds light on some of the brain regions that contribute to the morphometricity, and some variations in clusters sizes : for instance, for High blood pressure, FreeSurfer exhibited 3 large clusters (**SFigure 19**, dark-red) located in the Left Putamen (N=2547 vertices), the Right Caudate (N=1006 vertices) and the Left Caudate (N=882 vertices).

Lastly, we built brain-based predictors that demonstrated significant prediction accuracy (R2 in [0.0;0.1]) for most traits and may be applied to independent samples, where the trait/disease was not collected or is not available.

Interestingly, `Maternal smoking` displayed the largest morphometricity across all processings (R2 in [0.04;0.1]), much larger than the association found with smoking status (R2 in [0.00001;0.0004]). Indeed, previous studies showed that prenatal exposure to maternal smoking can lead to problems in cognitive development (IQ for instance[37]) as well as abnormal brain development[38]. Moreover, maternal smoking may reflect not only prenatal toxin exposure but also postnatal environmental factors such as second-hand smoke, increased stress, or other behavioral correlates of maternal smoking, further strengthening the associations we detected[39]. Some previous research showed some association between smoking exposure during pregnancy and smaller grey matter volume in the inferotemporal and parahippocampal regions, and with smaller surface area in the parahippocampal and postcentral regions [40]. These results were partially consistent with our findings, as all processing showed replicating clusters in the right hippocampus and two processing in

the amygdala (**Figure 8**). We also identified some regions in the metathalamus and calcarine sulcus which is less common but could be involved in addiction[41].

	FSLVBM (Volume-based)	**FSLANAT (Volume-based)**	**CAT12 Volume (Volume-based)**	**CAT12 Surface (Surface-based)**	**FS All Modalities (Surface-based)**
Computational cost	~2h from raw nifti Bottleneck: Study specific template	~2h30 from UK Biobank segmented images Bottleneck: Study specific template	~2h from raw nifti Use prespecified template Standalone pipeline	~2h from raw nifti Uses Prespecified template Standalone pipeline	~10h if done from raw nifti But processing provided by the UK Biobank. ~20 mins Enigma-shape on top of FreeSurfer processing to extract subcortical measurements
Brain measurements	181,544 voxel-wise measurements Requires defining GM-mask Same MNI space than Harvard-Oxford atlas Different MNI space than Julich atlas	184,637 voxel wise measurements Requires defining GM-mask Same MNI space than Harvard-Oxford atlas Different MNI space than Julich atlas	192,483 voxel-wise measurements Requires defining GM-mask Different MNI space than Harvard-Oxford and Julich atlas	299,881vertex wise measurements Only cortical thickness measurement	654,002 vertex- wise measurements No cerebellum
Sensitivity to image confounders	64(49-76) % Contaminated	63(49-75) % Contaminated	55(36-71) % Contaminated	21(18-38) % Contaminated	55(37-61) % Contaminated
Association with body size	52(43-59) %	52(43-59) %	43(33-46) %	15(12-16) %	34(30-41) %
Morphometricity of traits of interest (median (min-max))	5.8 (0.3-35)	5.8 (0.0-36)	3.0 (0.1-32)	1.5 (0.0-13)	4.2 (0.0-33.7)

FWER using Bonferroni	4.1%	5.3% Small fraction of small positive associations due to voxel distributions	2.8%	0.5% Large number of vertices + large sum R2 = overly stringent	2% Large number of vertices, overly stringent
SumR2/Skewness/ **Kurtosis**	SumR2=42 Skewness=0.36 Kurtosis=3.0	SumR2 =48 Skewness=0.35 Kurtosis=3.1	SumR2=240 Skewness=0.24 Kurtosis=3.3 Important correlation among voxels	SumR2=1471 Skewness=-0.076 Kurtosis=3.5 Some heavy left tails (negative skewness) Very important correlation among vertices	SumR2=105 Skewness=1.1 Kurtosis=4.9 Some heavy right tails and leptokurtic (‘thin bells’) distributions
Number of significant clusters/voxels **Optimal threshold**	Voxels=4,277 Clusters=692	Voxels=3,945 Clusters=676	Voxels=13,674 Clusters=635	Voxels=7,077 Clusters=190	Voxels=10,936 Clusters= 544
% replicating rate with (and without) size one clusters	Clusters=18% (35%)	Clusters=17% (33%)	Clusters=19% (29%)	Clusters=22% (23%)	Clusters=10% (23%)
Proportion of clusters of size one and their replication rate (R=-)	60% R=7.2%	63% R=8.5%	45% R=7.4%	9.5% R=7.1%	61% R =2.1%
BLUP prediction R2 (%) **(median (min-max))**	0.21(0-8.9)	0.40 (0-9.5)	0.32(0-10.4)	0.12(0-3.9)	0.20(0-5.7)

Table 1 : Summary dataframe of all metrics used to compare processing

Limitations

Our results suggest that it is important to include imaging and body size covariates when performing brain imaging analysis in the UKBiobank. However, the list of recommended covariates may be refined in the future, as more confounders are evaluated. In addition, some processings seem to capture more information on average (e.g FSL and FreeSurfer). However, the best processing may depend on the trait of interest, and our results cannot be extrapolated to all traits and disorders.

We performed our analyses on the UK Biobank, and we do not know whether our results would generalize to other studies that have acquired images of different quality, or used different scanners and acquisition protocols, that may influence the amount of correlation or the distribution of measurements. In particular, one should know that UKBiobank is likely the most homogeneous research dataset in terms of image acquisition and it would be important to see whether our results would generalize to other datasets where MRI scanner are more varied or acquisition parameters are not harmonized. Furthermore, the heterogeneity of clinical routine datasets is usually even larger and, again, it remains to be studied which processing would be the most reliable in such context.

Our primary focus was on ensuring robustness of the findings -the ability to detect consistent results across different processing- in published results-, while also addressing reproducibility and replication. To encourage future work, we will make our code and summary statistics openly available, and the processed data will be returned to the UK Biobank for others to reproduce and extend our analyses. Additionally, we evaluated replication and prediction in an independent dataset within the same cohort.. That said, testing these findings in other cohorts with different scanners, phenotypes and datasets would be an interesting future research, offering the opportunity to further validate and expand upon our work.

We only include MRI processing pipelines that used default options or off-the-shelf implementations, even though many variations exist. For instance, FreeSurfer may be used with T1w only, CAT12 processing with a specific DARTEL (instead of the default IXI one) or using different resolution of the output map for voxel-based processing (different than 2mm voxels resolution). Thus, we cannot claim that one software or brain representation is superior to another, at this stage, as we have only explored some of the main branches of the multiverse. However, the results we reported here could be used to benchmark and guide the development of

processing pipelines. For example, processing that increase the morphometricity against our reference that used the same software are likely to lead to improved discoverability and brain-based prediction. More generally, processing that capture more morphometricity and less processing-specific signal would contribute to more robust results in neuroimaging analyses.

The interpretation of the grey-matter density in voxel-based representation remains a challenge. Unlike cortical thickness, which has a clear interpretation[42], grey-matter density is a more abstract measure, and its biological significance is less well-understood[2]. This makes the results harder to interpret from a biological perspective, which may justify using surface-based representations.

[1] S. Marek *et al.*, « Reproducible brain-wide association studies require thousands of individuals », *Nature*, vol. 603, n° 7902, p. 654‑660, mars 2022, doi: 10.1038/s41586-022-04492-9.
[2] J. Ashburner et K. J. Friston, « Voxel-Based Morphometry—The Methods », *NeuroImage*, vol. 11, n° 6, p. 805‑821, juin 2000, doi: 10.1006/nimg.2000.0582.
[3] G. Antonopoulos, S. More, F. Raimondo, S. B. Eickhoff, F. Hoffstaedter, et K. R. Patil, « A systematic comparison of VBM pipelines and their application to age prediction », *NeuroImage*, vol. 279, p. 120292, oct. 2023, doi: 10.1016/j.neuroimage.2023.120292.
[4] R. Botvinik-Nezer et T. D. Wager, « Reproducibility in Neuroimaging Analysis: Challenges and Solutions », *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 8, n° 8, p. 780‑788, août 2023, doi: 10.1016/j.bpsc.2022.12.006.
[5] N. Bhagwat *et al.*, « Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses », *GigaScience*, vol. 10, n° 1, p. giaa155, janv. 2021, doi: 10.1093/gigascience/giaa155.
[6] « Choice of Voxel-based Morphometry processing pipeline drives variability in the location of neuroanatomical brain markers | Communications Biology ». Consulté le: 22 mai 2024. [En ligne]. Disponible sur: https://www.nature.com/articles/s42003-022-03880-1
[7] J. Carp, « On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments », *Front. Neurosci.*, vol. 6, oct. 2012, doi: 10.3389/fnins.2012.00149.
[8] A. E. Fürtjes, J. H. Cole, B. Couvy-Duchesne, et S. J. Ritchie, « A quantified comparison of cortical atlases on the basis of trait morphometricity », *Cortex J. Devoted Study Nerv. Syst. Behav.*, vol. 158, p. 110‑126, janv. 2023, doi: 10.1016/j.cortex.2022.11.001.
[9] B. Couvy-Duchesne *et al.*, « A unified framework for association and prediction from vertex-wise grey-matter structure », *Hum. Brain Mapp.*, vol. 41, n° 14, p. 4062‑4076, oct. 2020, doi: 10.1002/hbm.25109.
[10] T. E. Nichols, « Multiple testing corrections, nonparametric methods, and random field theory », *NeuroImage*, vol. 62, n° 2, p. 811‑815, août 2012, doi: 10.1016/j.neuroimage.2012.04.014.
[11] A. Fry *et al.*, « Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population », *Am. J. Epidemiol.*, vol. 186, n° 9, p. 1026‑1034, nov. 2017, doi: 10.1093/aje/kwx246.
[12] F. Alfaro-Almagro *et al.*, « Confound modelling in UK Biobank brain imaging », *NeuroImage*, vol. 224, p. 117002, janv. 2021, doi: 10.1016/j.neuroimage.2020.117002.
[13] B. Grinde, « Sexual Behavior in Modern Societies: An Interdisciplinary Analysis », *Sex. Cult.*, vol. 25, n° 6, p. 2075‑2091, déc. 2021, doi: 10.1007/s12119-021-09865-2.
[14] M. Draps *et al.*, « Gray Matter Volume Differences in Impulse Control and Addictive Disorders—An Evidence From a Sample of Heterosexual Males », *J. Sex. Med.*, vol. 17, n° 9, p. 1761‑1769, sept. 2020, doi: 10.1016/j.jsxm.2020.05.007.
[15] « Sex Differences in Gray Matter Changes and Brain-Behavior Relationships in Patients with Stimulant

Dependence - PubMed ». Consulté le: 22 mai 2024. [En ligne]. Disponible sur: https://pubmed.ncbi.nlm.nih.gov/26133201/
[16] K. J. A. Johnston *et al.*, « Genome-wide association study of multisite chronic pain in UK Biobank », *PLoS Genet.*, vol. 15, n° 6, p. e1008164, juin 2019, doi: 10.1371/journal.pgen.1008164.
[17] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, et S. M. Smith, « FSL », *NeuroImage*, vol. 62, n° 2, p. 782‑790, août 2012, doi: 10.1016/j.neuroimage.2011.09.015.
[18] C. Gaser, R. Dahnke, P. M. Thompson, F. Kurth, E. Luders, et A. D. N. Initiative, « CAT – A Computational Anatomy Toolbox for the Analysis of Structural MRI Data », 13 juin 2022, *bioRxiv*. doi: 10.1101/2022.06.11.495736.
[19] R. Dahnke, R. A. Yotter, et C. Gaser, « Cortical thickness and central surface estimation », *NeuroImage*, vol. 65, p. 336‑348, janv. 2013, doi: 10.1016/j.neuroimage.2012.09.050.
[20] R. A. Yotter, R. Dahnke, P. M. Thompson, et C. Gaser, « Topological correction of brain surface meshes using spherical harmonics », *Hum. Brain Mapp.*, vol. 32, n° 7, p. 1109‑1124, juill. 2011, doi: 10.1002/hbm.21095.
[21] B. Fischl, « FreeSurfer », *NeuroImage*, vol. 62, n° 2, p. 774‑781, août 2012, doi: 10.1016/j.neuroimage.2012.01.021.
[22] B. A. Gutman, S. K. Madsen, A. W. Toga, et P. M. Thompson, « A Family of Fast Spherical Registration Algorithms for Cortical Shapes », in *Multimodal Brain Image Analysis*, vol. 8159, L. Shen, T. Liu, P.-T. Yap, H. Huang, D. Shen, et C.-F. Westin, Éd., in Lecture Notes in Computer Science, vol. 8159. , Cham: Springer International Publishing, 2013, p. 246‑257. doi: 10.1007/978-3-319-02126-3_24.
[23] « Shape matching with medial curves and 1-D group-wise registration | IEEE Conference Publication | IEEE Xplore ». Consulté le: 22 mai 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/6235648
[24] F. Zhang *et al.*, « OSCA: a tool for omic-data-based complex trait analysis », *Genome Biol.*, vol. 20, n° 1, p. 107, mai 2019, doi: 10.1186/s13059-019-1718-z.
[25] K. Amunts, H. Mohlberg, S. Bludau, et K. Zilles, « Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture », *Science*, vol. 369, n° 6506, p. 988‑992, août 2020, doi: 10.1126/science.abb4588.
[26] N. Makris *et al.*, « Decreased volume of left and total anterior insular lobule in schizophrenia », *Schizophr. Res.*, vol. 83, n° 2‑3, p. 155‑171, avr. 2006, doi: 10.1016/j.schres.2005.11.020.
[27] J. Diedrichsen, J. H. Balsters, J. Flavell, E. Cussans, et N. Ramnani, « A probabilistic MR atlas of the human cerebellum », 2009.
[28] G. Grabner, A. L. Janke, M. M. Budge, D. Smith, J. Pruessner, et D. L. Collins, « Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults », *Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, vol. 9, n° Pt 2, p. 58‑66, 2006, doi: 10.1007/11866763_8.
[29] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, et J. Lancaster, « A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM) », *NeuroImage*, vol. 2, n° 2, p. 89‑101, juin 1995, doi: 10.1006/nimg.1995.1012.
[30] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, et J. C. Gee, « A reproducible evaluation of ANTs similarity metric performance in brain image registration », *NeuroImage*, vol. 54, n° 3, p. 2033‑2044, févr. 2011, doi: 10.1016/j.neuroimage.2010.09.025.
[31] J.-D. Tournier *et al.*, « MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation », *NeuroImage*, vol. 202, p. 116137, nov. 2019, doi: 10.1016/j.neuroimage.2019.116137.
[32] M. R. Sabuncu *et al.*, « Morphometricity as a measure of the neuroanatomical signature of a trait », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, n° 39, p. E5749-5756, sept. 2016, doi: 10.1073/pnas.1604378113.
[33] B. Couvy-Duchesne *et al.*, « Parsimonious model for mass-univariate vertexwise analysis », *J. Med. Imaging Bellingham Wash*, vol. 9, n° 5, p. 052404, sept. 2022, doi: 10.1117/1.JMI.9.5.052404.
[34] F. Alfaro-Almagro *et al.*, « Confound modelling in UK Biobank brain imaging », *NeuroImage*, vol. 224, p. 117002, janv. 2021, doi: 10.1016/j.neuroimage.2020.117002.
[35] T. E. Nichols et A. P. Holmes, « Nonparametric permutation tests for functional neuroimaging: A primer with examples », *Hum. Brain Mapp.*, vol. 15, n° 1, p. 1‑25, oct. 2001, doi: 10.1002/hbm.1058.
[36] P. Fang *et al.*, « Changes in the cerebellar and cerebro-cerebellar circuit in type 2 diabetes », *Brain Res. Bull.*, vol. 130, p. 95‑100, avr. 2017, doi: 10.1016/j.brainresbull.2017.01.009.
[37] S. Ak et S. M, « Cigarette smoking during pregnancy », *Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob.*, vol. 10, n° 2, févr. 2008, doi: 10.1080/14622200701825908.
[38] R. D. Eiden, D. S. Molnar, D. A. Granger, C. R. Colder, P. Schuetze, et M. A. Huestis, « Prenatal tobacco exposure and infant stress reactivity: role of child sex and maternal behavior », *Dev. Psychobiol.*, vol. 57, n° 2, p. 212‑225, mars 2015, doi: 10.1002/dev.21284.
[39] M. D. Cornelius et N. L. Day, « The Effects of Tobacco Use During and After Pregnancy on Exposed

Children », *Alcohol Res. Health*, vol. 24, n° 4, p. 242‑ 249, 2000.
[40] M. O. Ekblad, P. Ngum, H. Merisaari, V. Saunavaara, R. Parkkola, et S. Setänen, « Maternal smoking during pregnancy negatively affects brain volumes proportional to intracranial volume in adolescents born very preterm », *Front. Hum. Neurosci.*, vol. 16, janv. 2023, doi: 10.3389/fnhum.2022.1085986.
[41] A. S. Huang, J. A. Mitchell, S. N. Haber, N. Alia-Klein, et R. Z. Goldstein, « The thalamus in drug addiction: from rodents to humans », *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 373, n° 1742, mars 2018, doi: 10.1098/rstb.2017.0028.
[42] B. Fischl et A. M. Dale, « Measuring the thickness of the human cerebral cortex from magnetic resonance images », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, n° 20, p. 11050‑ 11055, sept. 2000, doi: 10.1073/pnas.200033797.