Parsimonious model for mass-univariate vertexwise analysis

Baptiste Couvy-Duchesne,[a,b,*] Futao Zhang,[a] Kathryn E. Kemper,[a] Julia Sidorenko,[a] Naomi R. Wray,[a,†] Peter M. Visscher,[a,†] Olivier Colliot,[b,†] and Jian Yang[a,c,d,†]

[a]University of Queensland, Institute for Molecular Bioscience, St. Lucia, Queensland, Australia
[b]Sorbonne University, Paris Brain Institute (ICM), CNRS, INRIA, INSERM, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France
[c]Westlake University, School of Life Sciences, Hangzhou, China
[d]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China

Abstract

Purpose: Covariance between gray-matter measurements can reflect structural or functional brain networks though it has also been shown to be influenced by confounding factors (e.g., age, head size, and scanner), which could lead to lower mapping precision (increased size of associated clusters) and create distal false positives associations in mass-univariate vertexwise analyses.

Approach: We evaluated this concern by performing state-of-the-art mass-univariate analyses (general linear model, GLM) on traits simulated from real vertex-wise gray matter data (including cortical and subcortical thickness and surface area). We contrasted the results with those from linear mixed models (LMMs), which have been shown to overcome similar issues in omics association studies.

Results: We showed that when performed on a large sample ($N = 8662$, UK Biobank), GLMs yielded greatly inflated false positive rate (cluster false discovery rate >0.6). We showed that LMMs resulted in more parsimonious results: smaller clusters and reduced false positive rate but at a cost of increased computation. Next, we performed mass-univariate association analyses on five real UKB traits (age, sex, BMI, fluid intelligence, and smoking status) and LMM yielded fewer and more localized associations. We identified 19 significant clusters displaying small associations with age, sex, and BMI, which suggest a complex architecture of at least dozens of associated areas with those phenotypes.

Conclusions: The published literature could contain a large proportion of redundant (possibly confounded) associations that are largely prevented using LMMs. The parsimony of LMMs results from controlling for the joint effect of all vertices, which prevents local and distal redundant associations from reaching significance.

Keywords: structural brain MRI; vertex-wise processing; linear mixed model; association; brain mapping.

Paper 21309SSR received Nov. 24, 2021; accepted for publication Apr. 26, 2022; published online May 20, 2022.

1 Introduction

Brain MRI scans can generate hundreds of thousands of vertex/voxelwise measurements per individual, which can be linked to other measured traits/diseases using mass univariate vertex/voxelwise association analyses. Results of association analyses (and subsequent

*Address all correspondence to Baptiste Couvy-Duchesne, b.couvyduchesne@uq.edu.au

†These authors contributed equally.

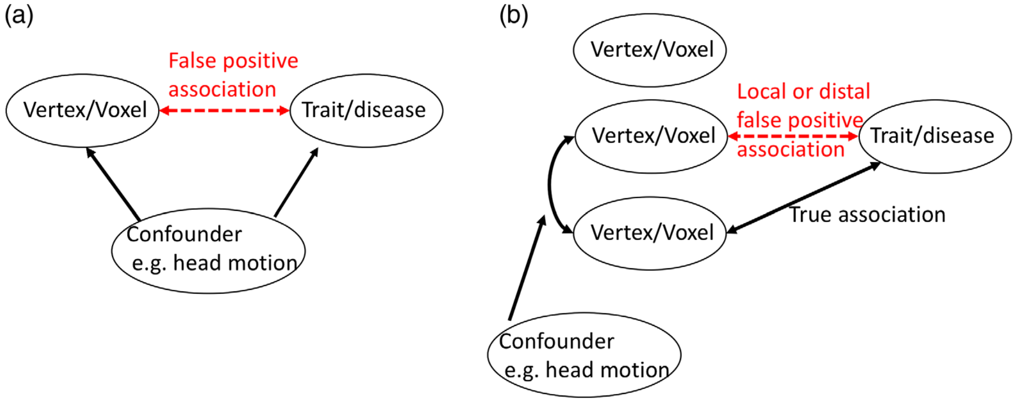
(a)
False positive association
Vertex/Voxel
Trait/disease
Confounder e.g. head motion
(b)
Vertex/Voxel
Vertex/Voxel
Vertex/Voxel
Local or distal false positive association
Trait/disease
True association
Confounder e.g. head motion

Fig. 1 (a) Illustration of the traditional confounding paradigm (and (b) of the confounding that may arise in association studies performed across correlated brain features. One-sided arrows represent a causal effect, and two-sided arrows a correlation.

follow-up analyses) can shed light on the brain networks or cell composition relevant for the trait/disease and may be leveraged for brain-feature-based phenotype prediction. However, brain measurements may exhibit a pattern of correlation, owing to factors (e.g., head size, MRI scanner/artifact,[1] or demographics[2]) that can generate confounded brain-trait associations. Induced local correlations with a true brain biomarker can generate a smear of association (i.e., a cluster of associated vertices), which may limit the precise localization of the directly associated regions. On the other hand, long-range vertex correlations caused or inflated by factors irrelevant to the trait of interest, may be more prejudicial, as they can yield distal false positives (Fig. 1).

Two approaches can be used to limit the inflation of false positives described above. One is to control for the confounders in the association testing, although it requires knowledge and measurement of the factors influencing (or more generally associated with) the covariance between brain measurements. Note that these factors can overlap with traditional confounders of neuroimaging studies (e.g., head size, age, sex, and head motion), and additional confounders are being identified as sample sizes increase.[3] Another correction strategy is to control for the other vertices in the association testing, to remove the signal that could be attributed to another brain vertex or region. The difficulty of such approach is that typically, the number of vertex/voxelwise measurements (p) far exceeds the number of participants (N) in the study. The $p \gg N$ paradigm implies that the marginal joint associations with all p vertices cannot be estimated in a single general linear model (GLM).

Statistically, the challenge of mass univariate vertexwise analyses resembles that of genome-wide association studies (GWAS) or methylation-wide associations studies (MWAS), which aim to identify genomic regions associated with a phenotype in the presence of correlated features (i.e., genetic variants or DNA methylation probes). Several studies have demonstrated that feature correlation [i.e., linkage disequilibrium (LD) or population structure in genetics] can result in inflated false positive rate (FPR),[4–6] even more so when the sample size increases.[5] This led GLMs to be replaced by linear mixed models (LMMs)[6–8] that co-vary out all features by fitting them as random effects. LMMs have been shown to better control the inflation of false positive associations arising from LD or correlation between probes and to minimize the occurrence of false positives in both GWAS and MWAS.[6,7,9]

LMMs are commonly used in neuroimaging to model longitudinal data.[10] Instead, we rely here on a formulation that allows fitting the high-dimensional brain image as a single random-effect. Such LMMs allow estimation of the overall degree of association between a trait and a high-dimensional brain image, coined "morphometricity" in the context of structural brain measurements.[11,12] Recently, we have shown that a single LMM framework was suited to estimate morphometricity in large datasets, to draw links between traits through their associations with similar brain structure (gray-matter correlation), and to build brain-based predictors.[12] The LMMs we propose here complement our previous work by identifying the vertices/voxels that contribute to the morphometricity and phenotype prediction.

Here, we sought to evaluate whether the inflation of false positives observed in omics data is also present in neuroimaging data. In the first part of the analysis, we performed extensive simulations of continuous phenotypes from real gray-matter data to quantify FPR as well as statistical power, mapping precision and prediction accuracy achieved from mass-univariate analyses. We compared the performances of the current state-of-the-art GLMs to that of LMMs inspired by omics association studies. In the second part, we sought to characterize the brain regions associated with real phenotypes (i.e., age, sex, BMI, fluid IQ, and smoking status) that previously exhibited significant morphometricity,[12] to confirm the results obtained on simulated traits. Our analyses relied on 14,451 MRI images collected by the UK Biobank (UKB), one of the largest brain imaging initiative.[13]

1.1 *Novelties and Contribution*

The novelties and contributions of our paper are as follows:

- We propose LMMs for brain mapping, inspired from those using in genetics, which aims at overcoming false positive issues found in standard analyses.
- By controlling for all brain measurements (fitted as a random effect), the LMMs remove redundant associations leading to more parsimonious results.
- We demonstrate that, compared with the current state-of-the-art, the LMMs minimize FPR while also maximizing power, mapping precision, and prediction accuracy.

2 Material and Methods

2.1 *Models of Mass-Univariate Vertexwise Analyses*

First, we considered five GLMs that differ in term of covariates used when estimating the association (b_i) between the trait and the i'th (standardized) vertexwise measurement ($\mathbf{X_i}$). They can be written under the form:

$$\mathbf{y} = \mathbf{Zc} + \mathbf{X_i}b_i + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is the vector of phenotype for the N individuals, $\mathbf{Z}$ is a matrix of size $N \times q$ of q covariates, and $\mathbf{c}$ is a vector of the q fixed effects.

The five GLMs are differentiated as follows: (1) GLM with no covariates (no covariates), (2) GLM including the most commonly used covariates in similar analyses: age, sex, and intracranial volume (ICV) (age, sex, ICV corrected), (3) and (4) GLMs including 5 and 10 principal components (PCs) of gray-matter variation, respectively (5 global PCs, 10 global PCs), (5) GLM including 10 PCs specific to the measurement type (cortical thickness, cortical surface, subcortical thickness, or subcortical surface area), referred to as "10 modality specific PCs." Gray-matter PCs capture the main axes of covariations between vertices, and we expect that by controlling for them we may be able to remove unmeasured or unknown factors contributing to long-range correlation between vertices (which might include demographics, MRI machine, head motion, software update, processing option, etc.). Note that PCs from genetic data are commonly used in GWAS to limit the FPR of GLMs analyses[14] but are rarely used in neuroimaging analyses. The difficulties of PC correction are to determine the optimal number of PCs, which controls for confounding effects without removing signals of interest. In practice, this may prove extremely difficult considering that the optimal number of PCs could depend on the trait/variable of interest, and that PCs are notoriously hard to interpret and have not been comprehensively investigated on these data. Thus, we arbitrarily chose two scenarios with the first 5 or 10 PCs. In addition, GLMs without covariates are also very rare, but worth considering to appreciate the effect of including covariates.

Finally, we considered three LMMs that can be seen as extensions of the previous approaches in that they further control for all vertexwise measurements. The first LMM model (LMM global BRM), analogous to the MOA (MLM-based omic association) model,[6] can be written as

$$\mathbf{y} = \mathbf{X_i}b_i + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2}$$

Here, $\mathbf{X}$ is the $N \times p$ matrix of all standardized vertexwise measurements, $\boldsymbol{\beta}$ is the $p \times 1$ vector of joint vertex-trait associations. $\boldsymbol{\beta}$ is a vector of random effects, allowing for $p > N$, with $\boldsymbol{\beta} \sim \mathcal{N}(0, \mathbf{I}\sigma_{\beta}^2)$, and $\boldsymbol{\varepsilon}$ is the error term assumed to follow $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma_{\boldsymbol{\varepsilon}}^2)$. $\sigma_{\boldsymbol{\beta}}^2$ and $\sigma_{\boldsymbol{\varepsilon}}^2$ are the variances of the random effects $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. The variance-covariance matrix for $\boldsymbol{y}$ is $\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{XX}'\sigma_{\boldsymbol{\beta}}^2 + \mathbf{I}\sigma_{\boldsymbol{\varepsilon}}^2 = \boldsymbol{B}p\sigma_{\beta}^2 + \mathbf{I}\sigma_{\varepsilon}^2$. Here, we regard $\mathbf{B} = \mathbf{XX}'/p$ as the brain relatedness matrix and $p\sigma_{\boldsymbol{\beta}}^2$ the morphometricity (proportion of phenotypic variance captured by all vertices).[15]

We considered a second LMM ("LMM with covariates") that includes known covariates (age, sex and ICV) fitted as fixed effects. Thus, we can separate the effect of the random effects from that of the known covariates on the results. The model becomes:

$$\mathbf{y} = \mathbf{Zc} + \mathbf{X}_i \mathrm{b}_i + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

Our third LMM ("LMM multi. BRM") includes 4 random effects $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$, each corresponding to a type of vertices (cortical thickness, cortical surface area, subcortical thickness and subcortical surface area).

$$\mathbf{y} = \mathbf{X}_i b_i + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{X}_4\boldsymbol{\beta}_4 + \boldsymbol{\varepsilon}. \tag{4}$$

This more general LMM allows the distribution of effect sizes to differ based on vertex type, rather than enforcing a single distribution over all types of measurements.[15] Note that each random effect takes up a single degree of freedom meaning that LMMs and GLMs have a comparable (large) numbers of degrees of freedom given the same sample size.

2.2 *Statistical Testing and Multiple Comparison*

We performed a χ^2 test of the association between a vertex $(\mathbf{X}_i)$ and the phenotype using that, for large sample size N, $(\frac{\mathrm{b}_i}{SE(\mathrm{b}_i)})^2 \sim \chi_1^2$ under the null hypothesis of no association. In each model (GLM or LMM), we accounted for multiple testing over the vertices using Bonferroni correction, thus setting a brain-wide significance threshold of $0.05/652{,}283 = 7.6\mathrm{e}-8$. We chose the straightforward Bonferroni correction over random field theory (RFT)[16] as RFT requires stationarity and a smooth mesh of vertex-wise residuals, which is unlikely to be the case here (we did not apply kernel smoothing on the data as it reduced the estimated morphometricity of the UKB phenotypes[15]). In addition, RFT is not currently implemented to be performed using residuals of LMMs or across several surfaces and type of measurements. Bonferroni correction is expected to be conservative under the null hypothesis (no association) because the correlations between vertices means that the effective number is tests lower than the number of tests conducted and used for the Bonferroni correction.

2.3 *MRI Image Processing*

MRI images were mostly collected in Cheadle (for 96% of the sample) and Newcastle using a 3T Siemens Skyra machine (software platform VD13) and a 32-channel head coil[13] (see Supplementary 1 in the Supplemental Material, for MRI sequence details).

We processed the T1w and T2 FLAIR images together to enhance the tissue segmentation in FreeSurfer 6.0,[17] which should result in a more precise skull stripping and pial surfaces definition. When the T2 FLAIR was not acquired or not usable, we processed the T1w image alone, though a recent report showed this results in systematic differences in cortical thickness.[18] This may represent a source of noise in the data, albeit it was limited in term of number of individuals (see quality control, Supplementary 1 in the Supplemental Material 1). We extracted vertex-wise data mapping cortical surface area and thickness ("recon-all" processing in FreeSurfer) and used the maximal resolution allowed by the software (fsaverage atlas - unsmoothed). In short, FreeSurfer segments the gray/white and gray/cerebrospinal fluid borders, which delimits the gray-matter. Surfaces are mapped onto a spherical atlas to align the cortical folding patterns of the individuals, and a tessellation is applied. Cortical thickness is calculated as the closest distance from the two gray-matter boundaries, for each vertex on the tessellated surface.[19]

Surface area is measured as the mean area of all faces that meet at a particular vertex, on the gray/white matter surface.[20] We previously showed that this cortical processing maximized the morphometricity for a wide range of phenotypes.[15] In other words, this cortical processing maximized the information retained by the processed MRI images. In addition, we applied the ENIGMA-shape processing,[21,22] where subcortical structures segmented in FreeSurfer are projected onto spherical atlases to quantify vertex-wise radial thickness and log Jacobian determinant,[21,22] which is analogous to a surface area.[23] This yielded a vertex-wise characterization of the hippocampus, putamen, amygdala, thalamus, caudate, pallidum, and accumbens. Overall, the imaging data used in the analyses comprised 652,283 vertex measurements per individual: 299,009 for cortical thickness, another 299,034 for cortical surface area, 27,120 for subcortical thickness and 27,120 for subcortical surface area.

In a post-hoc analysis, we also utilized smoothed cortical data (surface based kernel with FWHM = 20 mm), in order to evaluate the robustness of our results to variation in the MRI processing.

2.4 *Main Sample for Simulation and Discovery*

Our final sample comprised 9890 adults with complete cortical and subcortical data, aged 62.5 on average ($\text{SD} = 7.5$, range 44.6 to 79.6) with slightly more (52.4%) female participants (see Supplementary 1 in the Supplemental Material for participant inclusion and exclusion). Of note, 341 participants did not have an exploitable T2 image.

We performed a stringent quality control (QC) to exclude one of each pair of individuals whose brains were too similar or dissimilar relative to most other individuals, resulting in 1228 exclusions (12.4% of the sample). The main reason for this exclusion was to prevent bias in the LMM estimates, although it should also remove individuals flagged as outliers by other QC criteria (e.g. 80.6% of the participants processed using T1w only, spike-like cortical parcellation in FreeSurfer)[12] (see Supplementary 1 in the Supplemental Material for more details on QC). Importantly, all analyses were performed on the same list of individuals (post QC) to ensure that performance of the models would be comparable.

2.5 *Independent Samples for Prediction and Replication*

Our first independent sample included an additional 4942 participants of the UKB with a T1w image (downloaded in May 2018, most participants also had an exploitable T2w). The final sample ($N = 4160$ after processing and QC) was on average 63.1 years old ($\text{SD} = 7.46$, range 46.1 to 80.3) with 52.1% of females.

In addition, we used the OASIS3 (Open Access Series of Imaging Studies) sample[24] to evaluate the generalizability of the prediction. The OASIS3 dataset gathers several longitudinal MRI studies conducted in the Washington University Knight Alzheimer Disease Research Center over the past 15 years. Our final sample included 1006 unique participants after processing based on T1w images and QC. When several visits were available for a participant, we selected the one with the most phenotypic information. Participants were 71.1 years old on average ($\text{SD} = 9.18$, range 42.6 to 95.7) and mostly female (55.5%). Almost a quarter of the participants (23.6%) had a diagnosis of Alzheimer's disease at the time of imaging.

2.6 *Mass-Univariate Analyses on Simulated Phenotypes*

2.6.1 *Simulation of phenotypic traits from real gray-matter data*

We simulated phenotypic traits from the UKB processed (standardized) gray-matter data, instead of relying on synthetic/simulated images. This approach ensures the vertexwise data retains a realistic correlation structure. In addition, our framework includes simulation of the phenotype under not only the null hypothesis (H0) that no vertex is associated with the phenotype but also the alternative hypothesis (H1) that a set of vertices are truly associated with the phenotype.

First, we randomly selected a set of associated vertices and drew their relative effects from a normal distribution. We then calculated the simulated phenotypes as a linear combination of the individuals' vertex values and noise.[6] We considered three scenarios that differ in terms of number of associated vertices and total association with the phenotype. This global association between gray-matter measurements and a trait has been coined morphometricity[11,15] and may be expressed as the proportion of the trait variance (R^2) captured by the vertexwise measurement. Our scenarios were: (i) 10 associated vertices accounting for a phenotype morphometricity of $R^2 = 0.20$ (i.e., 20% of the trait variance); (ii) 100 associated vertices with $R^2 = 0.50$; (iii) 1000 vertices with $R^2 = 40\%$. For each scenario, we simulated 100 phenotypes.

In follow-up analyses, we simulated phenotypes using the same parameters, this time restricting the associated vertices to a single type of measurement. This allowed evaluation of the specificity of each type of measurement, which possess a unique correlation pattern. In addition, this ensures our phenotypes were not associated with cortical vertices only, which represent 90% of the vertexwise measurements.

To evaluate the effect of smoothing on our results, we simulated phenotypes from smoothed brain maps. For the ease of computation, we restricted the analysis of smoothed data to the case of 10 associated vertices ($R^2 = 0.2$). We kept the same associated vertices (and weights) as in the previous simulation from unsmoothed data. Finally, we randomly simulated 100 "null" traits, to evaluate the calibration of the models under the null hypothesis of no association. All simulations were generated using the OSCA software.[6]

2.6.2 *Inflation of test statistics*

First, we compared the empirical distribution of χ^2 statistics to the expected distribution, which is assumed to follow a $\chi^2(1)$ for non-associated (null) vertices. We considered the ratio of empirical over expected median χ^2, known as the inflation factor (λ), which is expected to be equal to one across nonassociated vertices. We also used the nominal FPR defined as the proportion of null vertices with p-values < 0.05 (expected to be 0.05). Correlation between associated and null vertices (e.g., due to confounding factors) typically result in an inflation of test statistics, which may cause null vertices to reach significance in mass-univariate analyses.

2.6.3 *Discoverability and mapping precision*

First, we quantified the model discoverability using the true positive rate (TPR) defined as the proportion or truly associated vertices reaching significance (after Bonferroni correction). Importantly, the TPR is dependent on the FPR, which can limit comparison across models (see Sec. 3.2.4). In addition, we quantified the mapping precision of mass-univariate analyses by reporting the median size of the true positive (TP) clusters. We defined TP clusters as sets of significant contiguous vertices of the mesh that contain a TP vertex.

2.6.4 *False positives and statistical power*

We reported the familywise error rate (FWER) defined as the proportion of replicates with at least one false positive vertex (null vertex significant after Bonferroni correction). In the presence of strong correlation between neighboring vertices, it is statistically difficult to separate a TP vertex from the flanking ones, thus we can expect an FWER greater than 5%. Hence, we also reported the cluster FWER defined as the proportion of replicates with at least one false positive cluster. FWER is more stringent than false discovery rate (FDR), implying that any false positives that remain after FWER correction would also be observed using FDR.

To account for the models' differences in FWER, we further reported the statistical power, defined as the TPR for a set risk alpha. We chose cluster FWER < 0.2, which was easier to achieve than the traditional FWER < 0.05, as we enforced comparable FWER by iteratively lowering the significance threshold, for each of the models (Appendix 2 in the Supplemental Material). The choice of risk alpha does not impact the relative performance of the models, and we can expect models best powered for FWER < 0.2 to also be best powered at other FWER levels.

Finally, we reported the proportion of false positive clusters out of all significant clusters (cluster FDR). We labeled false positive clusters, the groups of significantly associated, contiguous vertices that did not contain a TP association.

In follow up analyses, we simulated associations on a single type of vertexwise measurements, to evaluate the probability of false positive (FWER) arising on the same type of measurements, other types of measurements as well as contralateral regions.

2.6.5 *Prediction from significant vertices*

We evaluated the prediction accuracy achieved from the brain regions reaching significance, in the different mass-univariate models. We used prediction as a meta-criterion to compare the model performances, as it is dependent on power, true and false positives, and association effect sizes. We selected the most significant vertex in each cluster and constructed a linear predictor using association weights [$\hat{b}_i$, see Eqs. (1) and (2)] estimated from the different mass-univariate analyses. Because some significant clusters might contain several independent signals, we also built predictors that included all significant vertices. We evaluated the prediction of in the independent UKB and OASIS3 samples.

2.6.6 *Mass-univariate analyses of UK Biobank phenotypes*

Next, we performed mass-univariate vertexwise analyses on five UKB phenotypes that showed significant replicated morphometricity:[15] age, sex, BMI, smoking status, and fluid intelligence. We used the raw fluid intelligence score provided by the UKB, a nonstandard test which has demonstrated some reliability in a test–retest analysis.[25]

For each UKB phenotype and model, we reported the number of significant vertices, number of significant clusters as well as their sizes. We defined significance using a Bonferroni significance threshold of $0.05/(652{,}283 \times 5) = 1.5\mathrm{e}^{-8}$, which accounts for the total number of tests performed. For those phenotypes, the true pattern of association is unknown which prevents evaluation of the FPR (or power) of the different approaches. However, false positives or redundant associations should not improve prediction accuracy. In this regard, we evaluated each GLM or LMM model in both the UKB replication and OASIS3 datasets. As above, we used linear predictors and reported the prediction accuracy (correlation) controlling for age, sex, ICV, and site. In OASIS3, we also corrected for clinical status (Alzheimer's disease and mild cognitive impairment).

3 Results

3.1 *Phenotypes Simulated Under H0*

We found that all GLM and LMM models behaved well under the null hypothesis, as indicated by no inflation of test statistic, FPR, or of FPR (FWER). As expected under a stringent Bonferroni correction, all approaches were conservative as indicated by FWER < 3% (Fig. S1 in the Supplemental Material).

3.2 *Phenotypes Simulated Under H1*

3.2.1 *Inflation of test statistics*

First, we quantified whether we could observe an inflation of test statistics on the vertices not associated with the simulated phenotypes. As expected in presence of correlation between truly associated and null vertices, we observed a global inflation of (median) test statistics when using GLMs (Fig. 2; Table S1 in the Supplemental Material). This was confirmed by an FPR greater than 5% for all GLM models even though controlling for covariates or PCs reduced the inflation of test-statistics compared with the "no covariates" GLM. In comparison, LMMs appropriately controlled the inflation of test statistics on null vertices ($\lambda < 1$ and FDR < 5%; Fig. 2; Table S1 in the Supplemental Material).

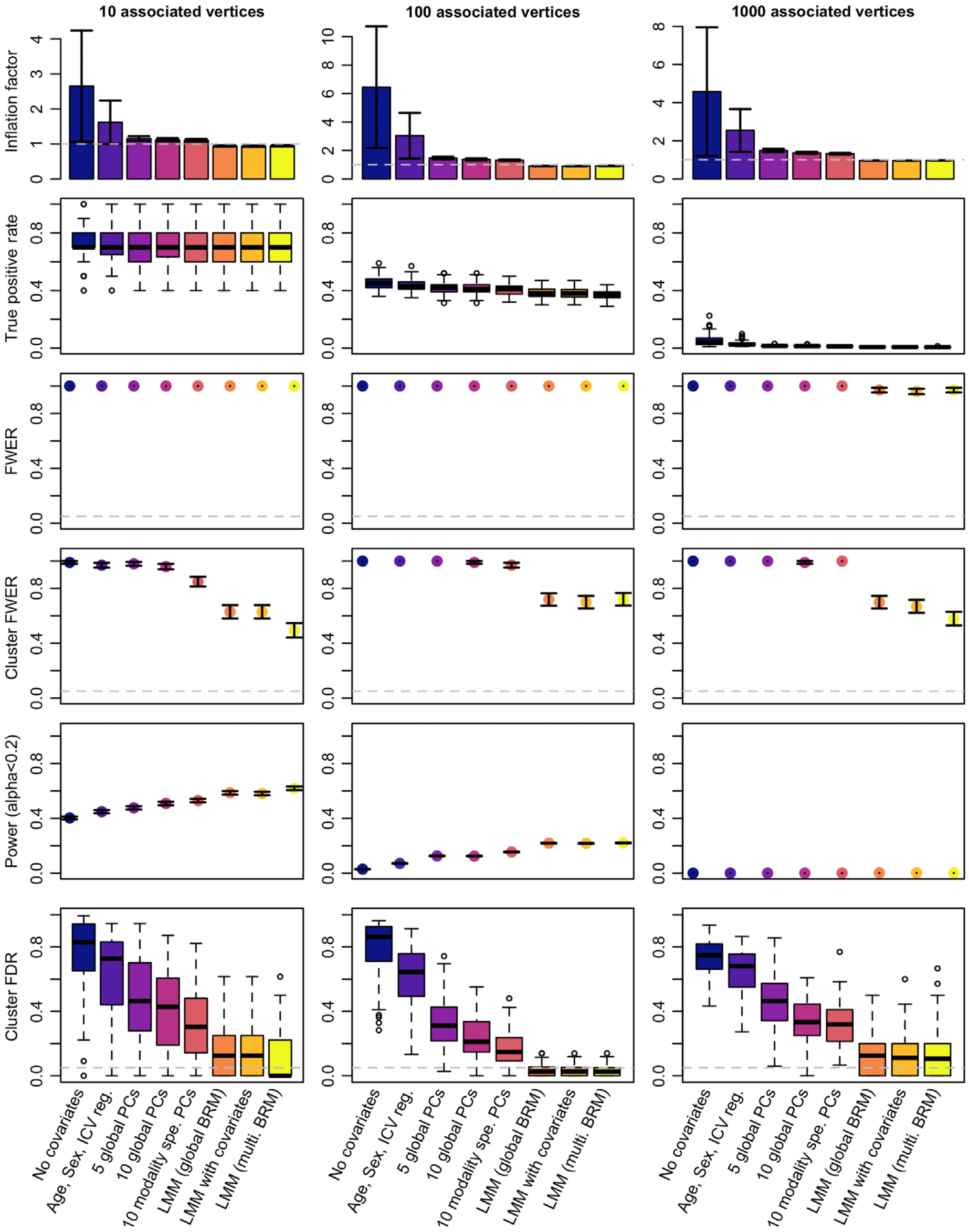

Fig. 2 Performance of GLMs and LMMs for mass-univariates vertexwise analyses: test inflation, statistical power, and FPR. The columns correspond to the different scenarios considered when simulating traits. We simulated 100 phenotypic traits for each scenario. Bars represent ±SE across the 100 replicates. Clusters are composed of groups of contiguous vertices each significantly associated with the phenotype (after Bonferroni correction). We labeled them as false positives if they did not include a TP association.

3.2.2 *True positive rate*

First, we confirmed that the TPR (after Bonferroni correction) was dependent on the scenarios that corresponded to different effect sizes for the vertices. For example, about 70% of the truly associated brain regions were detected in the case of a simple trait (10 associated vertices each accounting for 2% of the phenotypic variance on average). On the other hand, <5% of the associated brain regions were identified for the most complex phenotypes (scenario 3, 1000 vertices each accounting for 0.04% of variance, Fig. 2; Table S1 in the Supplemental Material).

Across all scenarios, LMMs exhibited a slightly reduced TPR compared to the GLMs (Fig. 2; Table S1 in the Supplemental Material). We investigated this result using phenotypes simulated from a single type of measurement. We found TPR of LMMs to be especially reduced on subcortical thickness and surface area (Fig. S2 in the Supplemental Material).

3.2.3 *False positives*

Here, we evaluated the occurrence of false positive vertices or clusters from our simulations. We found that every single simulation yielded at least 1 false positive vertex after Bonferroni correction (FWER = 1, Fig. 2). We noted that the FWER of 0.97 (SE = 0.02) found for LMMs in the scenario of "1000 associated vertices," came from three simulations returning no significant associations.

When evaluating the results at a cluster level, we found that using GLMs almost always resulted in one or more false positive cluster (Fig. 2; Table S1 in the Supplemental Material), leading to cluster FWER > 85%. Cluster FWER was reduced to 49% to 72% using LMMs (Fig. 2; Table S1 in the Supplemental Material). Despite this improvement, no model ensured a cluster-FWER below 5%. LMMs also minimized the proportion of false positive clusters (cluster FDR), compared with the GLM approaches. At the extreme, more than 70% of the significant clusters were false positives using GLMs without covariate. This reduced to about 60% when controlling for age, sex, and ICV and further reduced to <17% using LMMs (Fig. 2; Table S1 in the Supplemental Material).

Next, we simulated phenotypes associated with a single type of measurement and reported the FWER for each type of measurement in Figs. S3–S6 in the Supplemental Material. This allowed evaluation of whether false positives could appear as a result of associations with vertices from other types of measurements. We found that using GLMs resulted in contamination of signal between all the different types of measurements, as indicated by FWER > 5% (Figs. S3–S6 in the Supplemental Material). In comparison, LMMs always minimized the probability of false positives appearing on nonassociated types of measurement. In particular, LMMs ensured that associations on the cortex did not inflate the FPR on subcortical structures, and vice versa (FWER < 5%).

3.2.4 *Statistical power*

We found that the models differ in terms of FPR, which limits the direct comparison of TPR. Instead, we reported the statistical power, which consists of the TPR for a fixed level of FWER (cluster FWER < 0.2). We found the LMMs to be more powerful than the GLMs (Fig. 2; Supplementary 2 in the Supplemental Material).

3.2.5 *Mapping precision*

We defined mapping precision as the median size of the TP clusters. LMMs led to a more precise localization of the associations by minimizing the size of TP clusters (whether we looked a clusters median or maximal size, Fig. 3; Table S1 in the Supplemental Material).

The median size of TP clusters was reduced by a factor >10 on subcortical measurements, and by a factor greater than two on cortical thickness when using LMMs (Table S1 in the Supplemental Material). Of note, positive clusters on cortical surface area were particularly small (most clusters were composed of a single vertex), independent of the model used, Fig. 3; Table S1 in the Supplemental Material). However, LMMs still offered a greater precision than the GLMs when considering the maximal cluster size (Table S1 and Figs. S3–S6 in the Supplemental Material).

3.2.6 *Prediction accuracy from significant vertices*

As a way of aggregating the previous metrics of performance, we compared prediction accuracy achieved from significant vertices, using the UKB replication sample. Across all models and scenarios, selecting the top vertex per significant cluster maximized prediction accuracy, compared with including all significant vertices. This was expected, as significant vertices from the

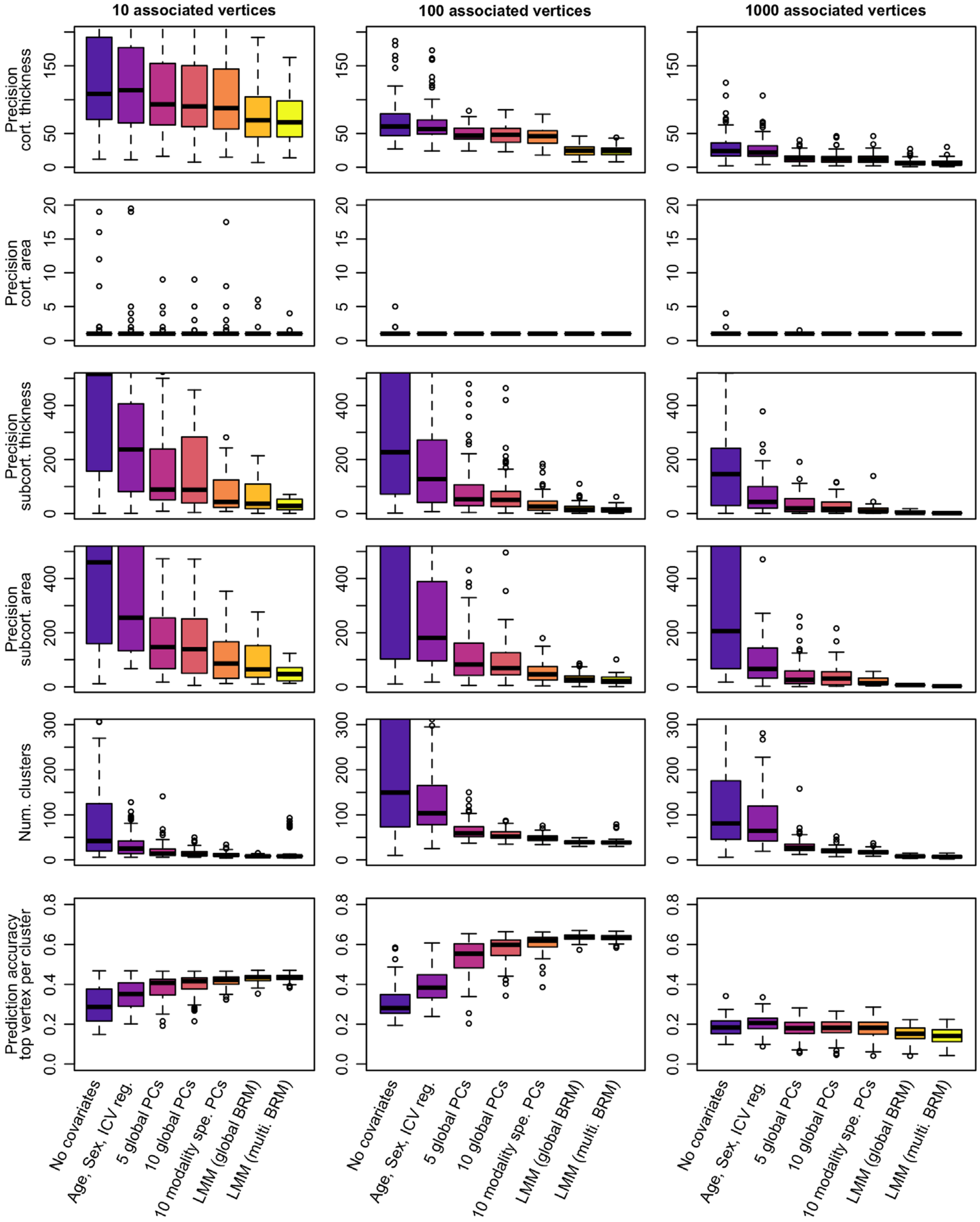

Fig. 3 Mapping precision and prediction accuracy from significant vertices between the different models of mass-univariate analyses. The columns correspond to the different simulation scenarios. We simulated 100 phenotypic traits for each scenario. Bars represent ±SE across the 100 replicates. Clusters are composed of groups of contiguous vertices each significantly associated with the phenotype (after Bonferroni correction). We labeled them as TPs if they included a TP association. (Mapping) precision refers to the median size of the TP clusters.

same cluster tag likely redundant information, leading to overweight the prediction signal coming from large clusters.

In simulation scenarios 1 and 2, we found that including more covariates in the GLMs resulted in greater prediction accuracy despite that predictors included fewer vertices (Fig. 3; Table S1 in the Supplemental Material). In addition, LMMs yielded marginally better prediction accuracy than the best GLM using even fewer vertices (Fig. 3; Table S1 in the Supplemental Material), consistent with observation from previous studies.[6,9] For the third simulation scenario, the prediction accuracy was comparable and limited for all models (Fig. 3; Table S1 in the Supplemental Material).

3.2.7 *Analyses using smoothed cortical surfaces*

We repeated the analysis using smoothed cortical meshes of surface and thickness (FWHM = 20 mm), which is more commonly used in the literature than unsmoothed meshes (Tables S4–S8 in the Supplemental Material). We sought to investigate how robust our results were to such variation of MRI processing.

Overall, smoothing did not change the results of the model comparison. LMMs resulted again in a reduced FPR (lower cluster FWER and cluster FDR) as well as reduced power (seemingly more important than in the unsmoothed case). LMMs maximized mapping precision and prediction accuracy, despite relying on fewer significant clusters (Fig. S7 in the Supplemental Material). Of note, performing analyses on smoothed data decreased the mapping precision, leading to TP clusters roughly 10 times larger on cortical meshes (Fig. 2; Fig. S7 in the Supplemental Material).

Data smoothing resulted in a large inflation of test statistic and FPR for GLMs (Fig. 2; Fig. S7 in the Supplemental Material), which is to be expected as smoothing increases the amount of correlation between vertices. We noticed that smoothing led to an increase of cluster FWER for the GLM with 10 PCs, while it decreased cluster FWER for the LMMs (despite the associated vertices and effect sizes remaining the same). This result warrants a more fined-grained evaluation of the associations. We can only hypothesize that the 20 mm (FWHM) smoothing can induce medium-range correlations (hence medium range false positives in GLMs) while it also increases local correlation which might aggregate false positive clusters in LMMs.

3.3 *Morphometricity of the Phenotypes*

First, we confirmed that the morphometricity estimates of our simulated traits matched the values chosen in simulations (Fig. S8 in the Supplemental Material). For the five UKB phenotypes, we also found consistent morphometricity using the three LMM models (Table S2 in the Supplemental Material), suggesting associations across all types of vertex measurements.

BMI and fluid intelligence exhibited large and moderate morphometricity [$R^2 = 0.51$ (SE $= 0.031$) and $R^2 = 0.17$ (SE $= 0.034$)] but only a limited association with age, sex, or the first 10 PCs from vertexwise data (adjusted R^2 with 10 PCs: $R^2 = 0.032$ for fluid intelligence, $R^2 = 0.033$ for BMI), which resembles the case of our simulations. Age and sex displayed high morphometricity [$R^2 = 0.83$ (SE $= 0.026$) and $R^2 = 0.99$ (SE $= 0.024$)] and large associations with the first 10 PCs (adjusted $R^2 = 0.41$ for age, $R^2 = 0.43$ for sex). Smoking status is a discrete variable (nonsmoker, former smoker, and still smoking) with a morphometricity of $R^2 = 0.12$ (SE $= 0.029$), and adjusted $R^2 = 9.2\text{e} - 3$ with first 10 PCs (Table S2 in the Supplemental Material). Note that the morphometricity estimates are slightly larger than the ones reported previously,[15] which had mean cortical thickness and area regressed out.

3.4 *Analysis of UK Biobank Phenotypes*

We sought to confirm the differences in model performance by applying them to real phenotypic traits. Using GLM without covariates resulted in many vertices and clusters reaching significance (Fig. 4; Table S2 in the Supplemental Material). Unsurprisingly, correcting for covariates that account for a large fraction of the phenotypic variance (see adjusted R^2 with covariates and PCs, Table S2 in the Supplemental Material), drastically reduced the number of associations in the GLMs. For example, correcting for 10 PCs in mass-univariate analyses of age and sex reduced the number of associated vertices by a factor 8 to 13, compared with the GLM without covariates (Fig. 4; Table S2 in the Supplemental Material). For smoking status, the number of significant vertices and clusters also dropped despite a negligible association with PCs (Fig. 4; Table S2 in the Supplemental Material). Similarly, for fluid intelligence, correcting for the top 10 PCs did not remove much of the trait variance over controlling for age sex and ICV (adjusted $R^2 = 0.030$ with age, sex, ICV, adjusted $R^2 = 0.034$ when further controlling for PCs) though it greatly reduced the number of associations. In addition, the more covariates we corrected for, the smaller the size of the associated clusters, suggesting they do remove confounding effects.

We found that across all phenotypes, LMMs resulted in a more parsimonious pattern of associations (Fig. 4; Table S2 in the Supplemental Material). Thus, using the LMM with a single

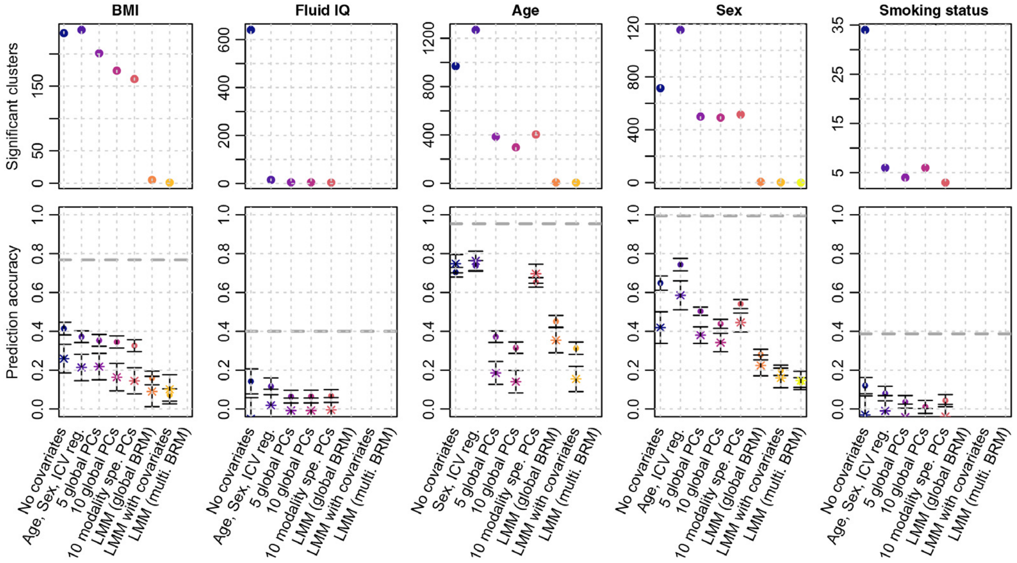

Fig. 4 Number of significant clusters and prediction accuracy for the real UKB phenotypes. Bars represent the 95% confidence intervals of the prediction accuracy (correlations). Dots indicate prediction accuracy in the UKB replication sample, while stars correspond to the prediction achieved in the OASIS3 sample. Prediction accuracy is reported controlling for age, sex (when pertinent), ICM, and site/machine. In the OASIS3 dataset, we further controlled for clinical status. The dashed lines correspond to the estimated morphometricity, which corresponds to the theoretical maximum prediction accuracy achievable from a linear predictor.

random-effect component, we identified five clusters associated with BMI, 8 with age and 6 with sex (Table S2 in the Supplemental Material). LMM with covariates yielded fewer associations, while LMM with multiple random-effects was the most conservative (Table S2 in the Supplemental Material).

Next, we compared the prediction accuracy achieved from the vertices reaching significance using each model (Fig. 4; Table S2 in the Supplemental Material). Predicting our traits of interest allows evaluation of how power and FPR of the different models may counterbalance each other. In addition, prediction into independent samples quantifies the generalizability of findings obtained in the different mass-univariate approaches. For BMI, we found that prediction accuracy from GLMs in the UKB replication sample was greater than that in the OASIS3 sample, which suggests that GLMs-based predictors capture information that is sample specific (e.g., the same confounders are more likely to be shared in the same cohort than across different cohorts). In contrast, the prediction accuracy from LMMs was comparable between the UKB and OASIS3 samples, pointing toward a better generalizability of the prediction. This suggests that the higher prediction accuracy in the UKB replication sample for GLM is likely to be driven by confounding factors shared between UKB data sets. The comparable performance of GLM and LMM seen on OASIS3 for BMI aligns with our simulations.

For age and sex prediction, prediction accuracy of LMMs was sometimes inferior to that achieved from GLMs, in particular those from the simplest models ("no covariates" and "age, sex, ICV"). Overall, prediction based on LMMs generalized well (comparable accuracy in the UKB and OASIS3), while the GLMs often displayed heterogeneous performances across the test samples (in particular for the GLMs with PCs, which may suffer from PCs being different between samples).

Regarding fluid IQ and smoking status, no LMM predictor was available, and the different GLMs resulted in comparable, albeit limited prediction accuracy.

3.5 *Description of Associated Regions*

We listed the significant associations identified using LMM (global BRM) in Table S3 in the Supplemental Material (Figs. S9–S11 in the Supplemental Material for Manhattan plots, and

Figs. S12–S16 for brain plots in the Supplemental Material). The significant associations were in the range of $R^2 = 0.5\%$ to 1%. Most associations were observed with subcortical volumes though the top cluster for sex was spatially located at the border of the lateral–orbitofrontal and medial orbitofrontal gyri (based on the Desikan atlas[26]). Out of the 85 vertices associated with age, sex, and BMI, 68 replicated in an independent UKB sample ($p < 0.05/85$, Table S2 in the Supplemental Material). In particular, 4/11 associations replicated for BMI, 43/47 for age, and 21/27 for sex. The replication rate was slightly lower in the OASIS3 dataset, where none of the vertices reached significance for BMI, 15/47 associations were replicated for age, and 12/27 for sex. Overall, the sign of the associations was consistent across the three datasets (Table S3 in the Supplemental Material).

4 Discussion

Using extensive and realistic simulations, we evaluated the statistical power, FPR, and precision of GLMs and LMMs for vertex-wise gray-matter association studies. In particular, we evaluated the different models in the context of big-data neuroimaging (large sample size but even greater number of correlated brain vertices).[27] We consistently found that using state-of-the-art GLMs resulted in a large number of false positive associations and clusters, whether we used smoothed or not-smoothed gray-matter surfaces. Thus, across all scenarios tested, more than 60% of the significant clusters were false positives using a standard GLM that controlled for age, sex, and ICV. In comparison, FDR was below 17% using LMMs, though still greater than the 5% expectation (Table S1 in the Supplemental Material, Fig. 2, and Fig. S7 in the Supplemental Material). In addition, we showed that unlike GLMs, LMMs could appropriately separate cortical from subcortical associations, even though signal contamination between thickness and surface still occurred (Figs. S2–S5 in the Supplemental Material).

Our results suggest that previously reported results from mass univariate vertexwise analyses obtained using standard GLM approaches could contain many redundant associations, some of which are likely to be false positives induced by confounding factors that cause correlation between vertices [e.g., Refs. 28 to 31, see also Fig. 1(b)]. Note that albeit redundant in term of association and prediction, some of the brain regions identified using GLM may correspond to indirect manifestations of the trait/disease of interest, which may be relevant to understand the dynamics of gray-matter structure. Importantly, the type 1 error (>5%) we observed in simulations also warns against taking for granted results from LMMs.

The increased FPR for GLMs has been well documented in omics association analyses studies (e.g., GWAS[8,14] or MWAS[6,9]) and has been attributed to proximal and distal correlations between features, caused by factors independent of the trait of interest (e.g., genetic ancestry in genetics,[14] cell composition of the biological sample, and smoking status in DNA methylation[6,32]). On the other hand, LMMs can reduce the probability of generating false positives, by fitting all other vertices as random effects which accounts for the complex correlation structure between vertices within and between individuals. In brain imaging, more work is needed to identify the factors that contribute to local and distal correlations between vertices, hence inducing a correlation between true associations and "null" vertices, beyond the usual covariates or confounders used in neuroimaging (e.g., MRI scanner/artefact[1] or demographics[2]).

LMMs yielded fewer TP associations, using the Bonferroni adjusted significance threshold (in particular for the simulated associations on the subcortical nuclei; Fig. S2 in the Supplemental Material). However, this result must be interpreted with caution as it may be partly due to a more stringent control of false positives, resulting in overall fewer vertices reach significance (Fig. 2; Table S2 in the Supplemental Material). To better compare the models performances, we estimated statistical power (i.e., TPR for a set FPR) and noted that the LMMs were more powerful than the GLMs (Fig. 2, Table S2 in the Supplemental Material, Supplementary 2 in the Supplemental Material).

Despite this, LMMs are known to suffer from a power reduction, which arises from the double fitting of the vertex of interest, once as fixed effect and again as a random effect [Eq. (2)].[7,33] For subcortical structures, the effect of double-fitting could be exacerbated by the high level of correlation between vertices. A workaround[7] is to exclude the candidate vertex (and vertices strongly correlated) from the BRM calculation,[33] though this requires computation of the BRM p times (complexity is $O(pN^3)$, where N is the sample size and p is the number of vertices), which becomes impractical for large sample sizes.[7,33] In comparison, the current LMM implementation makes our analysis scalable to samples sizes of tens of thousands [computational complexity of $O(pN^2 + N^3 + pN)$].[6] It should be noted that restricted maximum likelihood estimation approach used in LMMs requires substantially more computational resources than the GLMs and thus requires the use of high-performance clusters.

Beyond power and FPR, we observed from simulations that LMMs could pinpoint the gray-matter association with greater precision (smaller clusters of TPs, Fig. 3). Lastly, we found that prediction achieved from clusters reaching significance in LMMs was on par with that from the best GLMs (Fig. 3), despite fewer vertices included in the predictor. This suggests a higher specificity of the LMMs. Overall, our simulations indicate that LMM with a single random effect currently offers a good trade-off between power and FPR. However, it still fails to ensure a cluster FWER below 5% (also reported on MWAS[6]), despite a stringent Bonferroni correction to account for multiple testing.

Next, we applied the mass-univariate vertexwise models to five real phenotypes of the UKB: age, sex, BMI, smoking status, and fluid IQ. As in the simulations, the LMMs identified fewer vertices and clusters than the GLMs (Fig. 4, Table S2 in the Supplemental Material). The LMM with multiple random-effect components was the most stringent (a single cluster of association), consistent with simulations which showed it had the lowest FWER and statistical power. In contrast, the LMM with a single random-effect component identified several cortical and subcortical associations with BMI, age and sex (Table S3 in the Supplemental Material). Most (12/19) of the top vertices in the associated cluster replicated in the UKB left out sample, and six replicated in the OASIS3 sample (Table S3 in the Supplemental Material). The lower replication rate in OASIS3 may be due to a lower power even though we cannot rule out that the same confounders might act similarly on the two UKB data sets. Overall, replication may be warranted to conclude about an association in future studies, considering the inflation of false positives (even when using LMMs, Fig. 2). The top associated vertices with age, sex, and BMI each captured <1% of the phenotypic variance, suggesting that many more small associations are likely to account for the full morphometricity of the phenotypes (Table S2 in the Supplemental Material). Our results echo the warning against the risk of small associations being confounded (e.g., by artifacts) in big-data neuroimaging,[27] which was confirmed by a recent exploratory study of putative MRI confounders in the UKB.[3] Note that LMMs can reduce false positive associations caused by correlations across and within the different types of measurements (Figs. 2 and 3). Finally, unlike in our simulations (Fig. 3), LMMs often resulted in lower prediction accuracy than GLMs in the UKB left out sample (Fig. 4). Nonetheless, prediction from LMMs generalized better in the OASIS3 dataset (Fig. 4).[24] This suggests that LMMs result in a more robust and parsimonious predictor, less sensitive to sample specific vertexwise patterns and confounders.

In the past years, many studies have been published on the association between gray-matter structure and our phenotypes of interest (see Tables S4–S8 in the Supplemental Material for a selective review of publications). Our simulation and empirical results suggest that some of these studies could report a substantial number of false positive or redundant associations. Nevertheless, due to the limitations outlined below, it is unclear which of these studies suffer from this issue and to which extent.

First, it has been shown in the omics literature that power of LMM may be reduced for phenotypes strongly associated with the covariation between features.[7,34] This is likely the case for age and sex as indicated by their strong association with the PCs calculated from vertex–wise data (Table S2 in the Supplemental Material). This may be an important limitation for phenotypes associated with a cascade of changes in gray-matter, for which LMM would be over conservative.

In addition, LMM assumes a normal distribution of random effects, which may not be realistic for all phenotypes studied. It is equivalent to assuming highly regionalized and specialized brain regions, each displaying a small association with the phenotype. Thus, LMM may be suboptimal under some architectures of association, such as if only a specific but sizable brain region is associated with the trait. Several models have been proposed to relax the LMM hypothesis, for example, to include large/outlying associations as fixed effects (stepwise LMM[35]), break down the feature list into sets of small and large associations (data driven approach: MOMENT[6]), or consider more complex distributions using Bayesian LMMs (Bayesian alphabet[34,36]). They remain to be evaluated in the context of vertexwise analyses. More simulations are warranted, to study other trait architectures, different trait distributions (e.g., skewed, discrete) or to evaluate more sophisticated models. Of note, we limited our trait complexity to 1000 associated brain regions, even if the true pattern of association might be more complex. Our simulations suggest that LMM outperform a GLM independently of the trait complexity, but also that larger samples are required to study traits with more complex architecture (Fig. 2). Our framework of simulation may be easily adapted for such investigations and offers the advantage of estimation of statistical power as well as FPR, which are not often reported at the same time.[37,38]

The nature of the gray-matter regions identified in our GLM analyses of real phenotypes (for which the truth is unknown) can be a matter of debate, which depends on the (also unknown) nature of the correlation between vertices. Two key scenarios can explain the correlation but the data currently available to us does not allow to differentiate between them. First, the correlation could be solely due to confounders [e.g., Fig. 1(b)], in which case the distal associations are false positives. Second, the correlation between vertices could reflect dynamic brain pathways relevant to the trait of interest. In this case, one could describe the GLM associations not found using LMM as redundant rather than false positives. Since we cannot differentiate between these two important causes of between-vertex correlation, we chose to label LMM models as parsimonious, until we understand better the effect of confounders on the vertices correlation structure as well as the longitudinal changes in gray-matter and their relationship with the phenotypes.

Finally, some additional limitations are worthy of note as they may limit the interpretation of mass-univariate vertexwise analyses (compared with GWAS results). First, gray-matter associations may be both causes or consequences of the phenotype studied, unlike GWAS findings, which can impact how to consider redundant associations. At one end of the spectrum are phenotypes such as age for which the direction of the causality is obvious (nothing causes chronological age). When describing which parts of the brain are affected by aging, one may be interested in reporting all associations, including all indirect and redundant. Though, there is no guarantee that those brain regions correctly map the brain pathway of ageing as they might also reach significance due to confusion factors. On the other hand, for many other phenotypes, the direction of causality is unclear (e.g., smoking, BMI) and one may prefer a more parsimonious and robust brain mapping. Second, gray-matter vertices are semiarbitrary features which may be defined and measured in different ways (e.g., different cortical meshes in FreeSurfer). For instance, the resolution of the cortical tessellation is arbitrary and thus so is the number of local vertices which are found to be significant. Hence, the results presented might differ if one were to use a different MRI processing or vertex definition (e.g., volume processing from SPM, coarser surface mesh). In addition, we used Bonferroni to control for multiple testing, although approaches based on RFT are more commonly used (Tables S4–S8 in the Supplemental Material).[16] RFT-based correction is reportedly less stringent than Bonferroni (at least for smoothed data, on which the RFT hypotheses are more likely to be met),[39] which suggests that RFT would also suffer from the inflation of test statistics that we reported.

Furthermore, we did not consider all possible covariates in GLM analysis, focusing on the more commonly used in previous analyses (age, sex, ICV, Tables S4–S8 in the Supplemental Material). More work is needed to evaluate the extended set(s) of covariates which have been recently proposed, from a large-scale study of the UKB data.[3] Finally, mass-univariate results may depend on the study sample used, which raises the question of generalizability into to samples from different age or ethnic groups or with different MRI qualities for instance.

In summary, we found that results obtained using the current state-of-the-art models (GLMs) used in MRI-trait association analyses likely suffer from a large inflation of false positive or redundant associations due to the unaccounted correlation between vertices. In contrast, LMMs allow to control for all vertices fitted as a random effect, which result in a more parsimonious, robust, and conservative characterization of the localized associations between a phenotype and gray-matter structure. However, LMM results should still be interpreted with caution as our simulations show that the FPR remains higher than the standard type 1 error of 5%, even after Bonferroni correction.

Disclosures

The authors declare no conflict of interest.

Acknowledgments

Informed consent was obtained from all UK Biobank participants. Procedures are controlled by a dedicated Ethics and Guidance Council,[40] with the Ethics and Governance Framework available at Ref. 41. 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 12505. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived by the OASIS team. This research was supported by the Australian National Health and Medical Research Council (1078037, 1078901, 1113400, 1161356, and 1107258), the Australian Research Council (FT180100186 and FL180100072), the Sylvia & Charles Viertel Charitable Foundation, the program "Investissements d'avenir" ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) and reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the European Union H2020 program (project EuroPOND, Grant No. 666992, the joint NSF/NIH/ANR program "Collaborative Research in Computational Neuroscience" (project HIPLAY7, Grant No. ANR-16-NEUC-0001-01), the ICM Big Brain Theory Program (project DYNAMO, project PredictICD), and the Abeona Foundation (project Brain@Scale). The OASIS3 data were provided by Principal Investigators: T. Benzinger, D. Marcus, J. Morris supported by NIH Grant Nos: P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352. We used R[42] (v3.6.2) for analyses not performed using OSCA[6] and for plots. We used the color-blind friendly R palette *viridis*,[43] *ukbtools*[44] to facilitate UKB phenotype manipulation, *Morpho* and *Rvcg*[45] to identify clusters, *rgl*[46] to generate brain plots of associations. Other packages used include, *dplyr*,[47] *readr*,[48] *rmarkdown*,[49] *matrixStats*,[50] *RcolorBrewer*,[51] *gridExtra*,[52] *ggplot2*,[53] *png*,[54] and *epuRate*.[55] We would like to thank the Research Computing Centre (RCC) at the University of Queensland for their support with high performance computing, data handling, storage, and processing.

Code, Data, and Materials Availability

Data used in this manuscript are held and distributed by the OASIS and UKB teams. We have released the code used in image processing and analyses to facilitate replication and dissemination of the results (https://baptistecd.github.io/Brain-Mapping-LMM/). Upon publication, we will also release, the summary statistics of mass-univariate analyses performed on the UKB phenotypes (https://cnsgenomics.com/content/data). Supplementary figures (Figs. S1–S16), tables (Tables S1–S7), and sections (Supplementary 1 and 2) may be found on the GitHub repository: https://github.com/baptisteCD/Brain-Mapping-LMM/blob/main/Supp._Article_JMI.pdf. Note: JMI does not require review of supplemental material; supplemental material may not have been reviewed.

References

1. A. A. Chen et al., "Mitigating site effects in covariance for machine learning in neuroimaging data," *Human Brain Mapping* **43**(4), 1179–1195 (2022).
2. M. Montembeault et al., "The impact of aging on gray matter structural covariance networks," *NeuroImage* **63**(2), 754–759 (2012).
3. F. Alfaro-Almagro et al., "Confound modelling in UK Biobank brain imaging," *NeuroImage* **224**, 117002 (2020).
4. L. R. Cardon and L. J. Palmer, "Population stratification and spurious allelic association," *Lancet* **361**(9357), 598–604 (2003).
5. J. Marchini et al., "The effects of human population structure on large genetic association studies," *Nat. Genet.* **36**(5), 512–517 (2004).

6. F. Zhang et al., "OSCA: a tool for omic-data-based complex trait analysis," *Genom. Biol.* **20**(1), 107 (2019).
7. J. Yang et al., "Advantages and pitfalls in the application of mixed-model association methods," *Nat. Genet.* **46**(2), 100–106 (2014).
8. A. L. Price et al., "New approaches to population stratification in genome-wide association studies," *Nat. Rev. Genet.* **11**(7), 459–463 (2010).
9. M. F. Nabais et al., "Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis," *NPJ Genom. Med.* **5**, 10 (2020).
10. J. L. Bernal-Rusiel et al., "Statistical analysis of longitudinal neuroimage data with linear mixed effects models," *NeuroImage* **66**, 249–260 (2013).
11. M. R. Sabuncu et al., "Morphometricity as a measure of the neuroanatomical signature of a trait," *Proc. Natl. Acad. Sci. U. S. A.* **113**(39), E5749–E5756 (2016).
12. B. Couvy-Duchesne et al., "A unified framework for association and prediction from vertex-wise grey-matter structure," *Hum. Brain Mapp.* **41**, 4062–4076 (2020).
13. K. L. Miller et al., "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat. Neurosci.* **19**(11), 1523–1536 (2016).
14. A. L. Price et al., "Principal components analysis corrects for stratification in genome-wide association studies," *Nat. Genet.* **38**(8), 904–909 (2006).
15. B. Couvy-Duchesne et al., "A unified framework for association and prediction from vertex-wise grey-matter structure," *Hum Brain Mapp.* **41**, 4062–4076 (2020).
16. T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Stat. Methods Med. Res.* **12**(5), 419–446 (2003).
17. B. Fischl, "FreeSurfer," *NeuroImage* **62**(2), 774–781 (2012).
18. H. Lindroth et al., "Examining the identification of age-related atrophy between T1 and T1 + T2-FLAIR cortical thickness measurements," *Sci. Rep.* **9**(1), 11288 (2019).
19. B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proc. Natl. Acad. Sci. U. S. A.* **97**(20), 11050–11055 (2000).
20. A. M. Winkler et al., "Measuring and comparing brain cortical surface area and other areal quantities," *NeuroImage* **61**(4), 1428–1443 (2012).
21. B. A. Gutman et al., "A family of fast spherical registration algorithms for cortical shapes," *Lect. Notes Comput. Sci.* **8159**, 246–257 (2013).
22. B. A. Gutman et al., "Shape matching with medial curves and 1-D group-wise registration," in *9th IEEE Int. Symp. Biomed. Imaging*, pp. 716–719 (2012).
23. G. V. Roshchupkin et al., "Heritability of the shape of subcortical brain structures in the general population," *Nat. Commun.* **7**, 13738 (2016).
24. P. J. LaMontagne et al., "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease," medRxiv 2019.2012.2013.19014902 (2019).
25. C. Fawns-Ritchie and I. J. Deary, "Reliability and validity of the UK Biobank cognitive tests," *PLoS One* **15**(4), e0231627 (2020).
26. R. S. Desikan et al., "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage* **31**(3), 968–980 (2006).
27. S. M. Smith and T. E. Nichols, "Statistical challenges in "Big Data" human neuroimaging," *Neuron* **97**(2), 263–268 (2018).
28. C. K. Tamnes et al., "Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness," *J. Neurosci.* **37**(12), 3402–3412 (2017).
29. S. J. Ritchie et al., "Sex differences in the adult human brain: evidence from 5216 UK Biobank participants," *Cereb. Cortex* **28**(8), 2959–2975 (2018).
30. S. R. Cox et al., "Associations between vascular risk factors and brain MRI indices in UK Biobank," *Eur. Heart J.* **40**(28), 2290–2300 (2019).
31. F. J. Navas-Sanchez et al., "Cortical morphometry in frontoparietal and default mode networks in math-gifted adolescents," *Hum. Brain Mapp.* **37**(5), 1893–1902 (2016).
32. A. E. Jaffe and R. A. Irizarry, "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," *Genom. Biol.* **15**(2), R31 (2014).

33. J. Listgarten et al., "Improved linear mixed models for genome-wide association studies," *Nat. Methods* **9**(6), 525–526 (2012).
34. L. R. Lloyd-Jones et al., "Inference on the genetic basis of eye and skin color in an admixed population via Bayesian linear mixed models," *Genetics* **206**(2), 1113–1126 (2017).
35. V. Segura et al., "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations," *Nat. Genet.* **44**(7), 825–830 (2012).
36. G. Moser et al., "Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model," *Plos Genet.* **11**(4), e1004969 (2015).
37. A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates," *Proc. Natl. Acad. Sci. U. S. A.* **113**(28), 7900–7905 (2016).
38. S. Noble, D. Scheinost, and R. T. Constable, "Cluster failure or power failure? Evaluating sensitivity in cluster-level inference," *NeuroImage* **209**, 116468 (2020).
39. K. J. Worsley, "An improved theoretical *P* value for SPMs based on discrete local maxima," *NeuroImage* **28**(4), 1056–1062 (2005).
40. "Ethics advisory committee," 10 February 2022, https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/governance/ethics-advisory-committee (accessed 9 May 2022).
41. "Ethics," 3 December 2021, https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics (accessed 9 May 2022).
42. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna (2012).
43. S. Garnier, "viridis: default color maps from 'matplotlib'" (2018).
44. K. Hanscombe, "ukbtools: manipulate and explore UK Biobank data" (2017).
45. S. Schlager, "Chapter 9 – Morpho and Rvcg – shape analysis in R: R-packages for geometric morphometrics, shape analysis and surface manipulations," in *Statistical Shape and Deformation Analysis*, G. Zheng, S. Li, and G. Székely, Eds., pp. 217–256, Academic Press (2017).
46. D. Adler and D. Murdoch, "rgl: 3D visualization using OpenGL" (2020).
47. H. Wickham and R. Francois, "dplyr: a grammar of data manipulation" (2015).
48. H. H. Wickham, J. Hester, and R. Francois, "readr: read rectangular text data" (2017).
49. J. X. Allaire et al., "rmarkdown: dynamic documents for R" (2018).
50. H. Bengtsson, "matrixStats: functions that apply to rows and columns of matrices (and to vectors)" (2019).
51. E. Neuwirth, "RColorBrewer: ColorBrewer palettes" (2014).
52. B. Auguie, "gridExtra: Miscellaneous functions for "Grid" graphics" (2017).
53. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer Publishing Company, Incorporated (2009).
54. S. Urbanek, "png: read and write PNG images" (2013).
55. Y. Holtz, "epuRate: a clean template for R Markdown documents" (2020).
56. C. Sudlow et al., "UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.* **12**(3), e1001779 (2015).

Baptiste Couvy-Duchesne is a CJ Martin fellow (National Health Medical Research Council, Australia) and INRIA researcher (National Institute for Research in Digital Science and Technology, France). He received his PhD in 2017 from the University of Queensland, working on complex-trait genetics and MRI brain imaging. Since, BCD has been working at adapting and applying some of the methods used in big-data genetics to the analysis of high-dimensional brain MRI. He is a member of the ENIGMA (Enhancing Neuro-Imaging Genetics using Meta-Analyses) and PGC (Psychiatry Genetics) consortia. He is working for the Paris Brain Institute and the University of Queensland. His research focuses on developing and applying statistical methods to analyze large-scale brain MRI data.

Futao Zhang is a postdoctoral researcher with the Institute for Molecular Biosciences (IMB) at the University of Queensland. He develops performant methods and software for the analysis of genomic and other large-scale datasets.

Kathryn E. Kemper is a postdoctoral researcher with the Institute for Molecular Biosciences (IMB) at the University of Queensland. Her research in statistical genetics focuses on examining the causes of variation within human populations for traits such as height and body mass index.

Julia Sidorenko is a postdoctoral research assistant with the Institute for Molecular Biosciences (IMB) at the University of Queensland. Her research in statistical genetics focuses on examining the causes of variation within human populations. She also processes and manages genetic and genomic datasets, including the UKBiobank.

Naomi R. Wray holds joint appointments at the Institute for Molecular Bioscience (IMB) and the Queensland Brain Institute (QBI) within the University of Queensland. She is a National Health and Medical Research Council (NHMRC) leadership fellow, a fellow of the Australian Academy of Science, and a fellow of the Australian Academy of Health and Medical Science. Her research focusses on development of quantitative genetics and genomics methodology with application to psychiatric and neurological disorders.

Peter M. Visscher joined the University of Queensland in 2011, where he is a professor of Quantitative Genetics. He is a laureate fellow of the Australian Research Council. He was elected a fellow of the Australian Academy of Science in 2010, a fellow of the Royal Society (London) in 2018, and a foreign member of the Royal Netherlands Academy of Arts and Sciences in 2018. His research is about genetic variation for complex traits (including quantitative traits and disease) in populations, with the broad aim to understand and quantify the causes and consequences of human trait variation.

Olivier Colliot is a research director at CNRS and the co-head of the ARAMIS Lab (www.aramislab.fr), a joint laboratory between CNRS, Inria, Inserm, Sorbonne University, and the Paris Brain Institute. He also holds a chair at the PRAIRIE Institute for Artificial Intelligence. He is a conference chair of the SPIE Medical Imaging Image Processing conference and an associate editor of *Medical Image Analysis* and *Frontiers in Brain Imaging Methods*. His research interests include machine learning, medical image analysis, and their applications to neurological disorders.

Jian Yang is a professor of statistical genetics at the School of Life Sciences, Westlake University, China. He received his PhD in 2008 from Zhejiang University, China, before undertaking postdoctoral research at the QIMR Berghofer Medical Research Institute in Australia (2008 to 2011). He moved to the University of Queensland (UQ), Australia, as a research fellow in 2012 and was reappointed as a senior research fellow and group leader in January 2014. He was promoted to be an associate professor in December 2014 and then a professor in 2017 at UQ. He joined Westlake University in 2020. His primary research interests are focused on understanding the genomic variations among individuals within and between populations and the links of genomic variations with health outcomes.