Linear Mixed Models Minimise False Positive Rate and Enhance Precision of Mass Univariate Vertex-Wise Analyse of Grey-Matter

Baptiste Couvy-Duchesne, Futao Zhang, Kathryn Kemper, Julia Sidorenko, Naomi Wray, Peter Visscher, Olivier Colliot, Jian Yang

LINEAR MIXED MODELS MINIMISE FALSE POSITIVE RATE AND ENHANCE PRECISION OF MASS UNIVARIATE VERTEX-WISE ANALYSES OF GREY-MATTER

Baptiste Couvy-Duchesne[1,2], Futao Zhang[1], Kathryn E. Kemper[1], Julia Sidorenko[1], Naomi R. Wray[1,], Peter M. Visscher[1,*], Olivier Colliot[2,*], Jian Yang[1,3,*]*

[1] Institute for Molecular Bioscience, the University of Queensland, St Lucia, QLD, Australia; [2] Institut du Cerveau et de la Moelle épinière, ICM, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, Inria, Aramis project-team, F-75013, Paris, France; [3]: Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China; * these authors contributed equally.

ABSTRACT

We evaluated the statistical power, family wise error rate (FWER) and precision of several competing methods that perform mass-univariate vertex-wise analyses of grey-matter (thickness and surface area). In particular, we compared several generalised linear models (GLMs, current state of the art) to linear mixed models (LMMs) that have proven superior in genomics. We used phenotypes simulated from real vertex-wise data and a large sample size (N=8,662) which may soon become the norm in neuroimaging.

No method ensured a FWER<5% (at a vertex or cluster level) after applying Bonferroni correction for multiple testing. LMMs should be preferred to GLMs as they minimise the false positive rate and yield smaller clusters of associations. Associations on real phenotypes must be interpreted with caution, and replication may be warranted to conclude about an association.

Index Terms— Grey-matter, mass univariate vertex-wise analyses, simulations, false positive, precision

1. INTRODUCTION

The recent availability of large MRI imaging cohorts (such as the UKBiobank) offers the opportunity to progress our understanding of the associations between phenotypes and grey-matter structure.

Mass univariate vertex-wise analyses (MUVA) aim to identify which of the ~650,000 cortical and subcortical vertices are associated with a trait. Current state of the art approaches rely on generalised linear models (GLMs, e.g. implemented in FreeSurfer[1]) though little is known about their power, false positive rate and spatial precision in large samples (sometimes referred to as "large degree of freedom problem"). Reports on several omics datasets (of similar size and complexity) have warned about increased false positive rate when performing mass univariate analyses on correlated features and large samples[2,3], which led us to evaluate standard GLMs against more robust linear mixed models (LMMs).

For realistic evaluations of the model performances, we simulated phenotypes based on real vertex-wise measurements from the UKBiobank (N~10,000). We used Bonferroni correction to account for multiple testing across the vertices, which we preferred over random field theory (RFT)[4] as it does not require hypotheses about smoothing of the grey-matter surfaces[4]. Permutation testing was too computationally costly to represent a viable option.

2. MATERIALS AND METHODS

2.1. Participants recruitment and MRI imaging

The UKB participants were unselected volunteers from the United Kingdom[5]. Exclusion criteria were limited to the presence of metal implant or any recent surgery and health conditions problematic for MRI imaging (e.g. hearing, breathing problems or extreme claustrophobia)[6]. T1w and T2 FLAIR MRI images were collected using a 3T Siemens Skyra machine and a 32-channel head coil[6].

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), with the Ethics and Governance Framework available at http://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf. 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.

2.2. Image processing

We processed the T1w and T2 FLAIR images together to enhance the tissue segmentation in FreeSurfer 6.0[1], hence a more precise skull stripping and pial surfaces definition. We retained the maximal image information by using the (fsaverage - unsmoothed) vertex-wise level data in the cortical surface and thickness analyses[7]. In addition, we applied the ENIGMA-shape processing[8,9] to the segmented images to extract radial thickness and log Jacobian determinant (analogous to a relative surface area) of the hippocampus, putamen, amygdala, thalamus, caudate, pallidum and accumbens[8,9]. The processed imaging data comprised 654,026 vertex measurements separable into 4 modalities (types of features): cortical thickness, cortical surface area, subcortical thickness or subcortical area).

2.3. Sample description and quality control (QC)

We considered the first 10,103 participants of the UK Biobank (UKB) imaging wave. Our final sample after processing comprised 9,890 adults with complete cortical and subcortical data, aged 62.5 years on average (SD=7.5, range 44.6–79.6) with slightly more (52.4%) female

participants. We used a stringent data-driven QC, which excluded 1,228 subjects (12.4%) who showed an outlying brain (+-5SD from the mean when looking at individual (pairwise) brain similarities). A more lenient QC may be applied in real data analysis to maximise the sample size.

2.4. Phenotype simulation and age at MRI

For realistic scenarios, we simulated continuous phenotypes from the (standardised) grey-matter data[2]. We selected randomly a fixed number of associated vertices and drew each effect size from a normal distribution. We considered 3 traits architecture that differ in term of number of associated vertices and total association R^2 (morphometricity): i) 10 associated vertices accounting for a morphometricity of R^2=0.10; ii) 100 associated vertices accounting for R^2=0.50; iii) 1000 vertices accounting for R^2=0.40. For a finer understanding of the results, we drew associated vertices on each modality independently and repeated each simulation 100 times.

We used age at MRI to validate our results on a real continuous phenotype.

2.5. GLMs for mass univariate analyses

Commonly used in the neuroimaging field are the GLMs without covariates ("uncorrected") or using standard covariates such as age, sex and ICV ("age, sex, ICV corrected"). Next, we varied the covariates by including the top 5 or 10 principal components (PCs) of the vertex-wise data ("5 global PCs", "10 global PCs"). We also considered 10 principal components specific to the vertex modality ("10 modality specific PCs"). The rationale is to correct for structure in the population in a data-driven manner. Thus, grey-matter PCs may be able to also remove unmeasured or unaccounted batch effects (e.g. software update, processing options) or factors showing large associations with grey-matter structure (e.g. height, body size[7]).

2.6. LMMs for mass univariate analyses

We considered 2 linear mixed models that are extensions of the previous approaches in that they explicitly model the population structure under the form of grey-matter similarities between any pair of individuals. The LMM models may be written as $\mathbf{Y} = \mathbf{X}\beta + \mathbf{b} + \mathbf{e}$ where $\mathbf{Y}_{N,1}$ is the phenotype considered with N the number of observations, $\mathbf{X}_{N,1}$ is a vector of vertex-wise measurement, β is the vertex-trait association we are trying to estimate, $\mathbf{b}$ is a random effect with $\mathbf{b} \sim \mathcal{N}(0, \mathbf{B}\sigma_b^2)$ and $\mathbf{e}$ is the error term with $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I}\sigma_e^2)$. σ_b^2 and σ_e^2 are the variances of the random effects $\mathbf{b}$ and $\mathbf{e}$ (residual). For the first LMM ("LMM global BRM"), $\mathbf{B}$ is the brain relatedness matrix (NxN matrix of variance-covariance between individuals[7]) calculated from all grey-matter vertices. This is similar to the MOA model implemented in OSCA[2]. The second LMM ("LMM multi. BRM") fits 4 random effect, one for each modality. For all models, we performed a χ^2 test of the association between each vertex and the phenotype using: $\left(\frac{\beta}{SE(\beta)}\right)^2 \sim \chi(1)$

2.7. Metrics of interest

On the null vertices (uncorrelated with the associated vertices), the empirical distribution of chi[2] statistics may be compared to the expected one using the ratio of median chi[2], known as the inflation factor (λ). Inflation factors close to 1 indicate no inflation of the test statistics.

We quantified the statistical power of the BWAS models using the true positive rate (TPR) after Bonferroni correction for multiple testing.

We measured the family-wise type I error rate (FWER) as the proportion of replicates with at least 1 false positive (FP) vertex (after Bonferroni correction). In presence of strong correlation between (neighbouring) vertices, it is statistically difficult to separate true and false positive and we can expect clusters of associations (thus large FWER). We evaluated how separable are the associations on the different modalities by reporting the FWER restricted to vertices on the non-associated modalities.

Beyond vertex-wise FWER, we reported the cluster FWER: proportion of replicates with at least 1 false positive cluster (contiguous FP vertices on the associated modality). We further reported the number and median size of the false positive clusters as well as the proportion of false positive clusters (cluster FDR).

For completeness, we evaluated whether TP and FP clusters could be separated by size (without excluding any true positive). We also assessed whether true positive clusters could pinpoint the correct cortical region (ROI-FWER, based on the Desikan atlas).

3. RESULTS

As expected in presence of population structure leading to widespread correlation between vertices, we observed a global inflation of tests statistics when using GLMs (**Figure 1** for scenario with 100 associated vertices). In comparison LMMs could control the inflation of test statistics on null vertices (**Figure 1**). When restricting the null probes to those from non-associated modalities we also observed an inflation in the linear case, which was well controlled by the LMM models.

LMMs had lower power than the linear models, as shown by reduced true positive rate, especially on subcortical volumes (**Figure 1**). However, the clusters of true positive identified using LMMs were much smaller than those identified by the other models (in terms of minimal, median and maximal cluster size). To note, the true positive clusters on surface area were overall small (median <3 vertices), though some large clusters were also observed. Thus, TP clusters in cortical surface area are nested in the correct cortical region (ROI-FWER<3%), though this was

not the case for cortical thickness where the larger TP clusters often overlapped several cortical regions.

All simulations with associated vertices on cortical thickness and subcortical structures yielded at least 1 false positive vertex (FWER=1). For associated vertices on cortical surface area, the FWER was minimised by using LMMs, though greater than 5%. More interestingly, we found that analysing the data using GLMs resulted in false positive associations on non-associated modalities in at least 20% of the replicates (**Figure 1**). Using LMMs minimised the FWER on other modalities, though it did not ensure an error rate<5% for all associated modalities and phenotype architectures. In particular, LMM (multi. BRM) could separate associations on the cortex and on subcortical nuclei (FWER<5%) but failed at separating associations on thickness and surface area.

Similarly, LMMs minimised the probability of observing false positive clusters on the associated modality but failed to ensure a cluster-FWER below 5% across all scenarios (**Figure 1**). For example, using the best LMM (multi. BRM), 40% of the simulations yielded a false positive cluster on cortical surface area but this rate was 10% on cortical thickness, 4.1% on subcortical area and 14% on subcortical thickness (**Figure 1**). The proportion of false positive clusters tended to be small with LMMs (FDR<5%) except for the scenario of 1000 associated vertices, which may be due to the low power and small number of TP clusters. Using LMM (multi. BRM) minimised the size of the FP clusters that typically comprised less than 3 vertices. However, in up to 30% of the replicates (depending on the simulation scenarios) TP and FP clusters could not be separated by size.

Using PCs in the GLMs resulted in an improvement over the use of standard covariates suggesting PCs can remove unaccounted factors associated with long-range vertex correlation. We did not observe any change in the results and conclusions after rank inverse normal transformation (RINT) of the vertices, which suggests the false positives are not caused by outliers.

We estimated the morphometricity of age at MRI to be R^2=0.83 (SE=0.026), and about half of this association resulted from large associations with global axes of grey-matter size/surface (R^2=0.41 with the first 10 global PCs). Consistent with this observation, "uncorrected" GLM yielded significant associations with 136,348 vertices (**Table 1**) from nearly all cortical and subcortical regions, while covarying the PCs removed signal of interest which resulted in a drop of number of associated vertices.

As in the simulations, LMMs minimised the inflation of test statistics (lambda<1), and resulted in fewer and smaller associated regions (**Table 1**). The 8 significant clusters pointed towards associations with subcortical thickness and surface of the thalamus, caudate and putamen. Note that no vertex reached significance for age at MRI using a LMM with multiple BRMs, which may be due to the reduced power of this model, especially on subcortical structures.

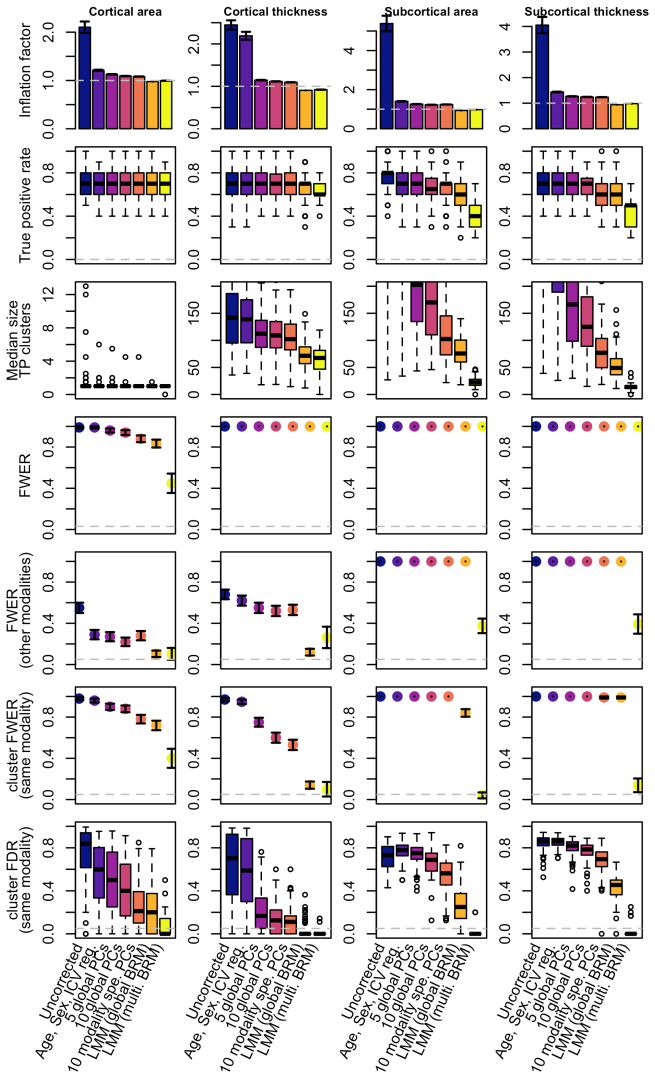

Figure 1: Main metrics summarising the MUVA results on simulated phenotypes (10 associated vertices). *The 10 associated vertices were selected in single modalities (labelled at the top of each column) to explain 10% of the phenotypic variance. Bars represent +- SE from 100 replicates.*

4. DISCUSSION

Our simulations revealed that using GLMs in mass univariate vertex-wise analyses of grey-matter structure resulted in a gross inflation of tests statistics and false positive rate. LMMs appropriately controlled the inflation of test statistics and offered a greatly reduced false positive rate, though they still failed at ensuring a vertex or cluster FWER below 5%. This is especially worrisome as the Bonferroni correction for multiple testing is considered overly stringent in presence of correlated features, though similar results have been reported on 'omics' datasets[2]. False positives were not attributable to outliers in the vertex wise measurements.

Within a single modality, false positive clusters were found in more than 5% of the replicates, for most scenarios,

though LMMs (especially multi. BRM) always minimised the false positive rate. Finally, most associated clusters found using GLMs may be false positive (e.g. up to 85% of the associated clusters were false positive using GLM with standard covariates, which dropped to below 26% when using LMM).

Table 1: Summary of mass-univariate analyses for age at MRI

	Adj. R^2 with sex & ICV	0.012
	Adj. R^2 with first 10 PCs	0.41
Uncorrected GLM	*N assoc. vertices*	136,348
	N assoc. clusters	970
	Max cluster size	22,358
sex, ICV GLM	*N assoc. vertices*	130,189
	N assoc. clusters	1,270
	Max cluster size	19,450
10 global PCs GLM	*N assoc. vertices*	16,772
	N assoc. clusters	297
	Max cluster size	894
Single random effect LMM	*N assoc. vertices*	**47**
	N assoc. clusters	**8**
	Max cluster size	**15**
	Morphometricity (SE)	**0.91 (0.021)**
Multiple random effect LMM	*N assoc. vertices*	0
	N assoc. clusters	0
	Max cluster size	NA
	Morphometricity (SE)	0.83 (0.026)

Beyond their lower FWER, LMMs had lower statistical power than GLMs, which is attributable to the double fitting of the vertex, both as a fixed and random effect[3,10]. However, LMMs always resulted in a more precise localisation of the true positive (**Figure 1**). Between the GLMs considered, including PCs was superior to fitting standard covariates to reduce FWER and maximise precision (at a small cost in power).

The empirical results obtained on age at MRI were consistent with our simulations. The limited number of findings using LMM (which jointly explain 28% of age variance) tend to suggest that the localised grey-matter associations with age may widespread (i.e. "complex architecture" with hundreds or thousands of associated regions).

Our simulation results may not hold for trait architectures not considered here (we assumed a normal distribution of effect sizes). Note that in presence of large/outlying phenotype-vertex associations, one may include them as fixed effects in the LMMs.

More work is needed to evaluate the FWER of MUVA using vertices or voxels derived from different processing (e.g. volume based processing of grey-matter), from different MRI images (e.g. diffusion weighted images), or when using non-normally distributed phenotypes (such as symptom scores).

In conclusion, none of the current methods for mass-univariate analyses appropriately control the FWER (at a vertex or cluster level) after applying Bonferroni correction for multiple testing. However, LMMs should be preferred to GLMs as they minimise the false positive rate and offer a more precise identification of the associated regions. Our simulations suggest that LMMs can, in part, separate long-range correlations induced by unobserved confounding factors from the meaningful short-range correlations. This arises from LMMs controlling for the individuals' pairwise brain similarities that capture unobserved factors (e.g. genetic, environmental, batch effects), which contribute to the correlations between vertices, hence the spread of the association signal. To note, LMMs may be overly conservative when the trait studies is itself associated with the correlation between vertices, such as when affecting different brain regions in cascade. Overall, associations on real phenotypes should be interpreted with caution, and replication or evaluation of prediction accuracy may be warranted to safely conclude about an association.

5. ACKNOWLEGEMENTS

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. It was also supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

6. REFERENCES

1 Fischl, B. FreeSurfer. *NeuroImage* **62**, 774-781, doi:10.1016/j.neuroimage.2012.01.021 (2012).

2 Zhang, F. *et al.* OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol* **20**, 107, doi:10.1186/s13059-019-1718-z (2019).

3 Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100-106, doi:10.1038/ng.2876 (2014).

4 Nichols, T. & Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research* **12**, 419-446 (2003).

5 Sudlow, C. *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 medicine* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

6 Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523-1536, doi:10.1038/nn.4393 (2016).

7 Couvy-Duchesne, B. *et al.* Widespread associations between grey matter structure and the human phenome. *bioRxiv*, 696864, doi:10.1101/696864 (2019).

8 Gutman, B. A., Madsen, S. K., Toga, A. W. & Thompson, P. M. in *Multimodal Brain Image Analysis: Third International Workshop, MBIA 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings* (eds Li Shen *et al.*) 246-257 (Springer International Publishing, 2013).

9 Gutman, B. A., Wang, Y. L., Rajagopalan, P., Toga, A. W. & Thompson, P. M. Shape Matching with Medial Curves and 1-D Group-Wise Registration. *2012 9th Ieee International Symposium on Biomedical Imaging (Isbi)*, 716-719 (2012).
10 Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**, 525-526, doi:10.1038/nmeth.2037 (2012).