Independent Multifactorial Association Analysis to analyze multiblock data: application in an Imaging Genetic study of Attention-Deficit/Hyperactivity Disorder

Natalia Vilor-Tejedor[1-3], *Silvia Alemany*[1-3], *Alejandro Cáceres*[1-3], *Mariona Bustamante*[1-4], *Jordi Sunyer*[1-3,5], *Juan R. González*[1-3]

[1]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[3]CIBER Epidemiologia y Salud Pública (CIBERESP), Barcelona, Spain.
[4]Center for Genomic Regulation (CRG), Barcelona, Spain
[5]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain.

Abstract

The main goal of Imaging Genetic studies is the assessment of the correlation between genomic and neuroimaging data to improve classification of individuals into disease status or infer genetic mechanisms associated with brain structure, function and behaviour. The joint analysis of genomic and neuroimaging data still represents a challenge to adequately evaluate complex behaviors such as Attention-deficit/Hyperactivity disorder (ADHD), due to the high dimensionality and the specific nature of the data. In this proposal, we present a novel multifactorial algorithm, referred as Independent Multifactor Association Analysis (IMFAA), to evaluate potential relationships of genetic and neuroimaging data to predict ADHD symptoms.

Keywords: Attention-Deficit/Hyperactivity Disorder, Imaging Genetics, Independent Multifactorial Association analysis.

1. Introduction

Neuroimaging information is used to understand neurodevelopment, cognition and behaviour at the brain level and how these processes can be affected by environmental exposures. The interpretation and understanding of these processes can be more accurate if genomic information is also considered. The joint analysis of genomic and neuroimaging data, a field known as Imaging Genetics, still presents methodological challenges in current biomedical research.

A common statistical strategy to assess potential associations in Imaging Genetic studies is the massive univariate linear method (MULM) in which extensive paired-wise correlations are performed between both data sets. This method has several limitations such as: (i) the multidimensionality of genomic and neuroimaging data, (ii) the requirement of well-powered large studies which are not usually accessible in this field, (iii) the infeasibility of most of

computational procedures and (iv) the impossibility to take into account complex multivariate relationships that may exist between genes and brain measurements.

To address these problems, several multivariate statistical approaches have been proposed, most of which are based on dimensionality reduction methods.

2. Original contribution

We propose a statistical methodology based on an extension of Multifactorial analysis [1], referred as Independent Multifactorial Association Analysis (IMFAA). This approach is designed to evaluate potential relationships between genomic and neuroimaging data. We evaluate the performance of IMFAA in the context of childhood Attention-Deficit/Hyperactivity Disorder (ADHD) symptoms.

3. Methods

MFA is an extension of Principal Component Analysis (PCA) used to integrate different sets of variables (e.g. tables) describing the same set of observations.

MFA is mainly comprised in four steps: First, a PCA of each data set is performed. Second, data sets are normalized by dividing by the square root of the first eigenvalue from PCA. Third, the normalized data sets are concatenated in a unique data set. Finally, a PCA is computed on the general data set to evaluate how much the whole set of variables contribute to the inertia extracted by a component.

Singular value decomposition in MFA is based on PCA. However, mathematically, PCA is adequate only if the data is normally distributed, linear and stationary.

Taking advantage from MFA, we propose an alternative algorithm, the IMFAA, in which the singular value decomposition is done by applying an Independent Component Analysis (ICA) [2]. As ICA maximizes non-Gaussianity between components, it searches for linear combinations of variables that optimize statistically independence. We then use the most relevant ICA components to test their association with the phenotype of interest.

IMFAA thus comprises five steps: In the first step an ICA of each data set is performed instead of PCA. In the second step, data sets are normalized by dividing by the square root of the first independent component from ICA. In the third step, data sets are concatenated as in step 3 of MFA algorithm. In the fourth step, an ICA is computed on the general normalized data set to extract general independent components. Finally, a regression analysis is computed to determine relevant predictors of ADHD symptoms.

4. Application on Attention-Deficit/Hyperactivity Disorder Imaging Genetic study

We apply IMFAA to investigate whether potential genetic variants identified through Genome-wide Association Analyses are related to neuroimaging characteristics in the context of ADHD.

Participants

This study used a sample drawn from the BRrain dEvelopment and Air pollution ultrafine particles in school childrEn (BREATHE) project. The BREATHE project is a longitudinal study conducted from January 2012 to March 2013 in 39 schools in Barcelona (Catalonia, Spain) to study the association between air pollution and cognitive development of school children [3]. From a total of 2,897 children participating in this project, behaviour and genomic data was available for 1,592 individuals and neuroimaging data was available for a subset of 135 individuals.

Measures

Child ADHD symptoms were collected using the ADHD criteria of *Diagnostic and Statistical Manual of Mental Disorders,* fourth edition (ADHD-DSM-IV), completed by teachers. ADHD-DSM–IV, consists of a list of 18 symptoms, assessing two separate symptom groups: inattention (9 symptoms) and hyperactivity/impulsivity (9 symptoms). Each ADHD symptom is rated in a 4-point scale of frequency from never or rarely (0) to very often (3). ADHD outcome was calculated as the sum of the score of each item. This continuous measure ranges from 0 to 54.

Genome-wide genotyping was performed using the HumanCore BeadChip WG-330-1101 (Illumina) at the Spanish National Genotyping Center (CEGEN). PLINK was used for the data quality control. The final genotyped data set consisted of 240,103 autosomal Single Nucleotide Polymorphisms (SNPs).

High-resolution 3D anatomical images were obtained using an axial T1-weighted three dimensional fast spoiled gradient inversion recovery-prepared sequence. All the anatomical images were visually inspected and subjects with poor quality images were discarded. Cortical and subcortical thickness measurements across the whole cortex were obtained using FreeSurfer software (http://surfer.nmr.mgh.harvard.edu/). The final data set consisted of 60 brain cortical thickness measures.

Statistical Design

A two-stage analysis was performed. In the first stage, we conducted a Genome-wide Association Study of 1,592 individuals in order to perform a selection of the nominal significant intragenic SNPs associated with ADHD symptoms. Zero-inflated negative binomial regressions

adjusted by age and sex were conducted. In the second stage, we evaluated the influence of the previous selected SNPs in brain morphology in the subset of 135 individuals who underwent structural magnetic resonance imaging (MRI) scanning. We applied MFA and IMFAA methods to assess potential relationships between genomic and neuroimaging data with ADHD symptoms.

Results

Preliminary results showed that genetic variations associated with childhood ADHD symptoms in general population are associated with cortical thickness of frontostriatal circuits. IMFAA increases validity, reliability and statistical power by integrating genomic and neuroimaging data across a novel multifactorial analysis approach.

5. Discussion

The main goal of Imaging Genetic studies in ADHD is the individual identification of genetic variants and neuroimaging correlates that are associated with ADHD symptoms. The integration of neuroimaging and genetic data provides strong evidences to advance the understanding of the underlying neurobiological mechanisms operating in the development of ADHD symptoms.

Nevertheless, research is still scarce in this field and results are still inconsistent. The use of well established methodologies in the context of Imaging Genetics, such as MFA, can help to bridge this gap. The improvement and rigorous application of known methodologies constitutes an important aspect in the analysis of small number of Imaging Genetic studies in ADHD. However, further research is needed to increase the statistical power and detect significant causal factors using multivariate analysis in Imaging Genetic studies.

6. Acknowledgments

Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2015 FI_B 00636), Generalitat de Catalunya. Silvia Alemany thanks the Institute of Health Carlos III for her Sara Borrell postdoctoral grant (CD14/00214). The research leading to these results has received funding from the European Research Council under the ERC Grant Agreement number 268479 – the BREATHE project. This research was also supported by grant MTM2011-26515 from the Ministerio de Economía e Innovación (Spain). The authors would particularly like to thank all the participants for their generous collaboration.

7. Bibliography

[1] Thurstone LL. *Multiple Factor Analysis*. Chicago, IL: University of Chicago Press; 1947.

[2] Jutten C, Herault, J. *Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture*. Signal Processing, 24 (1991), pp. 1–10.

[3] Sunyer J, Esnaola M, Alvarez-Pedrerol M, Forns J, Rivas I, López-Vicente M, Suades-González E, Foraster M, Garcia-Esteban R, Basagaña X, Viana M, Cirach M,Moreno T, Alastuey A, Sebastian-Galles N, Nieuwenhuijsen M, Querol X. *Association between Traffic-Related Air Pollution in Schools and Cognitive Development in Primary School Children: A Prospective Cohort Study*. PLoS Med. 2015 Mar 3;12(3):e1001792. doi: 10.1371/journal.pmed.1001792. Collection 2015 Mar. PubMed PMID: 25734425; PubMed Central PMCID: PMC4348510.