Independent Multiple Factor Association Analysis for Multiblock Data in Imaging Genetics

Vilor-Tejedor, Natalia; Ikram, Mohammad Arfan; Roshchupkin, Gennady V.; Cáceres, Alejandro; Alemany, Silvia; Vernooij, Meike W.; Niessen, Wiro J.; van Duijn, Cornelia M.; Adams, Hieab H.; More Authors

DOI
10.1007/s12021-019-09416-z

Publication date
2019

Document Version
Final published version

Published in
NeuroInformatics

Neuroinformatics (2019) 17:583–592
https://doi.org/10.1007/s12021-019-09416-z

ORIGINAL ARTICLE

Independent Multiple Factor Association Analysis for Multiblock Data in Imaging Genetics

Natalia Vilor-Tejedor[1,2,3,4,5] · Mohammad Arfan Ikram[6] · Gennady V. Roshchupkin[7,8] · Alejandro Cáceres[3,4,5] · Silvia Alemany[3,4] · Meike W. Vernooij[6,7] · Wiro J. Niessen[7,8,9] · Cornelia M. van Duijn[6] · Jordi Sunyer[3,4,5,10] · Hieab H. Adams[6,7,8] · Juan R. González[3,4,5]

Published online: 22 March 2019

Abstract

Multivariate methods have the potential to better capture complex relationships that may exist between different biological levels. Multiple Factor Analysis (MFA) is one of the most popular methods to obtain factor scores and measures of discrepancy between data sets. However, singular value decomposition in MFA is based on PCA, which is adequate only if the data is normally distributed, linear or stationary. In addition, including strongly correlated variables can overemphasize the contribution of the estimated components. In this work, we introduced a novel method referred as Independent Multifactorial Analysis (ICA-MFA) to derive relevant features from multiscale data. This method is an extended implementation of MFA, where the component value decomposition is based on Independent Component Analysis. In addition, ICA-MFA incorporates a predictive step based on an Independent Component Regression. We evaluated and compared the performance of ICA-MFA with both, the MFA method and traditional univariate analyses, in a simulation study. We showed how ICA-MFA explained up to 10-fold more variance than MFA and univariate methods. We applied the proposed algorithm in a study of 4057 individuals belonging to the population-based Rotterdam Study with available genetic and neuroimaging data, as well as information about executive cognitive functioning. Specifically, we used ICA-MFA to detect relevant genetic features related to structural brain regions, which in turn were involved, in the mechanisms of executive cognitive function. The proposed strategy makes it possible to determine the degree to which the whole set of genetic and/or neuroimaging markers contribute to the variability of the symptomatology jointly, rather than individually. While univariate results and MFA combinations only explained a limited proportion of variance (less than 2%), our method increased the explained variance (10%) and allowed the identification of significant components that maximize the variance explained in the model. The potential application of the ICA-MFA algorithm constitutes an important aspect of integrating multivariate multiscale data, specifically in the field of Neurogenetics.

H. H. Adams and J. R. González co-last authors

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12021-019-09416-z) contains supplementary material, which is available to authorized users.

✉ Natalia Vilor-Tejedor
natalia.vilortejedor@crg.eu

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology., C. Doctor Aiguader 88, Edif. PRBB, 08003 Barcelona, Spain

[2] BarcelonaBeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain

[3] Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain

[4] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[5] CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

[6] Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands

[7] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands

[8] Department of Medical Informatics, Erasmus MC, Rotterdam, the Netherlands

[9] Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands

[10] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

Keywords Data integration · ICA-MFA · Imaging genetics · Modelling · Neurogenetics

Introduction

Current biomedical research increasingly combines high-throughput data. For instance, sequencing technologies produces omics data at different levels of cellular components. In addition, magnetic resonance imaging produces vast amounts of data (e.g. structural, functional and connectivity) with even more complex features and broader dimensions (Luo et al. 2016). The analysis of this type of data presents different challenges, even more so if we consider the combined analysis of both sources, field referred as imaging genetics.

A common strategy to investigate potential associations between neuroimaging features and genetic data is based on performing massive marginal linear models in which extensive pair-wise correlations are computed (Hoogman et al. 2014). However, this strategy has important limitations such as (i) the inability to exploit the multidimensionality of data and synergistic effects between variables and (ii) the requirement of a large number of subjects for well-powered inferences.

Joint multivariate methods have the potential to better capture the complex relationships that may exist between different biological levels and significantly reduce the number of statistical tests, accounting for the multiple testing correction problems (Liu and Calhoun 2014). Multiblock methods are an alternative to address problems regarding marginal analyses (Kawaguchi et al. 2017).

Multiple Factor Analysis (MFA) is one of the most popular methods for analyzing multiple sets of variables measured on the same observations (Husson et al. 2011). MFA aims to provide common factor scores and measures of discrepancy between blocks of variables (Abdi et al. 2013). However, singular value decomposition in MFA is based on PCA, which is adequate only if the data is normally distributed, linear or stationary. Also, including strongly correlated variables can overemphasize the contribution of the estimated principal components (Lever et al. 2017). To overcome these problems, we propose a novel method called Independent MFA. This method is an extended implementation of MFA, where the component value decomposition is based on Independent Component Analysis (ICA) that does not assume multivariate normality and linearity (Hyvärinen 2013).

The main advantages of the proposed method are that (1) it is applicable if there is correlation between variables within structures, (2) it increases the variability explained by the data components and, (3) it performs a feature selection considering the correlation data structure of variables.

This article is organized as follows. In section 2, we propose an extension of MFA, referred to as Independent Multiple Factor Analysis (ICA-MFA). ICA-MFA incorporates ICA as a generalization of PCA decomposition. ICA-MFA also incorporates a feature selection based on a meaningful independent component regression (ICR). In section 3, we explore the performance of the algorithm. We evaluate and compare the performance of ICA-MFA with both the MFA method and traditional univariate analyses in a simulation study. In section 4, we applied and compare ICA-MFA, MFA and univariate analysis in an imaging genetics study using data from the Rotterdam Study (Ikram et al. 2017). The main results of simulations and real data analyses are discussed in the final section of the paper.

Method

MFA is a multivariate version of Factorial Analysis (FA) and an extension of PCA used to integrate m different sets of variables (in a matrix format), $X_1,\ldots,X_m$ on the same set of observations. MFA is mainly comprised of three steps:

First, a PCA of each data set is performed via single value decomposition (SVD). The SVD of a given IxJ rectangular data matrix X is its factorization into three matrices

$$X = U\Gamma V^T \text{ such that } U^T U = V^T V = I,$$

where U is a IxL matrix of the normalized left singular vectors, V is a JxL matrix of the normalized right singular vectors and is the LxL diagnal matrix of the L singular values, L being the rank of the decomposed matrix X, and U and V being orthonormal matrices.

Second, data sets are normalized by dividing all the elements of each table X_i by the corresponding explained variance of the first singular vector, given by the inverse of the first squared singular value $_{i,\,(1,\,1)}$.

Finally, the normalized data sets are concatenated into a unique data set and 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.

To address problems related with the single value decomposition in PCA (orthogonality assumption and multivariate normal distribution of the variables in each dataset), we present a statistical methodology based on an extension of MFA, referred as ICA-MFA. This approach is designed to evaluate potential relationships between sources of data based on ICA decomposition and ICR that is used to link MFA results with an outcome of interest.

Independent Component Analysis (ICA)

ICA aims to find a linear representation of non-Gaussian vectors such that the estimated vectors are statistically independent (Comon 1994). ICA decomposition is similar to PCA

model, but while the PCA identifies linear combinations of the original variables such that the covariance between the derived variables is zero, ICA identifies statistically independent variables. The PCA identifies linear combinations of the original variables such that the covariance between the derived variables is zero. As independence implies null covariance and not vice versa, it is a stronger condition that can better reflect the intrinsic properties of mixed signals.

The ICA decomposition of X is given by

$$X = AS,$$

where S and A are matrix of independent components and mixture matrix, respectively. Independent component regression (ICR) is similar to Principal Component regression (PCReg), with the difference that ICR uses independent components S, and the coefficient matrix, A, obtained by ICA in the regression analysis instead of the principal components and matrix of scores obtained by PCA decomposition. Hence, since X can be described by its coefficient matrix, A, the multiple linear regression equation between A and the matrix of components, can be defined as in PCReg (Bair et al. 2006).

Independent Multifactorial Analysis (ICA-MFA)

We propose a multiblock framework to evaluate relationships between two rectangular data matrices collected on the same set of observations. Although the method is described considering two data sets (genetic data and imaging features), the procedure can be extended to K matrices.

Consider an imaging genetics study where $N_{n,\ k}$ and $G_{n,\ p}$ denotes blocks of neuroimaging and genetic data, where n is the number of individuals, k is the number of neuroimaging-based features (i.e. brain volumes, …) and p denotes the number of genetic variants (i.e., SNPs, genetic scores, structural variants, …). The proposed algorithm comprises five steps (Fig. 1):

Step 1. *Computing ICA decomposition on each block of variables*: An ICA decomposition of each block of variables is performed in order to search linear combinations of variables that optimize statistical independence. Let us assume c independent components $\Phi_1, \ldots, \Phi_c$. Therefore, by definition, the joint probability density function (pdf) is factorizable as the joint product of c terms. Then, we obtain a set of observation signals for each dataset respectively, X_{N_j}, X_{G_z}, $j = 1,\ldots,k$ and $z = 1,\ldots p$, that are mixtures of the original variables. Then, we can model the mixing process decomposing into a linear mixture $x_1, \ldots, x_c$, of c independent components for each dataset,

$$\begin{cases} X_N = A\phi_{N_c} \\ X_{\mathrm{G}} = B\phi_{G_c} \end{cases}$$

where A, B are the associated mixing matrices. Notice that x vectors are understood as column vectors; thus the transpose of x (x^T), is a row vector.

Then, after estimating the matrices A, B, we can compute its inverse, W_N, W_G and obtain the independent components simply by:

$$\begin{cases} \phi_{N_c} = W_N \cdot x_N \\ \phi_{G_c} = W_G \cdot x_G \end{cases}$$

Step 2. *Normalization of each data table*: data sets are scaled by dividing all of its elements by the square root of the first independent component from those obtained in step 1, following the same strategy as in MFA.

$$\begin{cases} Z_N = N_{nxk}\sqrt{\phi_{N_1}^{-1}} \\ Z_G = G_{nxp}\sqrt{\phi_{G_1}^{-1}} \end{cases}$$

Step 3. *Concatenation of data sets*: The normalized datasets, Z_N and Z_G are concatenated into a complete dataset denoted by C.

$$C = Z_N \vee Z_G$$

Step 4. *Compute an ICA on the generalized dataset C:* ICA decomposition is performed on the concatenated Table C to extract a vector of i ($i = 1, \ldots, C$),independent imaging genetic components, Φ_{i},

$$C = \Sigma \cdot \Phi_{\mathrm{i}}$$

where Σ is the associated mixing matrix with elements σ_{yt}.

Step 5. *Feature Selection:* Finally, an ICR through a hold-out validation is computed to determine relevant features related to our outcome of interest (dichotomous or quantitative trait), Y, and the total amount of variability explained for those features,

$$Y_j = \Phi_{ji}^{\mathrm{T}}\beta_i + \varepsilon_j,$$

where j denotes the j-th sample ($j = 1, \ldots, N$), i the number of components ($i = 1, \ldots, C$), β the effect of each independent component, and Φ^{T} denotes the transpose, so what $\Phi_{\mathrm{j}}^{\mathrm{T}}\beta$ is the inner product between vectors Φ_{j} and β. In particular,$\Phi_{\mathrm{j}0} = 1$, and the corresponding element, β_0,is the intercept.PC.

Selection of A Priori Independent Components

Given a set of candidate number of independent components, $c = \{c_1, c_2, \ldots, c_l\}$, we compute l-times an ICA, specifying a different number of a priori components in each computation. For each pair of components, we calculate the log likelihood ratio

representing the relative likelihood of a correlation between both independent components. The number of selected components is the value, which minimizes the mean log likelihood of components relatedness.

Simulation Study

Simulation Design

We performed a simulation study to compare the variability explained by ICA-MFA, MFA, and univariate linear regression models for a quantitative trait. We use the *PhenotypeSimulator* package from GitHub (https://github.com/cran/PhenotypeSimulator). *PhenotypeSimulator* functions fit a linear model with the genotype as the explanatory variable and the phenotype as the response variable, including the effect of additional covariates and random noise.

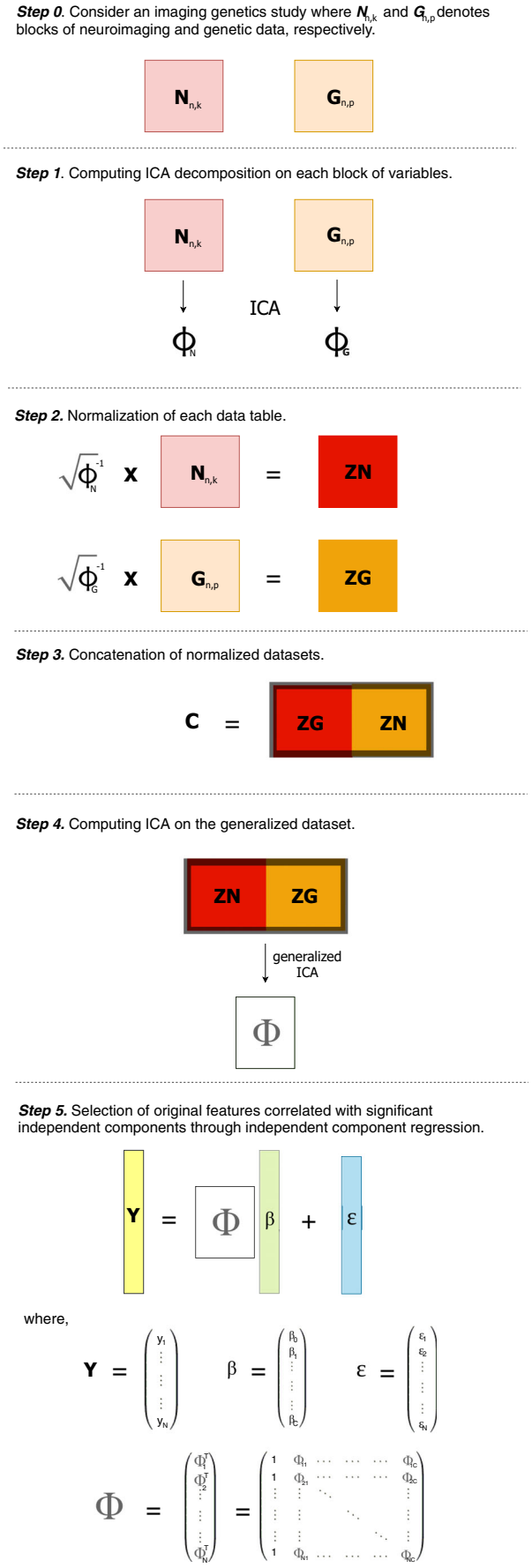
We simulated datasets with different sample sizes N, (N = 1000, 3000), a quantitative outcome Y mimicking disease score, n_{SNPs}variables representing the genotypes of a set of Single Nucleotide Polymorphisms (SNPs) (n_{SNPs}=*10, 100, 1000*), a set of causal SNPs in each simulated set of genetic variants (nc_{SNPs} =*10, 100, 1000*), and n_b=15 image variables representing different brain structures. Genetic effects were simulated as the matrix product of genotype matrix, $N \times nc_{SNPs}$, and effect size matrix $n_{SNPs} \times n_b$, assuming a Linkage Disequilibrium (LD) structure, an additive genetic model and allele frequencies of 5, 10, 30 and 40%. Allele frequencies were uniformly sampled and used to simulate individual genotypes by drawing values from a binomial distribution with 2 trials. Information was summarized using non-standardized allele codes (i.e; 0, 1, 2). Brain structure effects were simulated as quantitative variables following a multivariate normal distribution. From realistic data (McCarthy et al. 2015; Table 2) we extracted the mean modulate values, μ_{n_b}, and standard deviations, σ_{n_b}, of 15 scanner-specific cortical thickness values for brain structures. Each source of data was scaled to explain a certain proportion of the entire outcome variance. We assumed that the proportion of variance explained by brain structure components was 30%, the total genetic variance 40%, and the proportion of variance of fixed genetic effects 2.5%. Moreover, single SNP effects were assumed of 1% of the total phenotypic variance.

In addition, to illustrate the general n<<p case (in our context $N << n_{SNPs}$), we included scenarios with sample sizes N = 50, 100, 500, n_{SNPs}=1000, and n_b=15.

In total, we simulated 9 different scenarios assuming combinations of the considered parameters. The information is summarized in Table 1.

Simulation Evaluation Performance

We compared the performance of each method by computing the variability of the outcome (Y). The variability of Y was



◀ **Fig. 1** Steps of Independent Multiple Factor Analysis (ICA-MFA) method. Legend: $N_{n,k}$ = matrix of neuroimaging data; $G_{n,p}$ =matrix of genetic data; ϕ_N = first independent component from $N_{n,k}$; ϕ_G = first independent component from $G_{n,p}$; Z_N= normalized $N_{n,k}$; Z_G = normalized $G_{n,p}$; C = concatenation of normalized datasets; Φ_i= matrix of independent imaging genetic components; Y = outcome of interest; β= effects of independent imaging genetic components; ϵ = error terms

calculated as the explained sums of squares (ESS) due to hold-out validation (caret R-package) through an independent component regression (ICR) for the ICA-MFA method. For the MFA method we used principal component regression (PCReg), and for univariate models, we used a linear regression. The ESS is defined as the sum of the squares of the differences of the predicted values and the mean value of the response value:

$$ESS = \sum_i \left(\hat{y}_i - \overline{y}\right)^2$$

We additionally compared the goodness of fit for the different models based on an analysis of deviance with the inclusion of the latent components. We compared each model with models that do not include any component through Akaike criteria information. Notice that lower numerical values of AIC statistic indicate a better fit of the model to the observed data (Akaike 1998). Moreover, we tested whether the extracted principal and independent components were significantly associated with *Y*.

Simulation Results

Results of the simulation studies are summarized in Tables 2 and 3. We observed that in all scenarios, ICA-MFA outperforms both MFA and univariate regression approaches in terms of variability explained. Specifically, ICA-MFA provided an increase of ~10% of variance explained in all scenarios. We also observed that the magnitude of variability explained does not depend on the sample size or the number of imaging genetic covariates included in the models (Table 2).

Moreover, ICA-MFA can handle scenarios where the number of features is smaller than the number of samples. Again, in these scenarios, ICA-MFA outperformed the MFA method and the univariate regression approach while maintaining the percentages of variability obtained for the previous scenarios (Table 3). Moreover, in all scenarios, independent components obtained from ICA-MFA provides a better goodness of fit based on the Akaike information criterion (AIC), and a better performance of prediction compared with components obtained from MFA (Figs. 2 and 3).

For reproducibility purposes, the data sets and scripts supporting the results of these simulation studies can be found at https://github.com/natvt8/ICA-MFA.

Table 1 Characteristics of the simulated scenarios

Scenario	N	n_SNPs	n_b	δ	h2
1	1000	10	15	0.4	0.025
2	1000	100	15		
3	1000	1000	15		
4	3000	10	15		
5	3000	100	15		
6	3000	1000	15		
7	50	1000	15		
8	100	1000	15		
9	500	1000	15		

N Number of individuals; *n_SNPs* Number of SNPs; *n_b* Number of simulated brain regions; *δ* Proportion of fixed variance from brain structure components; *h2* Proportion of variance of fixed genetic effects

Application to Real Dataset: Executive Cognitive Function

We applied ICA-MFA and MFA methods on a subset of imaging genetics data from the Rotterdam Study (Ikram et al. 2017).

Study Population

The Rotterdam study is a prospective population-based cohort study comprising of 14,926 middle aged and elderly individuals, investigating the determinants and consequences of age-related diseases in older adults. Genotyping was performed on 11,496 individuals and 5691 unique participants underwent brain magnetic resonance imaging (MRI).

From the total of 14,926 participants in the Rotterdam Study, genotypic, neuroimaging data and executive function

Table 2 Percentage of variability explained depending on the sample size, N, and the number of single nucleotide polymorphisms

	ICA-MFA*	MFA**	univariate***
N = 1000			
n_SNPs = 10, n_b = 15	10.64%	0.32%	0.05%
n_SNPs = 100, n_b = 15	9.69%	0.59%	0.05%
n_SNPs = 1000, n_b = 15	10.05%	0.25%	0.05%
N = 3000			
n_SNPs = 10, n_b = 15	13.53%	0.17%	0.04%
n_SNPs = 100, n_b = 15	11.58%	0.10%	0.04%
n_SNPs = 1000, n_b = 15	10.04%	0.08%	0.04%

Scenarios N >> n_SNPs (from 1 to 6)

N Number of individuals; *n_SNPs* Number of SNPs; *n_b* Number of brain structures; *ICA-MFA* Independent Multiple factor Association method; *MFA* Multiple factor Analysis

Table 3 Percentage of variability explained depending on the sample size, N, and the number of Single Nucleotide Polymorphisms

	ICA-MFA*	MFA**	univariate***
n_SNPs = 1000, n_b = 15			
N = 50	10.12%	3.70%	~0%
N = 100	9.00%	2.00%	~0%
N = 500	8.80%	0.99%	~0%

Scenarios N < < n_SNPs (from 7 to 9)

N Number of individuals; *n_SNPs* Number of SNPs; *n_b* Number of brain structures; *ICA-MFA* Independent Multiple factor Association method; *MFA* Multiple factor Analysis

were available for 4057 individuals (mean age 64.7). [Fig. S1]. The Rotterdam study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. A written informed consent was obtained from all participants.

Executive Cognitive Function

Executive cognitive function was assessed with the Letter-Digit Substitution test (LDST: Jolles et al. 2017). The LDST asks the participants to make as many letter-digit combinations as possible in 60 s, following an example that shows the correct combinations. Normative LDST have been well established for adults (van der Elst et al. 2006). In the Rotterdam Study all LDST were administered by trained investigators in quiet rooms. The test took no longer than 30 min to complete and a stopwatch was used for the control of time (Hoogendam et al. 2014). LDST were assessed from 1997 to 1999 and consecutive follow-up examinations every 3 to 4 years have been conducted until now. For analytical purposes, in this study, we selected those executive function measurements closest to the brain MRI performed in the study participants.

Image Acquisition, Processing and Selection

Magnetic Resonance Imaging (MRI) scanning was done on a 1.5-T MRI scanner (Signa Excite II; General Electric Healthcare, Milwaukee, WI, USA). The MRI protocol included a high-resolution axial T1-weighted 3-dimensional fast radio frequency spoiled gradient recalled acquisition in steady state with an inversion recovery prepulse (FASTSPGR-IR) sequence (repetition time [TR] = 13.8 ms, echo time [TE] = 2.8 ms, inversion time [TI] = 400 ms, field of view [FOV] = 25 cm2, matrix = 416 × 256, flip angle = 20°, number of excitations [NEX] = 1, bandwidth [BW] = 12.50 kHz, 96 slices with slice thickness 1.6 mm 0-padded to 0.8 mm). All slices were contiguous. According to the Rotterdam Study standard acquisition protocol images were resampled to 512 × 152 × 192 voxels (voxel size: 0.5 × 0.5 × 0.8 mm3). The T1-weighted MRI scans were processed using a model-based automated procedure of Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) (Fischl et al. 2004) to obtain segmentations and volumetric summaries of subcortical structures and thickness of the cerebral cortex. This procedure automatically assigns a neuroanatomical label to each voxel in an MRI volume based on probabilistic information obtained from a manually labeled training set. This yielded intracranial volume (ICV) and gray and white matter volumes for cerebellum and cerebrum. Further details of the MRI protocol can be found in (Ikram et al. (2015). For the purpose of this study, we included all subcortical structures pre-processed using Freesurfer, excluding those labels that correspond to non-brain regions. In addition to the subcortical volumes of each hemisphere (right/left), we have included the total volume of the structure (being the sum of the volume of the region in

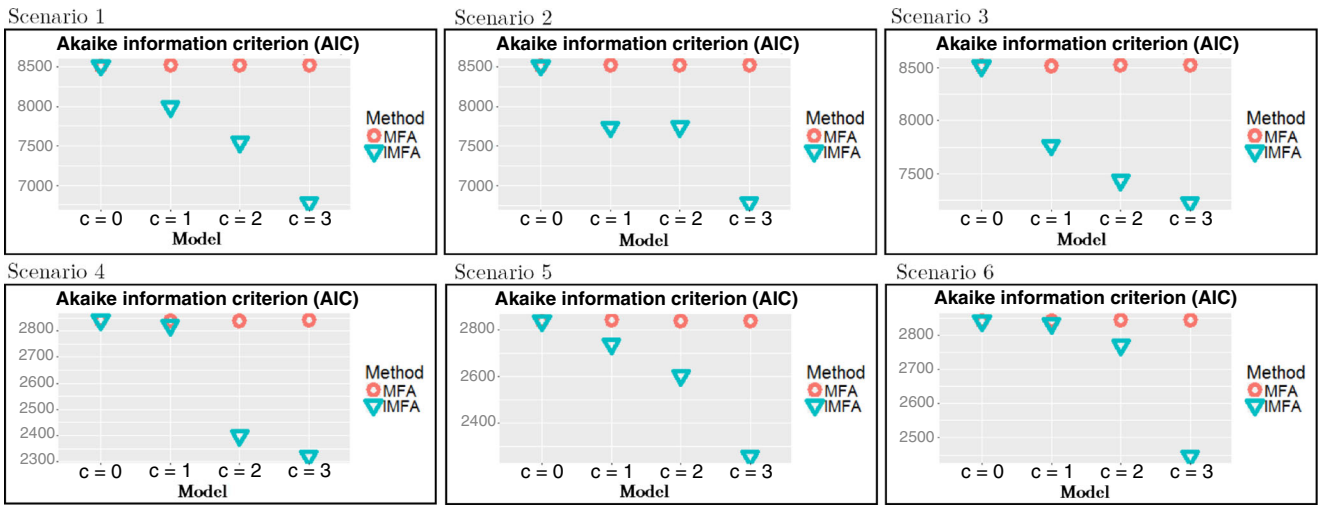


Fig. 2 Comparison of the Goodness of fit based on the Akaike information criterion (AIC), considering the inclusion of c = 0, 1, 2 and 3 components for ICA-MFA and MFA methods. Scenarios N > > n_SNPs (from 1 to 6)

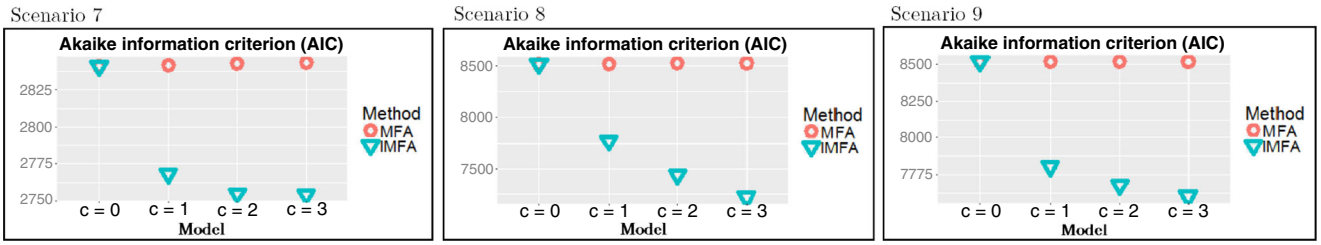


Fig. 3 Comparison of the Goodness of fit based on the Akaike information criterion (AIC), considering the inclusion of c = 0, 1, 2 and 3 components for ICA-MFA and MFA methods. Scenarios N << n_SNPs (from 7 to 9)

each hemisphere). The study included 39 subcortical structures [Table S1].

Genotyping Acquisition and Genetic Variant Selection

The Rotterdam Study consist on three subcohorts, which were genotyped with the 550 K (cohort 1), 550 K duo (cohort 2) and 610 K (cohort 3) Illumina arrays. Samples with a call rate below 97.5%, gender mismatch, excess autosomal heterozygosity (>0.336), duplicates or family relations and ethnic outliers were excluded. Genetic variants were filtered by Hardy-Weinberg equilibrium ($P<10^{-6}$), allele frequency (excluding minor allele frequency (MAF < 0.001)) and SNP call rate with a minimum of 98%. Genotypes were imputed using MACH/minimac software to the 1000 Genomes phase I version 3 reference panel (all populations). Among the variants imputed, a total of 9 loci recently associated with Attention-Deficit/Hyperactivity Disorder (ADHD) in an independent meta-analysis from Demontis et al. (2018), at a genome-wide threshold of significance ($P<10^{-8}$) were pre-selected [Table S2]. Moreover, we constructed a genetic risk score (GRS) by multiplying the number of risk alleles by their reported odds ratio (after natural logarithm transformation) for the disease, and summing this weighted allele score of each variant up into a disease risk score for ADHD.

Results

Variability Explained by each Component

ICA-MFA identified three independent components (Φ) that pass significance criteria Φ_1 ($P = 2.22E\text{-}94$); Φ_2($P = 9.03E\text{-}08$); Φ_3($P = 2.71E\text{-}88$), explaining approximately 18% of the global variance of executive function, while the first three principal components (PCs) from MFA were only able to detect 1% (Table 4). Specifically, we show how the increment in the variability explained for executive function varies by only 1% when going from incorporating a main component to three main components in the MFA procedure. For ICA-MFA, the amount of variability explained is around 9% considering one IC, increases to 10% when considering two components, and reaches 18% when including all three components. It seems then that the first and third components would be the most representative in the quantification of the total variability of executive function.

Contribution of Variables to each Dimension

Figures from S2 to S4 show those variables contributing the most to the definition of the three dimensions of ICA-MFA. Variables that contribute the most to the first dimension are lateral ventricle volumes, cerebellar cortex, cerebellum, white matter, and hippocampus volumes. Variables that contribute to the second dimension are white matter and gray matter volumes, and also cerebellar cortex. Finally, variables that contribute to the third dimension are gray matter, cerebellar cortex, lateral ventricles and corpus callosum volumes. For this third dimension we additionally appreciate the contribution of three genetic components, rs1427829 (*DUSP6/POC1B*), rs4858241 (*Intergenic*) and rs9677504 (*SPAG16*). Moreover, none of the dimensions of ICA-MFA are characterized by the influence of the genetic risk score.

Discussion

The aim in the field of imaging genetics is to find relations between genetic data and imaging phenotypes using large

Table 4 Variability explained by the first three principal component (PC1, PC2, PC3; MFA) and by the first three independent component (IC1, IC2, IC3; ICA-MFA) of the global variance of executive cognitive function

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	PctExp
PC1	1	994	994	20	7.31E-06	0.495
PC2	1	14	14	0.27	5.99E-01	0.0068
PC3	1	15	15	0.3	5.80E-01	0.0075
Residuals	4053	199,911	49			99.49
IC1	1	18,194	18,194	448	2.22E-94	9.05
IC2	1	1165	1165	29	9.03E-08	0.58
IC3	1	16,937	16,937	417	2.71E-88	8.43
Residuals	4053	164,638	41			81.94

Df Degrees of Freedom; *Sum Sq* Sum of Squares; *Mean Sq* Mean Squares; *F value* F ratio; *Pr(>F)* Pvalue; *PctExp* Percentage of variability explained

datasets; these relations often have a small effect size (Abi-Dargham and Horga 2016; Medland et al. 2014). In order to increase statistical power, new methods and technologies for data reduction are being considered. We developed a new method, referred to as ICA-MFA, which better explains variability in multifactorial analyses than conventional methods. Our method incorporates independent component decomposition instead of the more common principal component analysis. Decomposing the data into its most important sources of variation holds the potential to discover unanticipated sources of signals with biological meaning, and generate new hypotheses. The proposed multifactorial method derives an integrated picture of the observations and the relationships between the groups of variables. This strategy takes into account the structure of the genetic data and imaging markers, reduces the computational burden posed by large amounts of data, and handles the case in which the number of features is smaller than the number of samples (not common in other IG strategies). Furthermore, most variants identified confer relatively small risk increments, and the amount of variability that is explained by these genetic components, expected for traits showing a polygenic architecture, is relatively small (Manolio et al. 2009). By taking advantage of the independent component decomposition, the proposed method outperforms multifactorial analysis and univariate regressions in a simulation study. Moreover, in a real life proof of principle study, the explained variability accounted for by our proposed method is higher, demonstrating the potential of the algorithm.

We explored the performance of the proposed algorithm on a subset of imaging genetics data from the population-based Rotterdam Study, in which we explored genetics and imaging features in relation to executive cognitive function in an adult population sample. Instead of independently performing univariate regressions, or applying MFA, we integrated the multimodal feature datasets applying independent component decomposition. The proposed strategy makes it possible to determine the degree to which the whole set of genetic and/or neuroimaging markers contribute to the variability of the symptomatology jointly, rather than individually. While univariate results and multimodal MFA combinations only explained a limited proportion of variability (less than 2%), the proposed ICA-MFA increased the explained variability (9%) and allowed the identification of significant independent components that maximize the variability explained. Though meant primarily as a proof of principle example, from a biological perspective, the results obtained in this real data sample provide new views of research on the characterization of the cognitive processes that underlie more complex symptoms such as ADHD (Curatolo et al. 2010; Purper-Ouakil et al. 2011). Moreover, results obtained could suggest an approximation of the joint affectation of genetic profiles and changes at the level of brain structure on executive cognitive function (Mueller and Tomblin 2012; Willcutt et al. 2005).

The potential application of the ICA-MFA algorithm on imaging genetics studies constitutes an important aspect of integrating imaging genetics data, especially in relation to neurodevelopment domains due to the small number of studies and inconsistency of the results (Durston 2010; Vilor-Tejedor et al. 2016).

In addition, notice that the presented method, like most common multivariate methods used in imaging genetic studies, is based on PCA-based dimensionality reduction techniques (Pearson 1901), which are often a good strategy to deal with high-dimensional data (Liu and Calhoun 2014; Sui et al. 2012; Vilor-Tejedor et al. 2018). In these methods, data are replaced by a summary that still captures as much information as possible from the original data. The information is captured in principal components that summarize the data. The amount of information explained may vary through the principal components. The main differences among these methods are the assumptions held and the dimension of the data space. PCA relies on the assumption that the obtained components (PCs) are linear combinations of the original variables, independent, and continuous (multidimensional normality). The PCs are, in addition, orthogonal to each other, which allows effectively explaining variation of original variables and may have a much lower dimensionality. However, while PCA deals with only one data space, X (where it identifies directions of high variance), canonical correlation analysis (CCA, Härdle and Simar 2007) proposes a way for dimensionality reduction by taking into account relations between samples coming from two spaces, X and Y. The assumption is that the observations from these two spaces contain some joint information that is reflected in correlations between them. Directions showing high correlations are thus assumed to be relevant. PCA and CCA are also closely related to partial least squares regression (PLS, Rosipal and Krämer 2005) because both methods aim to define a linear relationship between a dependent variable/set of variables, Y, and predictor variables, X. Hence, the goal is to determine which aspects of a set of observations (e.g., imaging data) are related directly to another set of data (e.g., genetic data, phenotypic data). PLS maximizes the covariance between latent variables of the two modalities, while CCA maximizes the correlation between them. Hence, PLS is a way to model multivariate responses and multiple features. Following this strategy, the reduced-rank regression (sRRR, Chen and Huang 2012) method takes a more general formation based on a multivariate linear regression from X to Y. It reduces the rank of the project matrix, which facilitates an efficient search of multiple markers that are highly predictive of multiple phenotypes. However, notice that the core computations of PLS, CCA and RRR all involve single value decomposition so that the latent variables or projection vectors within one modality (genetic or imaging) are orthogonal to each other. In contrast, ICA emphasizes that latent variables (components) are maximally independent from each other,

which can be optimized through many forms of statistical measures, including minimization of mutual information and maximization of non-Gaussianity. One extension of ICA methods applied to imaging genetics is parallel ICA, which simultaneously maximizes both the independence of components and the correlations between projection vectors of the two modalities (Liu et al. 2008b). Following this strategy, our proposed method, ICA-MFA, allows accounting for the independence of components and for the relations between projection vectors of n spaces. Moreover, ICA-MFA allows integrating both numerical and categorical groups of variables and including supplementary groups of the data that need to be added to the analysis (e.g., integration of multi-omics data). However, when data are scarce as compared to the dimensionality of the problem, it is important to regularize the problem to avoid overfitting. This is provided, for instance, in the regularized CCA (RCCA, Cruz-Cano and Lee 2014) algorithm, and it is a possible potential extension to our methodological proposal. Hence, further research may greatly benefit from the development of multivariate approaches which represent a potential form to increase the statistical power to detect significant causal factors in multiblock data analysis (Meyer-Lindenberg 2012).

Information Sharing Statement

The latest version of our algorithm ICA-MFA can be accessed over GitHub: github.com/natvt8/ICA-MFA. Scripts for the simulation study are as well maintained on GitHub: github.com/natvt8/ICA-MFA. Moreover, the used dataset in this paper came from the Rotterdam Study project.

Acknowledgements Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2017 FI_B 00636), Generalitat de Catalunya – Fons Social Europeu. This work has been partially supported by a STSM Grant from EU COST Action 15120 Open Multiscale Systems Medicine (OpenMultiMed) and Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP). Further support was obtained through the Ministerio de Economía e Innovación (Spain), grant MTM2015-68140-R. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

Silvia Alemany thanks the Institute of Health Carlos III for her Sara Borrell postdoctoral grant (CD14/00214).

The generation and management of GWAS genotype data for the Rotterdam Study are supported by the Netherlands Organization of Scientific Research NWO Investments (no. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/ Netherlands Organization for Scientific Research (NWO) project no. 050-060-810. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. This research is supported by the Dutch Technology Foundation STW (12723), which is part of the NWO, and which is partly funded by the Ministry of Economic Affairs. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project: ORACLE, grant agreement No: 678543).

Compliance with Ethical Standards

Conflict of Interest None.

References

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*(2), 149–179. https://doi.org/10.1002/wics.1246.

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nature Medicine, 22*(11), 1248–1255. https://doi.org/10.1038/nm.4190.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle (pp. 199–213). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_15.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association, 101*(473), 119–137. https://doi.org/10.1198/016214505000000628.

Chen, L., & Huang, J. Z. (2012). *Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection in Multivariate Regression*. Retrieved from http://www.stat.yale.edu/~lc436/Chen_Huang_2012_JASA.pdf

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9.

Cruz-Cano, R., & Lee, M.-L. T. (2014). Fast regularized canonical correlation analysis. *Computational Statistics & Data Analysis, 70*, 88–100. https://doi.org/10.1016/J.CSDA.2013.09.020.

Curatolo, P., D'Agati, E., & Moavero, R. (2010). The neurobiological basis of ADHD. *Italian Journal of Pediatrics, 36*(1), 79. https://doi.org/10.1186/1824-7288-36-79.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., et al. (2018). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics, 51*(1), 63–75. https://doi.org/10.1038/s41588-018-0269-7.

Durston, S. (2010). Imaging genetics in ADHD. Retrieved September 3, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/20206707.

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage, 23*, S69–S84. https://doi.org/10.1016/j.neuroimage.2004.07.016.

Härdle, W., & Simar, L. (2007). *Applied Multivariate Statistical Analysis* *. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?10.1.1.233.897&rep=rep1&type=pdf

Hoogendam, Y. Y., Hofman, A., van der Geest, J. N., van der Lugt, A., & Ikram, M. A. (2014). Patterns of cognitive function in aging: The Rotterdam study. *European Journal of Epidemiology, 29*(2), 133–140. https://doi.org/10.1007/s10654-014-9885-4.

Hoogman, M., Guadalupe, T., Zwiers, M. P., Klarenbeek, P., Francks, C., & Fisher, S. E. (2014). Assessing the effects of common variation in the FOXP2 gene on human brain structure. *Frontiers in Human Neuroscience, 8*(473). https://doi.org/10.3389/fnhum.2014.00473.

Husson, F., Lê, S., & Pagès, J. (2011). Exploratory multivariate analysis by example using R. CRC Press. Retrieved from https://www.crcpress.com/Exploratory-Multivariate-Analysis-by-Example-Using-R/Husson-Le-Pages/p/book/9781439835814

Hyvärinen, A. (2013). Independent component analysis: Recent advances. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 371*(1984), 20110534. https://doi.org/10.1098/rsta.2011.0534.

Ikram, M. A., van der Lugt, A., Niessen, W. J., Koudstaal, P. J., Krestin, G. P., Hofman, A., Bos, D., & Vernooij, M. W. (2015). The Rotterdam scan study: Design update 2016 and main findings. *European Journal of Epidemiology, 30*(12), 1299–1315. https://doi.org/10.1007/s10654-015-0105-7.

Ikram, M. A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., Klaver, C. C. W., Nijsten, T. E. C., Peeters, R. P., Stricker, B. H., Tiemeier, H., Uitterlinden, A. G., Vernooij, M. W., & Hofman, A. (2017). The Rotterdam study: 2018 update on objectives, design and main results. *European Journal of Epidemiology, 32*(9), 807–850. https://doi.org/10.1007/s10654-017-0321-4.

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (2017). The Maastricht Aging Study: Determinants of cognitive aging. Retrieved from http://www.np.unimaas.nl/maas

Kawaguchi, A., Yamashita, F., & Alzheimer's Disease Neuroimaging Initiative. (2017). OUP accepted manuscript. *Biostatistics, 18*(4), 651–665. https://doi.org/10.1093/biostatistics/kxx011.

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of significance: Principal component analysis. *Nature Methods, 14*(7), 641–642. https://doi.org/10.1038/nmeth.4346.

Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics, 8*(29). https://doi.org/10.3389/fninf.2014.00029.

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights, 8*, 1–10. https://doi.org/10.4137/BII.S31559.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747–753. https://doi.org/10.1038/nature08494.

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J.-A., Coman, I. L., & Kates, W. R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience, 9*(379). https://doi.org/10.3389/fnins.2015.00379.

Medland, S. E., Jahanshad, N., Neale, B. M., & Thompson, P. M. (2014). Whole-genome analyses of whole-brain data: Working within an expanded search space. *Nature Neuroscience, 17*(6), 791–800. https://doi.org/10.1038/nn.3718.

Meyer-Lindenberg, A. (2012). The future of fMRI and genetics research. *NeuroImage, 62*(2), 1286–1292. https://doi.org/10.1016/j.neuroimage.2011.10.063.

Mueller, K. L., & Tomblin, J. B. (2012). Diagnosis of ADHD and its behavioral. *Neurologic and Genetic Roots Topics in Language Disorders, 32*(3), 207–227. https://doi.org/10.1097/TLD.0b013e318261ffdd.

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559–572. https://doi.org/10.1080/14786440109462720.

Purper-Ouakil, D., Ramoz, N., Lepagnol-Bestel, A.-M., Gorwood, P., & Simonneau, M. (2011). Neurobiology of attention deficit/hyperactivity disorder. *Pediatric Research, 69*(5 Part 2), 69R–76R. https://doi.org/10.1203/PDR.0b013e318212b40f.

Rosipal, R., & Krämer, N. (2005). Overview and recent advances in partial least squares. In *Notes in Computer Science* https://doi.org/10.1007/11752790_2.

Sui, J., Adali, T., Yu, Q., Chen, J., & Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods, 204*(1), 68–81. https://doi.org/10.1016/j.jneumeth.2011.10.031.

van der Elst, W., van Boxtel, M. P. J., van Breukelen, G. J. P., & Jolles, J. (2006). The letter digit substitution test: Normative data for 1,858 healthy participants aged 24–81 from the Maastricht aging study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology, 28*(6), 998–1009. https://doi.org/10.1080/13803390591004428.

Vilor-Tejedor, N., Cáceres, A., Pujol, J., Sunyer, J., & González, J. R. (2016). Imaging genetics in attention-deficit/hyperactivity disorder and related neurodevelopmental domains: State of the art. *Brain Imaging and Behavior, 11*, 1922–1931. https://doi.org/10.1007/s11682-016-9663-x.

Vilor-Tejedor, N., Alemany, S., Cáceres, A., Bustamante, M., Pujol, J., Sunyer, J., & González, J. R. (2018). Strategies for integrated analysis in imaging genetics studies. *Neuroscience and Biobehavioral Reviews, 93*, 57–70. https://doi.org/10.1016/j.neubiorev.2018.06.013.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57*(11), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006.