Proposal form

SSH Open Competition L 2023

Deadline **Tuesday 13 February 2024, 14:00 hrs CET.**

Title

Linking brain function and structure to phenotypes: does more data in fixed sample size lead to higher replicability?

Scientific proposal

1. State of the Art

One of the longstanding mysteries in science is what drives the profound variation in behaviour like treatment response, clinical symptoms, and other phenotypes. The search for the neural underpinnings of these individual differences has been one of the key motivating forces behind neuroscience research in the last couple of decades. **Brain-wide association studies (BWAS)** found preliminary links between individual brain function and structure and various phenotypes [e.g., 1]. In such studies, brain measures include structural MRI, resting-state fMRI, or task fMRI activations whereas phenotypes include physical or mental health, cognition, or personality measures.

Mass univariate BWAS analyses usually link a single brain feature to a single behavioural phenotype and repeat this procedure for all single brain features. Such analyses can test up to thousands brain voxels or regions individually. In contrast, ***multivariate*** **BWAS** integrates all features across brain regions into a single **predictive model** for a specific phenotype. In the multivariate approach, more data is used in a single statistical analysis. In detail, suppose we have, say, 2000 voxel wise measurements of brain structure, such as *cortical thickness*, and one phenotype, such as *working memory*. In the mass univariate approach this would lead to 2000 separate analyses, whereas in the multivariate approach this would lead to only 1 analysis with 2000 predictor variables (features) in a regression model (e.g., a support vector regression or a ridge regression). Substantively, the multivariate BWAS approach is biologically much more realistic, as brain cells and regions tend to operate in tandem instead of individually. Statistically, the multivariate approach takes into account that the voxel wise measurements are correlated, whereas the univariate approach completely neglects those correlations, and therefore the multivariate approach has better power and reliability [2]. Furthermore, multivariate analysis often uses **prediction models**, that are considered favourable [3], whereas the univariate approach uses correlations.

Many scientific disciplines, including neuroscience, behavioural science, and medical science, showed an increased interest in reproducibility and **replicability** [4]. Reproducibility refers to the question whether the same results can be obtained from the same data with the same methods, whereas replicability refers to the question whether the same results can be obtained from new data with the same methods [5]. The reproducibility and replicability of scientific findings is fundamental to their validity and utility in guiding further research, technological development, and treatment development. Threats for replicability include small sample sizes, measurement error in the brain features and phenotypes, and questionable research practices [5].

Recently the endeavour of linking brain and phenotypes through brain-wide association studies has come under fire [6] suggesting that **finding replicable brain-behaviour associations requires thousands of individuals**. In their study, [6] used data from large scale consortia studies: the Adolescent Brain Cognitive Development (ABCD) study with 11,874 participants, the Human Connectome Project (HCP) with 1,200 participants, and the UK Biobank (UKB) project with 37,735 participants. The data in these studies arise from diverse populations. Results from diverse populations are usually better generalizable to data from new populations than results obtained from homogenous samples. From these large data sets, subsamples were created and analysed, and it was verified whether a significant result obtained in one subsample (the *discovery sample*) replicated in another subsample (the *replication sample*). This procedure of subsampling and analysing was repeated for different subsample sizes from small to large. Such a subsampling procedure gives an upper bound estimate of replicability, as the discovery sample and replication sample come from the same population. To replicate the (significant) result in data from another population would be even more difficult. [6] found that small samples often resulted in false positives, false negatives, and a low probability of replication. Furthermore, estimated effect sizes obtained in the discovery sample could often not be replicated unless thousands (> 2000) of participants are included. Thus, effect-size estimates from the discovery sample tend to be inflated. This inflation is especially true for the mass univariate approach.

Following up on this study, [7] argued that the mass univariate approach should not be used for reasons outlined in [2] and showed replicability can be attained with smaller sample sizes for multivariate BWAS than for mass univariate BWAS. Furthermore, [7] showed for the **multivariate BWAS** that the way of estimation for the effect size matters: Typically, the standard in-sample estimation approach is used, whereas a **cross-validation** approach should be used. Cross-validation mimics the discovery-replication process by repeatedly breaking up a data set in two parts, one to estimate (or discover) and one to validate (or replicate) [8]. Using a cross-validated effect size measure can results in lower sample size requirements for replicability [7]. However, still large sample sizes are required (> 1000).

As the multivariate BWAS approach leads to improved replicability compared to the mass univariate approach, we come to our **first conclusion, using more data in a single analysis leads to better replicability for fixed sample size.** Furthermore, a prediction approach is to be preferred over a correlational approach. **Second conclusion,** we should refrain from in-sample effect size measures **and make use of cross-validation instead.** Yet, cross validation alone is not sufficient. Cross-validation makes a more honest assessment of the effect size, but is in itself susceptible to overfitting, that is, selecting the model that best cross-validates probably gives an inflated effect size [9,10].

One of the explanations for finding low replicability rates is **measurement error** in both the brain features as well as the behavioural phenotypes [11,12]. Measurement error and reliability are related, if there is more measurement error the measurement is less reliable; if the measurement is reliable there is not much measurement error. Already in 1910 Spearman showed that reliability of measurements attenuates (that is, dampens or lowers) estimated associations. There has been ample research in the reliability of brain features [e.g., 11], but recently [12] also showed that measurement error in the behaviour phenotypes have a large effect on the effect size and replicability.

One way to deal with measurement error is to develop better measurement instruments. Alternatively, we could also look at the statistical analysis method. The statistical methods used in multivariate BWAS studies assume no measurement error in the predictors (brain features) [13, 14, 15]. Regression models (including machine learning methods like support vector regression, ridge regression, or gradient boosting regression) assume the predictors to be measured without error. For multivariate BWAS, the brain features are not measured without error. Furthermore, the cross-validation procedure employed by [7] assumes that the variables are measured without error [16].

We come to our **third conclusion, statistical methods usually employed for multivariate BWAS use noisy data but do not handle the measurement error properly.**

2. Research Aims

As discussed above, multivariate BWAS analyses are not replicable unless very large sample sizes are used. Brain studies are costly and large sample sizes (> 1000) are therefore not affordable for most research groups. Can we handle the brain data and phenotype measurements more intelligently such that we need smaller sample sizes to obtain replicable results?

With this proposal we aim to develop a break-through in this fundamental problem by developing and testing new statistical procedures that use more information and accommodate measurement error. In what follows, we first outline how to use more data, whereafter we outline how to deal with measurement error.

A common conception is that larger sample sizes are required to detect smaller effects, that is more participants. When the effects are subject to measurement error even larger sample sizes are required. In other words, the common conception is that more participants are needed. We hypothesize that it is not necessarily the number of participants (sample size) but **the number of data points that needs to be increased** to reliably detect a small, possibly noisy, effect. It has already been shown that with fixed sample sizes the multivariate BWAS leads to improved replicability compared to the mass univariate approach [6,7]. In multivariate analysis, the number of data points in an analysis is much larger than in the mass univariate approach.

In a more detailed but abstract way, consider $\boldsymbol{X}$ to be the (set of) brain feature(s) in our statistical analysis and $\boldsymbol{Y}$ to be the (set of) phenotypical outcome(s). The size of $\boldsymbol{X}$ is N (the number of participants) by P (the number of predictors), whereas the size of $\boldsymbol{Y}$ is N by R (the number of responses). In the mass univariate approach, every analysis has $P = 1$ and $R = 1$, therefore the number of data points is $2N$. In the multivariate approach $P = 2000$, say, and $R = 1$, therefore the number of data points is $2001N$, or more generally $(P + 1)N$. The number of data points in a multivariate analysis is much larger (i.e., $1999N$ more data points) than the number of data points in a univariate analysis. The multivariate analysis is better replicable. Can we further increase the number of data points and does this lead to improved replicability?

One way of increasing the number of data points is to increase the amount of brain modalities. From a **structural scan**, researchers usually compute a measure like cortical thickness to associate with the phenotype (working memory). However, many other measures can be computed from the same scan. In [17], we also computed cortical area, cortical curvature, grey matter density, subcortical volumes, and hippocampal shape besides cortical thickness to discriminate between Alzheimer's Disease patients and healthy controls. Multiple feature sets (which we call **modalities**) are available once we have performed a structural scan. These modalities do not develop in isolation in the brain. Measures of cortical thickness are intrinsically related to, for example, measures of cortical curvature. During childhood and adolescence, the different modalities develop together, and one modality cannot be seen independent from another. The modalities are biologically related and as such using them together in a statistical analysis makes substantive sense. For example, [18] showed that voxel based morphometry, cortical folding, and cortical thickness complement each other in showing neurodegenerative changes related to Parkinson's disease. These modalities are related to each other, and we can argue that phenotypes depend on the interplay between these modalities. In [17] we showed that combining the modalities lead to improved cross-validated classification performance. Here we investigate whether combining modalities also results in better replicability.

Similarly, from a **functional resting state scan**, researchers usually compute functional connectivity matrices and associate these to the phenotype. However, from the same scan we can compute many other feature sets. In [19], we also compute functional connectivity dynamics, functional connectivity states, graph metrics, functional connectivity with resting state networks, functional connectivity with the Hippocampus, eigenvector centrality, the amplitude of low frequency fluctuations and the fractional amplitude of low frequency fluctuations to classify Alzheimer Disease patients. Once we obtained a functional resting state scan for a participant, all these modalities are available, we just need to derive the measures from the raw data. Like for modalities of structural scans, these modalities are related to each other, and we can argue that phenotypes depend on the interplay between these modalities. In [19] we showed that combining the modalities lead to improved cross-validated classification performance. With the current proposal, we address the question whether combining modalities also improves replicability for fixed sample size.

The **phenotypes** are usually assessed outside the scanner and are therefore less costly to obtain. Furthermore, often several phenotypes are measured together. In the ABCD study, for example, the following measures of cognition are collected: vocabulary, attention, working memory, executive function, processing speed, episodic memory, reading, fluid intelligence, crystallized intelligence. Instead of linking structural and functional brain features to one of them, we might link those brain measures to all in a multivariate way. We can thus obtain a single regression model with multiple (i.e. multivariate) outcomes. Note that we also have multiple features, such that a so-called double multivariate model is obtained [20]. Statistically, these outcome variables are often correlated and therefore a combined analysis that takes into account these dependencies will have better replicability. Substantially, [21] already argues that multivariate models reflect social reality more accurately, that is, sytems of brain regions do not explain single phenotypes in isolation, but a multitude of phenotypes [22,23]. A double multivariate analysis also better controls Type I errors and therefore should increase replicability.

The first central hypothesis for the current proposal is that using more data from each participant leads to improved replicability. As just discussed, once we obtained a structural or resting state scan these data are available. The measures just need to be extracted from the raw data.

As we saw in the previous section, current statistcal approaches to BWAS do not take into account that the predictor variables have measurement error. To deal with measurement error, the field of psychometrics has developed latent variable models [24,25,26] like principal component analysis, factor analysis, latent class analysis, structural equation modelling, and so forth. These models explicitly take into account that the predictor and response variables are noisy representations of underlying true variables.

Neuroscience research sometimes use such latent variable models. Commonly, a two-step procedure is used, where first principal component analysis is applied to the brain features and in a second step the extracted components are used to predict the phenotype. Although this two-step procedure takes into account measurement error, the approach is suboptimal. A main drawback is that the components are constructed with no account of the prediction problem and hence may miss the components that are relevant in predicting the outcome(s). This is especially true when the number of predictor variables is huge, as is for example the case with BWAS data [27]. Other techniques, closely related to principal component analysis and multivariate multiple regression, are redundancy analysis [28] also known as **reduced rank regression** [RRR; 29, 30] and **principal covariate regression** [PCovR; 31, 32]. These techniques are targeted towards finding underlying latent variables that predict the outcome well. Whereas reduced rank regression is specifically targeted towards multiple outcomes (phenotypes) principal covariate regression can be applied both for a single outcome as well as for multiple outcome variables.

The second central hypothesis for the current proposal is that using these latent variable models to BWAS leads to improved replicability.

One aspect that needs to be considered, is that for PCovR and RRR the number of underlying latent variables must be determined. The common approach is to fit models in a range of dimensionalities and select the optimal one, where researchers choose a criterion for optimality. As argued above, for the BWAS applications cross-validated error should be used as a criterion to make replicability most likely. A relatively new approach is to select the dimensionality by including a penalty in the loss function. The nuclear norm penalty is a penalty that can be used to select the dimensionality [33, 34, 35], where the penalty parameter is tuned using cross-validation. Including such a penalty leads to automated dimensionality selection.

In this study, we will develop these nuclear norm penalized latent variable models (PCovR and RRR) with multiple brain modalities or multiple phenotypes for brain-wide association studies and test whether these new approaches lead to increased replicability.

3. Proposed methodology

The first step is to develop the new statistical analysis method for BWAS. In this **derive and program** step, the latent variable model must be written down and assumptions must be made explicit. The loss function for estimation of the parameters needs to be defined (including the penalty) and an algorithm needs be developed. This requires mathematical and numerical reasoning. Once the algorithm is developed, it needs to be implemented in software (R, Matlab, Python, or C++) and tested. This requires programming and computational skills. In developing the software, special attention is needed for the parameters that we like to replicate. Furthermore, a cross-validation routine together with a prediction function needs to be developed for the new latent variable model.

The second step in both studies is the replication and extension of the **large-scale simulation studies** performed in [6,7] and [12]. We will exactly follow the set-up of these large-scale simulation study, but besides the multivariate methods used in the original papers, we will include our new latent variable methods. Our analysis will focus on the question whether the new statistical models, that include more data and adjust for measurement error, lead to improved replicability of the effects.

In the third step, the newly proposed methods will be **applied to new data** acquired in the new Gravitation program Growing Up Together in Society (GUTS), of which the Principal Investigator is one the work package leaders.
The first goal of this application is to further investigate replicability. Therefore, we will use the GUTS data to estimate the model. We will then test whether the results can be replicated in the ABCD data, the UKB data, and the HCP data. This replicability is more difficult to obtain than the replicability studied in step 2, as in the studies from [6,7] replicability is investigated with data from the same population, whereas in this study, replicability is investigated for data from a different population.

The second goal of this application to GUTS data is to provide empirical researchers a **tutorial** on how to apply the new methods in practice. A step-by-step analysis pipeline will be described using open data, such that researchers can follow these steps with their own empirical data. The tutorial will also include recommendations about sample size, pre-processing and transformation of variables, and interpreting the results of the analysis.

4. Design of the Project

The project is composed of two studies, one focussing on **more brain modalities**, the other on **more phenotypes**. Compared to existing multivariate BWAS studies, we use much more data, that is, more modalities or more phenotypes. Substantively, the use of multiple modalities or multiple phenotypes is more realistic, as these characteristics do not exist in isolation. Statistically, the use of multiple modalities or multiple phenotypes is also more realistic, and relationships between the modalities or phenotypes are taken into account into a more robust outcome. For each study, we will recruit a PhD student. The two studies run in parallel, such that, for example, the replication of the large-scale simulation studies of [6,7] are performed at the same time. The students can thus collaborate, and the projects can reinforce each other. Below, we describe each PhD project in more detail.

4.1. *More brain modalities study.*

For the study with multiple brain modalities, we can apply PCovR. Having multiple modalities, we can distinguish two approaches. In the first approach, the data of the different modalities are concatenated into one (very) large data set and analysed. This is the approach we used in [17,19] using a regression model. Recently, such multiple modal analyses have also been performed with PCovR [27,36,37], where [36,37] showed how to search for common (i.e., across the modalities) and distinctive (i.e., modality specific) components by including sparsity penalties. This first approach (just described) falls under the framework of **low-level fusion** [38]**,** that is the data are fused before the statistical analysis.

Alternatively, for every modality a PCovR analysis can be performed and using a second analysis the outcomes can be combined using a so-called meta-learner. This second approach falls under the framework of **high-level fusion** [38]**,** that is the results of various analyses are fused, not the data. We developed such high-level fusion techniques for binary classification [39,40,41]. High-level fusion has some advantages above low-level fusion [39], better determination of the modalities that are important and increased computational speed. Furthermore, high-level fusion makes handling of missing modalities much easier [42]. We develop high-level fusion methods for PCovR analyses and investigate the level of replicability.

In PCovR analyses the dimensionality of the latent space should be determined. In lower dimensionality, less parameters need to be estimated, which usually results in more stable results, that is improved replicability. This determination is usually done with model selection statistics as information criteria or the convex hull approach [43,33]. Recently, dimension selection has also be performed using the nuclear norm penalty [33,34,35] which transforms the discrete problem of finding the optimal number of dimensions in a continuous problem of finding the optimal regularization parameter. The optimal regularization parameter can be determined using a cross-validation approach. For the cross-validation approach, we need to take the lessons of [16] into account such that the procedure is valid also when there is measurement error.

In this study, both the low-level fusion and high-level fusion techniques are developed for BWAS and implemented in a software package. The techniques will include the nuclear norm penalty for dimension selection and cross-validation for selecting the optimal parameters.

The simulation studies of [6,7] are replicated and extended with the low-level and high-level fusion PCovR analyses. Furthermore, as in [12] we will manipulate measurement error in the brain features and phenotypes by adding noise. Because both [6,7,12] made the code of their studies public on a github page to make the analyses reproducible, we can make strong comparisons to the results described in these papers. These extended simulation studies also investigate how many extra modalities are needed to optimize replicability and whether this number depends on the reliability of the modalities, the strength of the relationship between modalities, and the sample size.

The newly proposed methods will be applied to data acquired in the new Gravitation program GUTS. This 10-year Growing Up Together in Society program has the long-term goal to use predictive modeling to test which combination of measures of multiple disciplines best predict societal outcomes, such as contribution to society, school success, mental health, and well-being. In the GUTS program, both structural and functional MRI scans are made for a relatively large set of participants. From the structural and functional scans a wide variety of modalities can be computed. We will link these modalities to a societal outcome measure using our newly defined methods.

Replicability of the results will be tested with data from different populations. Therefore, we will use the GUTS data to estimate the model. We will then test whether the results can be replicated in the ABCD data, the UKB data, and the HCP data. Whereas the simulation studies of [6,7] (that we replicate and extent in step 2) focus on replicability in the same population, this study investigates replicability in another population, which is usually more difficult as characteristics of the population differ. We will also estimate the models in each of these large data base and verify the replicability in the GUTS data.

The application to GUTS data also has a second goal, that is, communicating an analysis pathway for the new methods for empirical researchers, such that those researchers can apply the methods on their own data. A step-by-step analysis pipeline will be described using open data, such that researchers can follow these steps with their own empirical data. The tutorial will also include recommendations about sample size, pre-processing and transformation of variables, and interpreting the results of the analysis.

4.2. *More phenotypes study.*

In this second study, instead of predicting one phenotype we will predict several phenotypes simultaneously using a double multivariate model [20, 45]. In this study, we focus on one brain modality, but with all features of this modality included. Researchers have considered double multivariate models before, but these used canonical correlation analysis (a correlation approach) whereas a regression approach should be used [3]. For this goal we can apply both PCovR and RRR, both are techniques that can deal with noisy brain features by finding latent variables that optimally predict the outcomes. The main difference between the two is that in RRR the latent variables are found that optimally predict the phenotypes, whereas in PCovR the latent variables are found that also describe the features well. Often, this leads to more stable results [27,31], i.e., a main component of replicability.

This second study on more phenotypes follows the study on more modalities closely, that is, the steps are approximately equal. In the first step in this project, for both RRR and PCovR analyses the dimensionality of the latent space should be determined. In lower dimensionality, less parameters need to be estimated, which usually results in more stable results, that is improved replicability. This determination is usually done with model selection statistics as information criteria or the convex hull approach. Recently, dimension selection has also be performed using the nuclear norm penalty [33,34,35] which transforms the discrete problem of finding the optimal number of dimensions in a continuous problem of finding the optimal regularization parameter. The optimal regularization parameter can be determined using a cross-validation approach. In the first step of this second study, we need to develop algorithms for RRR and PCovR including the nuclear norm penalty for dimension selection. The algorithm will be implemented in a software package including a prediction function and a cross-validation function. For the cross-validation approach, we need to take the lessons of [16] into account such that the procedure is valid also when there is measurement error.

In the second step, the simulation studies of [6,7] are replicated and extended but now adding the conditions with multiple phenotypes. Furthermore, as in [12] we will manipulate measurement error in the brain features and phenotypes by adding noise. This extended simulation study investigates how many extra phenotypes are needed to optimize replicability and whether this number depends on the reliability of the phenotype measure, the correlation between the phenotypes, and the sample size.

In the third step, the newly proposed methods will be applied to data acquired in the new Gravitation program GUTS. As described above, this program has the long-term goal to best predict societal outcomes, such as contribution to society, school success, mental health, and well-being. Here, we focus on the prediction of the different societal outcome measures together (i.e., multiple phenotypes). Replicability of the results will be tested with data from different populations. Therefore, we will use the GUTS data to estimate the model. We will then test whether the results can be replicated in the ABCD data, the UKB data, and the HCP data. Whereas the simulation studies of [6,7] (that we replicate and extent in step 2) focus on replicability in the same population, this study investigates replicability in another population, which is usually more difficult as characteristics of the population differ. We will also estimate the models in each of these large data base and verify the replicability in the GUTS data.

The application to GUTS data also has a second goal, that is, communicating an analysis pathway for the new methods for empirical researchers, such that those researchers can apply the methods on their own data. A step-by-step analysis pipeline will be described using open data, such that researchers can follow these steps with their own empirical data. The tutorial will also include recommendations about sample size, pre-processing, transformation of variables, and interpreting the results of the analysis.

5. Summary

To summarize, brain-wide association studies investigate the neural underpinnings of individual differences in behaviour. Recently, BWAS came under fire as these studies require huge sample sizes to be replicable. We investigate whether the sample size requirements can be lowered when we use new, better, statistical analysis approaches. These new methods, to de (partly) developed use more data (which are available once the participants are scanned) and use latent variables to deal with measurement error.

b. Scientific and/or societal impact

Increasing the reproducibility and replicability of research is vital to its advancement and to its ability to translate findings into integrative theories and clinical interventions [5]. The reproducibility and replicability of scientific findings is fundamental to their validity and utility in guiding further research, technological development, and treatment development. Reproducibility and replicability are therefore central to the return on investment of public money dedicated to scientific research [5].

Understanding differences in behaviour is key in many scientific disciplines. In ***marketing***, researchers are interested why some persons react positively to an advertisement while others negative. In ***education***, researchers wonder why some children benefit more from practicing themselves, whereas others benefit from class instructions. In ***medicine***, researchers wonder why treatments affect some patients but not others. In ***psychology***, researchers investigate why some people react aggressively and others passively in different circumstances. Neuroimaging tries to find the biological, brain basis for such individual differences and has been adopted in all above fields to understand the differences in behaviour.

Reproducibility and replicability directly affect the validity of research. If the results of an enquiry cannot be reproduced using the same data and same methods, the investigation is worthless. Replicabilty is about the question whether the same or similar results are obtained when the same investigation is run with a different set of participants and therefore questions the generalizability of results obtained in an investigation. Generalizability of results is important, because in general we are not only specifically interested in the participants in our research but in a wider population. We like the results to be valid for a general customer, for a general school child, for a general patient, not only for the participants in a study. Recently, it turned out that so-called brain-wide association studies, that aim to cross-sectionally relate individual differences in human brain structure or function to cognitive or mental health phenotypes, need huge sample sizes to be replicable. However, as such studies are usually expensive, huge sample sizes are typically not affordable.

In this project, we aim to show that the sample size can be reduced by using new, more advanced, statistical analysis methods. **These advanced methods use more data** from each single participant, data that is available but that researchers choose not to take into consideration or do not know how to take into consideration simultaneously. From a structural MRI scan, for example, researchers often compute a set of features, such as cortical thickness, that is subsequently linked to the phenotype of interest. However, once the structural scan is made many more feature sets can be computed. Together, these feature sets tell more about the brain structure than any single feature set. Biologically, using the feature sets together is therefore more realistic. Furthermore, these feature sets are related and therefore can reinforce each other, leading to better statistical properties. Similar arguments are in place for functional MRI scans. Instead of focusing on one phenotype, multiple phenotypes can be considered simultaneously. Certain behaviours do not develop in isolation and taking sets of outcome measures together is closer to the social reality. As, such phenotypes often are collected outside the scanner they might be relatively cheap to collect.
The advanced methods we will develop also take into account that the computed feature sets have measurement error. In brain-wide association studies researchers are aware of the burden of measurement error, but till today often fail to take this into account in their analysis.

If we can succesfully show that we are able to decrease the number of participants to obtain replicable estimates, this influences all scientific fields mentioned at the start, that are, marketing, educational science, medicine, psychology, and many more. Smaller sample sizes are needed to obtain replicable results and therefore studies become less costly. Furthermore, the results of these analyses will be closer to biological or social reality, taking into account more information, and therefore might open up new theories about brain-behaviour associations.

In the large interdisciplinary gravitation program GUTS, we plan to communicate our results to the researchers in this consortium. We further plan collaborations with these researchers to apply the new methods on other data sets. We plan to present our results at annual meetings of the consortium but also write tutorials such that the new methods can easily be used and communicate these to the outside world. The PI will develop a workshop that can be given as a pre-conference workshop for researchers who like to obtain replicable results and therefore want to apply our new methods. This workshop will be advertised for conferences like the Organization for Human Brain Mapping and NeurIPS, but also for methodological Winter and Summerschool at Universities.

Although, we target our new methods for brain-wide association studies, the newly developed penalized principal covariate regression and reduced rank regression models will also be applicable to other types of data. Whereas in our proposal, we focus on multiple modalities derived from brain scans, the multiple modalities might also derive from other data sources. In alll cases, we ontain so-called multiblock or multiview or multimodal data and developing and application of methods for such data is an active reseach area, see for example [38]. With our porposal we add to scientific literature, and it might results in new methods that can also be of interest in, say chemometrics.

As we develop new statistical methods, for which we develop new algorithms, also the fields of data science, statistics, artificial intelligence, pattern recognition, psychometrics, econometrics, chemometrics, and biometrics are influenced by the results of this project. All these scientific disciplines develop new ways of analysing data. Results from one discipline often influence researchers in other disciplines. The international federation of classification societies (IFCS) is bringing together researchers from all these disciplines. We plan to present our results in the biennial conferences of this society as well as conferences of the daughter societies (Dutch/Flemish classification society, of which the PI is the current president), but also related conferences like the European Conference on Data Analysis, that has similar goals.

c1. Work plan and planned deliverables

Max 1 page.

Timeline of both PhD projects

The two PhD projects will run in parallel. As detailed above, the two projects have many elements in common, although they have separate goals. We like to recruit two PhD students for 0.8 fte for 5 years. The timeline for both PHD project can be found in Table 1.

Table 1: *Outline of PhD studies.*

	Year 1				Year 2				Year 3				Year 4				Year 5			
Literature	x	x																		
Develop		x	x	x	x	x	x													
Simulation							x	x	x	x	x	x								
GUTS												x	x	x	x	x	x	x		
Dissertation																	x	x	x	x

Literature = Literature study; Develop = Work out the new latent variable models plus the corresponding software; Simulation = replication and extension of the simulation studies of [6,7,12]; GUTS = application of the new method to GUTS data and verifying replicability in ABCD, UKB, and HCP; Dissertation = writing of the PhD thesis.

Deliverables PhD student 1

- Manuscript with software package for early and late fusion of multiple modality BWAS studies.
- Manuscript describing the simulation studies and their results.
- Manuscript where the new methods are applied to the GUTS data and where we test replicability in the existing cohorts (ABCD, UKB, HCP) and vice versa.
- Tutorial on how to use the new methods using the GUTS data.
- PhD dissertation

The software and data used in each of the subprojects will be made publicly available in a github-repository to make our analyses reproducible. All manuscripts will be made publicly available as pre-prints. Furthermore, the manuscripts will be submitted to journals for publication.

Deliverables PhD student 2

- Manuscript with software package for RRR and PCovR analyses of multiple phenotype BWAS studies.
- Manuscript describing the simulation studies and their results.
- Manuscript where the new methods are applied to the GUTS data and where we test replicability in the existing cohorts (ABCD, UKB, HCP) and vice versa.
- Tutorial on how to use the new methods using the GUTS data.
- PhD dissertation

The software and data used in each of the subprojects will be made publicly available in a github-repository to make our analyses reproducible. All manuscripts will be made publicly available as pre-prints. Furthermore, the manuscripts will be submitted to journals for publication.

Deliverables PI

The PI will develop a workshop for disseminating the results of the two PhD projects. In this one day workshop, the new statistical methods for handling multiple modalities and multiple phenotypes will be explained. Using open data and open software, an analysis pipeline will be presented and the participants have to practice the analysis steps. Also the topic of reproducibility and replicability is discussed and reasons why our methods increase both.

d. Reference list

Max 2 pages.

[1] Kanai, R., Rees, G. The structural basis of inter-individual differences in human behaviour and cognition. *Nat Rev Neurosci* **12**, 231–242 (2011). https://doi.org/10.1038/nrn3000

[2] Woo, CW., Chang, L., Lindquist, M. *et al.* Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci* **20**, 365–377 (2017). https://doi.org/10.1038/nn.4478

[3] Bzdok D, Varoquaux G, Steyerberg EW. Prediction, Not Association, Paves the Road to Precision Medicine. *JAMA Psychiatry.* 2021;78(2):127–128. doi:10.1001/jamapsychiatry.2020.2549.

[4] National Academies of Sciences, Engineering, and Medicine 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

[5] Botvinik-Nezer, R., & Wager, T. D. (2023). Reproducibility in neuroimaging analysis: Challenges and solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 780-788.

[6] Marek, S., Tervo-Clemmens, B., Calabro, F.J. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022). https://doi.org/10.1038/s41586-022-04492-9

[7] Spisak, T., Bingel, U. & Wager, T.D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* **615**, E4–E7 (2023). https://doi.org/10.1038/s41586-023-05745-x

[8] de Rooij M, Weeda W. Cross-Validation: A Method Every Psychologist Should Know. *Advances in Methods and Practices in Psychological Science*. 2020;3(2):248-263. doi:10.1177/2515245919898466

[9] Traut, N., Heuer, K., Lemaître, G., Beggiato, A., Germanaud, D., Elmaleh, M., ... & Varoquaux, G. (2022). Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *NeuroImage*, *255*,

[10] Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, *77*(5), 534-540.

[11] Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychological Science, 31(7), 792-806. https://doi.org/10.1177/0956797620916786

[12] Gell, M., Eickhoff, S., Omidvarnia, et al. The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions (2023). bioRxiv 2023.02.09.527898; doi: https://doi.org/10.1101/2023.02.09.527898

[13] Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

[14] Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

[15] Datta, and Zou. "Cocolasso for high-dimensional error-in-variables regression." *Annals of Statistics* 45, no. 6 (2017): 2400-2426.

[16] Datta and Zou (2020) A Note on Cross-Validation for Lasso Under Measurement Errors, Technometrics, 62:4, 549-556, DOI: 10.1080/00401706.2019.1668856

[17] de Vos, F., Schouten, T.M., Hafkemeijer, A., Dopper, E.G.P., van Swieten, J.C., de Rooij, M., van der Grond, J. and Rombouts, S.A.R.B. (2016), Combining multiple anatomical MRI measures improves Alzheimer's disease classification. Hum. Brain Mapp., 37: 1920-1929. https://doi.org/10.1002/hbm.23147

[18] Pereira, J. B., Ibarretxe-Bilbao, N., Marti, M. J., Compta, Y., Junqué, C., Bargallo, N., & Tolosa, E. (2012). Assessment of cortical degeneration in patients with Parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. *Human Brain Mapping*, *33*(11), 2521-2534.

[19] de Vos, F. Koini, Schouten, Seiler, van der Grond, Lechner, Schmidt, de Rooij, Rombouts (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease, NeuroImage, 167, 62-72. https://doi.org/10.1016/j.neuroimage.2017.11.025

[20] Rosenberg, M.D., Finn, E.S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat Neurosci* **25**, 835–837 (2022). Https://doi.org/10.1038/s41593-022-01110-9

[21] Larry J. Fish (1988) Why Multivariate Methods are Usually Vital, Measurement and Evaluation in Counseling and Development, 21:3, 130-137, DOI: 10.1080/07481756.1988.12022895

[22] Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology, 82*(5), 920–930. https://doi.org/10.1037/a0035628

[23] Genon, S., Eickhoff, S.B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat Rev Neurosci* **23**, 307–318 (2022). https://doi.org/10.1038/s41583-022-00584-7

[24] Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

[25] Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

[26] Hwang, H., & Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling*. CRC Press.

[27] Van Deun, K., Crompvoets, E. A., & Ceulemans, E. (2018). Obtaining insights from high-dimensional data: sparse principal covariates regression. *BMC bioinformatics*, *19*, 1-13.

[28] Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, *42*(2), 207-219.

[29] Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, *5*(2), 248-264.

[30] de Rooij, M. A new algorithm and a discussion about visualization for logistic reduced rank regression. *Behaviormetrika* (2023). https://doi.org/10.1007/s41237-023-00204-3

[31] De Jong, S., & Kiers, H. A. (1992). Principal covariates regression: part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, *14*(1-3), 155-164.

[32] Wilderjans, T. F., Vande Gaer, E., Kiers, H. A., Van Mechelen, I., & Ceulemans, E. (2017). Principal covariates clusterwise regression (PCCR): Accounting for multicollinearity and population heterogeneity in hierarchically organized data. *Psychometrika*, *82*, 86-111.

[33] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

[34] Fazel, M. (2002). *Matrix rank minimization with applications*(Doctoral dissertation, PhD thesis, Stanford University).

[35] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, *11*, 2287-2322.

[36] Park, S., Ceulemans, E., & Van Deun, K. (2021). Sparse common and distinctive covariates regression. *Journal of Chemometrics*, *35*(2), e3270.

[37] Park, S., Ceulemans, E., & Van Deun, K. (2023). Logistic regression with sparse common and distinctive covariates. *Behavior Research Methods*, 1-32.

[38] Smilde, A. K., Næs, T., & Liland, K. H. (2022). *Multiblock data fusion in statistics and machine learning: Applications in the natural and life sciences*. John Wiley & Sons

[39] van Loon, Fokkema, Szabo, de Rooij. Stacked penalized logistic regression for selecting views in multi-view learning, Information Fusion, 61, 2020, 113-123, https://doi.org/10.1016/j.inffus.2020.03.007.

[40] van Loon, W., Fokkema, M., Szabo, B., & de Rooij, M. (2020). View selection in multi-view stacking: choosing the meta-learner. *arXiv preprint arXiv:2010.16271*.

[41] van Loon, W., de Vos, F., Fokkema, M., Szabo, B., Koini, M., Schmidt, R., & de Rooij, M. (2022). Analyzing hierarchical multi-view MRI data with StaPLR: an application to Alzheimer's disease classification. *Frontiers in Neuroscience*, *16*, 830630.

[42] van Loon, W., Fokkema, M., & de Rooij, M. (2022). Imputation of missing values in multi-view data. *arXiv preprint arXiv:2210.14484*.

[43] Wilderjans, T.F., Ceulemans, E. & Meers, K. CHull: A generic convex-hull-based model selection method. *Behav Res* **45**, 1–15 (2013). https://doi.org/10.3758/s13428-012-0238-5

[44] Bulteel, K., Wilderjans, T.F., Tuerlinckx, F. *et al.* CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behav Res* **45**, 782–791 (2013). https://doi.org/10.3758/s13428-012-0293-y

[45] Gvaladze, S., Vervloet, M., Van Deun, K., Kiers, H. A., & Ceulemans, E. (2021). PCovR2: A flexible principal covariates regression approach to parsimoniously handle multiple criterion variables. *Behavior Research Methods*, 1-21.