Fakultät für Medizin

Dimensionality reduction and unsupervised learning techniques applied to clinical psychiatric and neuroimaging phenotypes

Riya Paul

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Claus Zimmer

Prüfende der Dissertation:

1. apl. Prof. Dr. Bertram Müller-Myhsok
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 28.01.2021 bei der Technischen Universität München eingereicht und durch die Fakultät am 13.07.2021 angenommen.

Table of content

Abstract

Psychiatric research is closely related to progress in the statistical field, mainly due to the etiologically heterogeneous, syndromal nature of psychiatric disease. Detecting stable patterns in high-dimensional datasets using multivariate machine learning analysis techniques can help to model this complexity. In this work, such techniques were applied and combined for the purpose of complexity reduction, using clinical symptom observations and high-dimensional structural neuroimaging data as a source of biological information.

Patients with major depressive disorder (MDD) differ with regard to the dynamics of their response to treatment, and early detection of difficult-to-treat depression is still a challenge. Yet, treatment response dynamics is mostly parametrized based on simple binary definitions of response and remission, likely missing intermediate classes. Here we suggest an alternative, data-based approach to identify treatment response clusters (TRC[s]) in MDD that could facilitate comparisons across cohorts and the development of treatment prediction algorithms. For this purpose, we analyzed a large, observational study (Munich Antidepressant Response Signature [MARS] study, 1017 patients) and a partly randomized interventional study (Genome-based Therapeutic Drugs for Depression [GENDEP], 809 patients). Symptoms in both studies were rated using the Hamilton Depression Rating scale (HRS) over 16 weeks or 12 weeks, respectively. We applied a finite mixture model with an integrated completed likelihood criterion for cluster stability evaluation to series of HRS sum scores of the MARS discovery sample (834 patients). This revealed seven TRCs ranging from fast and complete response (4.9 weeks to discharge, 94% remission) to a slow and incomplete response (10% remission at week 16). Even neighboured TRCs differed strongly with regard to established response and remission definitions. Internal validity and generalizability of the TRCs was investigated by applying the model coefficients to the MARS validation sample (236 patients) and the GENDEP sample where patients could be well assigned to the TRCs, with differences in the cluster sizes in GENDEP expectedly mirroring the different study design and sample. As external validation we used random forests as a regressor and classifier to relate the TRCs to predefined set clinical baseline items, identifying personality items, life events, episode duration, and specific psychopathological features as most contributing predictors. Prediction accuracy improved when cluster-derived slopes were used rather than individual slopes, supporting that a true complexity reduction has been achieved. Eventually, to integrate previous knowledge that single symptom items are co-correlated we clustered these using a parameterized finite Gaussian mixture model, breaking up the baseline HRS sum score into four sub-scores that may serve as basis for a more differentiated longitudinal clustering.

In the second project, we aimed at developing an atlas-like brain parcellation based on the covariance structure of grey matter (GM) volume maps gained by segmentation and spatial normalization of structural magnetic resonance images, similarly as in voxel based morphometry (VBM). Preliminary work suggested that this covariance structure harbors spatial modules known from functional imaging or from fiber tracking experiments. The implication of such parcellation is that VBM data could then be broken down into 'functional' units that align, for example, with known neurodegenerative diseases but that are also sufficiently sparse for genetic association analyses or secondary machine learning applications. Advanced implementations of an Infomax type of independent component analysis (ICA) that also provide measures of component stability through iterations (ICASSO) were used. The dimensionality of the ICA was optimized by combining agglomerative hierarchical clustering (HC), systematic re-agglomeration to lower dimensions and similarity analyses of these re-agglomerations. As discovery sample, a public repository dataset (563 healthy adults, age 20-83 years) and as replication sample a local (Max Planck Institute of Psychiatry) sample (566 healthy adults, age 18-83 years) was available. After state-of-the-art VBM preprocessing, Combat-based removal of site/scanner effects, non-linear intersubject coregistration, Jacobian modulation and spatial smoothing of GM voxel maps ($1.5 \times 1.5 \times 1.5$ mm^3, isometric Gaussian kernels FWHM 6 mm and 10 mm), influences of age, squared age, sex and their interactions as well as total intracranial volume were regressed out on a voxel-basis. Respective concatenated multi-subject 4D datasets restricted to GM (479384 voxels) were forwarded to ICASSO. The requested ICA dimension k was varied between 20 and 445 in steps of 25. The proportion of stable components for each k was explored, and a maximum Z value achieved per voxel (from all ICs or only stable/unstable ICs separately) was calculated, suggesting volatile anatomical location of unstable components. To determine which range of k delivered a stable parcellation and by this indirectly estimate the number of true data sources, we performed agglomerative HC of each ICA solution and secondarily re-agglomerated this solution to k' in steps of 5 between 10 and 540, for all $k' < k$. We then compared the similarity of all possible pairs of re-agglomerated component sets per k' and averaged these for each k'. As alternative, only re-agglomerations of neighbored ICA dimension k were compared. Both averaging techniques converged on about 150 components to which higher parcellations could be re-agglomerated with relatively highest spatial stability. This dimension estimate was confirmed in the replication dataset, with anatomical similarity of the respective optimal solutions. The dimension estimate was also stable for both smoothing settings that, however, differed slightly regarding their spatial pattern. Discrete versions of parcellations were visualized with and without consideration of component stability, and a fuzzy display of parcellation borders using the ambiguity of a voxels' component membership was developed. Overall, ICA and HC allowed for stabilizing a solution with about 150 covariance-based GM

components with a first indication of generalizability. In-depth assessment of smoothing strategies and exploration of a saturation effect with larger sample size may lead to a sample-independent, fully generalizable solution.

In conclusion, model-based clustering of longitudinal clinical symptom observations, after appropriate feature engineering, proved effective for identifying treatment response subgroups in depression, with usefulness of the model for new samples and an additional potential lying in an extension to data-driven symptom clusters instead of the total symptom severity. Random forests as supervised classification and regression method supported the validation. Conceptually, model-based clustering followed by a supervised learning step for validation could be a useful framework for other longitudinal observations. In the neuroimaging project, a 3000:1 dimension reduction of voxel wise GM volume information to anatomically plausible, network-like volume modules was achieved by ICA whereby the dimensionality question was solved by HC, systematic re-agglomeration of components and similarity analyses. Here, the embedding of different unsupervised learning techniques into a novel framework proved effective for developing a data-driven atlas parcellation.

Zusammenfassung

Psychiatrische Forschung ist inzwischen eng mit der Weiterentwicklung statischer Methoden verbunden, durch die überwiegend als Syndrome definierten, jedoch ätiologisch heterogenen psychiatrischen Erkrankungen abgebildet werden können. Für eine Modellierung solcher komplexen Muster sind multivariate Verfahren und maschinelles Lernen besonders geeignet, da sie latente Muster automatisch erkennen können. In der vorliegenden Arbeit wurden solche Techniken mit dem Hauptziel einer sinnvollen Reduktion von Datenkomplexität kombiniert und in zwei Projekten auf verschiedene Datenstrukturen angewendet: zum einen auf klinische (Verlaufs-)Beobachtungen und zum anderen hochaufgelöste anatomische Neuroimaging-Daten.

Patienten mit einer Depression (major depressive disorder [MDD]) sprechen erfahrungsgemäß sehr heterogen auf eine Behandlung an, wobei die möglichst frühe Erkennung einer drohenden Therapieresistenz eine besondere klinisch Bedeutung hat. Die Therapieantwort wird hierbei in Studien bisher meist als binäre Variable (z. B. Teilansprache [Response], Remission) codiert, so dass Zwischenstufen nicht ausreichend abgebildet werden. Als Alternative können durch einen datengetriebenen Ansatz Therapie-Antwort-Cluster (treatment response clusters [TRC]) ermittelt werden, die dann wiederum den Vergleich zwischen Kohorten und die Entwicklung von Prädiktions-Algorithmen erleichtern. Hierfür wurden eine klinische Beobachtungsstudie (Munich Antidepressant Response Signature [MARS] Studie, 1017 Patienten) und eine teilkontrollierte klinische Interventionsstudie (Genome-based Therapeutic Drugs for Depression [GENDEP], 809 Patienten) analysiert. Die Symptomschwere wurde in beiden Studien über 16 bzw. 12 Wochen durch die Hamilton Depressionsskala (Hamilton Depression Rating Scale, HRS) beurteilt. Auf den entsprechenden HRS Zeitreihen des MARS Discovery Samples (834 Patienten) wurde ein finite mixture model geschätzt, wobei die Anzahl der Cluster durch das integrated completed likelihood Kriterium ermittelt wurde. Es ergaben sich 7 abgrenzbare TRCs, die das gesamte Spektrum zwischen einer schnellen und meist vollständigen Therapieansprache (4.9 Wochen Klinikaufenthalt, Remission in 94%) und einer sehr langsamen oder ausbleibenden Therapieansprache (Remission in 10% nach 16 Wochen) abbildeten. Die TRCs, auch direkt benachbarte, unterschieden sich stark in Bezug auf herkömmliche Response- oder Remissionskriterien. Die Übertragung des Modells auf die MARS Validierungs-Stichprobe (236 Patienten) und das GENDEP-Validierungs-Stichprobe ergab, dass auch neue Fälle auf die TRCs projiziert werden konnten; andere relative Proportionen der TRCs bei GENDEP waren hierbei Ausdruck der unterschiedlichen Studienpopulationen bzw. -designs. Im Sinne einer externen Validierung wurden die TRCs durch einen Random Forests Algorithmus mit 50 klinischen Prädiktorvariablen vorhergesagt, wobei Persönlichkeitsmerkmale, Life Events, die Dauer der depressiven Episode und spezifische Einzelitems des HRS besonders beitrugen. Die

Prädiktionsgenauigkeit war bei Verwendung der modell-basierten Dynamik höher als bei Verwendung der individuellen Dynamik, was für eine echte Komplexitätsreduktion spricht. In einer Weiterentwicklung wurden die Gesamt-HRS-Scores wiederum durch modellbasiertes Clustering (parameterized finite Gaussian mixture model) in vier Symptomcluster aufgebrochen, durch die das longitudinale Clustering weiter differenziert werden kann.

Im zweiten Projekt wurden anatomische Daten in Form von Volumenkarten der grauen Substanz (grey matter [GM]) analysiert, um eine Atlasparzellierung auf der Basis der strukturellen Kovarianz zu entwickeln. Derartige GM-Karten werden in der voxel-basierten Morphometrie (VBM) verwendet und basieren primär auf Segmentierung und exakter räumlicher Koregistrierung. In der Literatur bestanden Hinweise, dass die GM-Kovarianzstruktur Netzwerken aus dem funktionellen Imaging oder Fiber-Tracking-Daten ähnelt. Durch eine stabile Kompentenstruktur könnten, i. S. eines Anwendungsziels, VBM-Daten auf 'funktionelle' Volumenmodule heruntergebrochen werden, die neurodegenerative Prozesse abbilden oder als Input für genetische Assoziationsanalysen und andere Sekundäranalysen dienen. Primär wurden Unabhängigkeitsanalysen (Independent Component Analysis [ICA]) mit Infomax zur Komponentenextraktion und ICASSO zur Stabilitätsbestimmung durchgeführt. Um die optimale Zahl der Komponenten einzugrenzen, wurde jede ICA durch ein (agglomeratives) hierarchisches Clustering (HC) analysiert, durch gezielte Reagglomeration auf eine niedrigere Komponentenzahl zurückgeführt und diese durch eine Ähnlichkeitsanalyse verglichen. Zur Verfügung standen das Discovery Sample (Public repository, 563 Gesunde, 20-83 Jahre) und ein lokales (Max-Planck-Institut für Psychiatrie) Replikationssample (566 Gesunde, 20-83 Jahre). Nach VBM-Präprozessierung mit nicht-linearer Koregistrierung, Jacobian'scher Modulierung, und räumlicher Glättung der GM-Karten ($1.5 \times 1.5 \times 1.5$ mm^3, isometrischer Gauß'scher Kernel FWHM 6 mm bzw. 10 mm) wurde eine multiple Regression zur Entfernung von Kovariateneinflüssen (Alter, Geschlecht, Alter-Geschlecht-Interaktionen, Schädelvolumen) durchgeführt. ICAs wurden dann auf 4D-Datensätzen (GM-Maske, 479384 Voxel) für 20 bis 445 Komponenten (Schrittweite 25) durchgeführt. Der Anteil der stabilen Komponenten wurde für jedes k ermitteln, außerdem pro Voxel der höchste durch eine Komponente erreichte Z-Wert, wobei sowohl alle Komponenten als auch nur stabile oder nur instabile Komponenten betrachtet wurden. Zur Bestimmung der optimalen Komponentenzahl, wurde ein agglomeratives HC, gefolgt von der gezielten Reagglomeration auf k' (10 bis 540, Schrittweite 5, für alle $k' < k$) durchgeführt. Für jedes k' wurde die räumliche Ähnlichkeit der aus den verschiedenen Quell-ICAs stammenden Re-agglomerationen anhand von 7 Ähnlichkeitsmaßen bestimmt, und für alle Permutationsmöglichkeiten oder nur für jeweils benachbarte ICAs gemittelt. Die Ähnlichkeit erreicht ein Maximum für circa 150 Komponenten in beiden Datensätze, bei auch hoher anatomischer Ähnlichkeit der Komponenten. Auch bei der höheren

Glättung (10 mm) blieb das Optimum stabil bei 150 Komponenten bei nur geringen anatomischen Verschiebungen. Die Komponentenlösungen wurden sowohl als diskrete Karten (exakt 1 Komponente pro Voxel, mit und ohne Gewichtung durch die Komponentenstabilität) als auch fuzzy dargestellt (hohe Ambivalenz der Komponentenzugehörigkeit als Trennlinie zwischen Modulen). Insgesamt konnte ein Kovarianz-basierter GM-Atlas erstellt werden, mit bereits akzeptabler Generalisierung. Weitere Analysen sollten auf den Einfluss der Glättungsstrategie oder der Sample-Größe fokussieren. Es lässt sich schlussfolgern, dass durch modell-basiertes Clustering auf longitudinalen Daten zur Symptomschwere während der Depressionsbehandlung unterschiedliche Therapieansprache-Cluster identifiziert werden konnten. Diese waren dahingehend stabil, dass auch unabhängige Patienten durch gut abgebildet werden konnten. Als weitere Entwicklung können ebenfalls datengetriebene Symptom-Cluster anstand der Symptom-Gesamtschwere verwendet werden. Eine externe Validierung gelang durch die Verknüpfung der Therapieansprache-Cluster mit klinischen Prädiktoren durch eine Random Forests Analyse, die hier sowohl als Klassifikations- als auch als Regressionsverfahren eingesetzt wurde. Die Kombination von modellbasiertem Clustering mit supervidiertem Lernen zur Validierung scheint übertragbar auf andere longitudinale Beobachtungsdaten. Im Neuroimaging-Projekt konnten voxelweise GM-Karten in ihrer Dimensionalität durch das ICA-Verfahren um einen Faktor 3000 reduziert und dabei anatomisch plausible, netzwerk-ähnliche Volumen¬module stabilisiert werden. Die optimale Komponentenzahl wurde hierbei durch hierarchisches Clustering, Re-agglomeration und räumliche Ähnlichkeitsanalysen ermittelt. Diese bisher nicht beschrieben Kombination aus unsupervidierten Lerntechniken ermöglichte die Generierung einer datenbasierten Atlasparzellierung.

List of figures

List of tables

1 Introduction

1.1 Major depressive disorder (MDD)

While major depressive disorder (MDD) is one of the leading causes of the disease's global burden. It is diagnosed as having an enduring low- or depressed-mood individual, anhedonia or reduced interest in enjoying life, a sentiment of guilt or worthlessness, lack of energy, poor concentration, appetite. In the 5th edition of the Diagnostic and Statistical Handbook (DSM-5) [1], a person must be diagnosed with MDD with five of the above symptoms, one of which must be a depressed mood or anhedonia, leading to social or occupational disability. To make an MDD diagnosis, the history of a manic and hypomanic episode should be omitted.

Diagnostic and Statistical Manual of Mental Disorders (DSM)

The Diagnostic and Statistical Manual of Mental Disorders (DSM) presents a common vocabulary in which physicians, experts and health authorities in the US communicate about psychiatric illness. The new DSM, 5th Review (DSM-5)[1], was issued in May 2013, marking the first substantial overhaul to diagnostic standards and ratings since the DSM-IV in 1994[2] In chapter V of the International Classification of Diseases (ICD), the World Health Organization (WHO) provided its own method for classifying psychiatric illnesses, primarily used for compensation and collect national and international medical statistics. Nevertheless, the ICD settled on a global consensus to implement explicit standards for diagnosing psychiatric illnesses after the 1982 Classification International Conference in Copenhagen, after the 1980 DSM-III Models[3]. This was followed in a decade of collaboration between DSM-IV developers from the American Psychiatric Association (APA) and ICD-10[4] developers from the WHO which was supported by a joint partnership between the National Mental Health Institute and the WHO (Sartorius N Principal Investigator. The WHO/Alcohol, Drug Abuse, and Mental Health Administration Joint Project on Diagnosis and Classification. Cooperative agreement U01MH035883, from the National Institute of Mental Health to the World Health Organization, 1983-2001). Although DSM is a US method for diagnosing mental illnesses, international interest in the handbook has flourished since the publication of DSM-III in 1980, in accordance with the use of the official ICD statistic code numbers. The DSM-5 is focused on explicit condition definitions, along with a large explanational text which is first stated in electronic version of this DSM, which constitutes a nomenclature of mental disorders.

Multifactorial, including molecular genetics, environmental and psychosocial causes were thought to have been the etiology of major depressive disorders. MDD was previously considered mainly due to the abnormalities in neurotransmitter particularly serotonin, norepinephrine and dopamine. Different counteractants, such as serotonin receptor selective inhibitors, serotonin-norepinephrine receptor inhibitors, dopamine-norepinephrine receptor inhibitors have been demonstrated in the treatment of depression[5]. MDD has a lifetime prevalence of about 5 to 17%, with an average of 12%. The prevalence rate is almost double in women than men[6]. The disparity has been regarded due to hormonal changes, childbirth effects, multiple psychosocial stressors in men and women. While the average age of onset is about 40 years, new studies indicated that the prevalence is increasing in young population due to the use of alcohol and the other substances.

Table 1.1: Diagnostic Criteria for Major Depressive Disorder (DSM)

A. Over the same two-week period, five (or more) of the following symptoms were present and represent a change from previous functioning: at least one of the symptoms is either (1) depressed mood or (2) loss of interest or enjoyment.[1]	1. Depressed mood most of the day, almost every day, as indicated by either a subjective report (e.g., feeling sad, empty, hopeless) or other observations (e.g., appearing tearful). (Note: An irritable mood can be present in children and adolescents.)
	2. Most of the day, almost every day (as indicated by either subjective account or observation), interest or pleasure in all, or almost all, activities decreased significantly.
	3. Significant weight loss in the absence of diet or weight gain (e.g., change in body weight by more than 5% in a month), or decrease or increase in appetite almost every day.
	4. Almost every day, insomnia or hypersomnia.
	5. Almost every day, psychomotor agitation or retardation (observable by others, not merely subjective feelings of restlessness or slowing down).
	6. Tiredness or loss of energy almost every day.
	7. Feelings of worthlessness (not just self-reproach or guilt about being sick) or excessive or inappropriate guilt (which may be delusional) almost every day.
	8. Almost every day, decreased ability to think or concentrate, or indecisiveness (either by subjective account or as others observe).
	9. Recurrent thoughts of death (not just fear of dying), recurrent ideation of suicide without a specific plan, or an attempted suicide or a specific suicide plan.
B. In social, occupational, or other important areas of functioning, the symptoms cause clinically significant distress or impairment.	
C. The episode is not attributable to a substance's physiological effects or another medical condition.	
D. Schizoaffective disorder, schizophrenia, schizophrenic disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders do not explain the occurrence of the major depressive episode better.	
E. There has never been a manic episode or an episode of hypomania. Note: If all manic-like or hypomanic-like episodes are substance-induced or attributable to the physiological effects of another	

medical condition, this exclusion shall not apply.

Treatment for MDD

Significant depressive symptoms can be managed by different types of treatment, including alteration of pharmacology, psychotherapy, intervention and life styles. Medications or/and psychotherapy are the initial treatment of MDD. Combination therapy, using both drugs and psychotherapy, is more effective than any medication alone [7,8]. In contrast with any other form of extreme major depression, electroconvulsive therapy has been shown to be more effective.

Table 1.2: Different types of FDA approved medication for the treatment of MDD

FDA approved medication for the treatment of MDD	Name	Details
Antidepressants	Selective serotonin reuptake inhibitors (SSRIs)	SSRI include fluoxetine, sertraline, citalopram, escitalopram, paroxetine, and fluvoxamine, the most widely prescribed antidepressants.
	Serotonin-norepinephrine reuptake inhibitors (SNRIs)	SNRI include venlafaxine, duloxetine, desvenlafaxine, levomilnacipran, and milnacipran, frequently used for depressed patients with comorbid pain disorders.
	Serotonin modulators	trazodone, vilazodone, and vortioxetine
	Atypical antidepressants	Atypical antidepressants include bupropion and mirtazapine. They are often prescribed as monotherapy or as augmenting agents when patients develop sexual side-effects due to SSRIs or SNRIs.
	Tricyclic antidepressants (TCAs)	amitriptyline, imipramine, clomipramine, doxepin, nortriptyline, and desipramine.
	Monoamine oxidase inhibitors (MAOIs)	tranylcypromine, phenelzine, selegiline, and isocarboxazid.
Psychotherapy	Cognitive-behavioral therapy, Interpersonal therapy	
Electroconvulsive therapy (ECT)	Acute suicidality, Severe depression during pregnancy, Refusal to	

	eat/drink, Catatonia, Severe psychosis	
Transcranial magnetic stimulation (TMS)	FDA-approved for treatment-resistant/refractory depression	
Vagus nerve stimulation (VNS)	FDA-approved as a long-term adjunctive treatment for treatment-resistant depression	
Esketamine	Nasal spray to be used in conjunction with an oral antidepressant in treatment-resistant depression	

Despite sequential combination or increase of treatment interventions, almost half the patients suffering from major depressive episode are not able to respond, regardless of the operational concept used for treatment-resistant depression (TRD)[9,10]. The detection of its subtypes is another challenge in MDD. We are aware that MDD is heterogeneous in such characteristics as clinical presentation, disease development, reaction to treatment, genetics and neurobiology[11]. This heterogeneity hampers progress in the detection and efficient treatment of the cause of MDD[12]. A number of studies have been conducted to classify data-driven subtypes of MDD using clinical questionnaires in order to address this issue. But either conflict or clusters linked to depression are identified by the results of these studies, which do not provide any definitive evidence for subtypes of depression[13,14].

1.2 Longitudinal clustering and clinical subtypes of major depressive disorder

Personalized, specific, stratified medicine is understood as a medicinal approach, in which patients are stratified by specialized diagnostic assessments depending on their condition subtype, risk, prognosis or treatment response. In precision medicine, an important question is how to model disease progression appropriately and thus determine the correct type and time for individual therapy. The inherent complexity of diseases that exhibit highly diverse clinical phenotypes may be missed by classical univariate clustering methods. Grouping of patients by symptom development therefore leads to the difficult question of how a clustering of a multivariate time series can be learned. In

machine learning and statistics, clustering is a fundamental and well-researched domain in general. The aim of clustering is to divide samples into clusters so that there is a higher degree of similarity of samples *within* a cluster compared with samples *between* clusters. Following Hastie et al. (2009)[15], clustering algorithms exist in three major categories: (i) mixing algorithms, (ii) mixing modeling, and (iii) searching for mode modes. A wide range of methods for different clustering issues is available within each of these three categories. The combinatorial cluster analyzes can be seen as a method for investigating (heuristically) the region of any possible group data arrangement and selecting that which fits to an Optimized Target function[16]. Combinatorial algorithms do not assume a basic likelihood model, but work directly with the data. Examples therefore include C-means, spectral and hierarchical clustering[17,18]. Mixing models assume that some probabilistic models can describe the data. One example for this is the Gaussian mixture clustering. In the search mode, the underlying multi-modal probability density is then attempted to be estimated directly. The medium-shift algorithm[18] is an important example here. Several techniques have been developed for the clustering of *multivariate* time series data[19]. In general, however, these approaches rely on a much longer time series than typically available in most longitudinal clinical observations. In addition, these methods are not suitable for many missing values that are often found in clinical data. Missing values in clinical data can be present for several reasons: (i) patients drop out of the study or trial, e.g. due to worsening symptoms; (ii) a diagnostic test is not performed (e.g. due to a lack of the patient's agreement), which could lead to a lack of information for whole variable groups.

Longitudinal data contain observations that are measured repeatedly over time. One way to analyze this type of data is to classify it, i.e. to divide it into *unique s*ubgroups. To achieve this, different methods have been proposed, including variants of k-means and various models based on mixture models[20]. While there are no generally accepted recommendations on which methods to use in a specific context[21,22], these approaches are regularly considered. The general idea behind clustering is to group individuals by their similarity to each other. For the concept of "similarity" several concepts can be employed, and basically they are built on the concept of distance, similarity itself or probability. For example, two subjects are considered similar in the majority of current approaches when each point has close trajectories. This approach takes local similarities into account, but not necessarily the overall shapes of trajectories. Two trajectories which has similar shape but can be may be assigned to different clusters. The direct result is that the mean of the group does not tell the shapes, but in many cases it is more important to observe the progression of a phenomenon rather than simply drawing inference from its occurrence. In those circumstances, one would prefer to divide people whose trajectories have similar forms regardless of their shift in time[23].

1.3 Neuroimaging

A neuroimaging method can be defined as any imaging technique that allows to obtain images of the human central nervous system structure or its function. Neurophysiological techniques such as electro-encephalography or evoked potentials, along with neuroimaging, can be subsumed as human brain mapping as they also help o map (i.e., localize) brain function. Such a method should ideally provide good correct spatial resolution of brain anatomy and good temporal resolution of functional changes. Ideally, for human studies, the method should be minimally invasive and repeatable. The way neuroimaging techniques address questions of functional neuroanatomy, particularly in the context of behavioral studies or clinical disorders, has strongly evolved over the last 20 years. Structural neuroimaging addresses the structure of the brain (for example, by demonstrating image contrast among between major tissue types cerebrospinal fluid (CSF), gray matter (GM) and white matter(WM)). Indirect measurement of the brain function is made possible through functional neuroimaging (e.g., neural activity).

a Computed Tomography (CT),

b Positron Emission Tomography (PET),

c Single Photon Emission Computed Tomography (SPECT),

d Magnetic Resonance Imaging (MRI),

Structural MRI including diffusion imaging and functional MRI including its main representative BOLD fMRI, has taken a lead role due to its low invasiveness, lack of radiation exposure and relatively wide availability. A great portion of neuroimaging research is thus based on MRI. Basic MRI researchers like Peter Mansfield and Paul Lautebur, who won the 2003 Nobel Prize for Physiology or Medicine, have developed the basic principle of magnetic resonance to an applicational status where it became immediately useful as key imaging technique in in many medical disciplines. In the neurological and psychiatric field, MRI produces high quality images of the brain macro- and microstructure and helps to map its functional status, hereby avoiding any ionizing radiation (X-rays) or radioactive tracers, only relying on magnetic fields and their interaction with the body tissue and high frequency radio waves. Generally, there is consensus that MRI does no harm to the organism unless taken to extreme field strengths (e. g. > 9 Tesla) or to the human fetal life stage where indications are very strict and safety data incomplete. By sensitizing the sequence to T1-, T2-, proton density or diffusion properties of the tissues by independent parameters, a whole range of tissue properties can be highlighted.

1.4 T1-weighted magnetic resonance (MR) images and basics of segmentation

T1-weighted MRI images

T1-eighted images (T1WI) in magnetic resonance imaging (MRI) represent one basic pulse sequence in MRI to showcase T1-properties of the tissue. This property is based on the principle of MRI and needs a short excursion: During MRI, the tissue is first exposed to a constant, strong magnetic field, that aligns the magnetic spin of all H+-atoms in parallel, or anti-parallel, to that field. The net vector of this magnetization, however, is not zero, but slightly positive. The outer, constant magnetic field further leads the spins of the H+-atoms to take a frequency (Larmor Frequency) linearly correlated with the strength of the B0-field (The B0 is the main static magnetic field in the MRI and is determined by Teslas (T). In clinical applications, the majority of the MRI systems are 1.5T, with an increasing number of 3T mounted. There have been 7T clinic scanners since 2017), multiplied by the gyromagnetic ratio (in units Mhz/T). This ratio is specific to the type of isotopes studied, and for H+ it is 42.58 Mhz/T. As soon as radiofrequency close the Larmor frequency is sent to the tissue, the energy is absorbed first, and the synchronously spinning vectors are desynchronized on one hand, and they lose their net vector magnetization on the other hand. After the HF application stops, the system develops back to the formerly state it was in the constant magnetic field. Here the T1-relaxation time is a measure for how fast the net magnetization vector comes back to its ground state in the direction of B0. This associated energy loss when the excited nuclei return to their lower energy state is lost to the surrounding nuclei. Phrased differently, the T1WI relies on this longitudinal relaxation of a tissue's net magnetization vector. The T2-time (more correct, the T2* time) represents the time needed until the spinning process is desynchronized again, and component orthogonal to the B0 field is falling back to zero again. T1-weighted image (also referred to as T1WI or "spin-lattice" relaxation time) is one of the basic pulse sequences in MRI and demonstrates differences in the T1 relaxation times of tissues.
In clinical and brain mapping practice, T1WI generally depict a good intensity contrast between grey matter (incl. the cortex), white matter, and CSF. T1WI are also sensitive to the enhancing effects of gadolinium[24]. The order of intensity of these three compartments is WM > GM > CSF. Fat tissue or sometimes calcifications can also appear bright on T1WI, so sometimes fat is be by fat saturation techniques. To create a T1WI, the magnetization is allowed to mostly recover before reading out the MR after a rather long[24] (TR). TR is the time between excitations. In essence, in brain imaging, a T1WI allows a proper assessment of the large macroscopic compartments of GM (including the cerebral cortex), white matter and CSF, and, in studies on inflammatory diseases is also useful to collect post contrast images. An example of a whole head T1WI is given in **Figure 1.1**. The typical acquisition time as such an image is about 8 to 15 minutes, depending on geometrical details, and typical resolutions of the original images are round 1x1x1 mm^3, or slightly smaller voxel size. There are other pulse sequences that emphasize abnormal brain tissue more than normal anatomy. By altering the

parameters of the sequence, mainly the repetition time (TR) and echo time (TE), more T1 or more T2-properties of the tissue can be extracted. Naturally, sequences vary in acquisition length depending on these parameters.

The characteristics of T1 weighted images are: a) Water as well as dense bone and air, such as CSF, have low intensity, b) fat and lipid tissues appear hyper intense (compared with cortex) which also explains the high intensity of myelinated white matter due to lipid component of myelin, c) Grey matter is typically in between the intensity of CSF and WM. d) In areas of (microscopically) mixed tissue, so when GM and WM are neighbored, such as in the thalamus or pallidum that contain white matter fibers, the intensity lies between GM and WM (so, e. g. light grey).

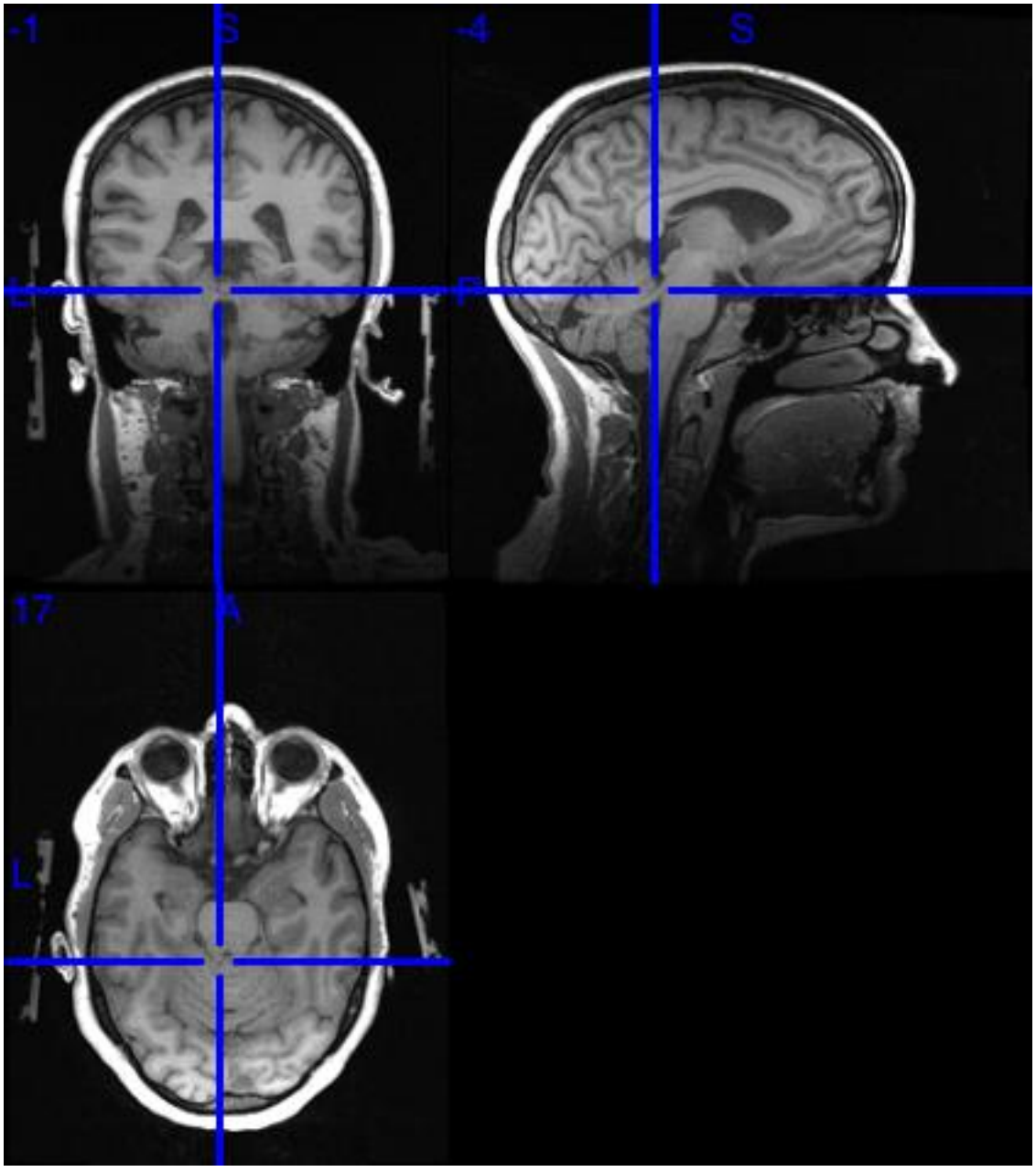


Figure 1.1: Example of T1-weighted image shown in three planes (sagittal, axial, coronal).

Note intermediate, greyish intensity of grey matter, high intensity of white matter, and low intensity of CSF (and w. g. air).

Brain tissue segmentation

As indicated, healthy brain tissue, or better the intracranial space, at an intermediate resolution typically generated in a 3 Tesla system, can be classified into three tissue types on the basis of a T1WI: GM, WM, and CSF **(Figure 1.2)**. Roughly, the distinction of the brain tissue in GM and WM is also corresponding to its first macroscopic, anatomical appearance when dissected post-mortem. Grey matter is the component consisting of neuronal cell bodies, neuropil (dendrites and unmyelinated axons), glial cells (astrocytes and oligodendrocytes), synapses, and capillaries. White matter is the tissue that contains the axons of the neurons and their myelin sheaths that together are responsible for signal transduction between neurons within the central nervous system. The intensity of white matter is due to the presence of fatty substances (myelin) surrounding the nerve fibers (axons). Cerebrospinal fluid (CSF) is a clear, colorless body fluid found in specific spaces in the brain (ventricles, aqueduct), around the cortex (external CSF spaces such as sulci and cisterns) and around the spinal cord. CSF is produced by ependymal cells in the choroid plexuses of the ventricles of the brain, and, after taking part in a circulation process, absorbed in the arachnoid granulations and extra spinally.

The segmentation of different tissue can be performed manually on a good quality T1 image, by selecting particular image intensity ranges encompassing the voxel intensities of the desired tissue type. However, this procedure is highly subjective. T1-weighted images, because of their good GM/WM/CSF contrast, have been widely employed for automated segmentation methods and brain morphometry. The voxel intensity in an image is one (and the most important) type of information we can use in the segmentation of tissue classes. The fact that most brains share many of the features in the spatial distribution of tissue classes gives us an additional type of information: that of location. This means that types of tissue are not randomly spread in the brain but are located relatively systematic and can be used by us to improve our segmentation. Tissue probability maps for various tissue classes made of a large number of brains registered into a common space thus represent very important source of information needed to define a voxel's tissue class. These maps to be considered the probability of a voxel belonging to a specific tissue class in a Bayesian context and are therefore often referred to as 'prior.' A large number of tissue segmentation algorithms use three category maps of probability: gray matter, white matter and CSF. It should be added that other classes are also used in these algorithms, such as 'skull', 'facial soft tissue', air space surrounding the head' etc. These additional classes help to reduce the ambiguity further.

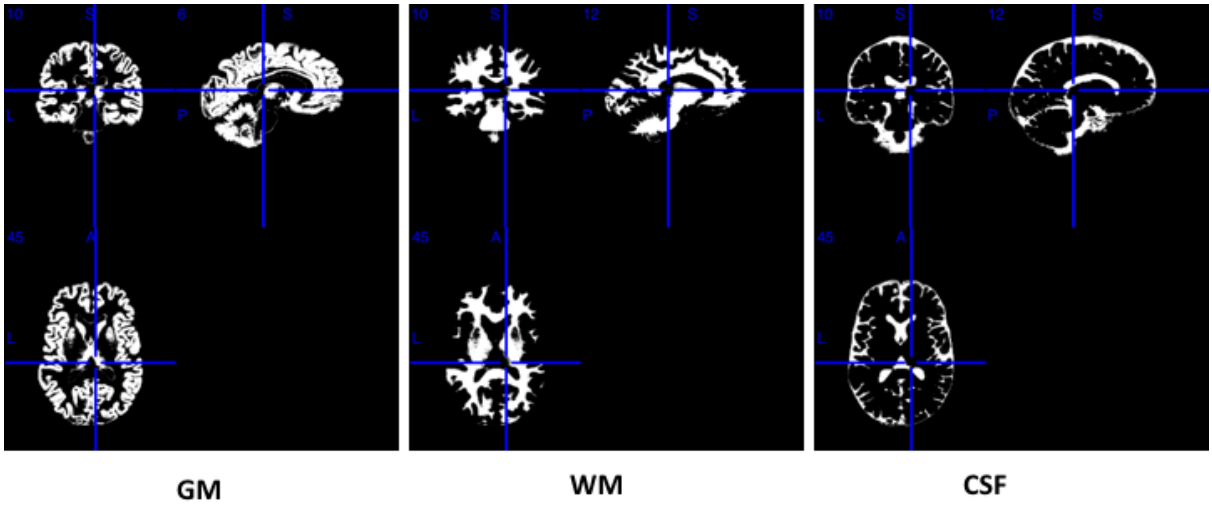


Figure 1.2: Segmented grey matter(GM), white matter (WM) and cerebrospinal fluid (CSF) images.

1.5 Unified segmentation and DARTEL registration

The use of tissue probability maps to segment the image of a subject poses a circular ('chicken-egg') problem. That is, the probability maps of the tissue are usually in a stereotactic standard space (in SPM this is the MNI space (Montreal Neurological Institute-Hospital). Therefore, it is necessary to perform a registration between the subject and the MNI spaces before using the tissue probability map. In turn, however, the tissue probability map gained for an individual needs the information from the MNI space prior. Thus, a segmentation algorithm that wants to make use of tissue priors needs to optimize both steps, registration between the individual's native space and the MNI space, and segmentation. Ashburner (2005) [25] managed to unify these steps, along with bias field correction as third element, into a single integrated generational model that is ref. to as *unified segmentation*. To achieve optimal local solutions for every process this model involves alternating tissue segmentation, bias field correction, and spatial registration. The tissue likelihood maps are registered to the data of the single subject, and the Bayes rule uses these priors in combination with tissue likelihood from voxel intensities to create native tissue likelihood maps[25].

The next challenge is to optimally co-register subjects of a study onto each other for later voxel-wise statistics. is important that this inter-related registration is as precise as possible. A sophisticated registration algorithm through exponential lie-algebra is available in the DARTEL algorithm[26] that is based on the output of the above-explained Unified Segmentation. DARTEL is currently not seamlessly integrated into the segmentation model and requires unified segmentation of the gray and white tissue maps. The current version of the unified segmentation typically does not segment the brain spinal fluid very reliably. Therefore, the DARTEL registration usually uses only gray matter and white matter tissue maps. DARTEL acts iteratively in up to six generations, each aligning the GM and WM

segments more and more to each other. The inter-subject co-registration in DARTEL can be started with pre-existing six generation TPMs of GN and WM (such as from the IXI sample), or these 6 generation of maps can be produced from the sample analysis itself. In the end, registered tissue maps are converted into standard size, and in order to conserve the amount of tissue in each structure the transformed maps are multiplied by the Jacobian deformations determinants[27,28]. By this step, the probability information (of the native space voxels) is transformed into a volume information in MNI space. Recent studies have shown that the performance of non-linear registration algorithms, such as unified segmentation and DARTEL, is better than the immediate output of the unified segmentations in MNI space[28,29]. It should be emphasized that the input to DARTEL is not the (already provided) MNI space GM and WM segments, but the native GM and WM (after a simple affine first pre-positioning referred to as import step).

1.6 Spatial normalization and voxel based morphometry (VBM)

Good inter-subject co-registration is the prerequisite to reliable VBM, because only then voxels with presumably the same anatomical 'origin' (or function) come to lie over each other across all subjects. The above-mentioned Unified Segmentation does provide MNI space and Jacobian modulated GM maps. For the spatial normalization, a Bayesian framework is used, using a prior knowledge of normal brain size variability, to estimate a maximum posteriori spatial transformation. The second step involves the differences in global nonlinear shapes modeled by a linear combination of smooth spatial basis functions. The nonlinear registration consists of evaluating coefficients of the base functions, which at the same time maximize the smoothness of the deformations and minimize the residual squared difference between images and template[30] After unified segmentation, there is reasonable agreement between the cortex shapes at the mesoscopic level of granularity – yet, DARTEL further improves the alignment *between* subjects. The better the voxel alignment is, the less shape differences remain – thus it becomes more important to apply the JM step after the full transformation through DARTEL - then, close-to-perfect co-registration is combined with individual volumetric information. One consequence of DARTEL over only Unified Segmentation is that due to the better geometric inter-subject co-registration, less spatial smoothing can be applied, and geometric result precision is thus higher. Generally, a relatively high resolution (1 mm or 1.5 mm isotropic voxels) is required for the spatially normalized images so that the gray matter extraction does not excessively interfere with a partial volume effect where voxels contain a mixture of different kinds of tissues.

Voxel-based morphometry (VBM) is a mass univariate approach for comparing the voxel-wise volume of tissue among populations of subjects[24–26,31]. Voxel-based morphometry has grown in popularity

since its introduction although there have been several criticisms[32]. Some studies have compared results of Voxel-based Morphometry analyses to manual measurements of particular structures and have shown relatively good correspondence between both techniques[32]. Although other software packages such as FSL can also be used, SPM is most widely used for Voxel-based Morphometry analyses. This is mainly due to the reliable, six tissue priors and the Unified Segmentation step combined with DARTEL-based improved spatial co-registration.

VBM implies a voxel-specific comparison between two groups of subjects of the local gray material concentration. Yet, all GLM-based models can be set up in the GLM framework of SPM, both with classical or Bayesian statistical inference. The normalized and Jacobian modulated GM images (sometimes, WM) are spatially smoothed to some degree (e. g. 6x6x6 mm^3 FWHM of a Gaussian kernel). As a rule of thumbs, the size of this kernel should be the *expected* size of areas of difference. This step leads to higher and more homogeneous spatial smoothness of the residual images, and to improved normal distribution of the residuals per voxel. The theory of Gaussian random fields can then be used to correct for multiple comparisons [33]as based on a certain smoothness the number of expected clusters emerging by chance can be estimated[34]. In **Figure 1.3**, the blue areas stand for voxel based comparison between different types of neurodegeneration (for example Alzheimer's disease, or semantic dementia), indicating areas with a deficit of grey matter volume[35]. This work will be explained in more detail as it bridges between disease-specific GM deficits and these areas resembling physiological structural and functional networks.

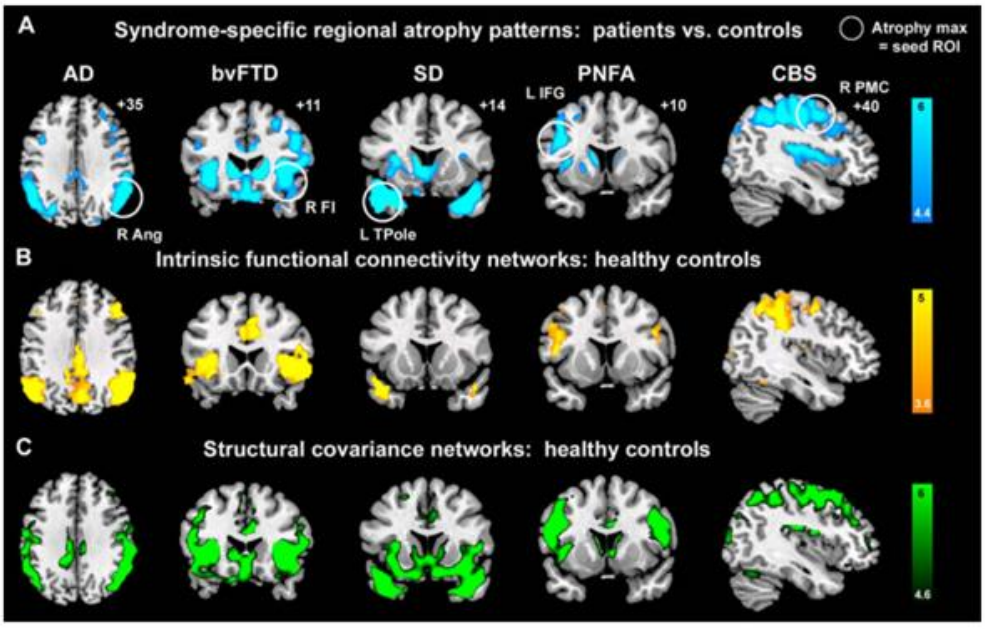


Figure 1.3: Combined VBM, resting state and morphological covariance study by Seeley.
[35]In this study, disease specific areas/patterns of GM deficits of five defined neurodegenerative disease (blue, here frontotemporal dementia) were compared with seed-based results of functional connectivity time series (of healthy subjects, see yellow result map) and morphological covariance maps (also of healthy subjects, see green result map). It was noted that these maps show high

spatial similarity among each other, as an indication that the pathological process may occur within a specific network and not randomly.

The modulation step in VBM is an essential step that is particularly important as soon as the spatial deformation (normalization) steps become very precise: In this case, the resulting voxel positions are very similar and most or all *individual information* is contained in the deformation field. Thus, this information has to be re-coded into the voxel value in the atlas space. For this re-scaling, the Jacobian determinant is calculated from the deformation field (see REF for vector mathematics of this). For example, a small area of e. g. 15 voxels of a brain structure in subject A (e. g., the hippocampus) with clear GM *probability* (>0.99), is stretched to 45 voxels in template space (so, expanded) – its resulting voxel values will thus be lowered (e. g. by some factor 3, so resulting *volume* values round 0.33). Another subject B's same original structure is sized 30 voxels, and it is also stretched to 45 voxels – here, the expansion is less strong, and the stretching facto only 45/30=1.5, so volume values will be round 0.66. Eventually, the 45 voxels are in the same geometric comparison, but for subject A, the lower original volume compared with subject B is still preserved. In essence, this Jacobian modulation step preserves the original volume information (that is a GM probability value) during the spatial transformation. Mathematically, the multiplicates of the original shape by the probability values is equivalent to the MNI shape by the modulated volume values of the corresponding voxels.

Grey matter voxel-wise volume based morphological covariance

Brain networks can be constructed based on similarity of GM volume between brain areas, which was named as the gray matter (GM) structural covariance network (SCN)[24,36–38]. Based on prior studies, the biological meaning of structural connectivity network (SCN) may link to coordinated GM growth during development[36], functional co-activation[36], axonal connectivity[39,40] and genetic factors[36,41,42]

Cortical thickness based morphological covariance

Structural covariance with the cortical thickness in two regions of the brain was suggested to reflect their synchronized maturation changes, possibly through axonal links forming and reforming over time. There is evidence for such structural covariance development models, but they have not yet been fully tested. Structural MRI networks are of qualitatively similar cost-effectiveness, small-world and modular properties to those for functional brain grids. Pairs of functionally interconnected regions may also have a strong structural covariance and highly correlated rate of adolescent anatomical change. The link between structural or maturational networks and functional networks, however, is not yet systematically investigated. The study now uses a structural MRI-data set from healthy young persons, scanned longitudinally at least 3 times over 6 years, for each of the 360 regional nodes for cortical thickness and mature change. The hypothesis was tested that structural covariation is related

to synchronized mature changes between the distributed cortical areas. Brain networks were compared on structural covariance and maturation changes.

1.7 Independent Component Analysis (ICA)

ICA is considered a multivariate extension of voxel-base morphometry in sMRI analysis. Most individual subjects are segmented and gray matter maps are organized into a matrix that is used to analyze and to create a maximum spatially independent map of the component of the ICA. These maps include areas with similar covariation of gray matter between subjects. The extent to which the data are 'expressed' can be recorded via the parameters of its loading. The multi-site data were used to study Volume Based Morphometry (VBM)[31]. A major advantage of this approach is that strong assumptions related to the use of atlases must not be made. The main difference between principal component analysis and independent component analysis are given below. a) PCA[43,44]chooses *orthogonal* vectors iteratively to explain most of the variance with the first few of them. ICA on the other hand does not have the orthogonality constraint of PCA but wants to emphasize statistical independence between components, b) ICA is a development out of PCA, c) PCA, yet, optimizes the covariance matrix ("second order statistics"), e. g. orthogonal, it where ICA optimizes ***higher order statistics*** (e. g. kurtosis), d) PCA finds uncorrelated components where ICA finds independent components.

Principal of ICA

We can use a statistical latent variables model to define Independent Component Analysis [45].
Suppose we observe n linear blends x_1, ..., x_n of n components[45]

$$\boldsymbol{x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n} \text{ , for all } \boldsymbol{j}. \qquad (1)$$

In the ICA model, we assume that both x_j mixture and s_k component is the alteration variable, not the proper time signals; The time index t has now been dropped. A sample of this random variable is then observed values $x_j(t)$, such as microphone signals in the cocktail party problem. We can assume that both the mixture variables and the independent components have a zero mean without loss in terms of generality: if this is not true, the observable x_i variables can then always be centered by subtracting the mean, meaning that the model is zero.

Instead of sums like the previous equation, it is convenient to use vector matrix notation. We have denoted the random vector $\mathbf{x}$ whose components are x_1, ..., x_n and also the random vector $\mathbf{s}$ with elements s_1, ... , s_n. Let us denote an element matrix $\mathbf{A}$ with elements a_{ij}. In general, bold lower case

characters show vectors, and bold upper-case characters indicate matrices. All vectors are considered to be column vectors; therefore, $\mathbf{x}^T$ or transposition of $\mathbf{x}$, is a row vector. The above mixing model is written using this vector matrix notation

$$\boldsymbol{x} = \boldsymbol{As} \quad \textbf{(2)}$$

The mixing model can also be written as

$$\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{a_i}\, \boldsymbol{s_i} \quad \textbf{(3)}$$

The statistical model in Eq.(2)[45] is called independent component analysis, or ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components s_i. The independent components are latent variables, meaning that they cannot be directly observed. Also the mixing matrix is assumed to be unknown. All we observe is the random vector $\mathbf{x}$, and we must estimate both $\mathbf{A}$ and $\mathbf{s}$ using it. This must be done under as general assumptions as possible. Equation no. 4 in the statistical model is referred to as independent analysis of components or ICA model. It describes how the observed data is produced by a process of mixing the components s_i, and thus is a generative model. The separate components are latent variables, which means they cannot be observed directly. It is also assumed that the mixing matrix is unknown. All that we observe was the $\mathbf{x}$ random vector, and $\mathbf{A}$ and $\mathbf{s}$ must both be estimated with it. The starting point for ICA is the simple assumption of statistically independent components s_i. We must also assume that there must be non-gaussian distributions on the independent component. But we don't assume in the basic model that these distributions are known (if they are well known the problem is significantly simplified.) We also assume that the unknown matrix of mixing is square. These assumptions can sometimes be relaxed. After we evaluate the matrix, $\mathbf{A}$ we can then calculate its inverse $\mathbf{W}$, say, and simply by:

$$\boldsymbol{s} = \boldsymbol{Wx} \qquad (\mathbf{4})$$

The properties of mixed signals: Mixed signal have certain properties. Those are a) independence, b) gaussianity and c) complexity.

1. **Independence**: If the source signals are independent, their mixture signals are not. This is because the source signals are shared between both mixtures.
2. **Gaussianity**: The histogram of mixed signals is bell-shaped histogram (Gaussian or normal). This property can be used for searching for non-Gaussian signals within mixture signals to

extract source or independent signals. The source signals **must** be non-Gaussian because the ICA model cannot estimate Gaussian Independent components.

3 **Complexity:** Mixed signals are more complex than source signals

Independent Component is closely linked with the method called BSS or Blind Source Separation (BSS)[46]. An "source" here is an original signal like a speaker in a cocktail party problem, i.e. an independent component. "Blind" means that we make little or no assumptions on the source signals on the mixing matrix. ICA is one method for performing blind source separation, perhaps the most commonly used. It would be more realistic in many applications to assume that the measured noise exists which means that a noise term is added to this model. We remove noise terms, since the estimation of the model without noise is sufficiently difficult by itself and seems to be sufficient for many applications. There are mainly two preprocessing before ICA. Those are a) demeaning and b) whitening. Whitening also consist of two steps, "Decorrelation" and "Scaling", There are three different kinds of ICAs., projection pursuit, infomax[47,48] and FastICA[45]

I Projection pursuit: Here the algorithm is trying to find ICs which maximize measures of **non-**Gaussianity such as *negentropy* or *kurtosis*.

II Infomax: Minimizing the mutual information.

III FastICA: Here algorithm is trying to maximizes non-Gaussianity by maximizing the negentropy for the extracted signals using a fixed-point iteration scheme.

Challenges of ICA

a When the number of sources (p) and the number of mixture signals (n) are equal, the matrix A is invertible.

b n < p (so less 'microphones' than sources): Over-complete problem; thus, A is not square and not invertible. Sometimes advantageous as it uses as few "basis" elements as possible; this is called sparse coding.

c n > p: number of mixtures is higher than number of source signals: Under-complete problem. This problem can be solved by deleting some mixtures using dimensionality reduction techniques such as PCA to decrease the number of mixtures.

2 Clustering of treatment response dynamics and prediction from clinical variables

The content of this chapter has been taken from the published paper and has been paraphrased (Paul, R., Andlauer, T.F.M., Darina, C., Hoehn, D., Lucae, S., Pütz, B., Lewis, C.M, Uher, R., Müller-Myhsok, B., Ising, M., and Sämann, P. G. (2019). **Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models**. *Transl Psychiatry* **9**, 187 (2019). https://doi.org/10.1038/s41398-019-0524-4). The figures and tables have been taken from the above mentioned published paper and have been cited.

2.1 Depression: a major disorder and a relentless burden

The most common psychiatric illness is major depressive disorder (MDD) which is a leading global cause of disability for years[49,50]. According to the World Health Organization (WHO), depression is a leading contributor to the global disease burden and related costs. The burden of disease is a concept developed in the 1990s by the Harvard School of Public Health, the World Bank, and the World Health Organization[51] to quantify premature death and health loss from illness, injury, and risk factors for all regions of the world. MDD[1]is an affective disorder mainly characterized by persistent feeling of sadness, loss of interest and energy, sleeping problems, loss of pleasurable activity, pessimistic thinking, feeling of guilt and worthlessness, and suicidality[52]. About 50% of individuals who commit suicide have before been diagnosed with a MDD or depression in the context of a bipolar disorder[53]. Because mood disorders underlie 50-70% of all suicides, effective treatment of these disorders on a national level can reduce this major complication of mood disorders.

The cost of depression, particularly the cost of lost workdays, is as great as or greater than the cost of many other common medical illnesses[54]. Depression is also associated with more impairment in occupational and interpersonal functioning in comparison to other common medical (somatic) illnesses. It has a large impact on maternal and child well-being, and eventually contributes to risk factor for child mortality, for example influenced by less breastfeeding. Depression during pregnancy is strongly associated with low birth weight and other sequelae for the child[54]. It is thus important to consider that the outcome of depression is significantly improved by early detection[54]. As soon as detected, a wide range of highly effective treatments including pharmacological treatment, somatic/biological therapies (such as light therapy, electroconvulsive treatment, sleep deprivation), and psychotherapeutic interventions are available. Antidepressant treatment and supportive psychological interventions are effective in about 80% of patients[54], but often, treatment is prolonged

and first treatment attempts are ineffective in about half of (out-) patients (Star*D Study[55,56]). Also, the number of trained health care providers (psychiatrists, psychologists, and psychiatric nurses) is limited, particularly when critically reviewed at the global level; here, rural areas seem less well covered compared with urban and semi-urban areas[57].

2.2 Treatment Response (TR) in MDD

In the context of depression, treatment response can be understood as a general term that describes the change of clinical symptoms under a certain treatment. Both self-rating (such as the Beck Depression Inventory) and clinician rating schemes (such as the Hamilton Disease [HAM-D] Rating scale)[58,59] can be used to quantify treatment response. As no reliable biomarkers exist so far, monitoring treatment response is indeed still mostly done clinically. For scientific studies, self-rating scales or rating scales used by trained raters are the standard. The HAM-D covers a time period of the last 7 days on which the patient is interviewed regarding his/her symptoms of depression. Different clinical rating scales of different degrees of reliability and validity are available[60]. The HAM-D rating scale is one of the most widely used ones in international studies and assesses a past time window of up to one week, hereby covering most domains that define MDD, such as depressed mood, suicidality, anhedonia, lack of drive, circadian symptom changes, and autonomous nervous system disturbances. Good test-retest and interrater reliabilities are achieved by properly trained raters[61–65]. In clinical studies, the rater should not be the treating psychiatrist to avoid any bias, both on the patient and the rater side.

Clinical response and remission are defined as follows in the context of MDD research: Response is defined as a minimum 50% decrease of the score on the respective rating scale compared to the baseline score. Remission is not defined by a relative improvement, but by an absolute threshold and can best be described as the practical absence of symptoms. For the HAM-D rating scale both thresholds of lower than 8 or 10 points have been suggested[61–65]. According to the same mentioned previous studies on the Munich Antidepressant Response Signature project, partial responses were defined as a reduction of at least 25% after two weeks, and a reduction of at least 50% after 5 weeks compared to the HAM-D_{21} (HAM-D rating scale with 21 items) at the time of admission. The 17-item version of the HAM-D (HAM-D_{17}) with values lower or equal to 7 was used as equivalent definition of remission in accordance with consensus criteria[66,67]. *Stable* response or remission is defined as response or remission for at least two consecutive weeks. *Treatment resistance* has been defined as a situation in which patients fail to respond in an appropriate time (6-8 weeks) to at least two trials using different antidepressants, as suggested by Souery[68]. Despite typically successful short-term therapies for acute episodes of major depression, many patients have experience relapses (early

return of symptoms within the expected duration of a current episode, e.g. between three to twelve months) or recurrence (new episodes). A recurrent depressive disorder has frequent episodes of depression without an independent history of mood and increased mania or energy. There should be at least one previous episode which would lasted for at least two weeks and has been isolated for a period of a minimum of two months by the new episode. At no time in the past has there been any hypomanic or manic symptoms. We have four different kinds of recurrent depressive episode: a) MDD with mild recurrent, b) MDD with moderate recurrent, c) MDD with severe recurrent without any psychotic features, and d) MDD with severe recurrent with psychotic features. Different psychotherapeutic interventions could be applied to prevent depression relapse and to reduce depressive symptoms[69]

2.3 Hamilton Depression Rating Score

In patients diagnosed with a major depressive disorder, the Hamilton Rating Scale for Depression (HAM-D) is a widely used scale for assessing depression severity by trained clinicians (medical doctors or psychologists or psychiatric nurses)[58,59,70,71]. The initial version of the HAM-D of the year 1960[10,11] consisted of 21 items: the HAM-D_{21}. The number of citations of the HAM-D exceeds 21.000 in the Scopus literature search system[72], demonstrating its wide use. Later, the developer Paul et al. (2019)[73] recommended to use only the first 17 items[11] as the other four ones were neither considered part of the disease or relatively rarely occurring or as features related to depression severity (i.e. diurnal variation, de-personalization or de-realization, paranoia and obsession or compulsive symptoms). Several other amendments were proposed over the years and a number of HAM-D versions have been developed and implemented.

One measure of reliability in clinical scoring is the inter-rater reliability. Here, generally a rater is a person who scores or measures a human or animal's performance, behavior or skill. If the observations of different raters differ significantly, then the rating items or rating technique could be unreliable and brings rater-dependency into the statistical analyses. A meta-analysis of the interrater reliability of the HAM-D was carried out by Trajković et al.[74] who reported a combined mean intra-class coefficient 0.92 which indicates an excellent level of inter-rater reliability. Unfortunately, only the 17-items version of the HAM-D of the year 1960-2008 was investigated. They included 409 articles in their analyses that had reported one of the following measures: the intra-class correlation coefficient, Pearson coefficient of correlation, Spearman coefficient of correlation rank, kappa coefficient or Kendall W. They have found a pooled mean intra-class correlation coefficient ICC of 0.92, suggesting an outstanding degree of reliability of the intra-class correlations. Unfortunately, they did not differentiate between the

various version of the HAM-D. It is a dynamic and individually different process to develop and recover from an episode of an MDD. Multiple additional symptoms may occur during an episode, as described by **chapter 1.1**, with each individual pattern and variability during episode[11,67], as during the treatment response also during the development of MDD, patients could go through sub-clinical phases with areas with preserved function. Thus, while consensus definitions of MDD include basic symptoms such as anhedonia and depressed mood[1]. For instance, in the initial phase, MDD can first start with sleep problems.

This diversity of symptoms may be underpinned by strong, inter-individual differences in psychosocial stress sensitivity – a major risk factor for MDD[75]. Up to 119[76] combinations of defined symptoms were estimated to be possible in order to meet the MDD criteria. Similarly, there are significant individual differences between patients in the regression of symptoms under treatment. However, both stable subgroups[77–79] and predictive clinical patterns[78–83] have been hypothesized and shown to exist. The predictive clinical pattern is important for the successful clinical management of MDD.

The main rule is that treatment should ideally lead to full recovery, as persistence of residual symptoms significantly increases the probability of a relapse[1]. This implies that postponement of therapeutic intensification or too late medication switching may further increase the risk of therapeutic failure and chronification[84]. Early treatment response (e. g., within two weeks) is strongly predictive of the longer course[85]. This is an important and long established correlation which also applies to patients receiving first-time antidepressant treatment[86]. Similarly, various psychopathological profiles (i.e. combinations of symptoms) may reflect differences in functional stress sensitivity and therefore may be predictive of the treatment response. For example, a patient who has severe anhedonia as a key symptom may respond to a dopaminergic system therapy particularly well[87].

Clustering and multivariate prediction

Despite the heterogeneous symptom profile of depression, treatment response classes are usually based *on* compound scores, which *then are subjected to relative criteria of change or absolute thresholding* (e.g. depression severity below a certain cut-off over a specific time period). Various multivariate statistical approaches have been employed to identify predictive patterns for such conventional treatment response classes[80,82,83]. Chekroud et al.[80]have used an elastic network to identify 25 of the 164 patient-reportable variables from Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study[55,56,79] which predicted a citalopram reaction. These values were used to train a machine learning model that could be validated in an external sample with significant but low

percentage of correct predictions (59.7%). Nie et al.[82] trained five different machine learning algorithms, using data from the STAR*D study, either in the complete set of 700 or two differently reduced set of clinical parameters (30 and 22, respectively), to predict treatment resistance or non-resistance in STAR*D (at week 12). In addition, they predicted the same outcomes in an independent validation study (at weeks 6). Early response markers carried the predictions, yet obtained only moderate exactness. Wardenar et al.[83] have reported significant improvements in the prediction of persistence and severity of depression through information on co-morbidities. However, although the classes of responses predicted here are largely rooted in the long-known importance of early response and full remission, these markers are not fully data based, so they do not represent all dynamic patterns contained in the data.

Clustering analysis can be useful here if the data space is to be dissected into sub-spaces based on features shared between subgroups and distinct between them[88]. Clustering analysis has so far been mainly used towards cross-sectional markers to identify subtypes based on clinical symptom profiles, [13,58,89,90] cognitive markers[91] or functional imaging markers[92], assuming that the clusters could stand for the distinct pathophysiological components. Here we are trying to cluster the treatment response dynamics based on the trajectories of the total severity of symptoms, i.e. the clinical development of patients over a defined observing time. Such five[93] or nine prototypical trajectories[94] based on 12 weeks of observation were before reported in longitudinal latent class analysis. More specifically, this study[93] showed a relatively limited prediction by 13 basic clinical items and polygenic scores made up of five hypothesis-based genetic variants based on literature. The second study[94] found weak association between the response trajectories and the type of the medication, but here no clinical predictors were investigated. Seven trajectories were discovered in another study based on one year of observation[95]. The limitation of these studies is low generalizability as data from single center studies were only used.

Study design overview

In performing this analysis, we aimed to create TRCs in a non-biased way. A second aim was to investigate to what extent the results of the first aim were clinically valid and equally important, generalizable. The main input for these analysis were the trajectories of the depression ratings. In order to verify the generalizability, two different cohorts were used differing in aspects of patient selection and treatment approach. Study number one was the prospective Munich Antidepressant Response Signature (MARS) cohort. This is a multicentric study with the MPI of Psychiatry, Munich, being the main hub[61]. By design it is an open and observational. Diagnosis included bipolar depression, major depressive disorders, and schizoaffective disorder, covering thus various aspects of the

depressive spectrum[61] Second cohort was Genome-based Therapeutic Drugs for Depression (GENDEP)[96,97], a multicenter study investigating pharmacogenomic and clinical aspects. It included patients with at least a moderately severe depressive episode treated with nortriptyline or escitalopram and monitored up to 12 weeks in a partially randomized design[98]. Both studies used the Hamilton Depression Rating Scale (HAM-D) to measure current symptom levels for most MDD domains such as depressed mood, suicidality and anhedonia, lack of drives, circadian symptoms changes and autonomous nervous system disorders, achieved strong test and interrater reliability[99].

From a methodological point of view, we intended to make optimal use of the repeated clinical ratings and applied a model-based non-linear longitudinal clustering technique to detect TRCs (also referred to as *clusters*) in MDD. More specifically, we employed a mixed model-based clustering algorithm in our discovery sample which as a subsample of the explained MARS cohort. Summarized briefly, this mixed model-based clustering algorithm assigns individuals to a TRC by borrowing information from all other individuals and, thereby, improves cluster stability, which often is critical for generalizability and clinical applications. Another advantage of this algorithm is that the optimal number of clusters is also estimated from the same dataset by a bootstrapping approach.
Finally, we assessed cluster stability empirically in a two-stage design in the second subsample of MARS and in the GENDEP sample. Last, we performed an indirect validation by exploring if clinical characteristics at baseline (MARS) can predict the detected TRC by using a multivariate random forest approach.

2.4 Samples and Methods

The respective Local Ethics Committees have approved both the MARS and the GENDEP study protocols. Prior to the participation, all participants gave their written informed consent. Patients of MARS have been admitted for treatment for different depressive disorders to a hospital MPIP in Munich, Germany and to a collaborative hospital in southern Bavaria and Switzerland. The study was initiated in the year 2000 with the aim of generating large longitudinal observations with weekly ratings and sociodemographic, psychopathological and biological data from in-patient patients with all kinds of depressive disorders, such as MDD, bipolar depression and schizoaffective disorder[61]. ICD10[100] diagnoses have been obtained from the patient interviews and clinical recorded patients of trained psychiatrists[61]. Of the 1286 patients available, the ones were eligible with a single episode of MDD (ICD-10-F32), N=373) or a recurrent (unipolar) depressive episode (ICD-10-F33 and N=698). Patients with bipolar depression (N=175), chronic depression (ICD-10 F34, N=3), or baseline HAM-D patients < 14 (N=37) were excluded from the analysis. Of these 1071 remaining eligible patients, 834

(recruited 2002-2011) patients formed the discovery sample and 236 (recruited 2012-2016) formed the replication sample. The split point between discovery and replication, represented an organization time marker related to genotyping activities which has no association with this analysis[73].

The age range was 18 to 87 years (see demographic and clinical details in Table 1) and all patients were European ethnicity. In order to optimize the plasma medicine levels, patients received therapeutic drug monitoring and were treated psychopharmacologically according to the doctor's choice. Depression symptoms were assessed weekly with 21-item version of HAM-D up to week six and then bi-weekly with the most recent assessment up to week 16 unless not discharged earlier. In the first six weeks, 7.1% of HAM-D scores were missing due to organizational reasons. We used linear interpolation to obtain complete time series where HAM-D scores for the first six weeks missing as well as for the two-weekly HAM-D scores skipped. 88% of discovery patients and 99% of the validation samples for MARS were discharged prior to the 16th week and thus the original time series for HAM-D had less than 17 data points.

The GENDEP study is a partially randomized, multicenter, depression[87] clinical and pharmacogenomic study, in which 826 subjects were registered between July 2004 and December 2007. The main inclusion criteria were the diagnosis of a major depressive episode of at least moderate severity as defined by DSM-V[1], ICD-10 criteria[100] and Schedules for Clinical Assessment in Neuropsychiatry (SCAN, version 2.1)[101]. Exclusion criteria included a degree of first-grade relativity of bipolar disorder, a history of an event hypomanic or depressive, incongruous mood psychotic signs, primary abuse of the drug, primary organic disorder, ongoing antipsychotic therapy or mood stabilization, and maternity or lactation. For 12 weeks, nortriptyline (50 to 150 mg / day) or escitalopram (10 to 30 mg / days) with clinically informed dose titrations were randomly assigned to patients eligible for both antidepressants. The other medications were non-randomly assigned to those patients with a history of adverse effects, non-response or contraindications to one of these drugs. The other antidepressant was given to patients who could not tolerate or did not have adequate improvements to the first antidepressant with an adequate dose within 8 weeks. Depression symptoms were assessed by psychiatrists and psychologists on a weekly basis up until week 12, using the 17-item version of the HAM-D[97]. All subjects ranged from 18 to 72 years, and all patients were of European ethnicity. A combination of 15 people with lack of data on all three baseline suicidal items and 809 patients with baseline HAM-D scores < 14 were excluded[97]. Demographic information is provided in **table 2.9**. Various biological aspects of treatment response[102,103] and schemes of psychopathological predictors have been reported from this study[94].

Mixed-Model based clustering algorithm

The course of the HAM-D score time series after logarithm (ln) transformation was described by using a mixed model approach (ln of [HAM-D score + 0,5]) to take into consideration information not only from each trajectory but also combining pathways from several patients to identify TRCs. The *FlexMix*[104,105] clustering algorithm in R (version 3.3) was used on the HAM-D trajectories of the MARS discovery sample for a first organization of HAM-D responses into TRCs. *FlexMix* offers a flexible fitting infrastructure with the expectation-maximization algorithm to cluster individual trajectories for finite mixing models. The algorithm oscillates between computing and maximizing the expectation of the log-likelihood. Thus estimates of the parameters of the TRCs are obtained. We ran the clustering model with 200 repetitions and later jackknifing approach with 1000 repetition was applied on the clustering model to finalize a stable clustering solution. The optimal number of TRCs is determined on the basis of the Integrated Completed Likelihood (ICL) model criterion[106].

Stability and repeatability of clusters

The coefficients of the discovery sample model were projected onto a second sub-sample of the same cohort, the MARS validation sample, to validate the stability and generalization of the clustering solution (N=236). The hypothesis was that patients could be categorized into defined TRC with about equal proportions and similar HAM-D-mediated cluster-specific courses as observed for the sample.

Validation of the clusters in an independent study

In addition, we projected the same clustering model onto 12-week HAM-D courses of the GENDEP sample, hypothesizing similar median HAM-D courses per class, yet, not necessarily similar cluster proportions. These, we hypothesized, might be different due to differences in the patient population of GENDEP and the different study design. For both projection experiments, the resulting proportions of classes were compared with the original distribution of the discovery sample using a χ^2 test. In order to assess the suitability of the clustering solution for the validation samples, posterior likelihood values, classification log-likelihoods, and also ICL values were calculated on the basis of the clustering model of the discovery sample. *Furthermore, projecting our clustering solution on 12-week GENDEP Sample, a procedure implying similar mean HAM-D trajectories by class. We found related cluster frequencies not necessarily identical likely due to differences in patient population and studies design.* The resulting class proportions were compared with the original distribution of the discovery sample using a χ^2 test. We systematically decreased the number of applied coefficients down to 1 for every observation interval and compared that classification to the classification based on all coefficients (i.e. the complete observation interval). *Pearson's correlation between the model-based slope values for*

the respective TRCs value was calculated in order to determine the true distance between any two solutions.

Clinical variables for predicting and validating TRCs

A random forest algorithm was then used as implemented in the R package Ranger[107] to detect associations between the MARS clinical variables and the previously achieved TRCs, thus implementing *multivariate analysis.*

The batteries of clinical predictors: **Table 2.1**, explains all 72 clinical variables. Their choice was based on two principles: first, the availability in both MARS subsamples and second, the choice of those variables, based on widely available measuring instruments (to allow for replication experiments). The main model (Model 0) included 50 clinical variables based solely on basic evaluations, covering fields of sociodemography, clinical diagnosis, history of the MDD, present episode, family history of mental health, basic laboratory information, events of life, the current psychopathology (SCL-90R Symptom Checklist),[108]and questionnaires concerning personality (Eysenck Personality Questionnaire [EPQ])[109], Tridimensional Personality Questionnaire [TPQ][110]). As random forest model data sets are required, missing data were filled by calculating the median for the overall sample (see additional Table S2 for details). Extended models are Model 1 that extended model 0 by including the 21 baseline HAM-D singe items that represent the baseline psychopathology; Model 2 extended the model 0 by HAM-D data from weeks 2 that can be regarded early response. Model 3 combined both two expansions (**Figure 2.1**).

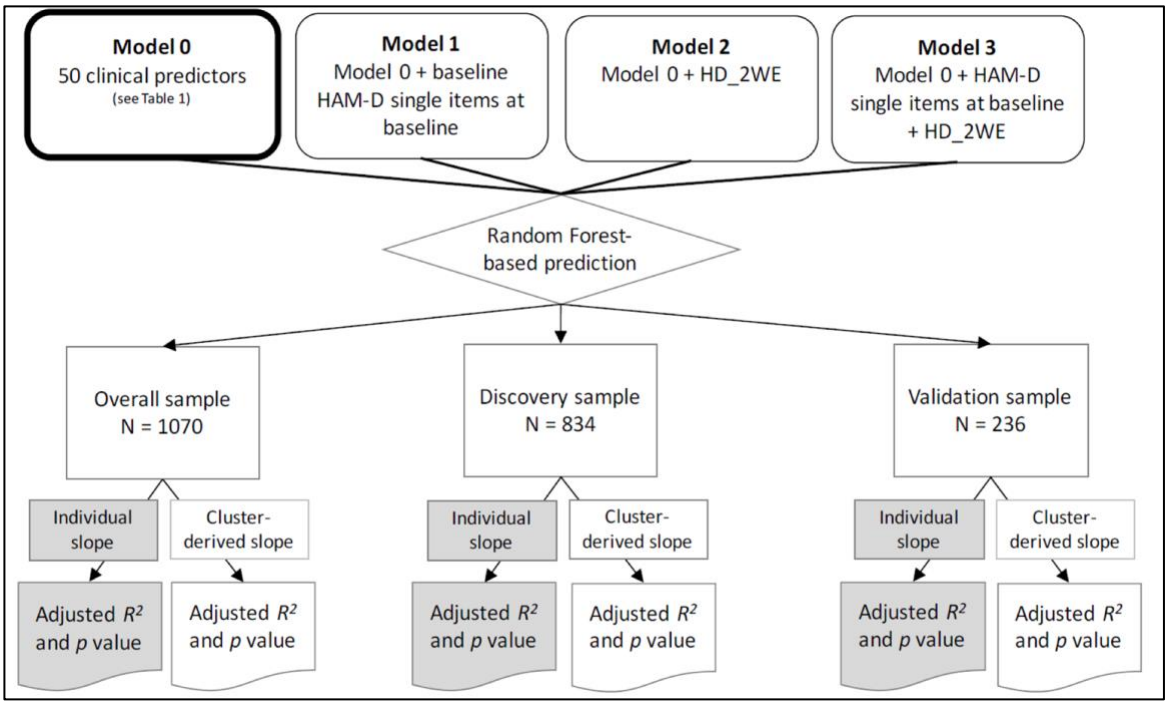


Figure 2.1: Specification of four clinical sets of predictors for the prediction model. *Model 0* (see methods and Table 2.1 for detailed specification of all clinical items), and three extended models (*model 1, model 2* and *model 3*). Random forest prediction models were estimated for all four sets of predictors with each two variants modelling the treatment response slope (individual slope vs. cluster-derived slope)[73]

Random forest-based prediction models

A fast implementation of random forests with high dimensional data is the key algorithm available in the Ranger package. In a random forest, every node is divided into a subset of predictors randomly selected at this node[111]. In order to control this process there were two parameters: prediction trees and search features to find the best characteristic (try). Mtry is the square root of D, which is the number of independent classification predictors. The predictions were achieved by aggregating forecast trees (i.e., the average for regression models and the majority for the classifying). In order to quantify the explained variance and the predictive quality of the entire model, we calculated adjusted coefficients for several correlations R^2 and corresponding p values. To characterize the characteristic importance of each variable, a permutation method was applied[112] (1000 permutations) which uses the distribution of measured meaning in a non-informative setting; the $p<0.05$ predictors are reported in more detail. Further, after Fisher's Z-transformation of their respective r values, R^2 differences between competing models were compared. The MARS discovery and validation sample were considered jointly in the multivariate prediction model. Two ways of HAM-D time series modeling were considered and compared for each set of predictors: first, the patient's individual response slope which was estimated using a simple linear regression of natural log (ln) transformed HAM-D values, and second, the slope estimated in the clustering model. The aim of the comparison was to find out

whether the clustering method would produce more meaningful and generalizable results. In addition, the accuracy values for classes were calculated on the basis of the respective confusion margins wherein the class of interest was defined as true class and the six other classes as false class. This was done by [true positives + true negatives], and by the other six classes as false class.

Table 2.1: Description of clinical items used for multivariate prediction models in the MARS cohort[73]

Category	Model	Short name	Variable description	Type	MARS discovery sample		MARS validation sample		*p*-value[a]
					Mean	SD	Mean	SD	
					%		%		
Socio-demo-graphic data	0	Age	Age at study inclusion (years)	*N*	48.26	14.02	45.48	14.99	0.008[c]
	0	k_sex	Gender (% female)	*D*	53.72 %		53.39 %		0.941
	0	spouse	Living with a partner	*D*	50.24%		38.56%		0.001[b]
	0	education	School years of education (university not considered) (years)	*N*	10.33	1.46	10.21	1.51	0.275
	0	training_retirement	Being in training/retirement *vs.* employment	*D*	25.42%		22.46%		0.393
	0	employment	Employment status: unemployed/ part time/full time	*N*	1.54	0.78	1.56	0.78	0.678
Diagnosis	0	ICD10	ICD-10 code for recurrent depressive disorders (F33) (%)	*C*	64.63%		66.95%		0.536
History of depressive disorder	0	age_on	Age at disease onset (years)	*N*	36.51	15.16	34.06	14.26	0.027[b]
	0	prev_epi	Number of previous depressive episodes	*N*	2.62	5.24	2.58	3.69	0.894
	0	s_history	Any suicide attempt before current episode	*D*	19.54%		9.32%		0.0001[c]
	0	psycho t_ history	Psychotic symptoms in any previous episode	*D*	11.15%		3.81%		0.0004[c]

Family history	0	Fam_history	Family history of any mental disorders	*D*	63.19%		64.41%		0.760
	0	Fam_F20_F25	Family history of schizophrenic disorders	*N*[d]	0.08	0.37	0.06	0.32	0.340
	0	Fam_F31	Family history of bipolar disorders	*N*[d]	0.05	0.31	0.08	0.37	0.348
	0	Fam_F32__F34	Family history of affective disorders (except bipolar disorder)	*N*[d]	0.88	0.95	0.87	0.96	0.857
	0	FA_X60	Family history of attempted suicide	*N*[d]	0.23	0.58	0.16	0.46	0.082
Information on current episode	0	index_d	Duration of the current episode (weeks)	*N*	34.54	58.74	32.19	51.58	0.577
	0	ATRQ_Score	ATRQ total score of treatment resistance for pre-medication	*N*	1.09	0.90	1.01	1.33	0.311
	0	s_current	Suicide attempt during the current episode	*D*	10.31%		2.54%		<0.0001[c]
	0	psycho t_current	Psychotic symptoms during the current episode	*D*	10.43%		2.97%		0.0001[c]
Basic medical and baseline laboratory data	0	Height	Body height (m)	*N*	1.72	0.09	1.72	0.09	0.545
	0	Weight	Body weight (kg)	*N*	25.65	6.07	25.94	5.28	0.504
	0	Bmi	Body mass index (m^2/kg)	*N*	25.34	4.41	25.94	5.27	0.075
	0	HR	Heart rate (1/min)	*N*	82.75	13.16	80.45	12.14	0.016[b]
	0	RRsys	Systolic blood pressure (mmHg)	*N*	125.78	18.10	128.04	17.44	0.088
	0	RRdia	Diastolic blood pressure (mmHg)	*N*	78.70	11.06	79.10	11.97	0.640

	0	cort_basal	Morning cortisol level (µg/l)	*N*	200.53	39.61	206.70	63.54	0.068
	0	TSH	Thyroid stimulating hormone level (µIU/l)	*N*	1.47	1.02	1.75	1.21	0.0005[c]
	0	fT3	Free T3 hormone level (pmol/l)	*N*	4.57	0.93	4.45	0.62	0.065
	0	fT4	Free T4 hormone level (pmol/l)	*N*	16.16	9.23	15.29	3.57	0.158
	0	CRP	CRP level (mg/l)	*N*	1.49	2.92	2.83	9.00	0.0002[c]
	0	HbA1C	HBA1c level (mmol/mol)	*N*	5.34	0.34	5.31	0.35	0.209
Life events	0	L-Event	Sum of life events	*N*	29.50	10.46	30.23	11.83	0.359
	0	wL-Event	Stress-weighted sum of life events	*N*	82.30	38.65	86.36	47.94	0.177
Baseline psycho-pathology	0	scl_som	Symptom checklist-90-R (SCL-90R) for somatization	*N*	0.97	0.64	0.99	0.64	0.488
	0	scl_comp	SCL-90R for compulsiveness	*N*	1.77	0.72	1.70	0.69	0.177
	0	scl_uncert	SCL-90R for uncertainty in social contact	*N*	1.30	0.77	1.33	0.83	0.630
	0	scl_dep	SCL-90R for depression	*N*	2.08	0.73	2.06	0.76	0.660
	0	scl_anx	SCL-90R R for anxiety	*N*	1.37	0.70	1.31	0.75	0.258
	0	scl_agg	SCL-90R for aggressiveness/hostility	*N*	0.77	0.60	0.86	0.69	0.046[b]
	0	scl_pho	SCL-90R for phobic anxiety	*N*	0.88	0.75	0.94	0.83	0.283
	0	scl_par	SCL-90R for paranoid ideation	*N*	0.92	0.72	0.99	0.82	0.218
	0	scl_psy	SCL-90R for psychoticism	*N*	0.83	0.55	0.80	0.54	0.507
Personality items	0	epq_neu	Eysenck Personality Questionnaire (EPQ)-RK neuroticism	*N*	6.85	2.50	6.84	2.73	0.938
	0	epq_psy	EPQ-RK psychoticism	*N*	1.92	1.24	2.16	1.40	0.010[b]
	0	epq_ext	EPQ-RK extraversion	*N*	5.20	2.97	5.07	3.03	0.567

	0	tpq_ha	Tridimensional Personality Questionnaire (TPQ) Harm avoidance total	*N*	20.63	5.58	20.27	5.92	0.386
	0	tpq_ns	TPQ Novelty Seeking total	*N*	13.07	3.81	14.04	4.41	0.001[b]
	0	tpq_rd	TPQ Reward Dependence total	*N*	17.75	3.30	17.50	3.84	0.318
	0	tpq_rd2	TPQ Reward Dependence - Subscale Persistence	*N*	4.81	1.70	4.88	1.86	0.623
HAM-D single items (baseline)	1, 3	HAM-D0_01-HAM-D0_21	21 HAM-D single items (baseline)	*N*	N/T[e]	N/T[e]	N/T[e]	N/T[e]	N/T[e]
Early partial response (at week 2)	2, 3	HD_2WE	HAM-D early partial response (≥25% reduction) after 2 weeks	*D*	N/T[e]	N/T[e]	N/T[e]	N/T[e]	N/T[e]

[a] Two-sided comparison between the MARS discovery and validation samples (Fisher's exact test and Fisher-Freeman-Halton test for dichotomous and categorical variables; Student's t test for **numerical** variables)

[b] Nominal significance (p<0.05)

[c] Significance after Bonferroni correction for multiple testing, here: p<0.05/50=0.001

[d] To allow optimal use in a parametric test, variables were coded as 0 (no relative affected), 1 (only second-degree relatives affected), and 2 (first-degree or first-degree and second-degree relatives affected).

[e] Not tested as these items were not part of model 0.

Abbreviations: N, numerical; *D*, dichotomous; *C*, categorical

Table 2.2: Percentage of imputed clinical batteries[73]

Category	Short name	Proportion of imputed values	Category (cont'd)	Short name (cont'd)	Proportion of imputed values (cont'd)
Sociodemo-graphic data	Age	0%	Life events	L-Event	41.49%
	k_sex	0%		wL-Event	41.86%
	spouse	2.61%	Baseline psycho-pathology	scl_som	19.53%
	education	5.23%		scl_comp	19.71%
	training_ retirement	3.73%		scl_uncert	19.62%
	employment	28.5%		scl_dep	19.43%
Diagnosis	ICD10	0%		scl_anx	19.62%

History of depressive disorder	age_on	4.39%		scl_agg	20.09%
	prev_epi	16.44%		scl_pho	19.71%
	s_history	12.14%		scl_par	19.81%
	psychot_history	4.20%		scl_psy	20.09%
Family history	Fam_history	2.80%	Personality items	epq_neu	37.47%
	Fam_F20_F25	1.78%		epq_psy	37.57%
	Fam_F31	1.78%		epq_ext	37.66%
	Fam_F32_ _F34	1.59%		tpq_ha	37.38%
	FA_X60	1.78%		tpq_ns	37.38%
Information on current episode	index_d	10.0%		tpq_rd	37.47%
	ATRQ_Score	21.96%		tpq_rd2	37.29%
	s_current	9.81%			
	psychot_current	0%			
Basic medical and baseline laboratory data	Height	2.89%			
	Weight	7.66%			
	Bmi	8.13%			
	HR	2.61%			
	RRsys	2.61%			
	RRdia	2.52%			
	cort_basal	62.99%			
	TSH	12.89%			
	fT3	35.42%			
	fT4	35.14%			
	CRP	54.29%			
	HbA1C	64.76%			

We have investigated not only treatment response dynamic clusters based on HAMD sum score over time points but also briefly investigated treatment response clusters in symptom space. This can be achieved by **developing symptom class specific response clusters** by (1) dissecting depression symptom space into factors (model based cross sectional clustering using *"mclust"* package) and (2) performing 3D-model-based clustering on trajectories of these factors over 16 weeks, similarly as in recent work (*Paul et al., Translational Psychiatry 2019*[73]), yet by a variation of the algorithm that considers several trajectories at the same time. Here we performed model based clustering *"mclust"* (R package) for the **Step I** (1) using Hamilton baseline 21 items as well as 17 items. Items 18 to 21 has lower score in general compared to rest of the item scores, to investigate that furthermore, we used

two different types of number of items for model based cross-sectional clustering using *"mclust"* package (Gaussian finite mixture modelling). We imputed 4% overall missing values in the Hamilton single items. To impute the missing value, we used *"imputeData"* package which is suitable algorithm for *"mclust"* packages to impute the missing data. Dataset here is being split into training and validation sets (50%-50%). Parameters for EM algorithm are tolerance *("tol"* $=1^{-05}$) and *"itmax"* which is a vector of length two giving integer limits on the number of EM iterations and on the number of iterations in the inner loop for models with iterative M-step ("VEI", "EVE", "VEE", "VVE", "VEV"), respectively ("*itmax"* = c(1000,10000)). Then in the **Step II we can calculate polygenic correlates of specific response clusters** by (1) correlating SNP sets from depression GWAS with symptom class specific model slopes and (2) calculating polygenic response scores (PRS) per symptom class in an independent sample. Finally, in the **Step III, we can integrate these PRS in prediction models per symptom class *(independent sample)*** by calculating individual PRS per symptom class (using the class specific weighting scheme from II) and including such PRS in multivariate prediction (Random Forest or Probabilistic neural network) models along with classical clinical baseline predictors. In current study we have mainly detected HAMD item specific clusters cross sectional and baseline items (HAMD -21 items and HAMD-17 items in baseline).

2.5 Results

Clustering of HAM-D time courses

The FlexMix clustering algorithm for any number of clusters k<4 or k>10 has not been used in the HAM-D time course for the sampled discovery sample. For k <=4 and k>=10, we have therefore evaluated the stability of the clusters in more detail using 200 algorithm repeats and 1000 repetitions using a jackknifing approach for each k. Seven clusters (**Figure 2.2 and Figure 2.3A**) found the lowest value of the ICL criterion, which represents the optimal model fit.

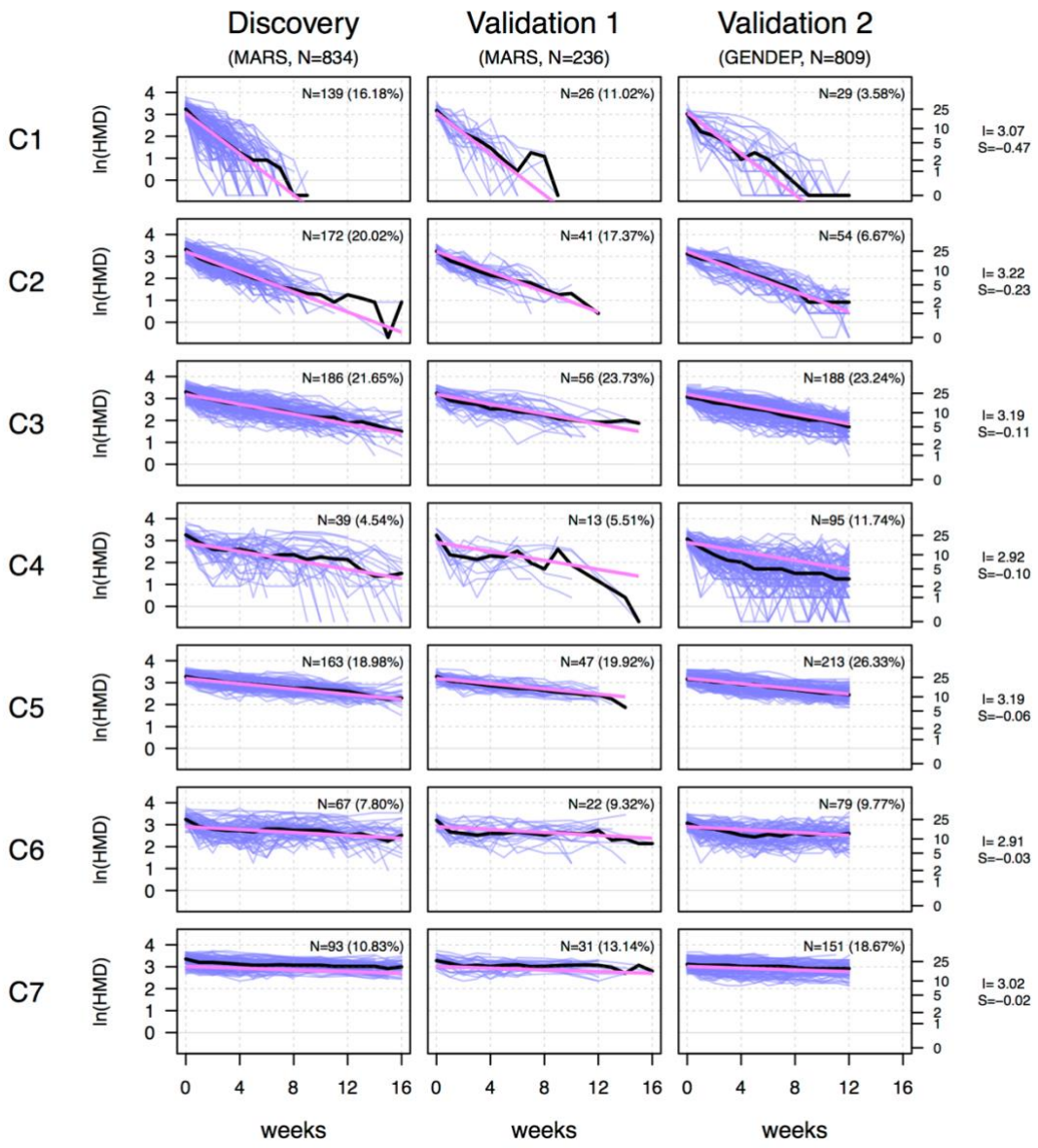


Figure 2.2: Resulting cluster shape characteristics and underlying natural logarithm-transformed HAM-D courses for the discovery sample and both validation samples[73]*.X axis:* observation time in weeks; *Y-axis:* natural logarithm-transformed HAM-D values (purple: raw values, black: cluster-specific median, pink: model-based linear fit). Slope and intercept values of all clusters are given on the right. Clusters are sorted from C1 to C7 according to the cluster-specific slope. Absolute and relative cluster sizes in all samples are given within the subplots. Green borders represent the limits in which 95% of HAM-D values of the discovery sample were contained. These were transferred to columns 2 and 3 to allow for comparison with the validation samples. *Abbreviations:* S, slope; I, intercept; ln, natural logarithm-transformed.

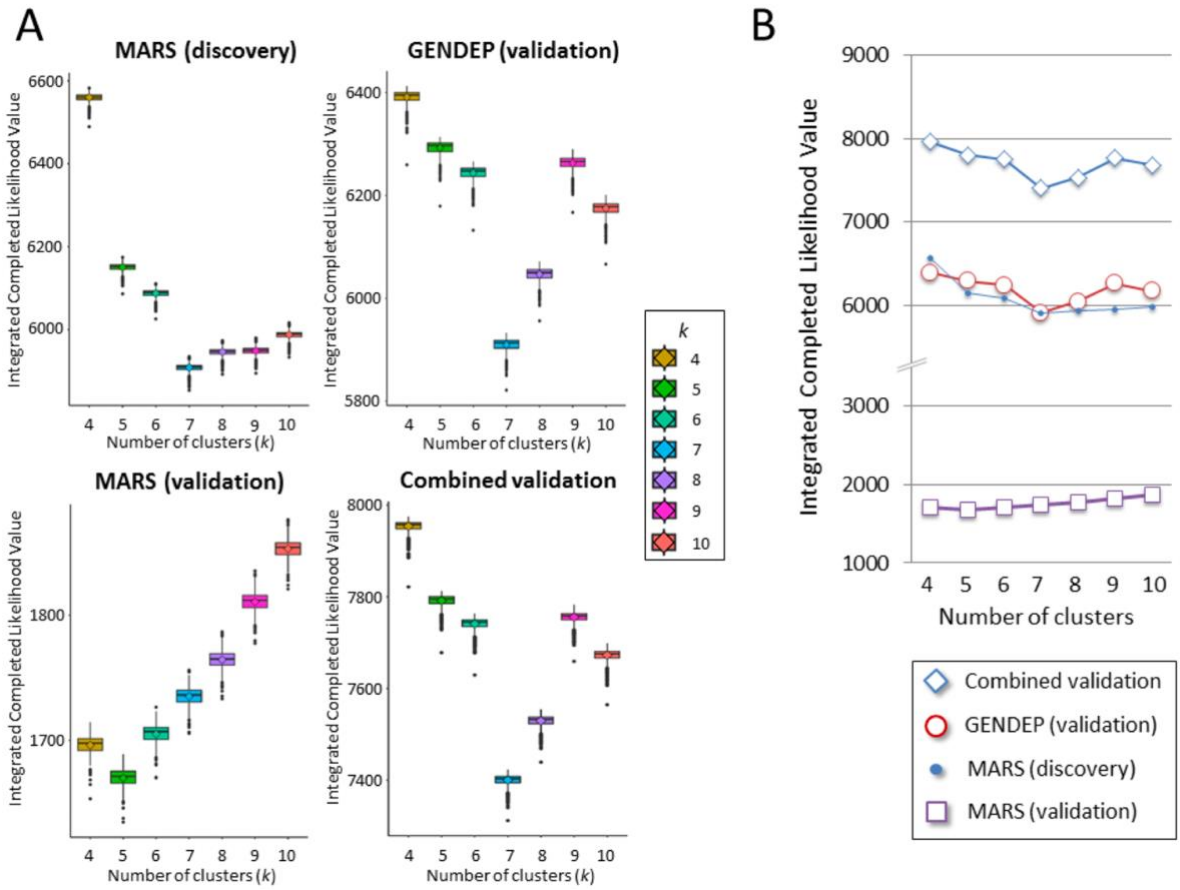


Figure 2.3: Integrated completed likelihood (ICL) values of the discovery sample and the validation samples. [73] (A) ICL values for clustering solutions between 4 and 10 clusters are plotted on scales adjusted to the respective result range. One data point represents one subject; open diamonds represent mean values; vertical lines represent one standard error of the mean; boxplots represent the median and quartiles. (B) For improved comparability, ICL mean values of the MARS discovery sample, the MARS validation sample, the GENDEP validation sample and the combined validation samples (MARS and GENDEP) are plotted on the same Y-axis.

The resulting TRCs (C1 to C7), sorted for their model-derived slopes, are shown in Figure 2.2. C1 displayed the most dramatic change in symptoms (fastest improvement), while C2 and C3 showed a decreased rate in symptom improvement. Cluster C4 represented a more dynamic and volatile pattern, while C5, C6 and C7 displayed a generally slow rate of improvement with C7 showing virtually no improvement over at least 16 weeks. As would be expected from these courses, the mean HAM-D baseline scores were significantly different from one cluster in ANOVA (p=0.009); the median HAM-D median of the episode differed considerably (ANOVA, $p=4.022\times10^{-116}$). The clustered slopes were weakly associated to HAM-D (r=0.09, p= 0.002) and to episode HAM-D (r=0.57, p=8.270) (**Table 2.3**).

Table 2.3: Baseline HAM-D and average HAM-D values per cluster (discovery sample) [73]

Cluster label	Baseline HAM-D [mean (SD)]	Average HAM-D across time series[b] [mean (SD)]	Comparison of neighboring clusters		
			Cluster pair	Baseline HAM-D [mean (SD)]	Average HAM-D across time series[a] [mean (SD)]

C1[a]	24.63 (6.08)	11.15 (3.74)	C1 *vs.* C2	0.006	2.10×10^{-9}
C2	26.35 (6.26)	14.15 (3.74)	C2 *vs.* C3	n. s.[d]	0.018
C3	26.10 (6.14)	15.47 (3.95)	C3 *vs.* C4	n. s.	0.001
C4	26.22 (6.65)	12.46 (3.53)	C4 *vs.* C5	n. s.	7.11×10^{-12}
C5	26.71 (4.86)	19.17 (4.07)	C5 *vs.* C6	n. s.	0.005
C6	25.10 (6.44)	16.23 (4.30)	C6 *vs.* C7	n. s.	4.25×10^{-8}
C7	27.04 (6.03)	21.80 (4.56)			
ANOVA (df=6)	$p=9.446\times10^{-3}$	$p=4.022\times10^{-116}$			
Linear correlation[c]	$r=0.09$, $p=2.497\times10^{-3}$	$r=0.57$, $p=8.271\times10^{-76}$			

[a] Sorting is by increasing cluster-derived slope, as in **Figure 2.2**.
[b] Average across all available HAM-D values of the time series until discharge.
[c] Correlation between cluster-derived slope and individual baseline HAM-D values (middle column) and individual average HAM-D across time (right column).
[d] n. s., not significant ($p>0.05$)

We have allocated clusters to patients of the two MARS and GENDEP validation samples, using coefficients of the model estimated in the discovery sample, to analyze whether TRCs represent stable and generalizable entities. **Figure 2.2** displays the respective cluster-specific median time courses along with boundaries that comprise 95 percent of the values for a sample discovery. **Figure 2.3B** displays different and combined ICL values for both samples of the validation analysis. The minimum ICL was observed for all samples, except the MARS validation study, for seven clusters. The latter showed a rather flat ICL profile, yet with a relative minimum at five clusters, most likely due to the relatively small sample size of about 30% compared with the MARS discovery and the GENDEP validation sample[73]. We also observed that the median HAM-D courses for the MARS validation sample were extremely similar to that of the discovery sample, with no different proportions detected for clusters ($X^2=6.157$, df=6, $p=0.40$). With an exception of C4, median of HAM-D trajectories were similar between discovery (MARS) and validation sample (GENDEP). This discrepancy in C4, appearing as lower average values compared to MARS, was triggered by some patients with high variability between week 4 and ~10 and HAM-D under the 95% threshold. The GENDEP clusters varied in proportion ($X^2=177.13$, $p=1.38\times10^{-35}$), displaying fewer fast responders (e.g. in C1, average 4.9 weeks to discharge) and higher number of slow responders (i.e. average of C7 20.8 weeks to discharge) compared with the samples of MARS discovery. Then, we evaluated to what extent a smaller number of sequential observations could surrogate for the full observation period. Here, we found that the correlation coefficient between the reduced and the cluster assignment based on the all variable data from week 0 to 4 was found to be almost linearly increasing. The MARS validation

and combination of MARS samples (0.96-0.98) were already strong at week eight (**Figure 2.4**).

The slope of GENDEP was typically below 0.77 at week 8 and was linear until its maximum.

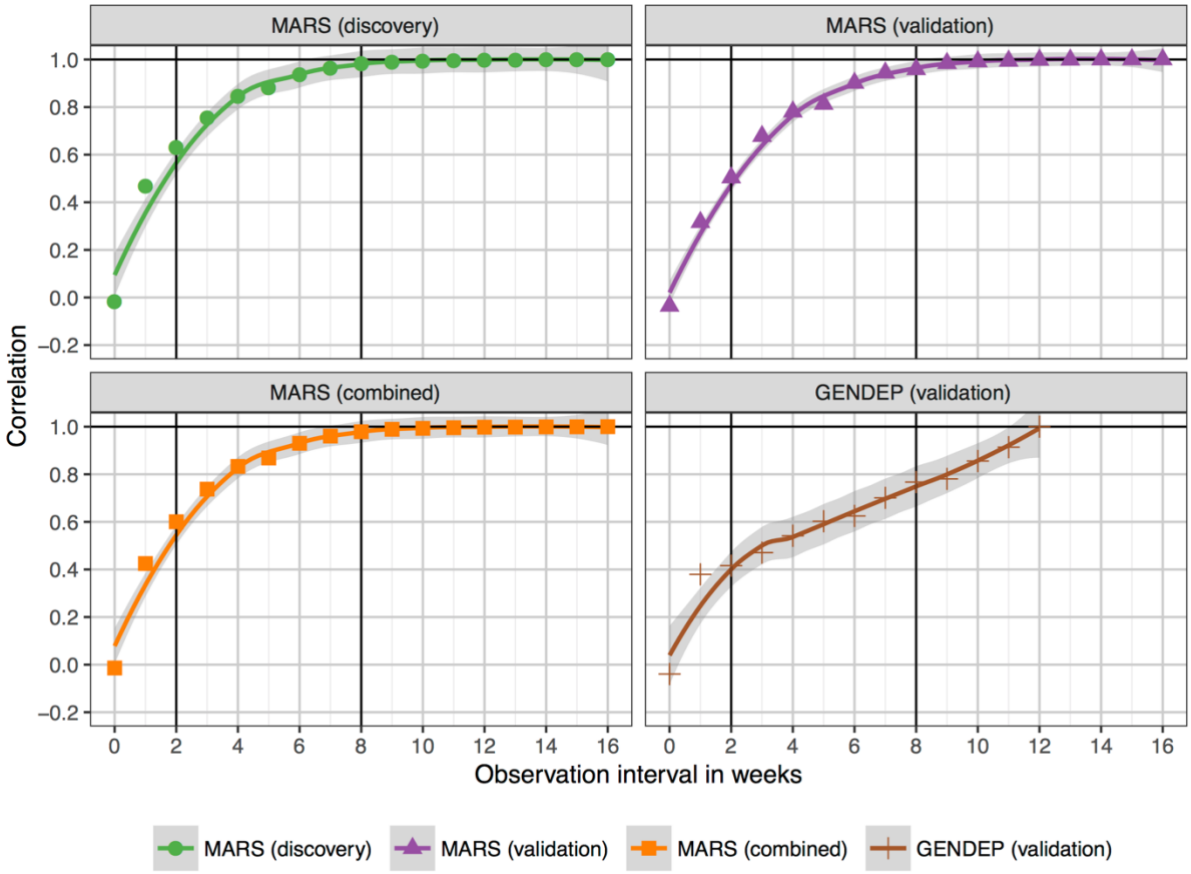


Figure 2.4: Prediction accuracy for reduced observation intervals. [73]Correlation of prediction result achieved from reduced observation intervals ranging from one observation (baseline HAM-D) to the full set of either 17 HAM-D values (baseline through week 16, for MARS derived samples) or 13 HAM-D values (baseline through week 12, for GENDEP sample). Pearson correlations were calculated between clusters predicted using the reduced and predicted with the full observation interval, using the model-based HAM-D slope of the respective cluster. Note that a positive linear correlation of ≈0.50 was reached at week 2 and a correlation of ≈0.96 (for the MARS samples) and ≈0.77 (for GENDEP) was reached at week 8.

Table 2.4: Association between TRCs, established response markers and psychopharmacological treatment in the combined MARS sample[73]

	Clinical item	ANOVA/ X^2 test[a] *p*-value	Cohen's *f*	Cohen's ω
Established response markers	Response (HAM-D reduction >50%) at discharge	4.16×10^{-65}	N/A	0.546
	Remission (HAM-D<10) at discharge	1.05×10^{-90}	N/A	0.609
	Weeks until discharge	4.59×10^{-84}	0.685	N/A
Psycho-pharmacological treatment[b]	Tricyclic antidepressants	1.75×10^{-5}*	0.177	N/A
	Selective serotonin reuptake inhibitors	0.071	0.104	N/A

	Selective serotonin and noradrenalin reuptake inhibitors	0.477	0.077	N/A
	Noradrenergic and specific serotonergic antidepressants	0.503	0.063	N/A
	Other antidepressants	0.012	0.126	N/A
	Antipsychotic medication	9.68×10^{-12}*	0.247	N/A
	Mood stabilizers	0.012	0.130	N/A
	Anxiolytic medication	1.44×10^{-7}*	0.197	N/A
	Sleep promoting medication	0.229	0.090	N/A

[a] X^2 test for categorical variables response and remission. ANOVA was applied if not otherwise specified, [b] Pharmacological treatment classes were binary coded (1:= applied/0 := not applied) every week and then aggregated across the entire hospitalization period, with "1" indicating treatment with the respective drug category during complete hospitalization, "0" indicating not applied at all, and values between "0" and "1" indicating the relative time under treatment with the respective drug category. Most patients were treated with several types of pharmacological treatments, *p-value robust towards Bonferroni correction for nine psycho-pharmacological classes.

Table 2.5: Comparison of established response markers between neighboring clusters in the MARS discovery sample [73]

	Cluster averages			Comparison of neighboring clusters			
Cluster	**Response** (>50% HAM-D reduction at discharge)	**Remission** (HAM-D<10 at discharge)	**Weeks until discharge**	**Cluster pair**	**Response** (>50% HAM-D reduction at discharge)	**Remission** (HAM-D<10 at discharge)	**Weeks until discharge**
C1	100.0%	100.0%	4.9	C1 *vs.* C2	n.s.[a]	7.38×10^{-6}	6.05×10^{-13}
C2	85.8%	98.7%	6.7	C2 *vs.* C3	1.78×10^{-5}	2.59×10^{-4}	4.99×10^{-18}
C3	76.1%	91.6%	11.3	C3 *vs.* C4	n. s.	9.13×10^{-3}	0.002
C4	88.2%	91.2%	14.0	C4 *vs.* C5	8.75×10^{-6}	1.58×10^{-16}	n. s.
C5	25.0%	57.5%	12.4	C5 *vs.* C6	n. s.	1.16×10^{-5}	4.19×10^{-4}
C6	59.3%	80.7%	16.8	C6 *vs.* C7	1.87×10^{-8}	1.22×10^{-9}	0.003
C7	12.9%	32.5%	20.8				

[a] n. s., not significant (p>0.05)

We then investigated if the TRCs would correlate with established response markers. Indeed, we found evidence for strong correlations of TRCs with such validated response markers (weeks before discharge, response [50% relative discharge symptom decrease], remission [HAM-D<10 discharge] Details on these results can be found in **Table 2.4**. These findings reached significance for about 80 per cent of the neighboring clusters, particularly for remission as a conservative criterion (**Table 2.5**). A further significant difference was found for three out of nine types of medication (benzodiazepines, tricyclic antidepressant and anti-psychotics) administered throughout the episode (**Table 2.4**)[73].

Predicting treatment response dynamic clusters from clinical characteristics

We evaluated whether clinical features could predict the attribution of patients to the TRCs. The analysis was exploratory and tested if the TRCs were associated with clinically plausible and previously reported markers, in the sense of a clinical cluster validation. In order to do this, four models were analyzed, yet we focused on model 0, comprising 50 baseline clinical components. HAM-D baseline items were included Model 1, and early partial response in Model 2. Model 3 represented both additions. All four models predicted treatment response in the combined MARS sample for both alternatives of modelling the slope (individual and cluster-derived) (both $p<2.17\times10^{-21}$, **Table 2.6**).

Table 2.6: Prediction characteristics of model 0 and the extended models 1-3 [73]

Model	Sample	Explained variance (Adjusted R^2)[a]		Overall model significance		Significance of the R^2 difference (p-value)[b]
		Individual	Cluster-derived	Individual	Cluster-derived	
Model 0	All	0.08	0.13	2.17×10^{-21}	1.53×10^{-33}	0.019
Model 0	Discovery	0.08	0.12	3.76×10^{-18}	1.54×10^{-24}	0.106
Model 0	Validation	0.06	0.19	8.71×10^{-5}	1.72×10^{-12}	0.009
Model 1	All	0.08	0.13	4.35×10^{-22}	1.49×10^{-34}	0.025
Model 1	Discovery	0.08	0.12	1.30×10^{-17}	2.06×10^{-24}	0.097
Model 1	Validation	0.10	0.20	7.35×10^{-7}	4.09×10^{-14}	0.047
Model 2	All	0.13	0.20	1.52×10^{-34}	3.42×10^{-54}	0.008
Model 2	Discovery	0.14	0.21	6.78×10^{-30}	8.43×10^{-45}	0.026
Model 2	Validation	0.07	0.20	3.64×10^{-5}	8.68×10^{-14}	0.008
Model 3	All	0.13	0.21	2.95×10^{-34}	1.53×10^{-57}	0.004
Model 3	Discovery	0.13	0.21	2.42×10^{-28}	1.71×10^{-46}	0.012
Model 3	Validation	0.11	0.21	2.76×10^{-7}	9.93×10^{-15}	0.050

[a] Adjusted R^2 coefficients indicate the explained variance and *p-value*s indicate the overall model significance,

[b] Based on Fisher's Z'-transformed *r* value

Two levels of performance (A and B) were observed, overall, for models with a cluster-derived slope: (A) Model 0 and 1, both of these explained 13% of this variance. The improvement over (A) was induced by adding the early part response element in Model 2 and Model 3 respectively, explaining the 20% and 21% of the variation; no additional effect for Model 3 that used the basic HAM-D items was noted as observed at the first comparison (A). Projections for both MARS subsample models ($p<1.30\times10^{-17}$ and $p<8.71\times10^{-5}$ for the discovery and validation sample, respectively) were also significant for all four models. It should be noted that the prediction analysis was entirely independent of the clustering procedure in the

MARS validation sample. On all models, the variance was significantly greater with the cluster-derived slope than with individual slopes (**Table 2.6**). Classification accuracies as calculated from cluster specific confusion matrices ranged between 75.0% and 95.2% (**Table 2.8** for details).

Table 2.7: Univariate comparison of significant predictors between TRCs (model 0, combined MARS samples) [73]

	index_d		scl_uncert		scl_psy		scl_pho		epq_neu		epq_ext		epq_psy		tpq_ha		wL-Event	
C1	22.27± 27.67	↓	1.04± 0.70	↓	0.63± 0.48	↓	0.64± 0.69	↓	5.17± 2.81	↓	6.42± 3.34	↑	1.91± 1.63	0	17.18± 6.23	↓	69.82± 33.76	↓
C2	28.46±5 4.70	↓	1.19± 0.70	↓	0.74± 0.46	↓	0.79± 0.65	↓	6.59± 2.52	↓	5.65± 3.04	↑	1.94± 1.13	0	19.81± 5.65	↓	77.92± 29.51	↓
C3	43.35±7 5.64	↑	1.41± 0.83	↑	0.84± 0.59	0	0.94± 0.79	0	7.43± 2.42	↑	4.95± 3.06	↓	2.00± 1.31	0	21.51± 4.86	↑	87.15± 44.22	↑
C4	17.23±1 4.19	↓	1.02± 0.86	↓	0.67± 0.52	↓	0.63± 0.71	↓	5.92± 2.81	↓	5.56± 3.08	↓	2.35± 1.44	↑	18.15± 6.55	↓	29.00± 12.94	↓
C5	38.67±6 0.82	↑	1.48± 0.76	↑	0.93± 0.50	↑	1.02± 0.75	↑	7.47± 2.19	↑	4.52± 2.39	↓	1.94± 1.10	0	21.92± 5.05	↑	86.46± 39.29	↑
C6	30.97±3 4.81	0	1.27± 0.74	0	0.86± 0.58	↑	0.94± 0.84	0	7.00± 2.05	0	4.76± 2.88	↓	2.10 ±1.27	↑	21.29± 4.73	↑	90.72± 45.12	↑
C7	43.46±6 3.86	↑	1.53± 0.79	↑	1.03± 0.63	↑	1.18± 0.86	↑	7.74± 2.02	↑	4.26± 2.39	↓	1.87± 1.12	↓	22.81± 4.90	↑	94.96± 50.82	↑
95% CI[b]	30.59; 37.46		1.26; 1.35		0.79; 0.85		0.85; 0.94		6.70; 7.00		4.99; 5.35		1.90; 2.05		20.21; 20.89		80.74;85.65	
Multivariate importance *p-value*	0.0210		0.0073		0.0208		0.0182		<0.0001		<0.0001		0.0445		<0.0001		0.0002	
ANOVA *p-value* (Cohen's *f* [c])	4.0×10^{-4} (0.153)		2.5×10^{-10} (0.233)		1.9×10^{-10} (0.234)		9.8×10^{-10} (0.227)		7.1×10^{-25} (0.355)		2.9×10^{-11} (0.243)		3.3×10^{-1} (0.081)		6.7×10^{-23} (0.341)		4.4×10^{-7} (0.196)	

[a] index_d: duration of the current episode; scl_uncert: uncertainty in social contact (SCL-90R); scl_psy: psychotism (SCL-90R); scl_pho: phobic anxiety (SCL-90R), epq_neu: neuroticism (EPQ-RK), epq_ext: extraversion (EPQ-RK), epq_psy: psychoticism (EPQ-RK), tpq_ha: harm avoidance total (TPQ), wL-Event: stress-weighted sum of life events. See **Table 2.1** for more details on the clinical items, [b] CI: confidence interval. Arrows indicate lower (↓), higher (↑), or within (0) positioning regarding the 95% CI of the respective parameter distribution, [c] Cohen's *f*: >0.10 and <0.25: small effect; ≥0.25 and <0.40, medium effect; ≥0.40: large effects

Table 2.8: Overview of classification accuracy (%) per class for all models.[73]

Model	Sample	C1[a]	C2	C3	C4	C5	C6	C7
Model 0	All	84.4	80.4	78.3	95.1	81.5	91.7	88.2
	Discovery	83.0	79.7	78.9	95.2	81,8	92.0	88.6
	Replication	89.0	82.2	75.4	94.5	79.7	90.7	86.9
Model 1	All	84.6	80.4	78.3	95.1	81.4	91.7	88.4
	Discovery	83.2	79.7	78.8	95.2	81.9	92.0	88.8
	Replication	88.6	82.6	75.0	94.5	79.7	90.7	86.9
Model 2	**All**	84.7	80.4	78.3	95.1	81.5	91.7	88.5
	Discovery	83.2	79.7	78.9	95.2	81.8	92.0	88.7
	Replication	89.0	82.2	75.0	94.5	79.7	90.7	86.9
Model 3	All	84.6	80.4	78.3	95.1	81.4	91.7	88.5
	Discovery	83.8	79.7	78.8	95.2	81.9	92.0	89.0
	Replication	88.6	82.6	75.4	94.5	80.1	90.7	86.9

[a] Classification accuracy defined as [true positives + true negatives]/[true positives + false positives + true negatives + false negatives])

Table 2.7 lists the 9 (out of 50) most significant predictors of model 0. For this ranking, we used a multivariate comparison of each single item with all other competing items[113]. The univariate associations of these items with TRCs (likelihood ratio test on a generalized linear model) were also analyzed. Both types of comparison therefore revealed the strongest effects for personality items neuroticism, extraversion and harm avoidance. In addition, we examined the cluster-specific averages of each clinical element compared with the 95 % confidence interval (CI) of the sample as a whole (**Table 2.7**): Fast improvement clusters (C1 and C2) showed lower average values for all predictors except for the personality trait of extraversion. The treatment resistance group C7 showed higher than average values, except for extraversion and psychotic personality elements. More generally, except for *extraversion*, there was a tendency that lower clinical scores (i.e., a shorter *index episode*, less *SCL-90R symptoms*, fewer *weighted life events*, and lower scores for the personality items *neuroticism* and *harm avoidance*) were found in clusters with good treatment response, and higher scores in clusters C6 or C7. Deviations from this pattern could point to nonlinear relationships or complex interactions, most obvious in the intermediate clusters C3 to C5 (see for example, weighted life events). The random forest algorithm had no demographic variables selected. We still did not overlook demographic that could influence the clustering, so we compared these between the clusters, especially of the MARS discovery sample, finding no significance differences. Age

was only significant in the GENDEP sample (**Table 2.9**). The three extended models are summarized in **Table 2.10** by significant predictors. In short, model 1 was characterized by priority setting for three basic HAM-D single items in comparison to model 0, model 2 was identified as a strong predictor, as anticipated, early partial response, along with smaller other shifts. Model 3 delivered a combined pattern with a basic HAM-D single item (somatic symptoms), early partial response and current psychotic symptoms as additional predictors.

Demographic variables	Differences between clusters (*p*-value)			MARS discovery & validation	All 3 samples
	MARS discovery	MARS validation	GENDEP		
Age[a]	0.135	0.251	3.463×10^{-05}	0.150 (C)[c] 0.006 (S)[d] 0.223 (C×S)[e]	0.008 (C) 1.268×10^{-21} (S) 0.048 (C×S)
Gender[b]	0.276	0.298	0.005	0.999 (C) 0.981 (S) 0.975 (C×S)	0.999 (C) 0.002 (S) 0.247 (C×S)
Spouse[b]	0.222	0.684	0.890	0.999 (C) 0.007 (S) 0.999 (C×S)	0.999 (C) 1.270×10^{-5} (S) 0.984 (C×S)
Education[b]	0.098	0.071	0.127	0.999 (C) 0.958 (S) 0.999 (C×S)	0.999 (C) 1.571×10^{-21} (S) 0.837 (C×S)
Employment[b]	0.404	0.788	0.062	0.999 (C) 0.986 (S) 0.999 (C×S)	0.999 (C) 1.390×10^{-5} (S) 0.903 (C×S)
Training/ Retirement[b]	0.365	0.022	0.915	0.999 (C) 0.635 (S) 0.775 (C×S)	0.999 (C) 0.015 (S) 0.460 (C×S)

Table 2.9: Demographic variables compared across clusters and samples[73]

[a] ANOVA used for continuous variables.

[b] Chi square statistics for categorical values.

[c] (C) stands for *p*-values for main effect of cluster.

[d] (S) stands for *p*-values for main effect of sample.

[e] (C×S) stands for *p*-values for the cluster-by-sample effect.

Table 2.10: Overview of significant predictor variables of all models (combined MARS sample) [114]

Category	Clinical item[a]	Model 0 *p*-value[bc]	Model 1 *p*-value	Model 2 *p*-value	Model 3 *p*-value
Personality items	epq_neu	<0.0001	<0.0001	<0.0001	<0.0001
	epq_psy	0.0444	0.0433	0.0192	0.0308

	epq_ext	<0.0001	<0.0001	<0.0001	<0.0001
	tpq_ha	<0.0001	<0.0001	<0.0001	<0.0001
Life events	L-Event	n.s	0.0220	n. s.	n. s.
	wL-Event	0.0002	<0.0001	0.0008	0.0002
Baseline psycho-pathology	scl_comp	n. s.	0.0304	n. s.	n. s.
	scl_uncert	0.0073	0.0020	0.0215	0.0099
	scl_pho	0.0207	0.0156	n.s	0.0327
	scl_psy	0.0182	0.0155	n. s.	0.0308
	HAM-D0_13[d]	not included	0.0327	not included	0.0045
	HAM-D0_14[d]		0.0258		n. s.
	HAM-D0_16[d]		0.0419		n.s
Information on current episode	index_d	0.0210	0.0186	n. s.	n. s.
	psychot_ current	n. s.	n. s.	0.0466	0.0492
Early response rated at week 2	HD_2WE	not included	not included	<0.0001	<0.0001

[a] epq_neu: neuroticism (EPQ-RK); epq_psy: psychoticism (EPQ-RK); epq_ext: extraversion (EPQ-RK); tpq_ha: harm avoidance total (TPQ); L-Event: sum of life events; wL-Event: sum of weighted life events; scl_comp: compulsiveness (SCL-90R); scl_uncert: uncertainty in social contact (SCL-90R); scl_pho: phobic anxiety (SCL-90R); scl_psy: psychotism (SCL-90R); index_d: duration of current episode; psychot_current: Psychotic symptoms during the current episode; HD_2WE: HAM-D early partial response (≥25% reduction) after 2 weeks. Categories/items that delivered no predictors at $p<0.05$ are not listed in the table.

[b] *p*-values are based on testing the respective single importance value against all other competing predictors (see methods for details).

[c] *p*-values for model 0 are identical with the *p*-values reported in **Table 2.6** and listed for direct comparison with the other models.

[d] baseline item HAM-D items: 13: somatic symptoms – general, 14: genital symptoms, 16: weight loss

Abbreviations: n. s., not significant (p>0.05)

We have performed Item clustering using model based approach and these item clusters can be used for detecting symptom based treatment response subtypes. We have investigated both types, HAMD-21 items and HAMD-17 items to detect robust item clusters. For cluster stability and to select optimal number of cluster solution we have also used the same ICL criterion with jackknife approach (1000 repetitions). We have detected 4 stable clusters with lowest ICL value for both HAMD-21 items and HAMD-17 items (**Figure 2.5 and 2.6**). We have performed jackknifing and calculated co-occurrence matrix for training (N=506) and test dataset (N=506) (**Figure 2.7 and 2.8**)

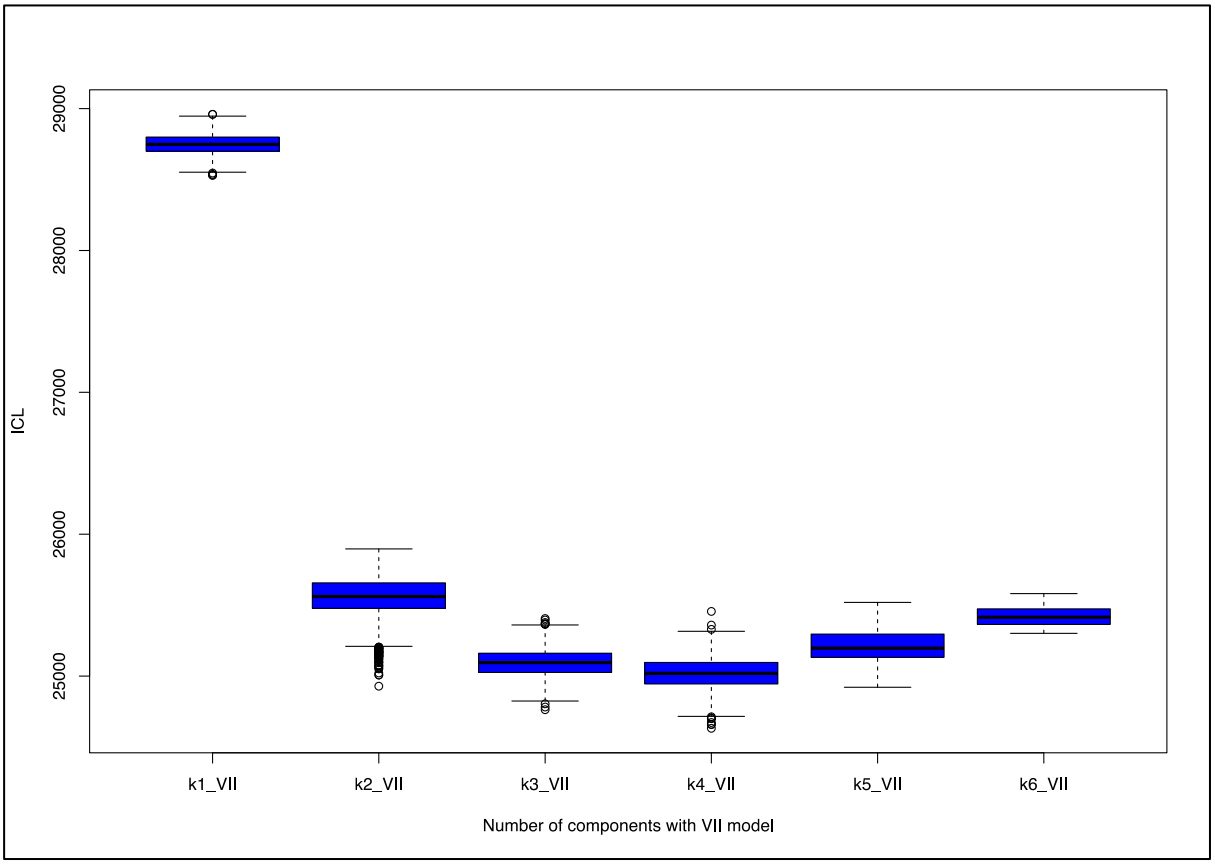


Figure 2.5: Overview of ICL value for all model using HAMD-single items train data set (N= 506) (baseline 17 items) and 1000 repetitions (jackknife). Here we have used the best model "VII" based on 8B and detected lowest ICL value for K=4.

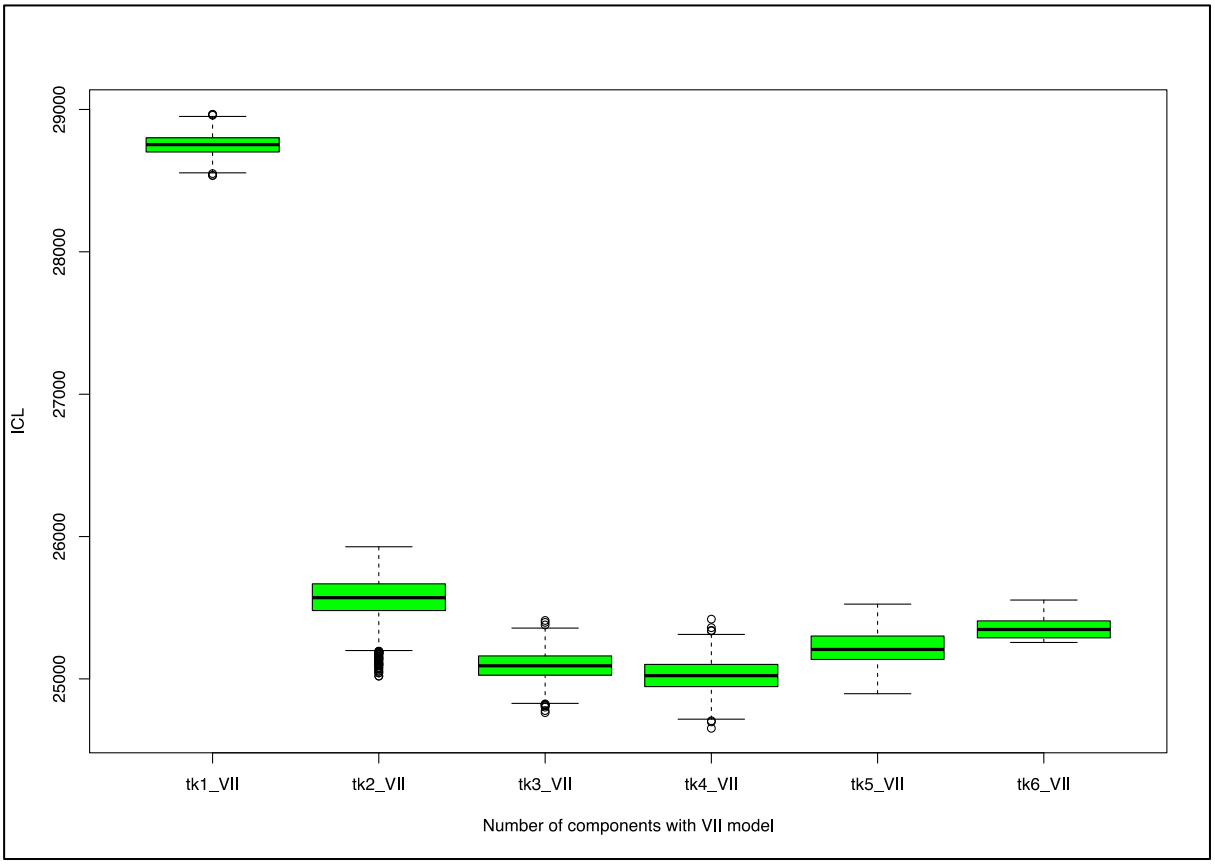


Figure 2.6: Overview of ICL value for all model using HAMD-single items test data set (N= 506) (baseline 17 items) and 1000 repetitions (jackknife). Here we have used the best model "VII" based on 8B and also detected lowest ICL value for K=4 (same as **Figure 2.5**)

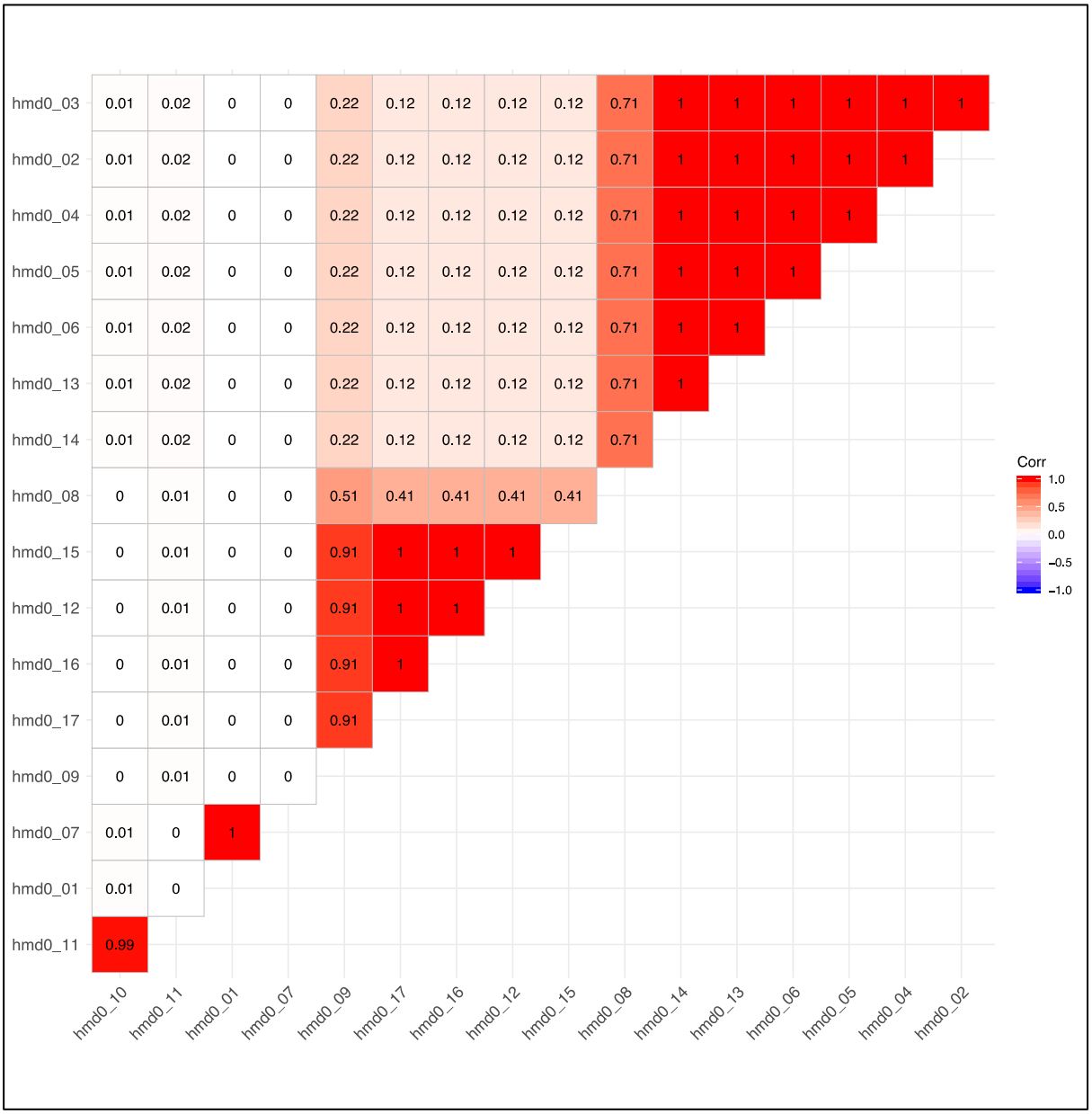


Figure 2.7: Overview of co-occurrence matrix using HAMD-single items train data set (N= 535) (baseline 17 items) and 1000 repetitions (jackknife) using the best model (k=4). Here we could visualize the item clusters which are based on number of (%) co-occurrence of items.

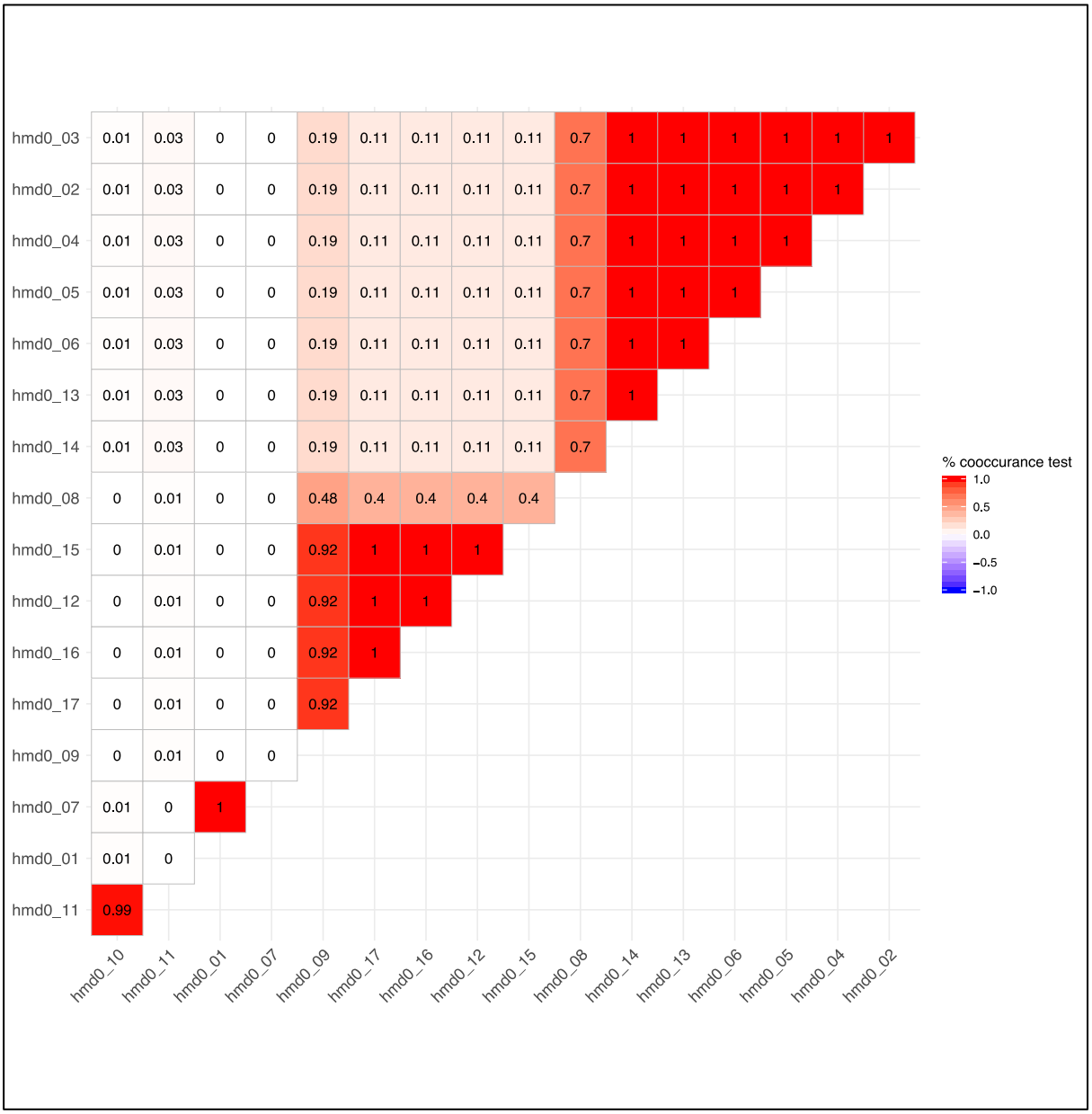


Figure 2.8: Overview of co-occurrence matrix using HAMD-single items test data set (N= 535) (baseline 17 items) and 1000 repetitions (jackknife) using the best model (k=4). Here we could visualize the item clusters which are based on number of (%) co-occurrence of items.

We have also investigated this four group of items for checking the reproducibility between train and test dataset for HAMD-21 items and HAMD-17 items. For example, we have found 4 clusters using HAMD 17 items: a) cluster 1 contains item related to depressed mood work and interest; b) cluster 2 contains items related somatic symptom general, genital symptoms, feeling guilt, suicide, insomnia-initial, insomnia-middle, insomnia-delayed; c) cluster 3 contains agitation, hypochondriasis, weight loss, insight; somatic symptoms gastrointestinal; d) cluster 4 contains anxiety psychic and somatic. We will briefly discuss about it in the **section 4** (General Discussion).

Table 2.11: Overview of single item clustering using HAMD-21 items for train and test datasets. Here we found four item clusters which has mismatch in the cluster membership between train dataset and test dataset.

Train Cluster-Using 21 items (N=535)	Test Cluster-Using 21 Items (N=535)	item
1	1	1 Depressed Mood
3	3	2 Feelings of Guilt
3	3	3 Suicide
3	3	4 Insomnia - Initial
3	3	5 Insomnia - Middle
3	3	6 Insomnia - Delayed
1	1	7 Work and Interests
3	3	8 Retardation
4	4	9 Agitation
2	2	10 Anxiety - Psychic
2	2	11 Anxiety - Somatic
4	4	12 Somatic Symptoms - Gastrointestinal
2	2	13 General Somatic Symtoms
2	2	14 Genital Symptoms
4	4	15 Hypochondriasis
4	4	16 Weight Loss
4	4	17 Insight
2	2	18 Diurnal Variation
4	4	19 Depersonalization and Derealization
4	4	20 Paranoid symptoms
4	4	21 Obsessional Symptoms

Table 2.12: Overview of single item clustering using HAMD-17 items. Here we found four item clusters with no mismatch in the cluster membership between train dataset and test dataset

Train Cluster-Using 17 Items (N=535)	Test Cluster-Using 17 Items (N=535)	item
1	1	1 Depressed Mood
2	2	2 Feelings of Guilt
2	2	3 Suicide
2	2	4 Insomnia - Initial
2	2	5 Insomnia - Middle
2	2	6 Insomnia - Delayed
1	1	7 Work and Interests
3	3	8 Retardation
3	3	9 Agitation
4	4	10 Anxiety - Psychic

4	4	11 Anxiety - Somatic
3	3	12 Somatic Symptoms - Gastrointestinal
2	2	13 Somatic Symptoms - General
2	2	14 Genital Symptoms
3	3	15 Hypochondriasis
3	3	16 Weight Loss
3	3	17 Insight

2.6 Discussion

We have used nonlinear, model based clustering[104,105] applied to symptom courses of 834 hospitalized patients and detected seven different TRCs. These classes were already visually distinct, ranged from rapidly, and unambiguously fast response to severe resistance to treatment. The average HAM-D decrease was significantly different between classes and classes correlated with establishing response markers, emphasizing that they reflect clinically important effects. Baseline severity was only weakly associated with the response slope over a small range of HAM-Ds, contradicting the intuition that a high initial disease severity is closely linked to a steep decrease in symptoms. The classification of 236 MARS patients and 809 GENDEP samples shows that these clusters can capture the response dynamics of the patients while at the same reflecting specific differences between the response profiles in the study[64].

Robustness and Reproducibility of the clustering solution

We detected similar cluster size and shape characteristics after projecting the discovery clustering model onto validation sample. (**Figure 2.2**). The consistency observed in this validation is superior to previous latent variables analyses, which neither used any machine learning nor produce stable symptom-based subtypes of depression.[11] However, one of the major differences that limits the comparability of these tests is that they were based on a cross-sectional symptom spectrum instead of on the trajectory of symptom changes, namely factors analyses, principal component analyses, and latent class analyses.

Here, we have applied a machine learning algorithms to identify MDD subtypes based on data gathered longitudinally over 16 weeks. Our findings show that the MARS cohort does indeed contain significant latent subtypes for MDD. One advantage of our approach might be the identification of the best model by using ICL criterion, which is more robust to the violation of some of the mixture model assumptions compared with the commonly used Bayesian

information criterion. The use of the ICL may therefore have led to an improved choice for the number of clusters and therefore to a more sensitive data partitioning [106,115]. In combination with the jackknife it allowed significance testing of the cluster solutions within themselves without recurrence to outside data, i.e. from another domain.

We have made another important observation: The use of slopes from the linear mixed model that characterizes each TRC led, in each model, to higher R^2 coefficients than the application of individual slopes, in particular in the validation sample (Table 2.6). This observation reinforces the validity of the detected response classes and emphasizes that the clustering algorithm actually assessed the individual information from HAM-D time courses. In addition, it stresses that the average slope of the class is a good approximation of response behavior, which contributes to denoising individual observations[73].

Clustering and Simulation of independent patient groups using reduced observation intervals

We analyzed two aspects to facilitate the translation of our clustering system into other cohorts and to understand the generality of our clustering solution. First, the clustering coefficients of the discovery sample were applied to an independent MARS subsample, and we found that the group plots with median HAM-D courses were classified into the same shapes as found in the discovery sample. The observation that the classes formed from the MARS validation sample were also similarly proportioned as in the discovery sample confirmed that a stable solution was found within the MARS cohort. The further projection to the GENDEP sample was also informative: With the exception of a small number of patients that exceeded the lower HAM-D boundaries of one (discovery) cluster the algorithm captured patients equally into the seven TRCs. However, the cluster proportions were different, and significantly more slow or non-responders were found than in the MARS cohort. This may be a hint towards the limited possibilities for changing and intensifying the treatment in GENDEP, which is inherent to its design. Also, generally different patient characteristics may be at the root of these proportion differences. The combination of these two observations leads us to conclude that the seven TRCs actually do reflect generalizable response patterns.

Different criteria for cluster stabilization might have led to different solutions, such as the longitudinal latent class analysis using the Bayesian Information Criterion identified nine clusters in GENDEP[94]. Comparability with our solution, however, is hindered by the use of a different scale of depression (Montgomery-Asberg Depression Rating *Scale* (*MADRS*)[116].

Secondly, during a simulation we reduced the observation period to see if the current clustering solution could also be of benefit to studies with shorter observation windows. In MARS subjects, we observed a correlation (Pearson's r) of 0.96, and r was 0.77 in the independent GENDEP study after eight weeks of HAM-D measurements (**Figure 2.4**). The rest of the improved prediction accuracy between weeks 8 and 12 was stronger in GENDEP. This indicated that eight week observations usually appear to be adequate, but the changes in sample characteristics expected to play a role, suggesting more observations recommended. The increased flexibility in the MARS study to repeatedly adjust treatment for the individual patient could be one reason for the difference between week 8 and 12. In general, we hypothesize that for observational studies the generalizability of our clustering solution could be more robust compared with controlled studies.[73]

Prediction of treatment response classes from clinical baseline features

The clinical usefulness of TRCs were further investigated by testing whether clinical basic characteristics in a multivariate model (random forest algorithm) can be predicted[117]. We did not conceptualize this analysis as a separate study and as an additional clinical validation of clusters that primarily represent statistical, data-driven entities. Several machine learning techniques had previously been used to predict treatment response in MDD[118–120]. Their models were still mostly trained for classic remission categories non-remission[80], treatment resistance[121], or persistence seriousness[83]. Briefly, 50 clinical baseline variables were reported to predict about 13 percent of TRC variance. While seemingly low, this is actually in the range of previous multivariate analyses that focused on the prediction of two outcome categories, reporting low to medium accuracy values from receiver operating characteristics analyses. These variable has been obtained through interviews, self-reporting of symptoms and standard physical or laboratory tests. This proportion of explained variance is apparently low, but it is in the range of previously reported multivariate analyses focused on the prediction of two outcome categories, reporting low to medium accuracy values from receiver functionality analyzes. [80,83,121]As an example for HAM-D measures, the clustering of these categories will show more fine grained and yet sparse and data-driven classification structures, as opposed to the use of predetermined cut-off criteria for these categories. Of our clinical predictors, nine have significantly higher weights than others: (i) duration of the index episode; (ii–iv) symptom checklist items psychosocial self-assurance; psychotic and phobic anxiety; (v–viii) personality traits: neuroticism, extraversion, psychoticism and harm avoidance, and, (ix), sum scores for life events (weighted for their straining impact). Although all items support the overall prediction, a review of these nine items strengthened the clinical validity in several ways:

Before starting antidepressant treatment, a longer period spent in depression has previously been identified as a negative predictor of outcome[122]. By contrast, for the current total episode, no consistent predictive value including periods with and without treatment has been determined[123,124]. As we did not quantified the time without medication in our cohort, we speculated that our current marker for the duration of the episode incorporated untreated period and gained significance through large statistical power. In addition, baseline symptom profiles contributed significantly to the model. Several reports highlighted that strong symptoms of anxiety during a depressive episode increase the risk of not achieving remission[125]. At least two out of the predictive symptom items (phobic anxiety, psycho-social self-assuredness, and psychoticism) reflected the aspects of anxiety, which were corroborated by high anxiety in MDD.

In a structural MRI analysis comparing MDD patients with high level of anxiety to patients with low level of anxiety, brain areas were retrieved that are involved in the processing of social issues[126] and that overlap with areas that predicted treatment response over six weeks in a MARS subsample[127] While the symptom checklists includes current disease state, personality questionnaires are designed to described a person's enduring personality traits. From the latter category, predictors such as harm avoidance and neuroticism were identified, both of which represent similar concepts of anxiety feelings and the fearful avoidance of new or upcoming challenges. This association has been reported before[128,129], and detecting this correlation with the TRCs represents an indirect validation of these data-driven entities. Extraversion is another personality trait: it has so far been found to mainly *protect* against chronic stress clinical symptoms[130]. We report a more clearly defined impact on treatment response, which is potentially facilitated by the random approach to forests that integrates multiple interactions. As reported[131,132], life events, especially early adverse events, represent episodes of long-term adjustment, stress and liability that increase the risks of MDD but also influence the chances of recovery[122–124,133,134]. Early childhood adversity information was only provided for one subsample (35%), and so analyses on this marker could not be performed. There is an expectation though, that having such childhood adversity markers, including the exact time point of their occurrence, could further improve prediction models.

In a previous representative MARS sample[61], previous treatment resistance was mentioned as powerful univariate biomarker for non-remission. Treatment resistance is usually defined as at least two unsuccessful trials with different antidepressants at sufficient doses at a minimum of six weeks[68]. In this study, the antidepressant treatment response questionnaire (ATRQ) encoded therapeutic resistance but showed no significant p-value significance (still a significant

univariate association [data not shown]). ATRQ based results could vary since the measure tends to underreport unsuccessful tests[135]. Likewise, in our study the BMI was not linked to TRCs. The BMI has previously been reported to be linked to remission rates [61] and treatment response[63]. One explanation for this is that the positive report[63] used a binary cut (25 kg/m^2) which could indicate a non-linear relation. Note that the number of previous episodes of depression – a lifetime burden surrogate – was also not predictive, confirming other negative reports [68].

Similarly, age at onset (AAO), was not predictive. AAO is often inversely correlated with the number of episodes. Regarding the AAO marker, we found a mixed report. We found some findings with no correlation[136,137] and some findings with an influence on the speed of the remission[138] or treatment resistance[139]. This variability can be explained through cached AAO's interactions with subgroups (as for co-morbid alcohol dependence)[140]. Cortisol was also not a predictive basis as a simple HPA-axis marker; stimulation tests are probably more sensitive, especially when longitudinally obtained[65] and as reported before. Due to the observational study design, however, the here detected TRCs may indirectly reflect either disease acuity (anxiolytic drugs), treatment escalations (i.e. tricyclic anti-depressants), or episode severity (antipsychotic drugs). TRCs indeed differed between the types of psychopharmacological treatment **(Table 2.4).** For biological markers, like in meta-analyses of brain structure, similar co-correlations between drug variables and disease severity have been reported[141,142]

We have further investigated two different strategies to improve our basic model 0 (**Table 2.1**), either by adding single baseline HAM-D items or by adding early response information after two weeks. It has not been possible to improve the model by inclusion of single baseline HAM-D elements. This may be due the current symptomatology having already represented in the symptom check list (SCL). This does, however, not mean that primary clustering of trajectories of a single item cannot lead to new results or different clusters. This conceptual change would increase the number of observations per case and may result in model instability. We see this as a worthwhile follow-up project that also breaks up the sum score and by this adds clinical elaboration. In all cases, including the early response (2 weeks) markers, improved the model which confirmed similar reports of observational and controlled studies[82,85,118–120,133,134]**(Table 2.7)**. There are several limitations of this study. First, important clinical variables such as neurocognitive results, more complex endocrine tests or neuroimaging markers, were not included, despite reporting that a trade-off needs to be made between higher statistical power through the use of large samples and the use of powerful, specific single predictors[65,143]. Second, in MARS there was no formalized evaluation of previous, non-pharmacological treatments, including psychotherapy, which prevented to probe the predictive value of these factors, even

if the psycho-pharmacological treatments were well-documented. Third, the MARS discovery and validation samples differed significantly in seven baseline items (**age at study inclusion (years), any suicide attempt before current episode, psychotic symptoms in any previous episode, suicide attempt during current episode, psychotic symptoms during the current episode, thyroid stimulating hormone level, Free T4 hormone level**) which could explain minor prediction results differences. However, these seven items did not overlap with the most informative predictors of model 0 or predictors emerging from the other models.

2.7 Conclusion

In summary, we have detected seven distinct classes of treatment response that are stable in two validation samples using model-based nonlinear clustering at clinical scores of a large cohort of MDD patients. In a multi-dimensional prediction analysis, 50 clinical variables with personality items, life events and the duration of the episode were predictors of these classes, with a special weight on baseline psychopathological characteristics. The construct and clinical validity of the here reported MDD (acute) treatment response classes support that their neurobiological basis (e. g. genetic underpinnings, imaging correlates) should be studied in more detail. In addition, the clustering system imprinted in the clustering coefficients may be useful to project other studies with HAM-D data into the same space.

3 Clustering of Source-Based Morphometry (SBM): Atlas Parcellation

3.1 Introduction

Structural Covariance Analysis (SCA) can detect anatomical structural patterns across different subjects. It is a general concept theoretically applicable to all structural brain measures that are systematically indexed, or, in the most straightforward form, aligned ('co-registered') in a stereotactic space, or more generally, in an anatomical space along with an anatomical labeling system. One system used in the mapping of brain features is Voxel-based morphometry (VBM): here, corresponding data points (volume elements = voxels) are registered in a standard space, mostly following segmentation steps[31,144]. Other than functional connectivity analysis, which is based on signal time-series analyses, SCA is not performed in individual data sets but across a group of individuals: thus, the concept of the connectivity analysis can be defined in the *subject* rather than the *time* dimension. This makes SCA particularly interesting for large cross-sectional samples. For example, 'seed analysis' was employed in an earlier VBM framework that measured gray matter density (GMD)[24] (note: GMD is similar to voxel-wise GM volume), enabling to reveal robust and symmetric volume characteristics of homotopic brain regions. Later work demonstrated a vital analogy between resting-state functional connectivity networks [145,146] and resting-state functional connectivity density hubs[147], MRI co-activation maps [148] and individually operated, morphologically operating fiber-based connectivity networks[149]. VBM-based SCA also includes the segmentation of smaller brain structures (e.g., hippocampus) using clustering [150] of voxels with similar covariance structures. GM voxels of the hippocampus that have a similar covariance with all other GM voxels of the brain are classified into one class. Critical processes like aging [35] seem to follow such SCA networks, and also pathological entities seem to follow these networks[151]. The analogies between functional connectivity and anatomical connectivity have often been reported and could reflect the 'firing together, wiring together' principle [152]. Neurons frequently involved in synaptic signal transmissions change their dendritic morphology up to a measurable extent through MRI morphometry. Aging may play a somewhat minimal role intrinsically, merely modifying existing neuroanatomy differences that reflect a combination of genetic factors and environmental influences[153]. Forest et al. [154]investigated that connectivity at some of the seed regions induces essential effects on their connected targets and that these effects are reflected in gene expression. Source-based morphometry (SBM) is an analysis strategy that combines VBM data (mainly of grey matter) with an independent component analysis (ICA)[155]. This combination allows the separation and

mapping of various 'sources' (or components) of voxels linked through a representative cohort by similar behavior., Source-Based Morphometry (SBM) has been applied compared with regulatory cohorts[156] to identify and compare gray-matter networks between disease groups and healthy subjects and patients, e.g., with schizophrenia. Guo et al. (2015) reported 6-7 SBM-based networks, including the posterior default mode net of 82 and 109 healthy young adults and their visual and auditory networks[157]. However, it was not systematically evaluated how many independent components (non-overlapping) exist that are stable and generalizable. Imaging data-driven parceling of brain atlas data as such is not new, and for this purpose, there have been several modalities and principles[158]. Naturally, these data-driven atlases challenge the principle that anatomic gold standard (AGS) methods are considered to define boundaries between functional units by cytoarchitectural, histochemical, gene expression, or post mortem fiber tracking experiments. On the other hand, macroscopy has generated parcellation that is widely used for human brain mapping (REF), in particular gyral and sulcal morphology. A grey matter VBM-based parcellation with its mesoscopic measurement scale is inter-positioned between a bottom-up 'microscopic/molecular' and a top-down 'topological' approach, as connections of both levels with VBM data have been demonstrated before[150].

In this study, we address the question of dimensionality (i.e., the number of 'true' components in the data) in a set of 563 healthy subjects with 3 Tesla T1 weighted images (T1WI). Before the study, pilot experiments with ICA as implemented in FSL (MELODIC) and VBM gray matter information had been performed, and plausible patterns were detected. The following steps are taken in this work:

First, we use a combination of the iterative ICA (ICASSO technique [159]), (agglomerative) hierarchical clustering of high-dimension ICA followed by re-aggregation and analyses of the similarity between aggregated solutions to estimate this.

Second, we present methods for building a parcellation that considers component stability and from which binary (discrete versions) and fuzzy border versions of an atlas can be generated.

Third, we repeat the whole procedure in an independent and similarly sized second data set for investigating the generalizability, again determining the dimensionality of the resulting parcellation and their spatial similarity with the original data set.

Fourth, we look at two methodological questions in the specific VBM context: the effect of spatial smoothing on the dimensionality of the ICA and removing covariate effects such as age on the formation process of components.

Lastly, we compare the resulting representative parcellation with several previously reported structural and functional brain parcellation in qualitative and quantitative ways[155].

3.2 Methods

Sample Description

As discovery sample, we used the IXI sample (http:/www.brain-development.org) which is publicly accessible was collected from 600 healthy participants (age 19.98-86.32 years, 56% females) from three separate scanner sites (one 1.5 Tesla [X] and two 3 Tesla sites [X, Y]). After exclusion of cases due to the lack of essential phenotypes (age, sex, ethnicity) or lack of images 563 subjects were included. Our replication sample were several combined healthy subjects' samples (**Table 3.1**) that counted 566 subjects when counted together (age 18-83 years, 54% female) and that were all scanned on a 3 Tesla research MRI scabber at the Max-Planck-Institute of Psychiatry (Neuroimaging Core Unit). All subjects gave their informed written consent.

Table 3.1: Details of subsamples of the replication sample. *TMEM:* Imaging Genetics Study directed towards effects of the TMEM gene group on anxiety related MRI tasks and sMRI; *PsyCourse:* Clinical MRI Study on healthy controls and patient (MDD, bipolar disease, schizophrenia) groups[160]; *BeCOME:* Biological classification of mental disorders: Ongoing deep phenotyping study performed at the MPIP [161]; *Imaging Stress Test Study*: Multimodal imaging study directed towards effects of Psychoscial Stress[162]; Switch-Box and Junior-Switch-Box: collaborative EU project (local PI Prof. J. Zihl) on cognitive reserve phenomena in elderly and young healthy subjects; IL-16-MS-Atrophy: Study directed towards atrophy effects in Multiple Sclerosis and inflammatory markers (cum Dr. S. Nischwitz).

Study	N	Age	Sex
TMEM	154	18-36	51.2%
PsyCourse	28	20-52	64.2%
BeCOME	115	19-66	66.9%
Imaging Stress Test Study	59	20-31	49.1%
Switch-Box	135	68-83	47%
Junior-Switch-Box	46	25-32	52.1%
IL-16-MS-Atrophy Study	29	20-61	41.8%

VBM-style preprocessing of T1WI for later SCA

We performed a VBM-style preprocessing to prepare the input dataset for the actual key analysis, the SCA to generate a data driven parcellation. 'VBM-style' refers to the fact that usually on the resulting smoothed GM images voxel-wise statistical analyses, with statistical inference on the voxel or the cluster level, are performed. **Table 3.1** gives an overview of these steps performed mainly in SPM (version SPM12[163]), a MATLAB based software package for image processing and statistical analysis.

Table 3.2: VBM-style preprocessing steps applied to the discovery (IXI) and replication sample (MPIP)

Processing Steps	Processing Details
1 **Visual quality Control (QC)**	Visual inspection of raw T1WI to verify AC-PC image orientation and corrupt for further processing in SPM12
2 **Unified Segmentation**	Unified Segmentation method (as described before) to derive grey (GM), white matter (WM) and cerebrospinal fluid (CSF) segmentatic six default templates in MNI space (parameter settings: lig regularization [0.001], FWHM of Gaussian smoothness of bias 60 iteration of the Markov Random Field cleanup procedure, cleaning 'thorough', warping regularization [0 0.001 0.5 0.05 0.02] regularization: European brains, smoothness 0 mm, sample distance 3 addition to native space results, (modulated) normalized (MNI space are also written out, but not used further.
3 **DARTEL import step**	Import step for later processing in DARTEL. This step performs a registration (driven by the whole head) for a first alignment of GM an with the MNI space.
4 **Iterative warping in DARTEL**	Generation of flow fields using DARTEL with 6-generation GM a templates in MNI space gained from the IXI sample (source: VBM8 (http://dbm.neuro.uni-jena.de/vbm8) (default DARTEL parameter s The resulting flow fields contain deformation information needed to tr images between the affine-registered position (after step 3) fully to MI The generation of this flow-field is based both on GM and WM channels.
5 **Writing out of warped GM with Jacobian Modulation**	Generation of spatially normalized GM images at a resolution of 1.5 × mm^3 through application of the flow-fields to the imported GM (dimensioned 121 × 145 × 121 voxels, 5th degree spline interpolation) Jacobian modulation (JM). JM preserves the original GM pr information which, after warping, can be interpreted as volume infor
6 **Smoothing**	Spatial smoothing using a Gaussian kernel sized (FWHM) 6 × 6 × 6 second set of smoothed images was produced with a kernel sized (FW × 10 × 10 mm^3.

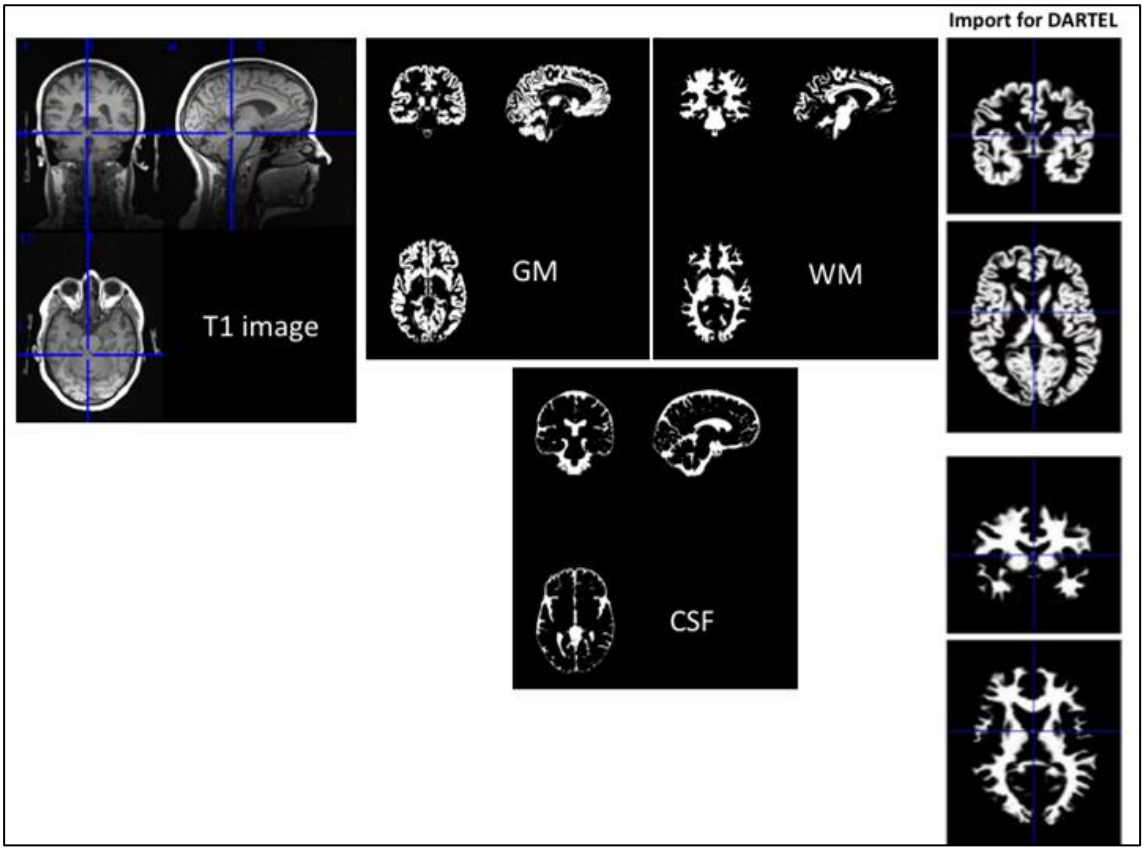


Figure 3.1: Early processing steps of Unified Segmentation and DARTEL import step.*Left:* Original T1WI in axial, sagittal and coronal view. *Middle:* native space GM, WM and CSF segmentations, in addition to two further non-brain compartments (skull/bone and soft-tissue). The 6th compartment is not displayed because it is usually not written out. *Right:* GM, and WM segments after affine registration with the template space. Note different bounding box and symmetric position. No Jacobian modulation is applied during this step, yet, the volume change during the affine registration is added up to the JM information later.

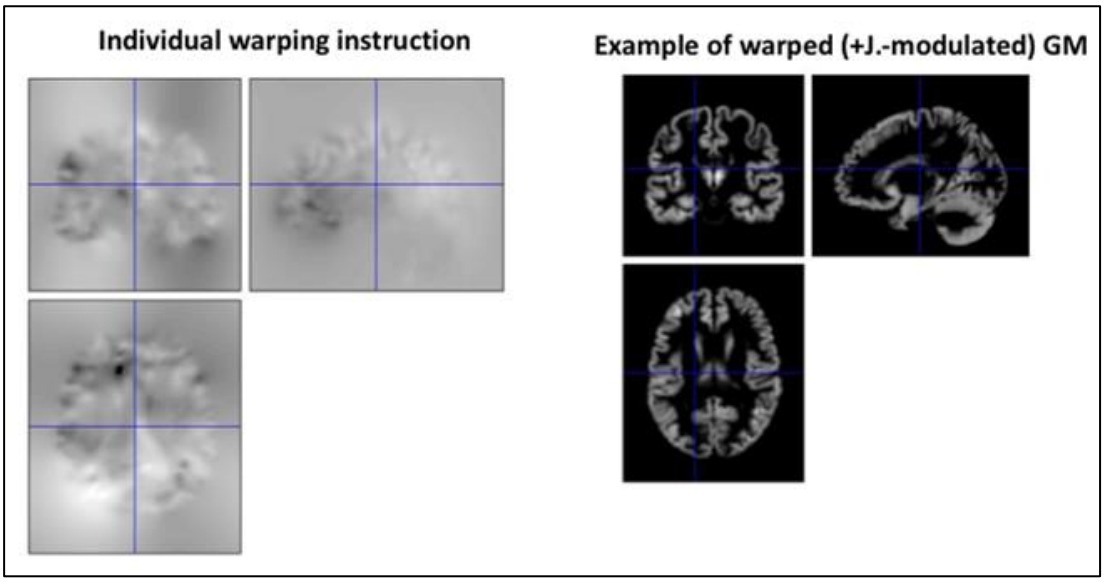


Figure 3.2: Example of a flow field and resulting warped and modulated GM image. Left: Example of a flow-field (as far as depictable in a triplanar image). Note complex global and local information. Right: GM segmentation after application of the flow-field. Note symmetric

position in MNI space), some individual gyrification pattern, and differences in local brightness (bright: 'condensed' areas for high GM volume, pale: expanded areas for low GM volume).

Calculation of intracranial volume

The intracranial volume (ICV) is an essential global measure to 'scale' morphometric analyses of local features because local brain measures are highly correlated with the skull's volume. Meanwhile, more complex than linear models have been detected in data and are used sometimes, but most analyses (as our) still base this step on the assumption of a linear relationship. Inaccurate ICV estimations could thus distort any further analysis results. The VBM toolbox [164] of Statistical Parametric Mapping [31]was used with default parameters to segments the voxels of T1-weighted brain volume into four classes, namely white matter (WM), gray matter (GM), cerebrospinal fluid (CSF) and other. No preprocessing or re-orientation was applied on the T1-weighted images in advance to estimate the ICV since manual intervention[165] to attain the method's automatic feature. ICV was defined as the sum of modulated, normalized GM, WM, and CSF maps, with each map being protected by an SPM12 default ICV mask. Thus, we calculated ICV based on the accurate segmentated data; that is, we summed up the values of the Jacobian-modulated, spatially normalized GM, WM, and CSF map of an individual. Theoretically, the same summing procedure could have been performed in native space. We preferred to perform it in MNI space, as here, by a default (template) ICV mask, the CSF compartment can be 'cut out' to prevent voxels being added in that are not genuinely intracranial CSF. Such wrong spaces are, for example (rarely): lower cisterns extending too far out due to soft (facial) tissue misclassified as CSF, parts of the transverse venous sinus, etc. Regarding the statistical approach, the ratio (or proportional) approach (reviewed in O'Brien study [166]). Still, the covariate approach is generally preferred because it is more flexible and borrows information from the other subject in the GLM, whereas the proportion method (i.e., dividing each voxel by ICV) can lead to error propagation, as two measures are combined.

Spatial Gaussian smoothing

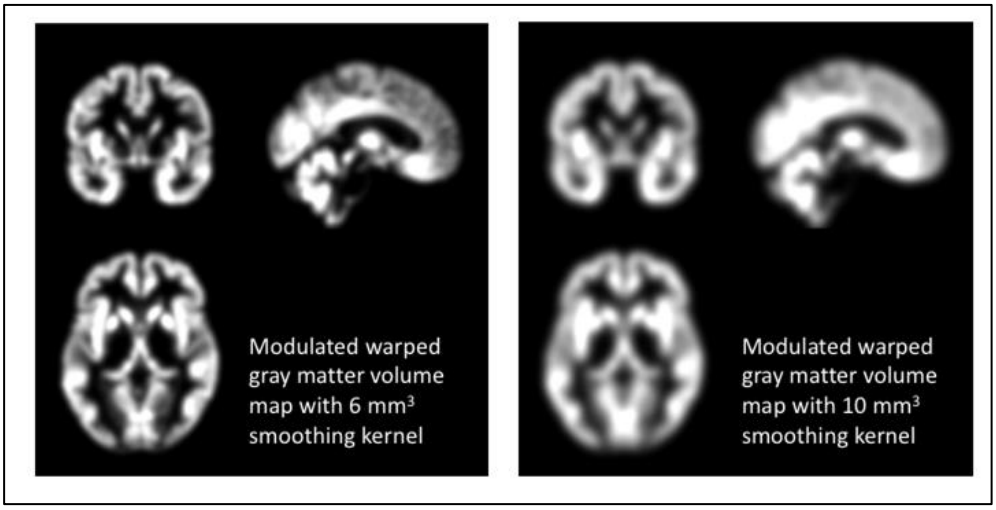


Figure 3.3: Warped and modulated GM image with 6 mm3 and 10 mm3 smoothing kernel. Left: Example of a modulated warped GM map with 6 mm^3 smoothing kernel, Right: Example of a modulated warped GM map with 10 mm^3 smoothing kernel

Jacobian modulated, warped GM volume maps are smoothed spatially in addition, replacing the voxels original value by a weighted average of the surrounding voxels. The number of voxels included in the averaging process at every point depends on the size of the smoothing kernel. A Gaussian isotropic kernel is generally used for this purpose with a full width of maximum of the Gaussian shaped kernel up to half to 6-12 mm (in all three directions, written for example as FWHM [6 6 6] mm^3). In VBM, usually isometric kernels are used; non-isometric kernels are sometimes used in functional MRI when the voxels are non-cubic. The smoothing level should be based on the accuracy of the co-registration and also be guided by the size of the anticipated regional differences among groups. Here we used 6 mm and 10 mm FWHM smoothing: 6 mm is the actually desired smoothing level suitable for DARTEL, as the latter provides excellent inter-subject alignment. The use of 10 mm represented a specific manipulation to investigate the effect of a higher smoothness on the ICA. Generally, spatial smoothing also increases the validity of parametric models by improving the normality of the residuals. Another effect of smoothing is also to reduce the number of independent spatial elements (referred to as 'resels' in VBM).

Concatenation of 3D volumes to a 4D volume

After smoothing, we concatenated all individual smoothed GM maps (dimension 121 x 145 x 121 voxels) to a single 4D file. This step was performed separately for the discovery and the replication sample. It is a mere technical step, not changing the information, as the ICA tools usually prefer to process 4D data.

Combat strategy and Residualization model

As an intermediate step, the 4D dataset as stored in the NIFTI format was transformed into a 2D-array (collapsing 3D spatial information into N-1-vector) with dimension [2122945, 563] (for the discovery) and [2122945, 566] (for the replication sample) in order to apply the *Combat* function[167,168] to it. Of this array, only the voxels within the GM mask (479384 voxels) were selected. Combat has been developed initially for genomic data to remove batch effects, but the method has been successfully used for different imaging data[167–169]. Notably, an array of covariates is also entered into the algorithm, and the goal is only to remove variance caused by the batch (here: three different MRI acquisition sites of the IXI sample (for the discovery sample) and six different original sub-studies of the MPIP sample (for the replication sample)) but preserve variance explained by the covariates.

ComBat algorithm[170] has recently been adapted for multi-site DTI modeling and elimination of site effects in batch-effect correction tools, widely used in genomics[171]. ComBat is an important harmonization strategy, eliminating undesirable site differences and preserving biological connections in the results. We used the ComBat algorithm[171] to harmonize gray matter maps collected from different scanners. We used two primary multi-site datasets: IXI, a three-site multicenter analysis, and MPIP, which has used a total of 6 scanners. We harmonize data with scanner and site impact elimination using ComBat while maintaining the availability associated with biology. We prove that Battle can also be used for integrating data sets for the study of life-length trajectories through many locations.

Suppose the data contain m batches containing n_i subjetcs within batch i for i = 1, . . . , m, for voxels g = 1, . . . , G . We assume the model specified in (2.1), namely,

$$\boldsymbol{\gamma}^{*}_{\boldsymbol{ijg}} = \boldsymbol{\alpha_g} + \mathrm{X}\beta\mathrm{g} + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \qquad (2.1)$$

and that the errors, ε, are normally distributed with mean zero and variance σ_g^2.

where $\boldsymbol{\alpha_g}$ is the overall data matrix (voxels * subjects) , X is a design matrix for sample conditions, and βg is the vector of regression coefficients corresponding to X. The error terms, ε_{ijg}, can be assumed to follow a Normal distribution with expected value of zero and variance σ_g^2. The γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g, respectively. The batch-adjusted data, $\boldsymbol{\gamma}^{*}_{\boldsymbol{ijg}}$

The algorithm contained three main steps:

Step 1. Standardize the data:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - \mathbf{X}\hat{\beta}_g}{\hat{\sigma}_g}$$

Step 2: Empirical Bayes (EB) batch effect parameter estimation using parametric empirical priors

$$\gamma^*_{ig} = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta^{2*}_{ig} \bar{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta^{2*}_{ig}} \text{ and } \delta^{2*}_{ig} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \hat{\gamma}^*_{ig})^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1}$$

Step 3: Adjust the data for batch effects

$$\gamma^*_{ijg} = \frac{\hat{\sigma}_g}{\hat{\sigma}^*_{ig}} (Z_{ijg} - \hat{\gamma}^*_{ig}) + \hat{\alpha}_g + \mathbf{X}\hat{\beta}_g$$

The covariates entered were: ICV, age, age2, sex, age-by-sex, age2-by-sex, and ethnicity for the IXI sample (as dummy variables, six levels). Age and age2 were centered before used for the interaction terms. **Table 3.3** lists the 6 and 12 terms of the covariate matrix for the Combat step for both samples. First, a 4D 'stacked' version of all individual smoothed GM images were generated, and the MATLAB implementation of Combat was used to remove site effects, separately of the discovery and replication sample, under preservation of 12 and 6 covariates, respectively. Second, a multiple linear regression model including an intercept term was defined in SPM12 (again separately for discovery and replication sample), estimated within the above-described mask area, and non-normalized residual images collected. Non-zero voxels were increased by a constant value of 100 to ensure positive values. 4D versions of these 563 (discovery) and 566 (replication) images were calculated as input for ICA.

Table 3.3 List of Covariates for the discovery, 6 lines for replication sample

Discovery (12 terms)	**Replication (6 terms)**
ICV	ICV
age	age
age^2	age^2
sex	sex
age-by-sex	age-by-sex
age^2-by-sex	age^2-by-sex
ethnicity (dummy coded): 1 to 6	**NA**

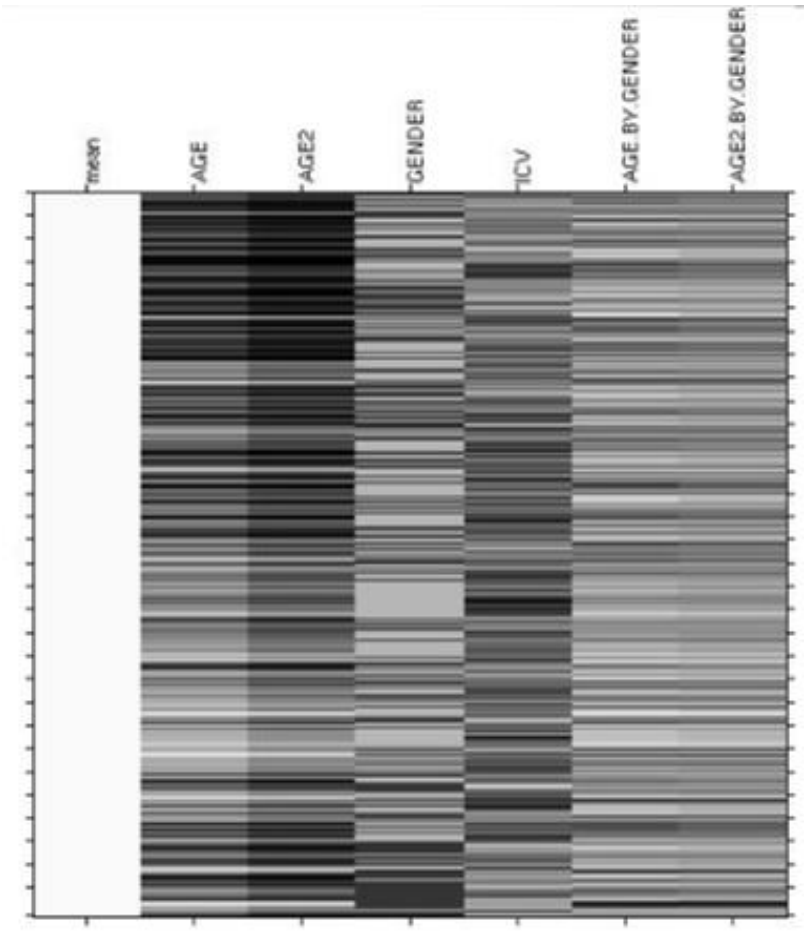


Figure 3.4: Exemplary residualisation model (replication sample). The GLM as visualized in SPM ('review model' function) with the vertical dimension representing the subject axis, and regressors labelled at the top. Note that additional dummy covariates were used to code to for 6 different ethnicity subgroups in the discovery sample (ICV, age, age^2, sex, age-by-sex, age^2-by-sex and ethnicity with 6 levels (dummy coded only for discovery*).*

Independent component analysis, SBM as implemented in GIFT and computational setup

Independent component analysis for biomedical signal analysis is a popular and essential multivariate statistical technique. Often, biomedical signals we can measure actually consist of mixtures of signals from various underlying sources that also bring in noise. ICA works by breaking up the mixed signals into underlying components. We have described Independent Component Analysis clearly in **section 1.7**.

Source-Based Morphometry (SBM) is a multivariate tool to study the gray matter differences between the patients and controls [156,172]. ICA is used on the subject images to determine the maximally independent sources. Basically, ICA decomposes data into subject loading coefficients

and component maps. It is similar to doing single subject single session analysis in the GIFT, except the time points are subject images.

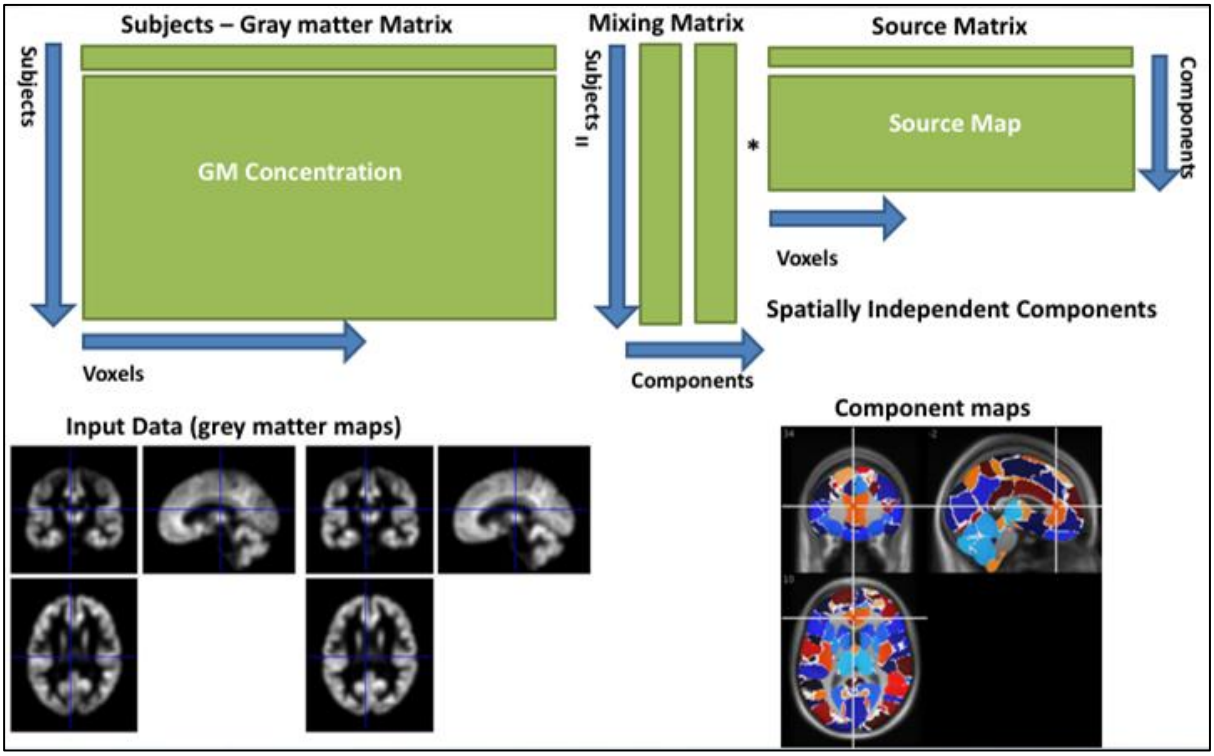


Figure 3.5: Source Based Morphometry using for discovery and replication data, where subject-by-gray matter matrix is decomposed into mixing matrix and source matrix. We have used the SBM/GIFT toolbox to perform the infomax ICA. ICA model in which the subject-by-gray matter matrix was decomposed into mixing matrix and spatially independent components. [156]

The following are the differences between the GIFT and SBM:

- Default mask used in the SBM includes voxels greater than or equal to 1% of the mean of the data.
- Detrending is excluded in the SBM toolbox whereas it is included in GIFT, which mainly dealing with time-series data of fMRI
- Component maps are stored with the suffix **group*component*ica** in Analyze or Nifti format. Subject component loading coefficients are stored with the suffix **group*loading*coeff** in Analyze or Nifti format.
- Batch template is provided in *icatb\icatb_batch_files\Input_sbm.m*. Specify *modality type* as 'smri' and enter the parameters similar to one subject one session analysis as in the GIFT. After entering the parameters, use *icatb_batch_file_run(inputFile)* at the MATLAB command prompt.

Percent variance utility can be used after running a group ICA analysis. Percent variance explained by the components in the data is calculated by doing a multiple Regression of the BOLD signal and the component. For the analysis of gray matter volume maps, the SBM toolbox

within GIFT [173] is a specific implementation of ICA and ICASSO, Of the several ways to extract the components (or sources), Infomax was found most suitable to estimate maximum spatially independent sources. The infomax principle has a close connection with maximum probability [45,46,48]. It is based on a neural network with nonlinear outputs to maximize the output entropy or information flow. The algorithm is based on the maximization of entropy and *presents a natural gradient form* for the computation of independent components[174].

As for the mathematical formulation, it can be interpreted as a neural learning method:

$$W(t+1) = W(t) + \eta(t)(I - f(s)s^T)W(t)$$

where η(t) is a learning-rate function and f(·) is a function related to the distribution nature (i.e. super Gaussian or sub Gaussian)[175].

As explained earlier, in batch mode, the SBM/GIFT tool was used to run ICA repeatedly over a range of predefined dimensions, from 20 to 545 components (in steps, 25, so [20:25:545]) one after the other, using the same input data (discovery sample with 6 mm and 10 mm FWHM Gaussian smoothing, respectively, and replication sample only with 6 mm). The number of components were estimated [176] from these sMRI datasets using the inbuilt MDL method first. Components are estimated for discovery and replication sample separately. The number of estimated components for the discovery sample and the replication sample, after Combat, but both with and without residualisation, is given in **table 3.4**.

Table 3.4 Estimated number of components for four different cohorts using MDL algorithm (GIFT toolbox)

Input	Estimated Number of Components
Discovery sample (6 mm^3) with combat + resdualisation	4
Replication sample (6 mm^3) with combat + resdualisation	5
Discovery sample (6 mm^3) with only combat	43
Discovery sample (10 mm^3) with combat + resdualisation	8

Infomax is a very popular ICA approach and produces more stable components compared to FastICA[177]. In general, ICA contains stochastic steps, so the resulting components are not completely deterministic, but can vary from run to run. Iterative ICASSO is an approach that delivers both 'centroid' versions of a component in addition to component stability measures[159,178].Random initialization (RandInit) was used for each of the 20 runs of a random initial value and 16 and 20 (default settings) were the minimum and maximum cluster size.

Further configurations (e.g., maximum number of steps) at the default values were kept. We extracted the component maps, and the stability index vector as main results of ICASSO. For each component, this stability index is calculated by ICASSO: It is mainly driven by spatial differences between recurrent components[179]. Reliable estimates (values close to 1) correspond to 'tight clusters' and unreliable ones do not point to any cluster (Example cluster number **69** in **Figure 3.6**)

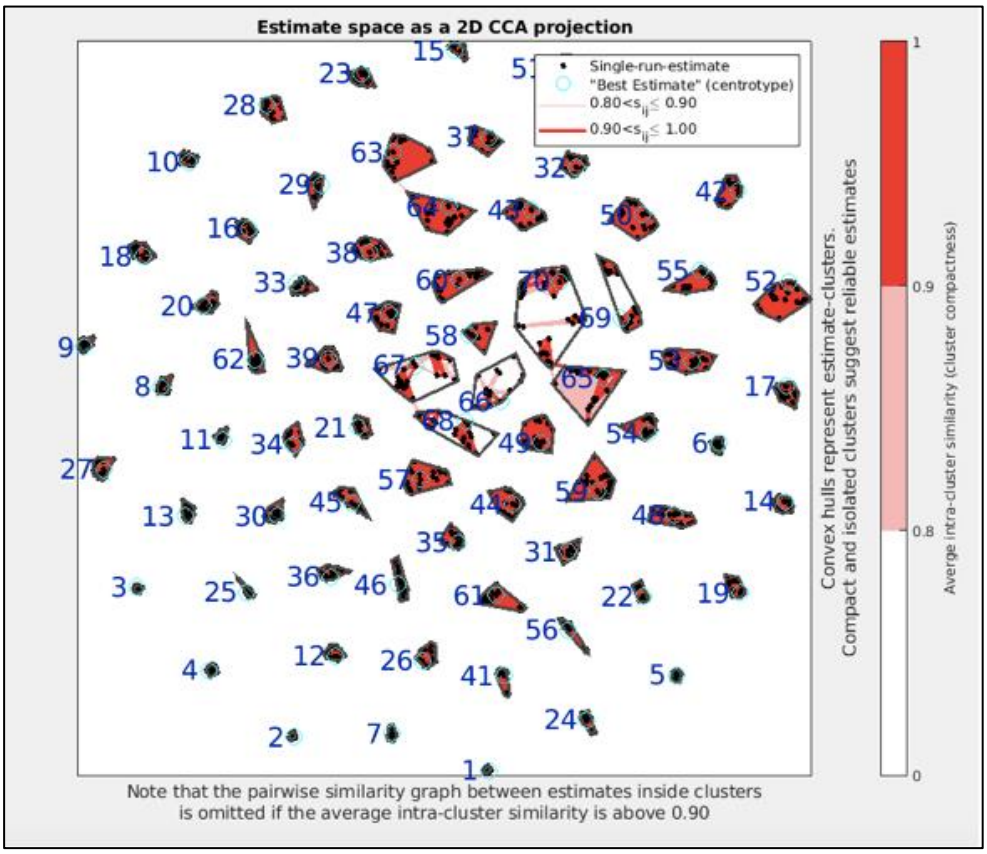


Figure 3.6: Estimated space as a 2d CCA projection to visualize reliable and unreliable clusters using stability index. Here estimates which acquired high intra-cluster similarity values (>=0.8) in multiple iteration formed compact and isolated clusters whereas estimates with low (<0.8) intra-cluster similarity values in multiple iteration formed volatile clusters.

Since ICASSO uses clustering of components and there is no constraint in ICASSO on the number of components within each cluster. A cluster containing more components than runs might combine components from different functional areas. Also the mixing coefficients of centrotype might come from different runs which might not be desirable as well. To avoid this stable run estimates are used. These estimates are calculated using stability index, minimum and maximum cluster size. After the ICASSO step is completed, subsequent group ICA analysis steps like Back Reconstruction, Scaling Components and Group Stats are run.

ICASSO uses group average link (AL) as default choice of agglomeration strategy. Himberg[180] introduced a conservative cluster quality index I_q that reflects the compactness and isolation of a cluster. It is computed as the difference between the average intra-cluster similarities and average extra-cluster similarities.

For stability and similarity measurement, we have calculated Stability index (Iq), and b) sR=Variable containing information about similarity measure, clustering, and projection. Cluster quality index (Iq) of the selected components from ICASSO-runs[179] were used to assess the repeatability of ICA components of interest and mean of all Iq was used to measure the overall stability of the whole ICA decomposition.

An Iq value of higher than 0.8 is often used to define a component as 'stable', but this is not a fix rule. The mixing matrix containing cluster centrotype based estimates from ICASSO was used to produce final IC maps.

Hu et al. (2019) [181] recently developed an improved stability index, which also accounts for a potential instability of the loading vector subject. As a post-hoc analysis, we used this algorithm to investigate the impact on the final data-driven atlas (parcellation) of the more sensitive stability estimation's spatial features. Iq is calculated in terms of inner similarity within one cluster and dissimilarity among different clusters. Himberg [179] used $\boldsymbol{Comp\_I_q}$ (Iq for the components) as a stability index. For example, if the number of components is *N*, and ICA decomposition is run *K* times with random initialization, *N*K* components are produced, and then, those components can be clustered. For ICA decomposition, each component is associated with each coefficient vector. N**K* components' memberships can be used to cluster the *N*K* coefficient vectors to obtain the coefficient matrix's stability. The parameter Iq can also be generated for the coefficient matrix. It is called $\boldsymbol{Coef\_I_q}$ in this study. If both the component and the corresponding coefficient vector are stably extracted, the ICA decomposition for the component and the coefficient vector is repeatable. As a result, the Iq of the ICA decomposition in this study is defined as:

$$\boldsymbol{I_q = Comp\_I_q \times Coef\_I_q}$$

Since the range of stability index of 0 to 1 so the multiplication does not change the range of stability evaluation. The probability of the stability of the coefficient matrix can be represented as $\boldsymbol{Comp\_I_q}$. The probability of the stability of the coefficient matrix can be understood as $\boldsymbol{Coef\_I_q}$ The multiplication can thus be represented as the overall stability[181].

Mask generation

Since we were primarily interested in grey matter coherence patterns, we defined a 1/0 mask for the analytical space based on two steps: First, all 563 smoothed GM images (FWHM 6 x 6 x 6 mm3) of the discovery sample were averaged into one 3D image, and this average image was thresholded at > 0.1 (resulting in 476219 voxels). Second, due to the known imperfect and potentially underestimated probability of GM in the basal ganglia area (particularly thalamus and pallidum), there was a risk to threshold away these areas. To counteract this, we produced a 'protective' 1/0 mask of the putamen, pallidum, caudate, and thalamus from the bilateral AAL atlas regions and added this onto the mask of step 1. This process increased the total number of voxels to eventually 479384. We first produced a mask using the same two-step, for the replication sample analysis, which resulted in a mask with slight differences in rim areas compared with the discovery sample mask (1279 additional voxels and 10634 missing voxels). However, as 99.73 percent of the discovery mask's voxels overlapped with the intersection of both masks, the discovery sample mask was also used for the replication sample. It covered the entire cortex, super-cortical and infratentorial, but spared the mesencephalon and most of the pons. These areas usually do not deliver reliable GM information, as they do not contain compact nuclei of GM, but a more diffuse mixing of GM and WM signal cannot be resolved using 3 Tesla MRI.

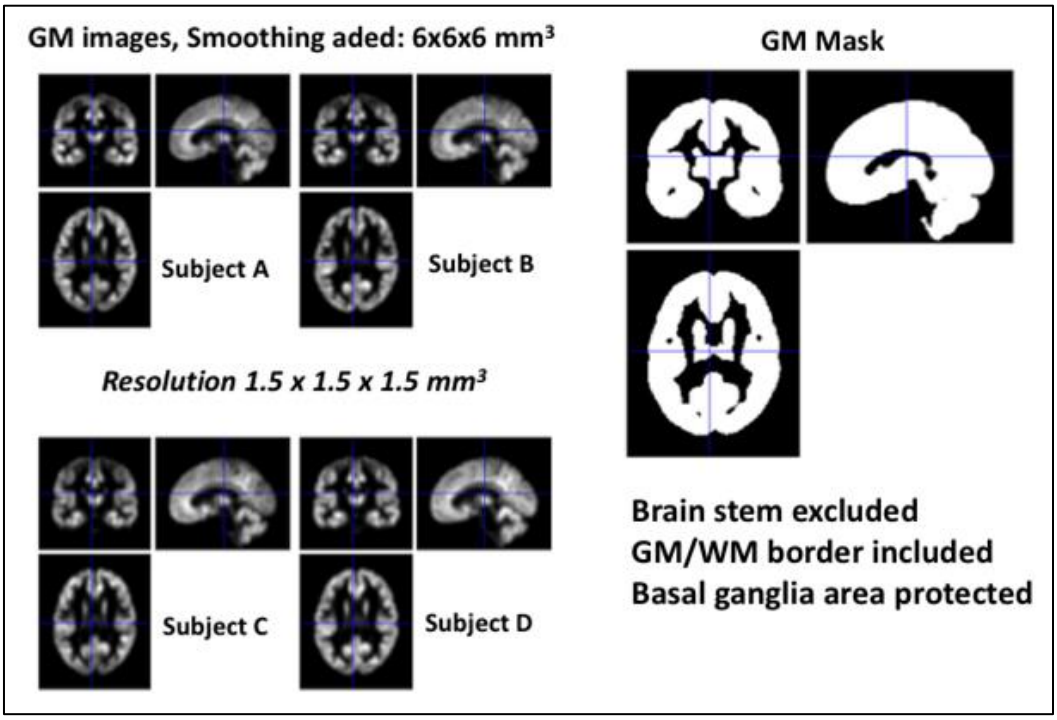


Figure 3.7: Exemplary smoothed GM maps of four subjects and tri-planar visualization of analysis mask. On the left, 4 Jacobian modulated and smoothed GM maps of subjects A-D are shown. Note very similar margins and geometry, but different intensities in corresponding areas

that represent different volume states. The right column shows the binary analysis masks that contained 479384 voxels as generated by the two step procedure (see text). The internal spared areas represent either WM or CSF areas.

Hierarchical clustering, re-agglomeration and convergence analysis for dimensionality estimation

ICA requires the definition of a number k of components expected. The estimation of the 'proper dimensionality' of the data set is independent of the ICA itself and usually carried out before. For its estimation, different methods and metrics have been suggested, for example [Minimum Description Length (MDL), Akaike's Information Criterion (AIC)[182], Bayesian Information Criterion (BIC)[183], Integrated Completed Likelihood (ICL)[106]. The number of sources can be estimated using the well-known Akaike's information criterion (AIC)[182] or the minimum description length (MDL) criterion[184]. These criteria have the following forms:

$$AIC(N) = -2V(ML - N)\mathcal{L}(\hat{\theta}_N) + 2\left(1 + NL + \frac{1}{2}(N-1)\right)$$

$$MDL(N) = -V(ML - N)\mathcal{L}(\hat{\theta}_N) + \frac{1}{2}\left(1 + NL + \frac{1}{2}(N-1)\right)\ln V$$

where *V* is the number of voxels, *M* is the number of subjects, £ ($\Theta\hat{}_N$) is the log of the maximum likelihood estimate of the model parameters (and is estimated from the data, e.g., sMRI data), *ML* is the number of time points following the first reduction stage, and *N* is the number of sources. The estimate for the number of sources is determined from the minima of the above functions with respect to *N* [185].

MDL is a popular approach for VBM based ICA to estimate the optimal number of independent components. We probed MDL in our data, too, as it is automatically calculated in the GIFT/SBM toolbox. Extracting the signal from noise is achieved using information theory metrics such as MDL[176,185,186]. We used the MDL criterion, a standard method for estimating the number of components from the aggregated dataset. The method makes a decision based upon the complexity or information content of the data.

However, the MDL approach is not robust and reliable in estimating the optimal number of the component which will cover the whole brain. This can cause disruption in the atlas parcellation and will develop an incomplete atlas. Here modified and extended a previously suggested

principle to estimate the dimensionality of a morphological, SCA-type of the dataset: The method has been initially referred to as 'ICA-by-block' in which ICAs with *n* components are performed on non-overlapping *B* blocks of the data are compared across blocks regarding their similarity. By a counting system of very high correlation values, the number of optimal components can be determined[187]. The principle could be rephrased in that an 'overstretching' of the estimated components are detected by the appearance of an over-proportional number of distinct components in the comparison scheme. Our modification and extension are as follows:

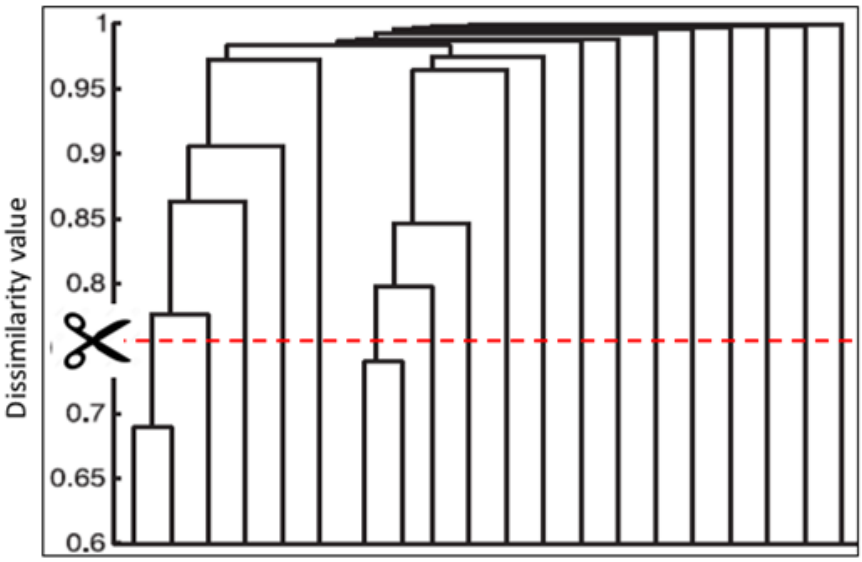


Figure 3.8: Exemplary hierarchical clustering tree and re-agglomeration to a lower number of components. Here, the 'cutting level' is indicated by the scissor and dashed red line. 'Re-agglomeration' means that 2 lower levels components (here: for example, the two leftmost leafs) are fused into one parent component. A cut at the shown level would result in 18 components of a maximum of 20 leafs. A higher cut would result in fewer agglomerated components. Functions exist that can determine the needed cut-level for a k'.

Assuming that N_{true} sources are contained in a dataset, these N_{true} sources could be directly estimated from running an ICA with N_{true} components, or by re-agglomerating components from an ICA run with an $N > N_{true}$. For this purpose, after performing the ICAs for a *k*, each of them was submitted to hierarchical clustering and re-agglomeration to *k'* (running from 10 to 545 in steps of 5) components were enforced for all available ICAs with dimension $k > k'$. This procedure follows the latent hypothesis that the closer k' is to N_{true}, the more similar with each other the re-agglomerated component sets of size *k'* are. The goal is thus to identify the row *k'* that delivers the highest similarity values across *k*.

For a quantification of the similarity of two component sets A and B (each of dimension *k'*), several metrics were developed and implemented in MATLAB:

- For each component of A, the maximum spatial correlation (Pearson correlation coefficient r, followed by Fisher's z transformation to r_z) with any component of B was identified and stored, along with the index of that maximally similar component of B. In addition, the second best match from the component set B was stored, as a high specificity of the best match would be reflected in a steep decrease of the similarity with the 2nd best match (i.e., a high difference value r_{best}-r_{2nd}); in turn, a low difference value would indicate a higher level of ambiguity of the choice. This resulted in a list as follows (e.g., for k'=20): A_1: best match with r_{max}=0.78 to B_3 (r_{2nd}=0.32); A_2: best match with r_{max}=0.97 to B_2 (r_{2nd}=0.12); A_3: best match with r_{max}=0.86 to B_{14} (r_{2nd}=0.44), etc., up to A_{20}.
- From the list of components of B that were selected as best matches, a vector was constructed that contained how often a component of B was chosen as best match ('pick count vector'): an ideal outcome would be a vector length *k'* only containing values of 1, i.e., each component of B has been chosen as best fit for a component of set A exactly one time. Deviations from this ideal vector contain zeros (a component was not picked as best match) or values larger than 1 (a component was picked several times as best match), both increasing the variance of this vector. This vector served as second basis for extracting the following seven overall similarity measures:

Table 3.5: Similarity metrics used for comparing agglomerations of the same k' constructed from different original k. The listed metrics can be used to compare any two set of ICs of the same dimensionality.

Metric #	Short name	Explanation	Range
1	'Mean Pearson'*	Mean of all (*k'*) highest Fisher's z transformed r-values comparing each component of A with each component of B	[0, 1] (optimum: 1)
2	'Median Pearson'*	Median of all (*k'*) highest Fisher's z transformed r-values comparing each component of A with each component of B	[0, 1] (optimum: 1)
3	'Delta 1st/2nd'	Mean of all difference between first and second best match for any $A_{1\text{-}k'}$ with two components of B	[0, 1] (optimum: 1)
4	'SD of PCV'	Standard deviation of the pick count vector	[0, *limit*] (optimum: 0)
5	'Percentage of PVC=1'	Percentage of values 1 in the pick count vector	[0, 100] (optimum: 100)
6	'Percentage of PVC=0'	Percentage of values 0 in the pick count vector	[0, *limit*] (optimum: 0)

7	'Percentage of PVC>1'	Percentage of values larger than 1 in the pick count vector	[0, *limit*] (optimum: 0)

In pilot studies we checked the robustness of these metrics to the order of the component sets A and B. A very correlated table of similarities was found when re-agglomerated component sets A and B were exchanged. Given the half matrix of *k*-by-*k'* with *k* > *k'* of re-agglomerated ICA solutions, we then faced the problem how to organize the comparisons within one row of *k'*: For example, for *k'* = 120, the available re-agglomerations from higher original values stem from ICAs with *k* of 145, 170, 195, 220, 245, 270, 295, 320, 345, 370, 395, 420, 445, 470, 495, 520, 545 and 558, i.e. 18 levels. All possible pair indices were formed under the constraint of $k_A < k_B$, resulting in (18^2-18/2) = 153 pairs of component systems compared using the resulting seven metrics Performing this procedure for all 108 levels of *k'*, this resulted in 9361 comparisons of pairs of component sets (>325 million component comparisons), requiring parallel computing. We noted that systematic gradients of the similarity values occurred in dependency of the position of a solution in the *k*-by-*k'* space. Due to this, we corrected for such effect by residualizing the entire vector against three values k_A, k_B, and *k'* using an in-house *R* script. Three aggregation schemes were considered for collecting information (per metric) from a single matrix field in the half-matrix of k-by-k':

First, to include as many measures as possible, comparisons of all pairs formed by a cell with other cells of the same line of k' were averaged (scheme 'permute'). In the example above, 18 values would have been averaged per cell.

Second, only immediately neighboring cells (scheme 'neighbor') were compared, similar to a sliding window to detect potential gradients in the horizontal direction per line of k'.

Third, comparisons of a cell were always performed against the most left positioned cell next to the diagonal, representing the re-agglomeration from the *k* closest to *k'* but skipping the original ICA (scheme 'diagonal').

The optimal number of components was determined separately for the discovery and replication set by calculating the mean value for each metric across a line of k' and creating profile plots for each metric across the k' axis. These plots were slightly smoothed, z-transformed for a uniform scale, and metrics #4, #6 and #7 inverted (to have local maxima as a uniform measure). 21 local maxima positions on the *k'* axis (seven metrics, three aggregation systems) were then determined and the median served as the basis for defining dimensionality.

Exploration of the relation of k with explained variance, component stability and anatomical location of stable and unstable component

The cumulative variance explained by each component of an ICA with dimension k was calculated. We also calculated the variance for each k separately for stable (Iq>=0.8) and unstable components (Iq<0.8) in order to understand the relation between *k*, the proportion of unstable components (explored at different Iq thresholds) and the variance explained by unstable versus stable components at standard Iq < 0.8 cutoffs. In order to understand if unstable components (again defined by Iq < 0.8) locate to similar areas with increasing *k*, i.e., if these areas are data/sample-dependent or dimensionality-dependent, we created maximum Z (max_Z-) maps of all stable and unstable components for the entire *k* range. For these maximum Z maps, a voxel was given the highest available Z value of all stable or unstable components, respectively. See figure 3.12 that illustrates schematically how maxZ (and maxLoc) images were calculated.

Compilation of atlas parcellation with discrete or fuzzy boundaries

The re-agglomeration system was used predominantly for estimating the dimensionality. Only original ICA runs without re-agglomeration were used to create discrete atlas systems for the discovery and replication sample. The assignment of a voxel to a component in the discrete atlas version was determined by identifying the component that provided the maximum Z-value for the voxel. To incorporate ICASSO-based component stability information, all component maps were weighted by the Iq value before this ranking-based decision. The refined ICASSO algorithm provided by a refined stability index (Iq*) integrating spatial stability for the component and subject-specific ranking stability was used for the respective optimal dimensions alone (in the discovery and replication sample)[177,181]. As the latter necessitated the use of an implementation of an ICASSO-based infomax outside of the GIFT toolbox (https://github.com/GHu-DUT/Tensor-clustering) we first compared similarity between the two component solutions and found close to perfect matches between the components. We thus considered the ICASSO implementations well-comparable, so the Iq* values could be used to weigh components instead of the classic Iq values[180]. This modified Iq* values are the multiplied form of component Iq and also coefficient Iq. The idea behind this approach is to cluster not only component matric and also coefficient matrix to produce more robust component maps (see section "**Independent component analysis, SBM as implemented in GIFT and computational setup"** for more details) [177,181]

We also tried to visualize how definite the boundaries between parcels were, besides creating discrete parcellation. This type of display is referred to as 'fuzzy border' approach. For this purpose, the difference between the highest Z value and the second highest Z value (after weighing with Iq or Iq*) were calculated per voxel. High values would represent strong distinction. The scale of this image was inverted and a 'web-like' image provided, in which the

most probable course of the boundary (‘likelihood valley') is displayed as brighter color. Both for discrete and fuzzy solutions, the ICA of the discovery and replication sample were repeated with a common suitable dimensionality.

Comparison of spatial similarity of the original and agglomerated component system of the discovery and replication sample

This comparison was made using four complementary analytical approaches due to the conceptual significance of this analytical analysis for reproducibility, replicability and generalization issues:

1. Parcellation of *k*-by-*k’* (half-) matrices were compared between samples using seven of the same metrics as the inter-sample comparisons. There has been no aggregation system here because only 1:1 comparison are made here. In other words, the original ICAs were included (*k* = *k’*) to draw a simple similarity profile for the 23 original ICAs, other than for the comparisons within the sample.
2. The 'mean Pearson' metric was re-calculated for fully pairwise aligned component systems in a variation of (1) to eliminate any influences of the asymmetric approach.
3. In the first place, the fuzzy boundary systems were inspected in direct opposition qualitatively for one common dimensionality (median of sample specific median optima).

Effect of smoothing level on dimensionality estimation and covariate residualisation on dimensionality estimation (discovery sample)

We repeated the whole analysis (multiple ICAs, hierarchical clustering, re-agglomeration, similarity analysis, determining local peak positions of profile plots) on the discovery dataset with larger Gaussian smoothing kernel of isometric 10 mm FWHM. The effect of skipping the residualisation of the covariates was also investigated by using the whole algorithm after removal of site effects on the discovery data set, but with no further correction for other covariates. In addition, for a common *k*, we calculated the variance of the subject loadings per component explained by age and (orthogonalised) age^2 in order to understand if structural covariance (and by this, component formation) is dominated by age effects.

3.3 Results

Proportion of stable components and explained variance over dimension *k*

The ratio of stable components plotted for different thresholds of Iq is represented for each k in **Figure 3.9** and **Figure 3.10**. We observed that the proportion of unstable components and k were inversely proportional to each other. In addition, and expected, the higher the Iq threshold was set, the lower the proportion of stable components was generally. For each *k*, **Figure 3.9** depicts the proportion of stable components calculated (Iq>=0.8), but plotted for our different samples: (i) discovery with 6 mm smoothing, (ii) replication with 6 mm smoothing, and (iii) discovery with 10 mm smoothing. Smoothing had a clear impact on the proportion of stable components. Smoothing with a larger kernel slowed the decline of the proportion stable components with higher *k.* Expressed differently, there was a general increment of the proportion of good components when smoothing factor was changed from 6 mm FWHM to 10 mm FWHM. **Figure 3.11** depicts the variance explained by all components and components split into stable (I_q >=0.8) and unstable ones (I_q <0.8). The total variance for stable components did not rise further after *k of* 300. The average maximum Z of all stable and unstable components is illustrated in **Figure 3.12,** which shows a steady increase of max-Z-values over *k*.

We have calculated the Iq values of classical ICASSO based on the spatial similarity of the repeatedly produced ICs (**Figure 3.13a**), but also the Iq values of the coefficient matrix, and based on both Iq values (according to Hu et al. 2019 and Zhang et al. 2018 algorithm), a tensor-clustering based 'combined' iq value (**Figure 3.13b**). Generally, voxels of components with low combined Iq values may have volatile characteristics, which means that their assignment to a specific component maybe unstable, or the pattern of how the subjects load on the component may be unstable. It was decided not to exclude components based on Iq value because including only those components with Iq>=0.8 during the atlas formation could break the brain coverage, leaving areas only components with even lower max-Z values would-be candidates. This means we used all components for the atlas formation but weighted Z-maps by their Iq value before the competition of the components for that voxel. In a volatile component, the component's Z-values would be low, and 'underdrive' a better component that then takes place for that voxel. We used both the classical, simple Iq value, and the tensor-clustering based one.

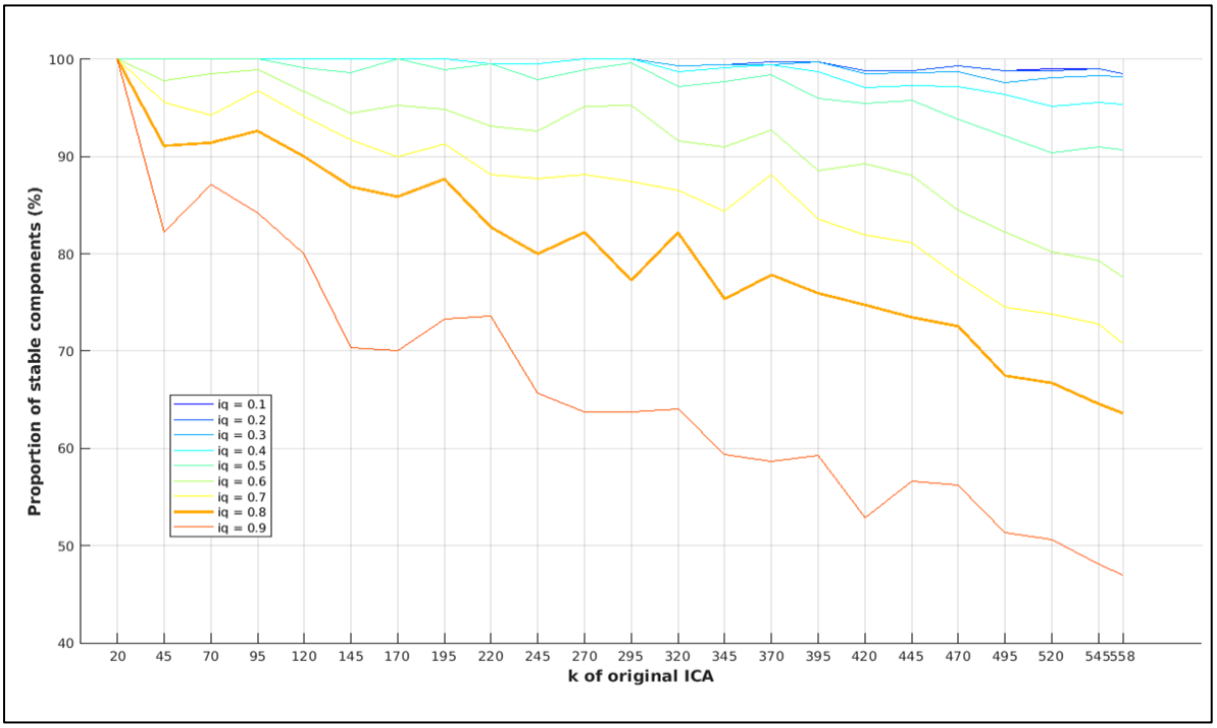


Figure 3.9: Proportion of stable components for different stability index (Iq) thresholds (0.1 to 0.9) for ICAs with different k. Note that the x-axis ticks represent all values of *k* used in this study, including the maximum of 558. Further note that the y-axis starts at 40% – at k=245, for an Iq-threshold of 0.8, there were still 80% stable components.

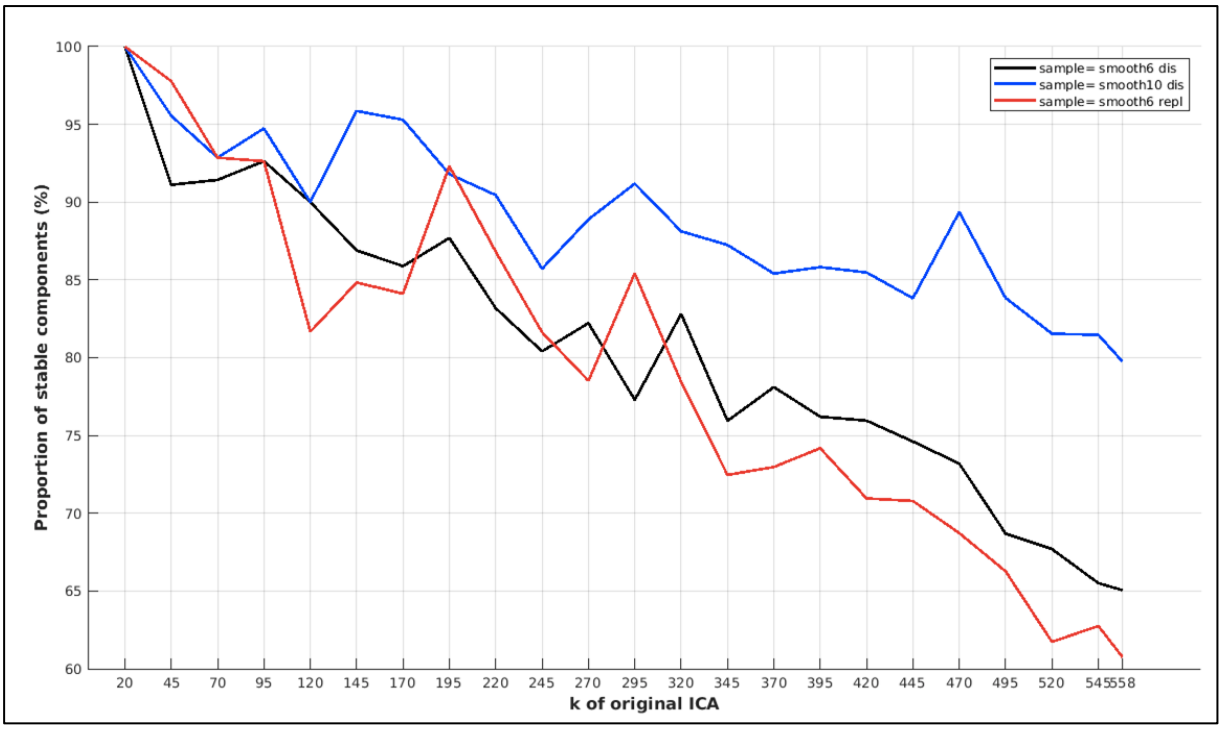


Figure 3.10: Proportions of stable components for a fix stability index (Iq) threshold (>=0.8) for ICAs with different k, plotted for different input data. Note similar course for the discovery and replication sample, and up-shifted values for the 10 mm smoothed discovery sample.

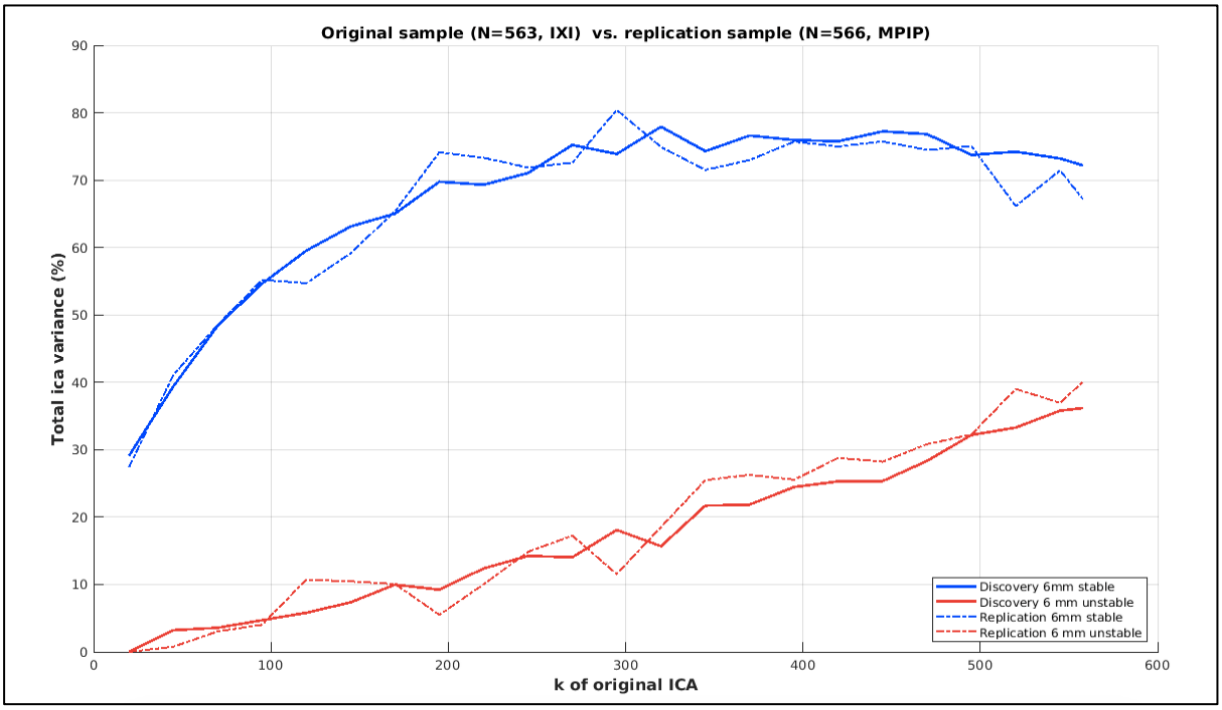


Figure 3.11: Comparison of explained variance using stability index (Iq) threshold (>=0.8) to stratify into stable and unstable components. Note some saturation for the cumulative variance of the stable components in both samples, and steady more linear increase of the variance explained by unstable components. Note a flat turning point round *k* of 300 from whereon variance seems to be shifted to unstable components.

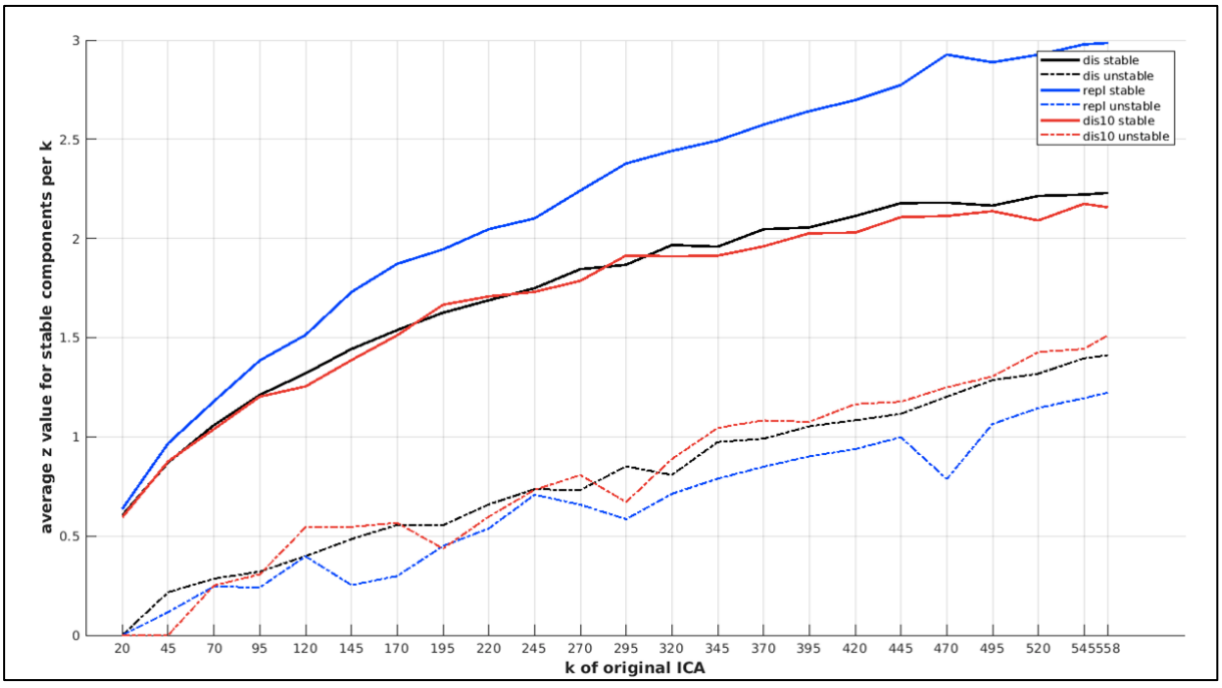


Figure 3.12: Average max-Z values of stable and unstable components. For this analysis, a max-Z-map was calculated for the stack of stable and the stack of unstable components. Per voxel, the maximum Z-value detectable in the respective stack was selected, and then an average calculated over the mask area. Interestingly, the replication sample showed a stronger

difference between stable and unstable components (higher positioned curve of stable components).

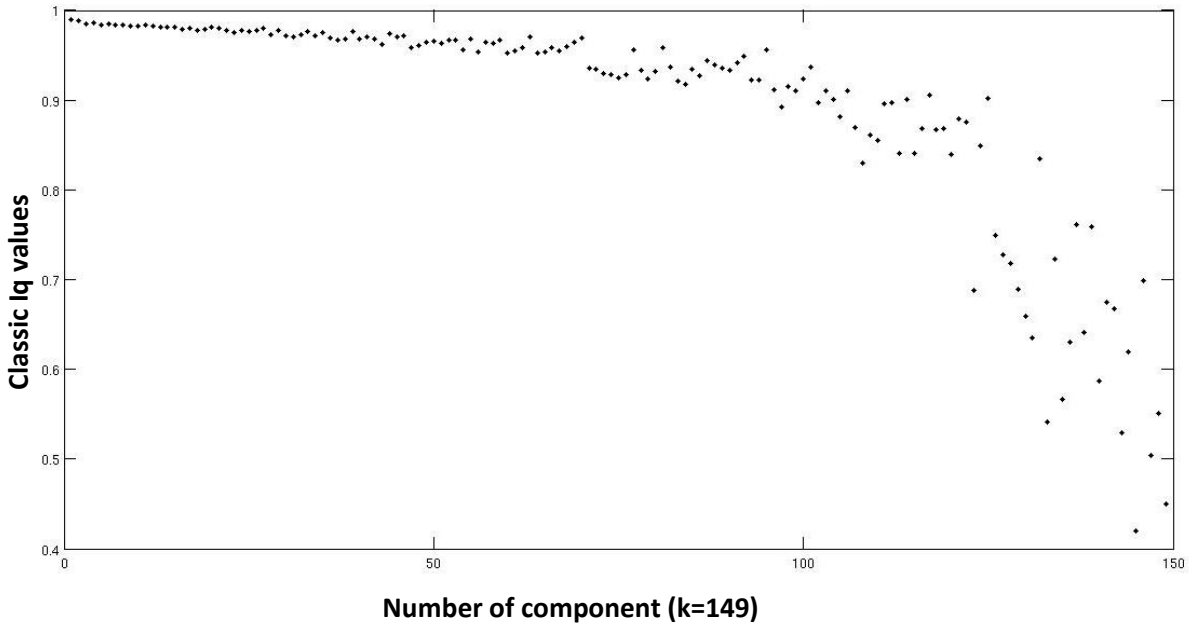


Figure 3.13: Standard stability index Iq plotted per component for the k=149 ICASSO of the discovery sample. Values were obtained from the GIFT/SBM software for *k* set to 149

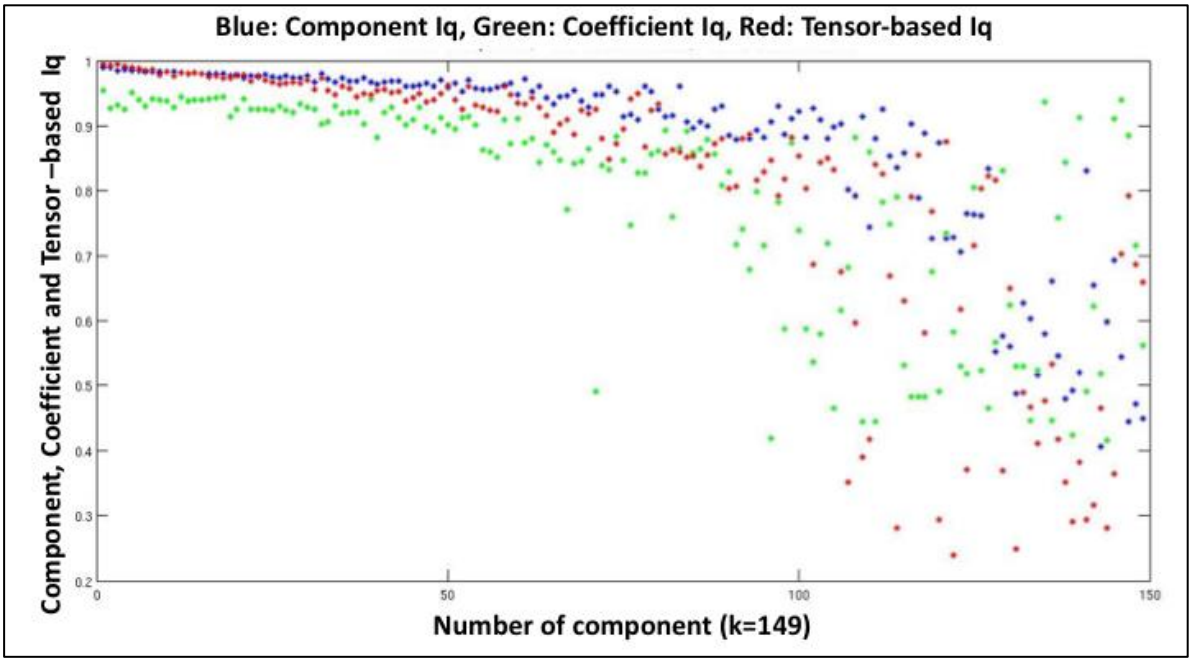


Figure 3.14: Component-related stability index (blue), coefficient-related stability index (green) and tensor based combined stability index (red). All values were obtained through an Infomax ICA implementation (2 Chinese REFS) (https://github.com/GHu-DUT/Tensor-clustering) with included tensor-based clustering for a combined Iq value. Note decline of about a third of the components. The tensor-based stability index seems to detect instability more sensitively.

Results of the similarity analysis of re-agglomerated ICAs

These result matrices served to decide if there is a promising range of *k'* for which re-agglomerations are most similar. **Figure 3.15 A-G** depicts seven half-matrices, each representing one similarity metric, of the discovery sample. Within the half-matrix, each field represents the average of similarity comparisons with all other fields (i.e., the aggregation scheme 'permute') of that same row *k'*. Corresponding results of the other two aggregation schemes ('neighbor', 'diagonal') are shown in **Table 3.4**. The rightmost column represents the average of the respective *k'* row and serves as the basis for the peak detection step. Note band-like maximum and mirrored minimum areas for inversely signed metrics. **Figure 3.16** shows the profile plot of those seven similarity metrics (for scheme 'permute'). Here, for the discovery sample a mean peak position of the 7 metrics (aggregation scheme 'permute') was located to k=143 (median 145). The replication sample results are plotted in **Figure 3.17 A-G** that again depicts the seven half-matrices, each representing one similarity metric. A very similar result pattern can be seen. **Figure 3.18** shows the respective profile plots with the mean of the 7 again being 151 (median 155). Similarly, we also investigated the similarity matrix and similarity profiles of the 10 mm smoothed discovery sample ('permute' aggregation scheme) and found that the profile showed a mean peak location at *k*=148 (median 145) (**Figure 3.19A-G and Figure 3.20**). The mean values reported here in the text are based on the raw values of **Table 3.6** using the 'permute' scheme.

The full result matrix hereof is given in **Table 3.6** that also shows mean, SD and median values of all 21 (7 metrics, 3 aggregation schemes) and overall mean, median and SD values for all 21 values. With the overall mean values of 147.6 (rounded 148) and 150.0 being very close, we defined k'=149 as the final dimensionality for visualization purposes and post-hoc analyses.

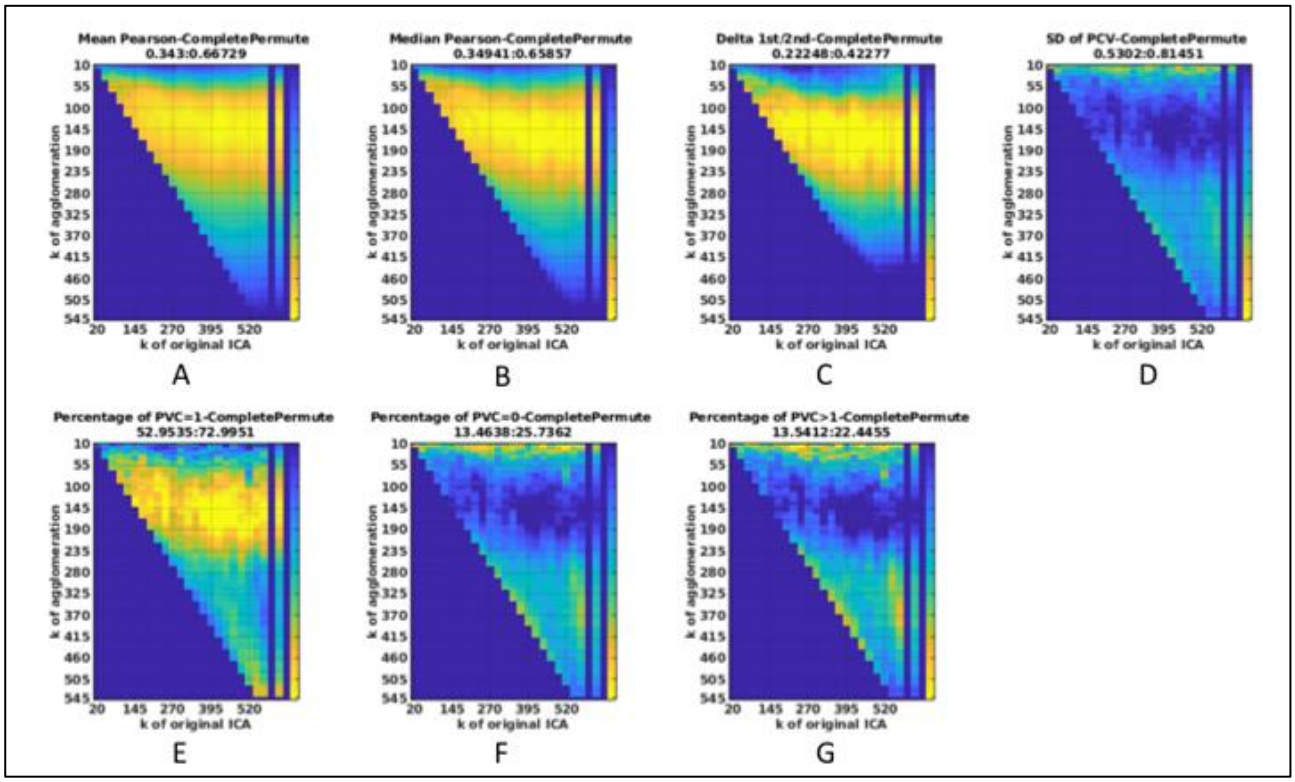


Figure 3.15: Half-matrices representing each one similarity metric of the discovery sample smoothed 6 mm. Within the half-matrix, each field represents the average of similarity comparisons with all other fields (i.e., the aggregation scheme 'permute') of that same row *k'*. The metrics are: A: Mean Pearson, B: Median Pearson, C: Delta 1st/2nd, D: SD of PCV, E: Percentage of PVC =1, F: Percentage of PVC=0, G: Percentage of PVC > 1. See table X for details on the metrics.

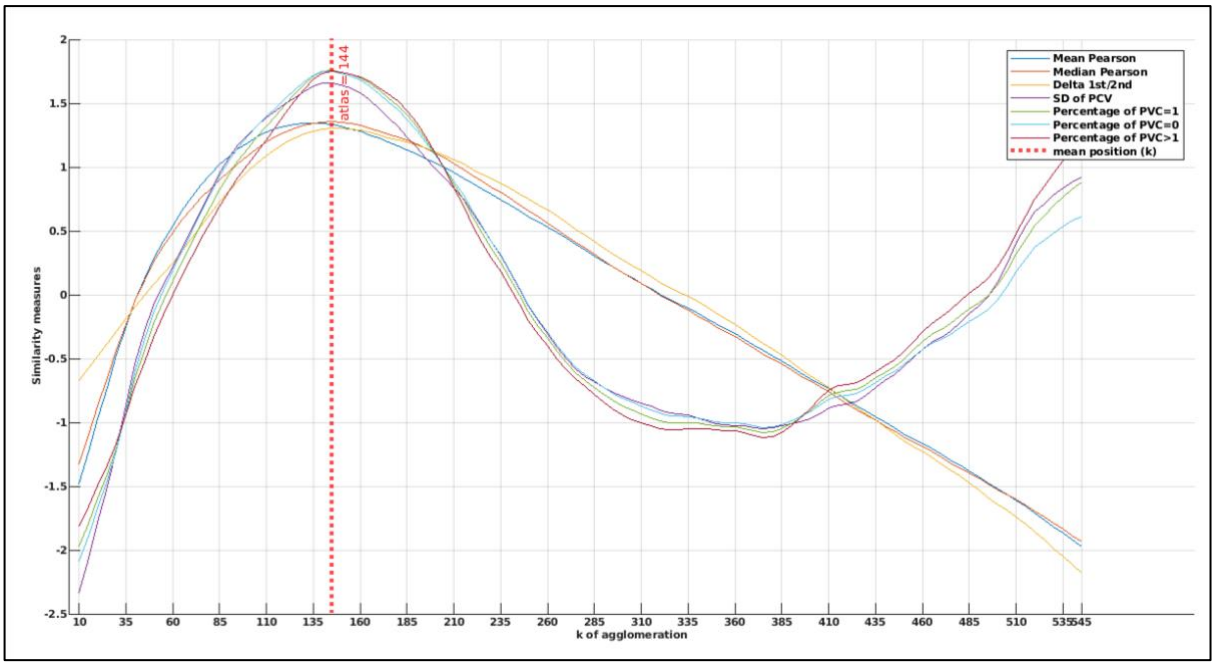


Figure 3.16: Profile plots of seven similarity metrics of the discovery sample smoothed 6 mm (aggregation scheme 'permute'). The vertical line shows the mean position of the seven peaks, locating to a rounded 144.

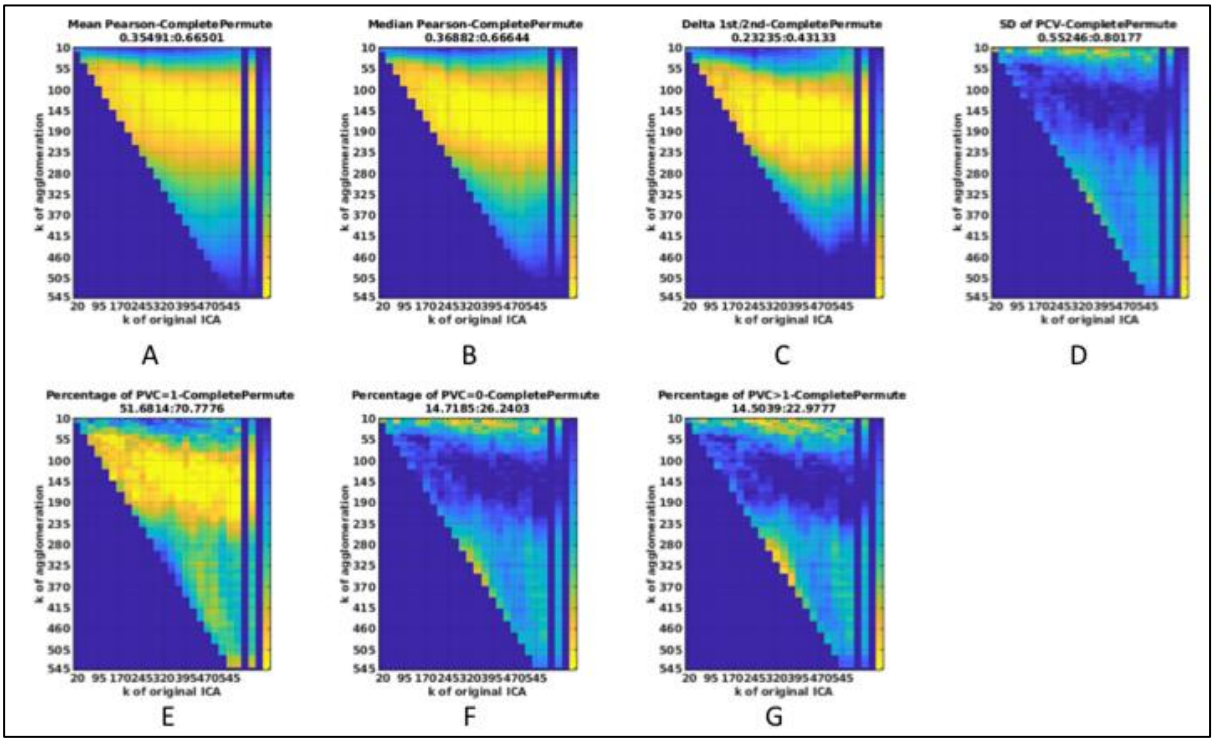


Figure 3.17: Half-matrices representing each one similarity metric of the replication sample smoothed 6 mm. Within the half-matrix, each field represents the average of similarity comparisons with all other fields (i.e., the aggregation scheme 'permute') of that same row *k'*. The metrics are: A: Mean Pearson, B: Median Pearson, C: Delta 1st/2nd, D: SD of PCV, E: Percentage of PVC =1, F: Percentage of PVC=0, G: Percentage of PVC > 1. See table X for details on the metrics**.**

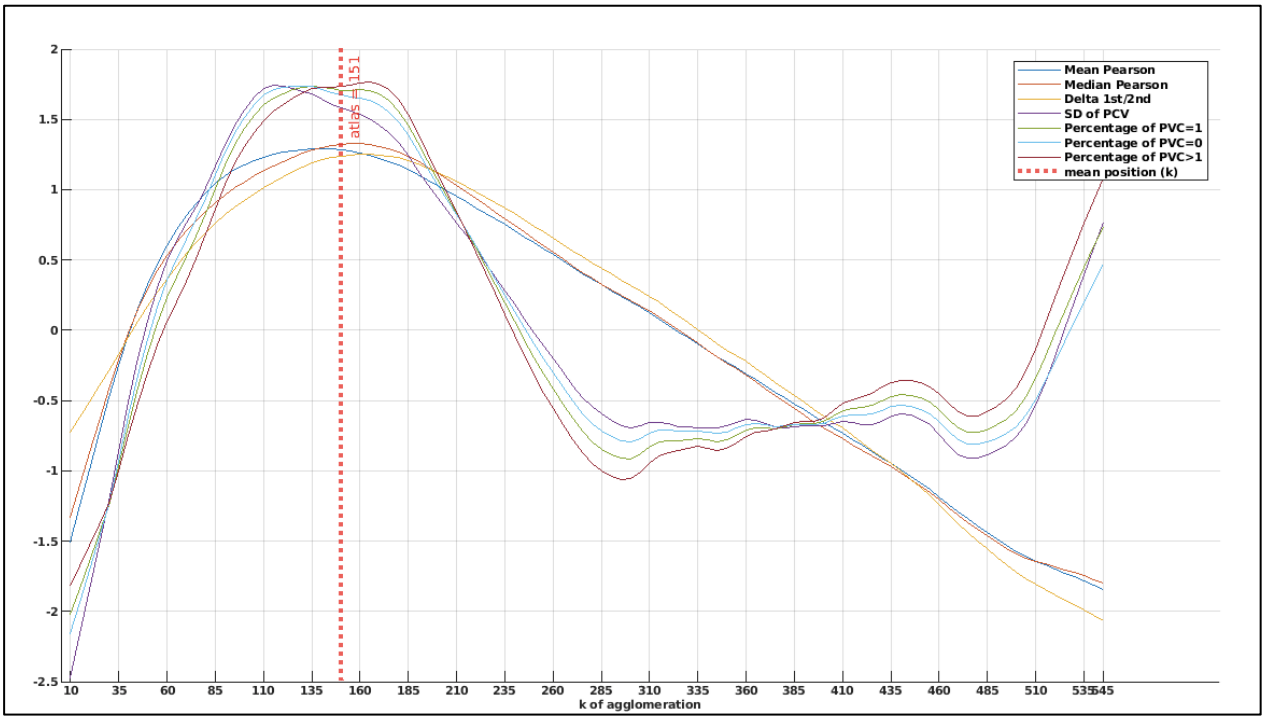


Figure 3.18: Profile plots of seven similarity metrics of the replication sample smoothed 6 mm (aggregation scheme 'permute'). The vertical line shows the mean position of the seven peaks, locating to a rounded 151.

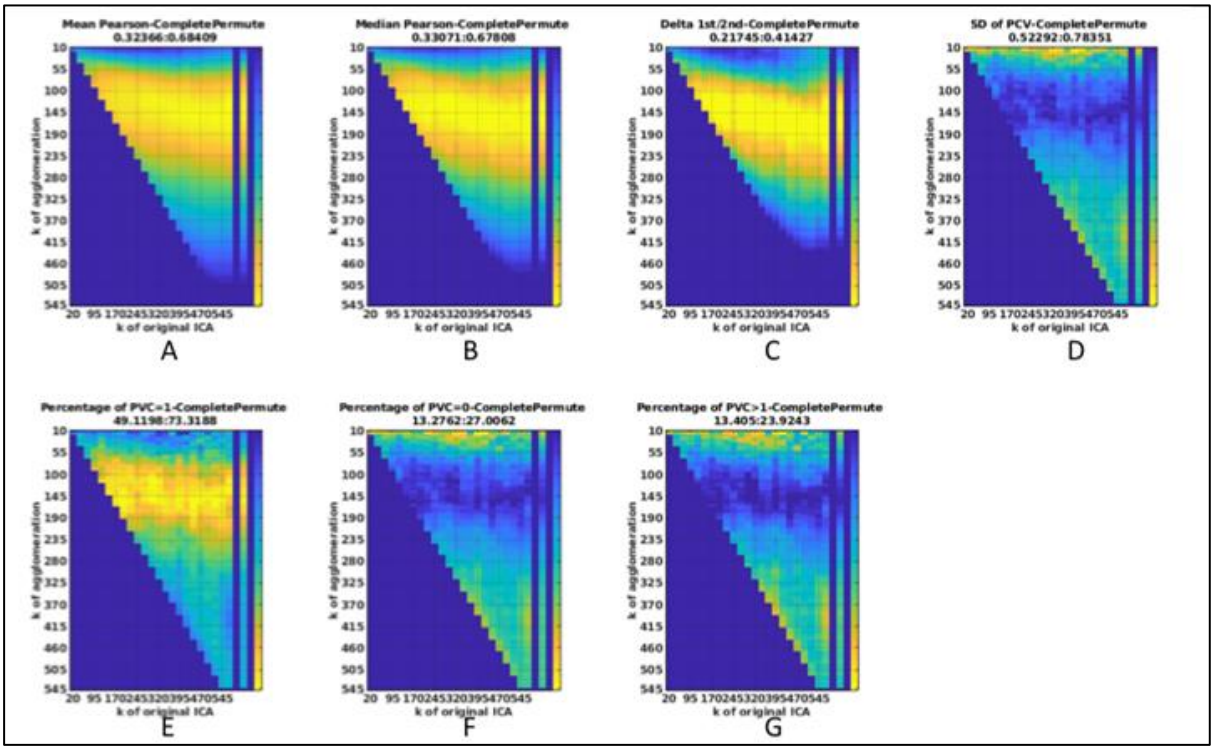


Figure 3.19: Half-matrices representing each one similarity metric of the discovery sample smoothed 10 mm. Within the half-matrix, each field represents the average of similarity comparisons with all other fields (i.e., the aggregation scheme 'permute') of that same row *k'*. The metrics are: A: Mean Pearson, B: Median Pearson, C: Delta 1st/2nd, D: SD of PCV, E: Percentage of PVC =1, F: Percentage of PVC=0, G: Percentage of PVC > 1. See table X for details on the metrics.

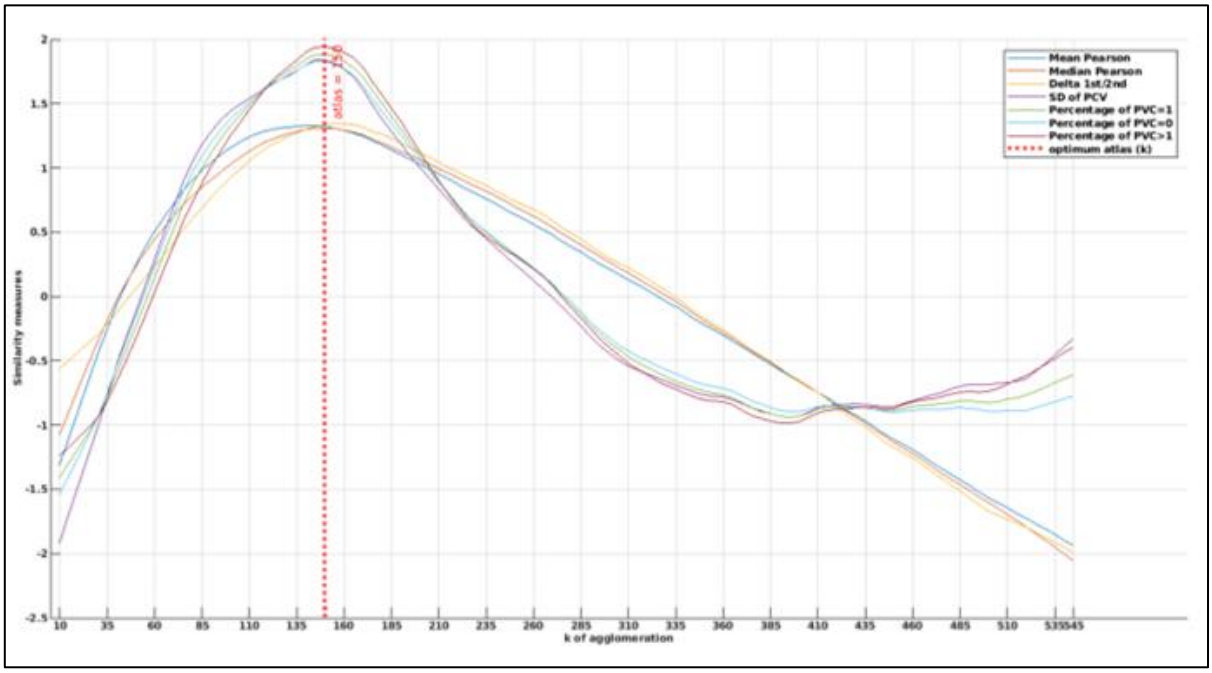


Figure 3.20: Profile plots of seven similarity metrics of the discovery sample smoothed 10 mm (aggregation scheme 'permute'). The vertical line shows the mean position of the seven peaks, locating to a rounded 150.

Table 3.6: Distribution of the peak positions of k' for each of the seven similarity metrics and three aggregation schemes for four samples. These samples were: discovery 6 mm, replication 6 mm, discovery 10 mm, and discovery 6 mm with combat-only)

Scheme	Discovery (6 mm)	Replication (6 mm)	Discovery (10 mm)	Discovery (6 mm) with Combat-only
Permute	135	140	140	135
Permute	140	155	145	140
Permute	135	135	140	110
Permute	145	155	150	145
Permute	145	175	160	170
Permute	150	135	140	115
Permute	150	165	160	165
Neighbour	155	175	165	170
Neighbour	120	135	140	105
Neighbour	140	115	150	155
Neighbour	180	175	150	175
Neighbour	140	130	150	110
Neighbour	145	135	150	165
Neighbour	175	180	150	180
Diagonal	135	130	160	115
Diagonal	145	130	150	160
Diagonal	175	180	145	175
Diagonal	135	130	150	115
Diagonal	145	165	150	170
Diagonal	175	180	150	180
Diagonal	135	130	165	120
Median (Permute)	145	155	145	140
Mean (Permute)	142.9	151.4	147.8	140
SD (Permute)	16.0	21.6	7.7	22.7
Median (Neighbour)	145.0	140.0	150.0	155.0
Mean (Neighbour)	147.6	150.0	150.0	146.4
SD (Neighbour)	16.0	21.6	7.7	22.7
Median (Diagonal)	145.0	140.0	150.0	155.0
Mean (Diagonal)	147.6	150.0	150.0	146.4
SD (Diagonal)	16.0	21.6	7.7	22.7
Median (Overall)	145.0	140.0	150.0	155.0
Mean (Overall)	147.6	150.0	150.0	146.4
SD (Overall)	16.0	21.6	7.7	22.7

Optimally dimensioned parcellation of the discovery sample with discrete and fuzzy boundaries

There are a lot of challenges for constructing the discrete atlas. Here, "discrete" means that one voxel is attributed to one and only one component. We have considered the degree of component stability in three different grades: a) no consideration of Iq values, b) using the conventional Iq value from ICASSO, c) tensor-clustering based Iq value that also considers instability of the subject loadings. We have considered an/the ICA solution with *k*=149 for comparing the discovery with the replication atlas with fuzzy boundaries where lower values indicating vague attribution (difference between first and second best Z-value) and higher values indicating clearer attribution. We have flipped the intensities due for visualization purposes. Here (**Figure 3.21**), broader bright bands indicate a larger area of uncertainty whereas thin bright lines indicate a certain distinction between bordering parcels. At the same time, dark areas indicate voxels/components with very low uncertainty.

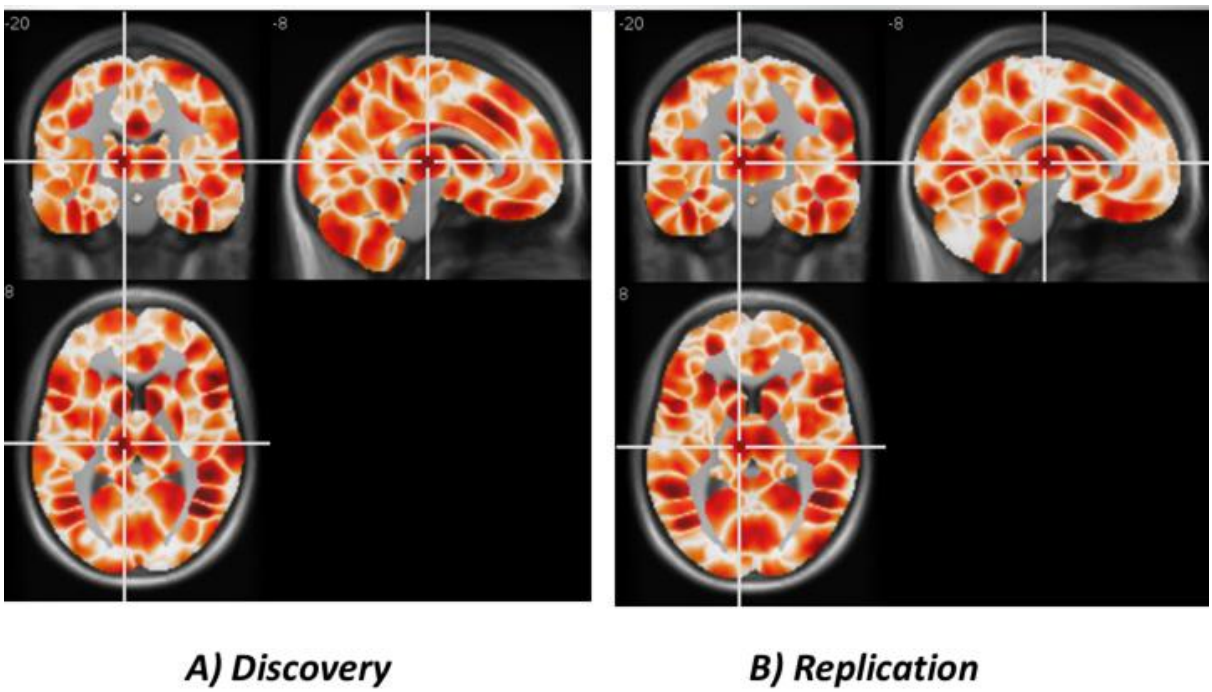


Figure 3.21: Discovery and replication sample atlas parcellation border images. In these depictions, we display high values for high ambiguity between the best and second best component Z-value, and low values for low ambiguity. Thin lines represent sharp, clear borders whereas broader bands represent areas more difficult to attribute to one parcel. Note that no separate parcels are shown in the strict sense, but rather ambiguity border information that underlies the attribution to discrete parcels.

Comparison between discovery and replication sample based parcellation

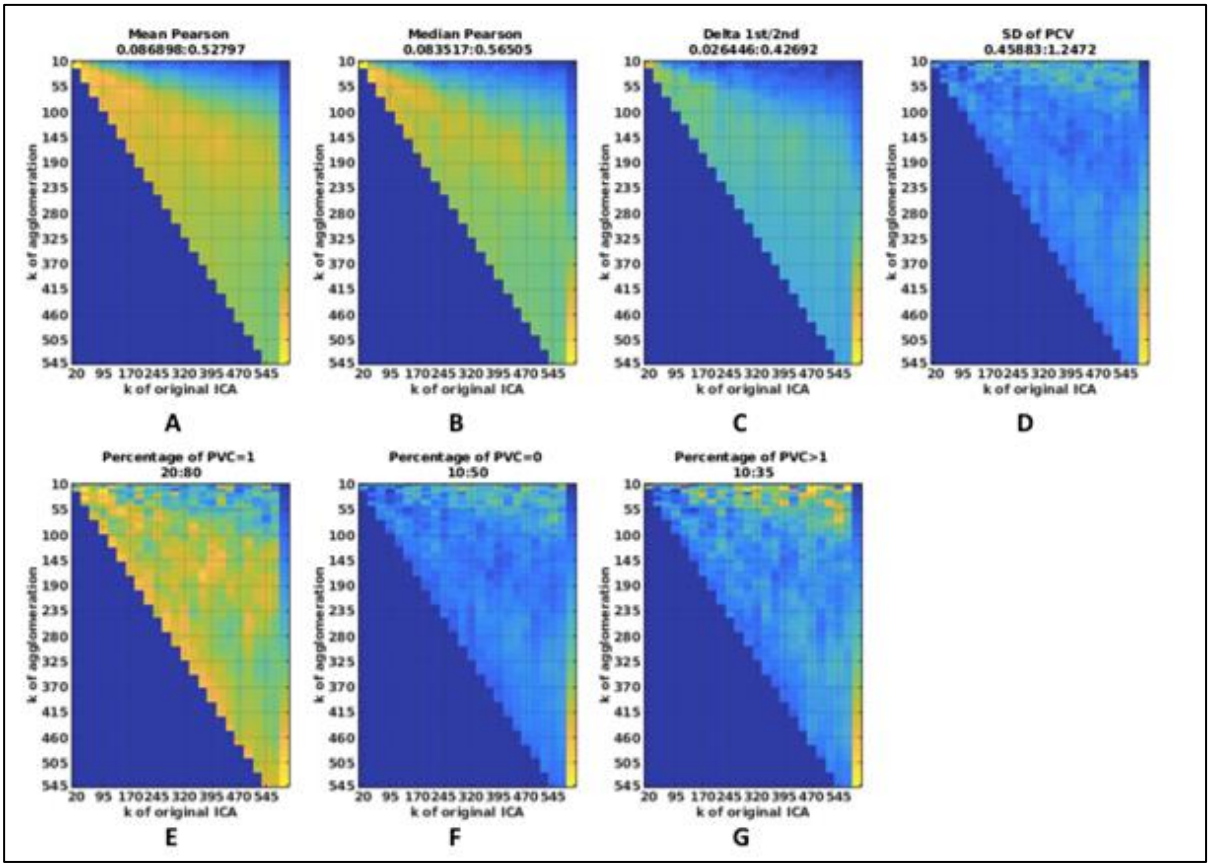


Figure 3.22: A-G Seven half-matrices, each representing one similarity metric, of a comparison of the discovery sample with the replication sample(A: Mean Pearson, B: Median Pearson, C: Delta 1st/2nd, D: SD of PCV, E: Percentage of PVC =1, F: Percentage of PVC=0, G: Percentage of PVC > 1).

We have also compared and investigated the similarity between discovery and replication sample with 6 mm smoothing. **Figure 3.22 A-G** shows the comparisons of corresponding parcellation of the *k*-by-*k'* (half-)matrix, including the original ICAs, using the same 7 metrics as between discovery and replication sample. The mean Pearson correlation between discovery and replication sample which assesses spatial component similarity (k=149 each) was 0.53 which is moderate. Most importantly, however, Figure 3.20 E depicts how many percent of the components in the replication atlas found exactly one corresponding component in the discovery atlas. This procedure accepts minor deviations between components as long as they bind together in comparison. Values were high throughout the diagonal line, dropped for agglomerations from higher parcellation, and showed a profile plot with a peak at 62 (**Figure 3.23**)

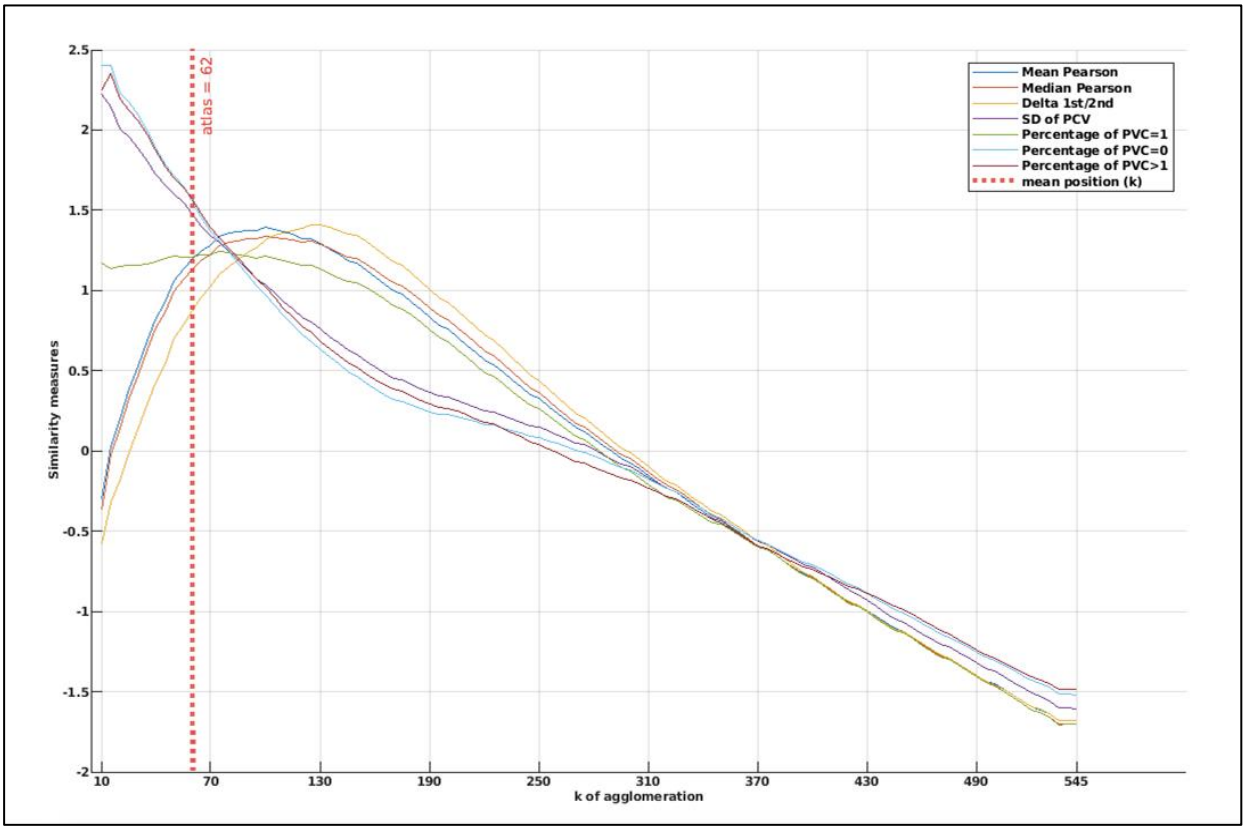


Figure 3.23: Profile plots of seven similarity metrics of the comparison between discovery and replication sample smoothed 6 mm. The vertical line shows the mean position of the seven peaks, locating to a rounded 62.

We have also visually compared the component similarity between discovery, replication with sorted and binarized resorted standard Iq values. Most of the components looks very similar in discovery and replication sample (**Figure 3.24** and **Figure 3.25**).

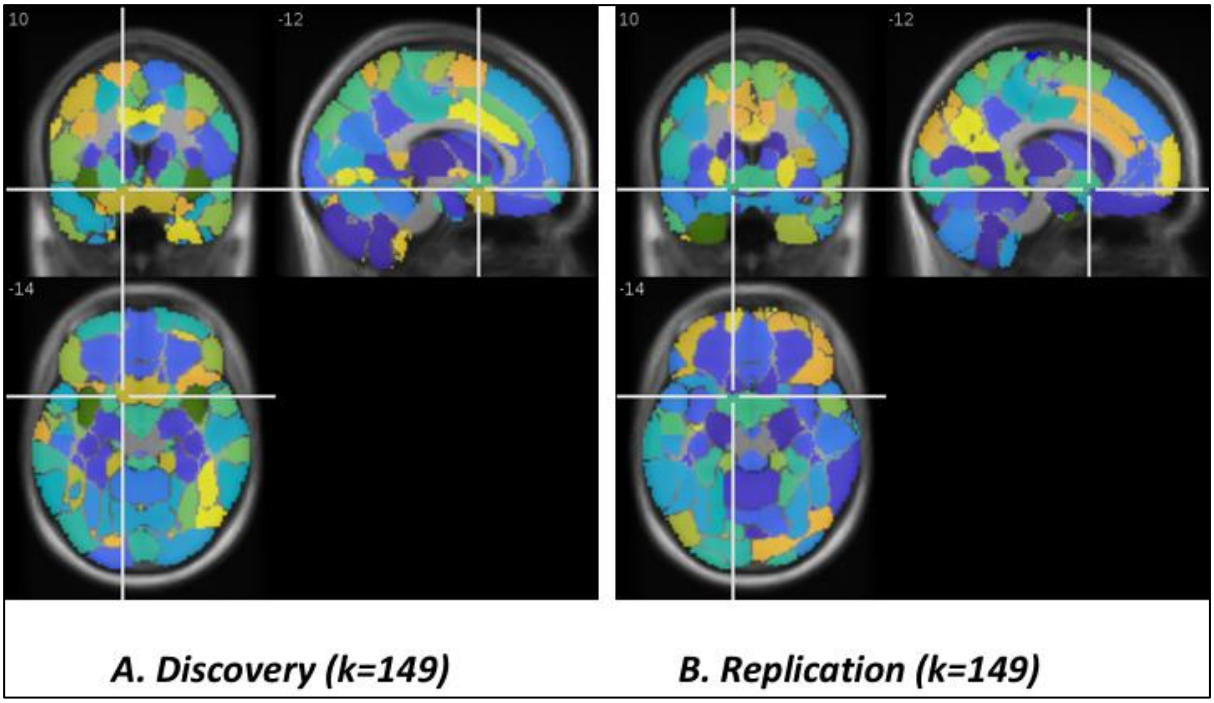


Figure 3.24: Discovery and replication sample atlas parcellation for k=149. Most of the components looks very similar between discovery and replication sample.

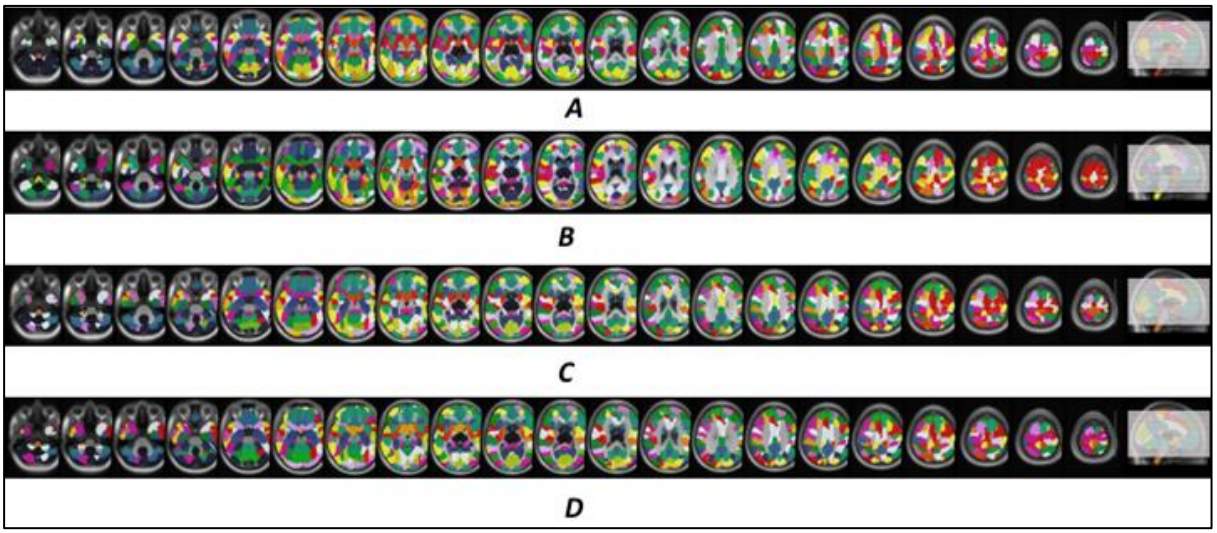


Figure 3.25: The representation of components maps for discovery and replication with a dimension of k=149 in a multi-sliced plot ;where A represents multi-sliced plot for discovery sample, B represents multi-sliced plot for replication, C represents multi-sliced plot for replication with sorted Iq values and then D represents multi-sliced plot with binarized and sorted Iq values.

Effects of residualisation on age (and other covariates ICV, sex, sex-y-age interactions)

Not performing the residualisation as explained before, led to lower Delta-values and seems to degrade the atlas quality. Voxel-wise age effects seem to put noise on the system.

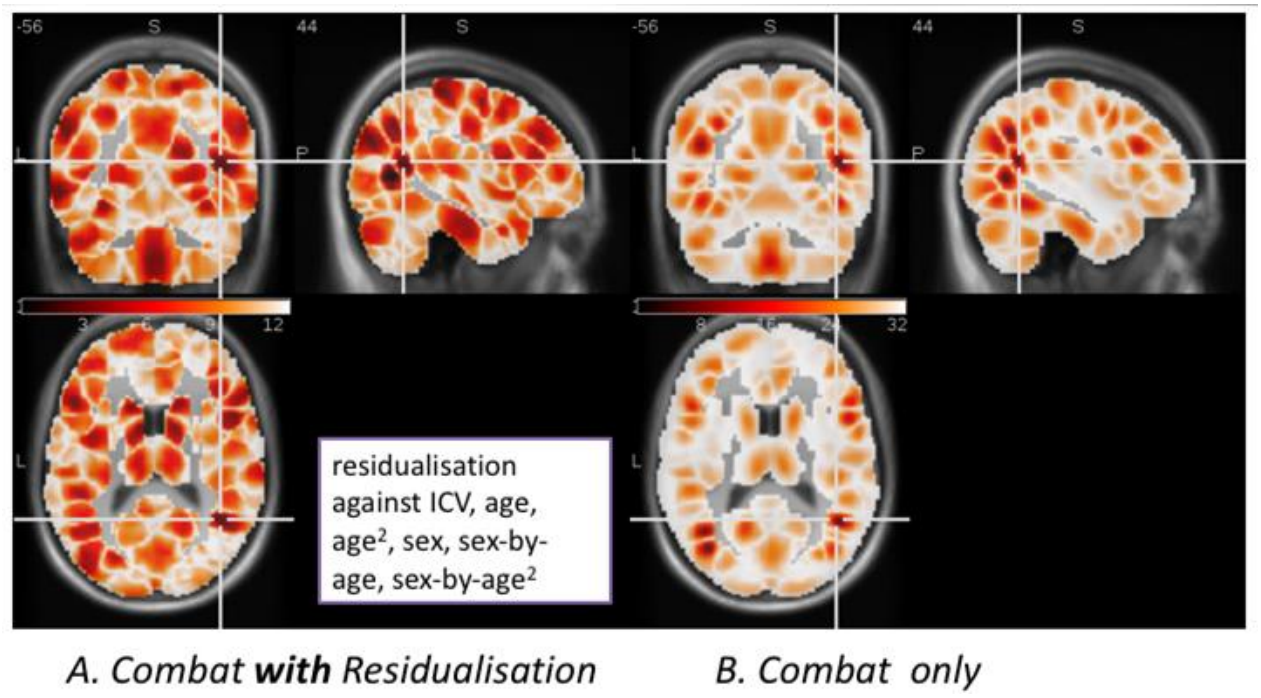


Figure 3.26: Effect of residualisation on component formation. Combat-only without residualisation demonstrates larger areas of ambiguity and generally higher values.

Effects of smoothing

The degree of smoothing applied (smoothing kernel 6 mm^3 and 10 mm^3) to data is inversely proportional to the range of the standard deviation of the position of number of components. 10 mm atlas is clearer so less number of small island around, and the borders sharper in some areas, so the atlas is more stabilized with higher smoothing kernel (**Figure 3.27**)

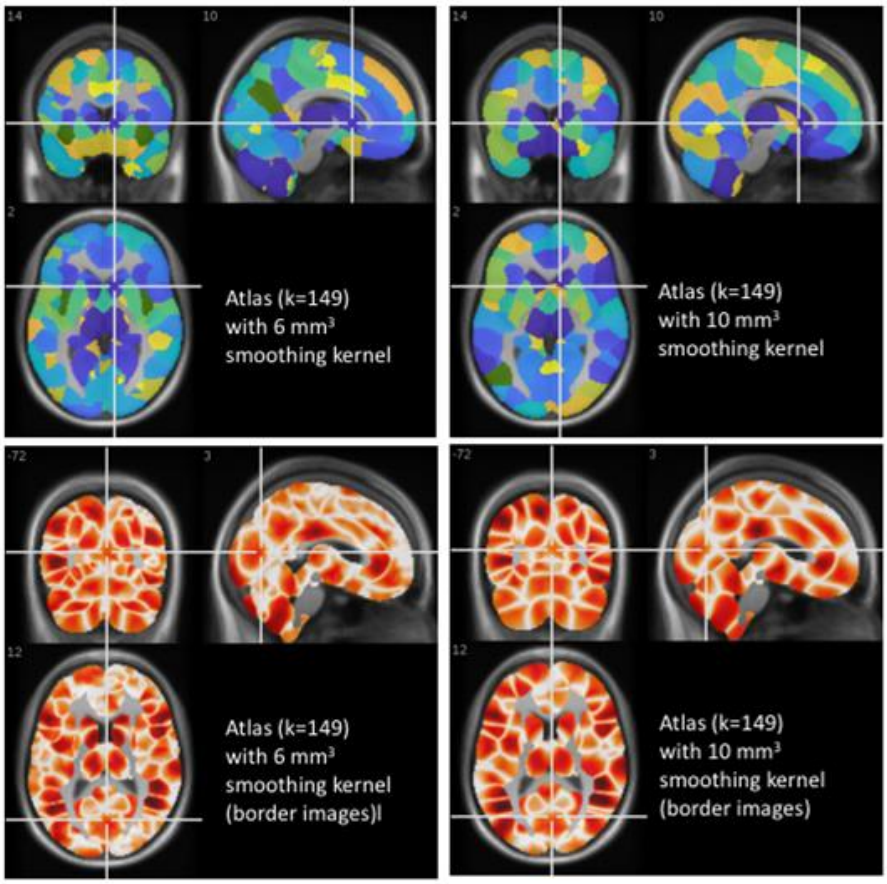
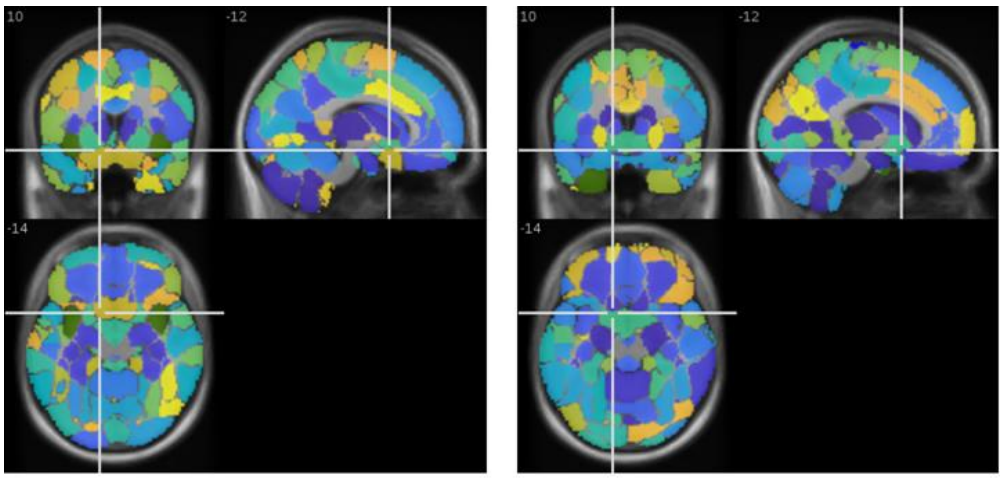


Figure 3.27: Effect of smoothing on component formation. Atlas and border images with higher smoothing kernel demonstrates smaller areas of ambiguity and less small changes around the atlas which helps to develop more stabilize parcellation framework.

Finally, we have constructed the atlas using K=149 and re-agglomerated them to 20 components to visualize each component with clear and separate color.



A. Discovery (k=149) *B. Replication (k=149)*

Figure 3.28: Atlas with k=149 parcels but re-agglomerated in 20 parcels for discovery and replication sample.

3.4 Discussion and Outlook

Summary of concept and results

We have built a data-driven parcellation framework and used it to dissect GM volume maps of the brain into a set of about 150 brain areas. As method to identify groups of voxels that behave similarly across the population we used ICA, more precisely, ICASSO, a variant of ICA that repeats several runs of ICAs to calculate stability values from this. As examples of the population we work on two about equally sized (563 and 566) sets of GM images of healthy subjects. The mixing matrix holds information of the contribution of individuals to the sources, and morphometric analyses using these values are referred to as source based morphometry (SBM) [156]. Still, we continued differently with the ICASSO solutions in that we performed across a wide range of components (20 to 558): We used this multi-ICASSO framework to feed a re-agglomeration step and then compare re-agglomerated components systems to find out the optimal number of 'true' components in the VBM-GM data. This computationally expensive approach was used in four different datasets to investigate and validate this optimal number, or at least range, of components.

Our main findings were as follows:

1. We detected a stable range of **145-150 GM volume components** that were purely ‚data-driven'(**Table 3.6, Figure 3.16, Figure 3.17, Figure 3.29),** identified by the combination of ICASSO, re-agglomeration and similarity comparisons.
2. This dimensionality median was stable across two samples and across two different spatial smoothing levels. Ranges were larger (still with a similar center) in the replication sample and when skipping the covariate correction step (see **Table 3.6**).
3. Their spatial patterns were reasonably similar, depending on the exact type of analytical approach. **(Figure 3.21)**: While re-agglomerations with k' very different from k were not similar, the similarity was higher when k' was close to k (see discussion below). Ignoring the hierarchical tree structure and focusing on the width of ambiguous border zone voxels revealed very good comparability.
4. Hierarchical clustering plus similarity analysis served as useful approach to determine the dimensionality in ICA. Though suggested in a simple form as 'blocked ICA', its systematic extension and combination with elaborate similarity comparisons has not been reported before.
5. Our approach allows to break down VBM-GM data in the sense of an anatomically interpretable dimensionality reduction. The approach may also be useful for non-

imaging related, high dimensional data such as genetic data, epigenetic (e.g., methylation) data and transcriptomics.

6 The residualisation step to remove linearly estimated effects of ICV, age, age2, sex and ethnicity had an impact on the border zones that – while still showing the same basic parcellation pattern - were less definite.

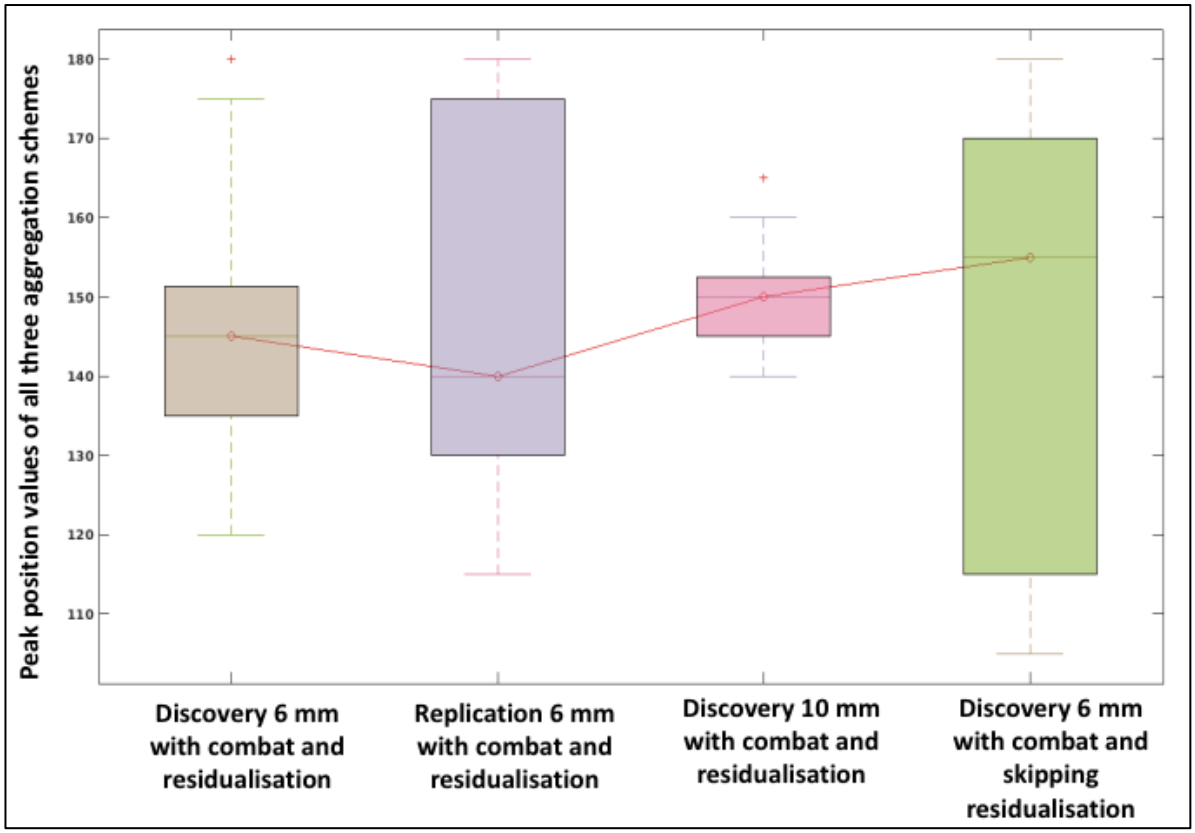


Figure 3.29: Distribution, median and quartiles of the optimal number of components of all 4 samples. Values per boxplot are the 21 peak position values (7 from each of 3 agglomeration schemes). Note smallest range in the Discovery 10 mm sample.

ICA variants and other clustering tools as alternative - defense of the choice made

We constructed and evaluated VBM-GM atlases using both the discovery and replication sample by employing ICA as main method to detect components in the data, followed by hierarchical clustering and similarity analyses. Various alternative clustering tools are available for dimensionality reduction and component analysis: K-means is perhaps the most widely used vector data clustering technique. It consists of an alternative optimization of (1) the assignment $u_{k-means}$ of samples to a cluster and (2) cluster centroid estimation. The inertia, i.e. the total of squared differences between samples and their representative cluster centroid is minimized. In K-means the clustering of sMRI data without explicitly considering their spatial structure, although before the clustering spatial smoothing, can provide for spatial regularization indirectly. K- Means clustering holds some inconvenience since it fails to

contemplate any information regarding spatial structure and is therefore sensitive to noise and other imaging relics, such as intensity inhomogeneity. It can also lead to a sub-optimal local solution due to poor initialization. Hierarchical clustering, which we used here only as a secondary methodology, can be also considered as an alternative[188]. These processes begin with every singleton {j} cluster xj voxels. Instead of measuring the Euclidean distance directly, it analyzes the variance of clusters. *At each iteration, a pair of clusters is selected according to optimal cluster selection criterion.* This process gives rise to a hierarchy of clusters that are depicted as binary tree, also often referred to as Dendrogram[188] in which each non-terminal node is related to the clusters of its two children. The variance-related approach to Ward's algorithm is used most commonly [189] among various hierarchical agglomerative clustering procedures [190]. Agglomerative clustering is the most general method of hierarchical clustering for grouping objects on the basis of similarities. Each object is viewed as a singleton group, which begins by the algorithm. Second, pairs of clusters will be successively fused into a large cluster, which will hold all objects, before all clusters are united. The effect is a tree-based image of the objects, known as Dendrogram. The ward algorithm also does not include any spatial structure information in the process. Although, we have used the agglomerative hierarchical clustering approach to fuse the higher component solution to lower component solutions but this can be alternative clustering approach for directly obtained brain parcellation[191]. For generating the component map, we have used Infomax ICA which not only contains all the spatial information but also provides robust solution[156]. Fixed-point based FastICA[192] and max mutual information based Infomax ICA [46,48] both have been widely used in many sMRI related studies but in our study we have used mainly due to its stability advantage[193].

Comparison with other brain parcellation methods

VBM based GM maps were used as input data and Infomax ICA along with hierarchical clustering based re-agglomeration scheme were used for finding components. In the following we will discuss other established brain atlases. A formal comparison between the here developed GM volume based data-driven parcellation and publicly available parcellation [194]is challenging in several ways: First, not all parcellation are available as volume representations but in a vertex/surface format. We focused on atlases available in volume space, due to VBM being clearly bound to a 3D-grid-framework. Second, several parcellation contain only cortical (with or without the cerebellum) but no subcortical areas. As our parcellation over the entire GM space, we compared the development of our atlas parcellation scheme with Glasser Atlas [195](based on multimodal and contains T1w images) and Shen atlas[196] (based on spectral clustering on resting state fMRI datasets)

Glasser Atlas (Glasser et al 2016): The complexity of the human brain cortex requires a map of its major subdivisions for many subsequent applications in imaging. Using the multi-modal MRI Human Connectome Project (HCP) images and an objective neuro-anatomatic semi-automatic approach, Glasser et al. presented 180 areas per hemisphere in a precisely aligned group average of 210 healthy young adults, using sharp changes in the cortical architecture, function, connections, and topography. They used T1w and T2w structural images, task-based and resting state-based fMRI images, diffusion-weighted images, and b0 field maps to generate a cortical parcellation generated from multimodal images of 210 adults from the HCP. They developed a semi-Automated quantitative method to detect transitions representing candidate's real boundaries, adapted to data on multimodal neuroimaging (T1w and T2w structural images, task-based and resting state-based fMRI images, diffusion-weighted images, and b0 field maps), based on gradient-based parceling, on two-dimensional cortical surface models. A trained machine-literature classifier was used to identify the multimodal fingerprint for each cortical area to enable automated delineation and identification of these subjects for new HCP topics and future studies. This classifier identified the occurrence, replicated group parcellation, 96.6 percent of cortical areas in new subjects, and could correctly find parcels in individuals with atypical parceling. Cortical regions were delineated concerning function, connectivity, cortical architecture, topography, and expert knowledge and meta-analysis results from the literature. The similarity was that both atlas parcellation framework was a model-driven approach and Glasser also included T1 weighted images in the atlas parcellation scheme. However, Glasser atlas focused on cortical areas but did not include any subcortical areas where our VBM-GM was based on grey matter maps and covered the full brain, including cortical and subcortical areas.

Shen Atlas (Shen et al. 2013): A group-based parceling approach was used to define network analysis nodes. They defined a number of nodes as the input to the first instance, and the purpose was to investigate network-theory-based analyses. The replicability of parcellation in every sub-unit was calculated and shown to be high among several groups of healthy volunteers. The proposed approach was then applied to real resting-state fMRI data, and the whole-brain parcellation results are shown along with reproducibility maps. For 200 subunits (102 L, 98 R), they used a spectral clustering approach to compute a volumetric group-wise parcellation based on an optimization process that guarantees functional homogeneity within each parcel that computed parcels are consistent across subjects. Volumetric parcels from the provided 1 mm sampled 268-parcel atlas are projected to the cortical surface. The similarity between the Shen atlas framework and our atlas framework was that Shen performed whole-brain parcellation to

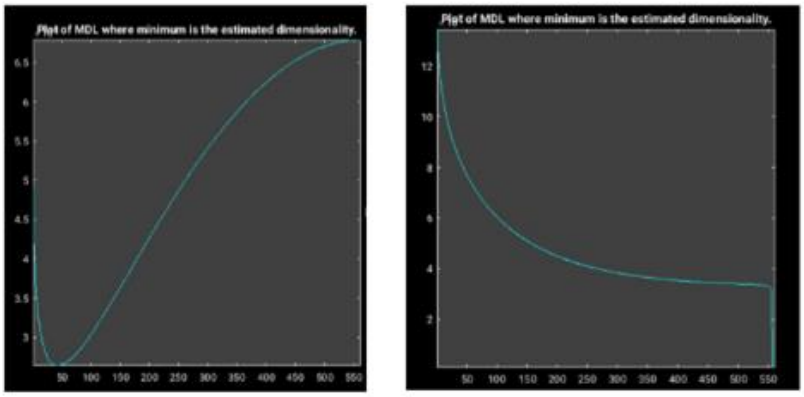
build the atlas, and our data-driven atlas was also done on the whole brain. Shen used a group-wise multigraph clustering algorithm to produce subunits with homogeneous temporal patterns and build a data-driven parcellation framework to finalize the atlas' optimal number the atlas. We have also built and data-driven atlas framework to finalize the number of parcels for our VBM-GM based atlas. However, Shen atlas was developed using resting-state fMRI data, and our data-driven atlas was build using grey matter maps.

Discussion of user-dependent settings in the pipeline smoothing levels

Smoothing effects: We have found significant differences in the median of the optimum component solution between 6 mm and 10 mm solution (both applied to the discovery sample), but the range of peak positions was clearly lower in the 10 mm smoothing version (**Figure 3.28**). However, increasing the spatial smoothing from 6 mm to 10 mm increased the proportion of stable components over the entire range of k **(Figure 3.27).** While comparing the atlas between 6 mm and 10 mm smoothing Comparison of we have found that 10 mm discovery showed more stability in border patterns than 6 mm smoothing (**Figure 3.27**), and the range for the standard deviation of the peak position of the atlas was also much narrower than 6 mm. Spatial smoothing increases the signal to noise ratio, which improved the stability in the brain atlas. So smoothing played an essential role in the data-driven atlas parcellation framework.

Finding the optimal dimensionality of ICA: alternative methods

A problem that comes naturally with clustering algorithms is choosing the number of clusters to be used in the model. We applied four methods among the methods considered as standard in the field to find the optimal dimensionality of an ICA. These are Minimum Description Length (MDL), Bayesian Information Criterion (BIC), Cross-validated likelihood, Bootstrap with similarity



A. Non-residualized Replication B. Residualized Replication

measurement

Figure 3.30: MDL plot for estimating number of component using replication datasets. We noticed a sharp minimum using non-residualized dataset. However, we don't see any sharp minimum after using the residualized datasets.

Component estimation using MDL from non-residualized and residualized replication dataset

In essence, we found that minima at 43 (SBM) for the discovery sample. After residualisation of the raw dataset, the MDL plots showed no local minimum but a sharp drop to zero (with high artificial minima reported). The pattern was similar for the replication sample, with 33 and 56 components for the raw data, slightly fewer after Combat, and no stable results after residualisation. Residualisation may cause to overfit the data, and they caused a wholly unreliable and deficient number of estimated components. We found a low number of estimated components using another preprocessing scheme, which might indicate the non-gaussian characteristics of the MDL algorithm were not entirely suitable for the datasets (**Figure 3.30**)

Table 3.7: Dimension estimate using MDL for discovery and replication samples in different versions. MDL was estimated in GIFT-SBM designed for VBM data.

Sample	Preprocessing	Method	Minimum	Comment
Discovery	None (raw)	MDL in sbm	**43**	Sharp minimum
	Combat	MDL in sbm	**43**	Sharp minimum
	Residualization	MDL in sbm	558	Sharp drop
	Combat & residualisation	MDL in sbm	4	Yet strange and with sharp drop at 550
Replication, N=564	None (raw)	MDL in sbm	**56**	Sharp minimum
	Combat	MDL in sbm	**22**	Sharp minimum
	Residualization	MDL in sbm	5	Not reliable / sharp drop at ~550
	Combat & residualization	MDL in sbm	5	Yet strange and with sharp drop at 550

Component estimation using ICL from non-residualized and residualized replication dataset

ICL: In separate experiments, we have also estimated the dimensionality using ICL. The **ICL criterion**[106] is an alternative to **BIC**. The Bayesian Information Criterion (BIC) is a method for estimating an optimal number of component. It is appropriate for models fit under the maximum likelihood estimation framework.

$$\boldsymbol{BIC} = \boldsymbol{-2 * LL + \log(N) * k}$$

Where log () has the base-e called natural logarithm, LL is the log-likelihood ration of the models, N is the number of subject and k is the number of parameters in the model. The score as defined above is minimized, e.g. the model with the lowest BIC is selected. ICL equals to BIC plus penalty.

Up to now it has been widely presented as a penalized likelihood **criterion**, which penalty involves an "entropy" term.

$$\boldsymbol{ICL = BIC + pM,}$$

where pM > 0 is the penalty applied to the likelihood L of model M.

ICA was not implemented in GIFT-SBM toolbox so we needed to calculated it outside of GIFT-SBM toolbox (**Table 3.8**)

Table 3.8: Dimension estimate using ICL for the discovery and replication sample. Note that again the raw samples and samples after, combat-only, residualisation and after both correction steps were used

Sample	Preprocessing	Method	Minimum	Comment
Discovery, N=563	None (raw)	ICL	**445**	Sharp minimum
	Residualization (raw)	ICL	**445**	flat
Replication, N=564	None (raw)	ICL	**420/445**	Sharp minimum
	Combat	ICL	**445**	Sharp minimum
	Residualization	ICL	**No minimum**	flat
	Combat & residualisation	ICL	**420/445**	flat

We have used PCA data matrix as input. After preprocessing, PCA was performed on data matrix [subjects × voxels] of the discovery ample. We have then used the score matrix (subjects-1 × subjects) obtained from PCA and fed this into "**mclustICL**"[197] function to calculate the ICL value. The estimated dimension was that of the relatively (locally) lowest ICL value.

Importance of reliability analysis using ICASSO

We showed how ICASSO could estimate the robustness of independent components. On the one hand, it is clear why reliability analysis is needed; any data findings should be based only on reliable components. On the other hand, reliability analysis provides an interesting "data mining" tool; it highlights some major stable components and suggests that the remaining unstable components can be excluded from the study. This reduced dimension can significantly lower the computational cost of the analysis and only consider robust and reliable components between dozens of components given by ICA. In principle, we can use ICASSO without reducing dimensions just as a weight for the components (multiplicative), where higher Iq will generate more weight for a robust component, whereas lower Iq creates less weight for the unstable component. PCA is the initial step in the ICA process. In practice, PCA significantly improves the quality of the ICA results as it reduces the level of noise. If we can already generate robust and reproducible components from PCA, then the overall reliability analysis in ICA will be more substantial and robust.

A novel approach in similarity analysis for atlas parcellation framework in terms of estimating the optimal number of components

As for the calculation load, the ICA with multiples run with hierarchical clustering in the current computational environment is the bottleneck. The computer load as a time consumption increases as a cubic function of the number of estimates concerning the current implementation. Therefore, more sophisticated implementation, e.g., ICA and estimation of number components using cross-validation, may be necessary for the larger number of voxels. We are still investigating the alternative clustering methods over ICA. Since we did not use the pre-estimated number of components by MDL (built-in GIFT) or ICL approach, we introduced a new similarity measurement pipeline with hierarchical clustering, re-agglomeration method, and seven different similarity metric to produce a full and sparse component map. Clustering in the high-dimensional signal space requires the determination of the number of clusters to be modeled. The theoretical problem is challenging to automatically determine optimal values for these parameters; ultimately, optimal values also depend upon application-specific and subjective considerations. Therefore, we use seven different similarity metrics to finalize the optimal number of components for our data driven atlas. Despite computational complexity, this similarity framework offered a stable solution over different cohorts.

Limitations

The time complexity of an algorithm is the total time required by the program to run till its completion. The time complexity of the algorithms is most often expressed by the big O-notation. The complexity of time is an asymptotic notation. This is most generally measured by counting the number of elemental steps performed by every algorithm to end execution. Time complexity in ICA is a major issue when we are considering reproducibility and repeatability of the components. First, we have performed 23 different ICASSO analyses for each of four samples. To increase the generalizability of the algorithm we have used 20 runs for each of the ICA solution.

The computational complexity of ICA (ref) is given as:

$$O(2md(d+1)n),$$

where d is the number of variables/dimensions, n is the number of samples, and m is the num ber of iterations. If we increase the sample size or we increase the repetitions, the computational complexity will be high[198,199]. Second, we have used similarity measurement

(Pearson correlation and six other measurements) to validate the component similarity. The computational complexity of Pearson correlation is

O(nlogn),

where, n is the number of samples. So, if we increase the sample size then the time complexity will increase. In our study we have computed 23 ICAs with 20 runs for each of them. So, all together we have performed 460 ICAs and then for the similarity comparison we have calculated more than 3 million component comparison analysis. This computational load required high performance computational system (HPC). Here we have used our HPC with several cores (>20) and 100 GB RAM for each of the run, along with parallel processing to compute the ICA, re-agglomeration and similarity measurement. All these calculations are only necessary in new studies, however, if a new sample-specific atlas should be produced from scratch. Possibly, as the range of optimal dimension seems to be round 150 components, for new samples sparser schemes could be set up based on this knowledge.

We used IXI sample which a multi-site cohorts including 1.5 Tesla and 3 Tesla scanner. Although we used Combat approach to remove the scanner effects from the data but 1.5 Tesla images are not fully comparable to 3 T datasets of IXI datasets and our replication cohorts. We still detected variability between discovery and replication which possibility indicated that with larger samples the atlas stabilizes more (e.g. both samples, 10 mm). Edge-preserving smoothing exists as alternative could be of advantage for smaller structures (e. g. hippocampus) that are smeared otherwise. Our atlas parcellation framework was only validated using health subjects but was not clinically validated if more sensitive to detect disease effect that e. g. a random parcellation, or other known atlases. Result atlas neither described nor compared anatomically to other atlas systems. This study was mainly focused on the algorithm or building model derived atlas parcellation scheme than its anatomical and clinical validity. Resulting parcellation were visualized and explored regarding symmetry features and similarity with known functional networks.

Outlook

Data-Driven atlas utilization for clinical comparisons

According to our goal we have built a GM map-driven atlas from healthy control samples. The resulting atlas with a dimension optimized for both samples can be written out in a discrete form and used in independent projects. It is ideally used in VBM-like studies because of this type of data best match the granularity of the atlas, but also altered functional patterns can be analyzed

using this atlas. . As indicated in the limitations, the use of a larger sample smoothed with 10 mm could reveal an even more generalizable atlas. The already developed parcellation, however, seems valid enough to study disease effects, aging effects, or to use it as sparse GM data representation for machine learning algorithms.

Data reduction technique for GWAS study of the whole brain:

This method can be also applied to voxel-wise GWAS study for the whole brain. Stein et al. (2010) have performed voxel-wise tensor-oriented morphometry to measure individual differences in brain structure at the voxel level in healthy subjects and then performed genome-wide association studies (GWAS) for each voxel. They discuss about this developed method (vGWAS) *by studying the most related variant in each voxel to address the multiple comparison problem and computer burden associated with unprecedented amount of genetic data.* In such scenario, our approach could help to reduce the dimension and generate component maps which will reduce the computational burden of the GWAS analysis and improve its interpretability.

Correction with functional connectivity analysis and fiber-tracking

This parcellation framework can be used to create ROI systems for fiber-tracking and functional connectivity analyses, for example to prove the „firing together/wiring together" hypothesis (Hebbian Theory).

Transfer of analysis framework to other voxel/vertex-wise measures

Here, voxel-wise GM volumes were studied. This measure, geometrically, is influence both by cortical thickness and surface area (what regards cortical areas), and also subcortical volumes have shape features that may be annihilated. Thus, the parcellation framework can be used in a similar way to map, for example, cortical thickness. Possibly, volume based networks differ from purely cortical thickness networks. As many studies use FreeSurfer to calculate regional cortical thickness and provide meaningful clinical results (ENIGMA-overview paper), a pure data-driven parcellation of the cortical thickness may be useful as an alternative parcellation compared with mostly used DESKANY atlas.

3.5 Conclusion

First, we have built a VBM-GM based data driven brain parcellation that contains cortical and

subcortical GM. Our data-driven approach that employed multiple runs of ICA with systematically varying dimension, re-agglomeration and similarity analysis to find an optimal dimensionality, helped to build such atlas without any prior assumption of the number of ROIs, or any other external knowledge. With a typical smoothing of 6 mm found a very similar dimensionality round 150 components across two larger samples of healthy subjects. We also compared the similarity of the resulting parcellation of the two samples as indicator of robustness and generalizability, finding well-matching component pairs across samples, with moderate spatial overlap of the components in detail, but very good overlap of the separation lines between components. More generally, we have developed a new, computationally enriched engine for multi-ICA, re-agglomeration and similarity measurements to define the likely dimensionality of VBM-GM data as typically used in clinical studies. Some elements of the pipeline may also be transferred to other clustering challenges (e.g., genomic data, or clustering the subject space).

4 General Discussion

4.1 Similarities and differences between the two projects

In both studies, the objectives were different; however, we applied the same approach to achieve the goals, which was unsupervised learning. So, the similarity between the two projects is the model derived or unsupervised learning approach. The unsupervised models we discussed in the first project used clinical variables to detect treatment response dynamic classes, and then we used supervised models to predict those classes of depression treatment response. A mixed model-based longitudinal clustering method includes or finds meaningful clusters within heterogeneous patient samples in the first project. In the second project, we used a clustering method to find robust components for developing a data-driven atlas parcellation framework. We have performed external validation in two studies, thus not risking that the model will fit well to the dataset used for discovery and replication and extrapolated to other patients. However, the difference between the two projects is that we have used longitudinal clinical datasets for the dissection of patient space in the first project, and as a result, we have detected seven treatment response subgroups. In the second project, we have used cross sectional imaging datasets to dissect voxel space into components/networks.

4.2 Extension of longitudinal clustering to multiple symptom development trajectories and to polygenic response scores

By using multivariate machine learning approaches, we investigated latent response subtypes by using the MARS dataset (N=1071 [61,73]). These subtypes were validated using a stable and conservative multivariate prediction model (random forest). In summary of the above presented study (chapter 2), model-based clustering identified distinct, clinically meaningful and stable TRD classes in MDD that were predictable from clinical baseline characteristics. As one outlook, conceptually, such clustering of patients with longitudinal Hamilton Disease Rating scale data can support larger studies on the neurobiology of treatment response. The latter might also apply to multicenter designs, as we showed that the transfer of the clustering model to an independent patient cohort worked well. Given the many clinical symptom profiles MDD may have, it is a certain limitation that we have investigated the treatment response dynamic characteristics only on the basis of the HDRS sum score, collapsing the information of individual items. Expressed differently, our approach assumes as a latent hypothesis, that differences between symptom profiles make no difference: For example, a patient with sleeping problems

and anhedonia as main problems with a sum score 21 is not handled differently than a patient with suicidality and anxiety, also reaching a sum score of 21. Still, both patient (types) may show different response dynamics. Thus, one future prospective of this approach could be to expand the clustering to the whole matrix of HAM-D single items implementing multivariate symptom-based clustering using the same model based robust clustering method. This extended data-driven approach would investigate the use of clusters of symptoms on the 21-item Hamilton Rating Scale for Depression (HAM-D-21) to define the symptom clusters and their development over the treatment period and respect, as first step, respect clinically distinct symptom profiles and their development over time. While a wide range of individual patient data was used in the current study already, such analysis and replication based on all single items may be challenging. Sample size also played an important role in terms of detecting robust and clinically meaning clusters. If the sample size is significantly low, then the detected clusters will not able to replicate in the different cohort. We have also performed jackknife with 1000 repetitions which is conceptually simpler than bootstrap and also performs well for confidence interval for pairwise agreement measures. The post-hoc nature of the analysis is further limited by studies of different protocols, which can be difficult to combine. The results of the cluster analysis may be affected by differences in study design, length, number of treatment arms or other factors. The seven-cluster solution may have been influenced by imbalances in the proportions of patients within the clusters and consequent differences in clinical features or contributing towards differences in efficacy results[73,200,201]. Thus we have performed a follow up approach will be to demonstrate that throughout the course of an episode symptoms develop in individual patterns. For these symptom (class) specific patterns we aim at calculating polygenic correlates, using results from large GWAS studies on depression. Eventually, this allows to calculate polygenic response predictors per symptom class. As a preliminary analysis we have developed model derived item clustering framework and found four stable clusters in symptom (item) space. We detected these items by dissecting depression symptom space into factors (model based cross sectional clustering using *"mclust"* package) and in future work these item clusters can be used to build symptom class specific response clusters by performing 3D-model-based clustering on trajectories of these factors over 16 weeks, similarly as in recent work[73], yet by a variation of the algorithm that considers several trajectories at the same time. These four clusters in symptom space were characterized by their baseline severity of core symptoms or anxiety-related symptoms. This will help use to detect not only the overall treatment response dynamic clusters but also symptom based treatment response dynamic clusters.

4.3 The problem of solution stability: when is an atlas parcellation final?

We present a data-driven framework for the development of a VBM-GM atlas with cortical and subcortical areas in order to obtain (1) provides a new perspective for the application of the data-driven scheme rather than traditional methods, (2) uses the large-scale heterogeneous data with multiple to achieve more specific probabilistic atlases than by using the single group and single site data in the form of individual variances, (3) demonstrates the advantages of using a large-scale scheme to generate robust data-driven VBM-GM atlases, (4) requires high computational resources for re-agglomeration of multi-ICA pipeline (from higher number of component map solution to lower number of component map solution) and similarity measurement between component solution. So, we found that there were spatial differences between atlases all though there were already very similar. The proposed method is able to be a baseline in many algorithms and applications for medical images because of its higher accuracy and low computational costs.

We found an influence of the smoothing kernel on the final atlas. Spatial smoothing decreased the degree of ambiguity clusters whereas the degree of robust and large cluster in the parcellation framework increased with the increasing full width at half maximum of the smoothing kernel.

We also aimed to investigate more about whether ICASSO with ICA will be robust enough to build data driven atlas in irrespective of cohort type. Instead of performing ICA multiple times and measuring inter cluster similarity coefficient for finding robust component we could also think of building nested cross validation framework with ICA where model could be evaluated on test data so that a cross-validation process with multiple folds and permutations. In each case, the model was learned by the training set (i.e., a random subsample of n fold of the data), including the estimation of clustering and the fitting of the model, the log-like value calculated from the test data is then summarized in parcels in order to give a unique amount (nested cross validation of log likelihood) in the test data. Finally select the optimal number of parcels for the based on the integrated completed likelihood obtained from the nested CV pipeline. However, in our current study, this approach will be highly expensive due to time complexity and lack of sample size for the replication datasets.

In this study for clustering, we have used seven different measures of cluster ‘goodness’ or quality to find optimal number of clusters (parcels) in the atlas framework. These type of measures allow us to compare different sets of cluster solution without reference to external knowledge and is called an internal quality which is used as a measure of ‘overall similarity’

based on the pairwise similarity of components. However instead of using seven different similarity metric we could also have calculated one similarity matric. We have use re-agglomeration scheme to fuse the component map from high dimension to lower dimension using hierarchical clustering and then measure the similarity using these fused atlases. In hierarchical clustering, a Dendrogram is the graphical representation of a cophenetic metric[202]. So alternative approach could be comparing the similarity between two Dendrogram (pairwise comparison) using cophenetic metrics and then finalize the optimal numbers cluster or optimal parcellation solution.

As a final atlas, we could provide it in two different ways; firstly, we could present the atlas solution by providing a range of downloadable atlases (k=140 to 150 parcels). Secondly, we could provide the pipeline to develop their atlas. The first way is very straightforward for the external user and has no computational cost. However, since the range of downloadable atlases is cohort-specific, it may not work correctly on an external cohort with different age ranges and sample sizes. Thus the second way will be to provide the complete pipeline to the user. This way will not be easier either for the external user, which involves computational cost and finding a cohort-specific optimal number of components. We will provide a range of downloadable atlases to pick the best one based on their study hypothesis as a immediate solution.

In a step towards introducing data-driven brain atlases to neuroradiology, we can think of various potential use of the atlases in clinical use: a) to detect an unknown intricate pattern in detecting disease groups effectively and label it; b) to cope with the data explosion. The usefulness of the atlas for automatic structure identification, localization, delineation, labeling and quantification, and reporting and communication potentially increases the interpreter's efficiency and confidence and expedites image interpretation in clinical diagnosis.

A data-driven Atlas is beneficial for brain aging research and may help diagnose Alzheimer's disease, as a study suggests. Most of the existing MRI atlases based on VBM-GM are based on young and medium-aging brains that do not span the large spectrum of ages. In our research, both for discovery and replication datasets, we used a larger aging spectrum that can be easily used to recognize improvements in patients' brain structure that may be an underpinning symptom of neurodegenerative disorder. The deterioration of brain tissue in an area of the brain called the medial temporal lobe is a crucial symptom of early Alzheimer's disease. These brain structure improvements are often subtle and difficult to recognize, but a detailed brain atlas powered by VBM-GM data could make their identification simpler. VBM-GM-based atlases with a wide-scale age range can diagnose brain damage in other disorders, including schizophrenia stable subjects. Data-driven brain atlases aim to support earlier detection of psychiatric and

neurological diseases that develop at different life stages. These parcellation can also be used for fiber-tracking and functional connectivity analysis to prove the "firing together/wiring together "(Hebb's hypothesis which will be useful in brain plasticity (Hebb's law)[203].

4.4 Conclusion

All the above-mentioned multivariate pattern analysis approaches; dimensionality reduction, independent component analysis, stability measurement of features, and unsupervised clustering algorithm will be very useful in detecting robust, meaningful signatures using the multimodal datasets.

In the first project, we used clinical datasets to detect seven robust and reproducible subtypes; however, we can also use disease brain features obtained from data-driven parcellation framework in the longitudinal clustering model and can detect robust brain imaging-based subtypes in psychiatric disease.

The fusion of project 1 and 2 would be useful to cluster (model-based clustering) patient space based on VBM component weights (unmixing matrix), and so ICA and SBM, in this case, will be a useful framework to get component weights using patient datasets and these components can be used for detecting disease-specific subtypes. We might face challenges to generate components using ICA and SBM based on healthy subjects and then project it on disease subjects because brain structure is different across healthy and disease datasets, so learning of components from healthy subjects and then predicting disease datasets will not be trivial. Consequently, the best approach will be to use the parcellation methodology to detect the atlas using disease dataset and extract the VBM component weights and perform clustering or use the atlas (based on the heathy subject) to extract brain volume feature for disease datasets and then perform clustering.

There is a possible dilemma between methodological development and clinical translation. The hypothesis or the purpose of the methodological development needs to be useful and applicable for clinical translation. In the clinical translation, one would be more interested in seeing the direct effect or biological and clinical signatures directly associated with disease detection (treatment response classes or changes in the brain structure across patients). If a methodological approach does not directly associate in terms of disease progression or

treatment response, doctors may not be very interested in using the clinical translation and disease diagnostic framework.

The treatment response dynamic subtypes obtained from the first project can be used for clinical purposes because they have been validated in a large sample cohort and require very few weeks (2 to 6 weeks) to correctly detect the treatment response group.

For psychiatric translation, we need unsupervised and supervised learning to find clinically meaningful and robust clusters or disease-specific signatures. We should not rely only on unsupervised learning, where we would need to validate clusters or features obtained from unsupervised learning. For psychiatric translation, we should opt for a hybrid approach to detect sub using unsupervised learning and validate those subgroups clinically using supervised learning. We should also use the semi-supervised methodology to develop and detect complex and sparse biological signatures from the multimodal high dimensional dataset. These sparse frameworks can also be applied in precision medicine's clinical translation for the early detection of psychiatric and neurological disorders.

References

1 American Psychiatric Association, American Psychiatric Association (eds.). *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed. American Psychiatric Association: Washington, D.C, 2013.

2 Bell CC. DSM-IV: Diagnostic and Statistical Manual of Mental Disorders. *JAMA J Am Med Assoc* 1994; **272**: 828.

3 Pichot P. [DSM-III: the 3d edition of the Diagnostic and Statistical Manual of Mental Disorders from the American Psychiatric Association]. *Rev Neurol (Paris)* 1986; **142**: 489–499.

4 Weltgesundheitsorganisation (ed.). The ICD-10 classification of mental and behavioural disorders. ... Diagnostic criteria for research. Geneva, 1993.

5 Bains N, Abdijadid S. Major Depressive Disorder. In: *StatPearls*. StatPearls Publishing: Treasure Island (FL), 2020http://www.ncbi.nlm.nih.gov/books/NBK559078/ (accessed 24 Nov2020).

6 Pedersen CB, Mors O, Bertelsen A, Waltoft BL, Agerbo E, McGrath JJ *et al.* A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry* 2014; **71**: 573–581.

7 Cuijpers P, Dekker J, Hollon SD, Andersson G. Adding psychotherapy to pharmacotherapy in the treatment of depressive disorders in adults: a meta-analysis. *J Clin Psychiatry* 2009; **70**: 1219–1229.

8 Cuijpers P, van Straten A, Warmerdam L, Andersson G. Psychotherapy versus the combination of psychotherapy and pharmacotherapy in the treatment of depression: a meta-analysis. *Depress Anxiety* 2009; **26**: 279–288.

9 Malhi GS, Parker GB, Crawford J, Wilhelm K, Mitchell PB. Treatment-resistant depression: resistant to definition? *Acta Psychiatr Scand* 2005; **112**: 302–309.

10 McIntyre RS, Filteau M-J, Martin L, Patry S, Carvalho A, Cha DS *et al.* Treatment-resistant depression: Definitions, review of the evidence, and algorithmic approach. *J Affect Disord* 2014; **156**: 1–7.

11 van Loo HM, de Jonge P, Romeijn J-W, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med* 2012; **10**. doi:10.1186/1741-7015-10-156.

12 de Vos S, Wardenaar KJ, Bos EH, Wit EC, de Jonge P. Decomposing the heterogeneity of depression at the person-, symptom-, and time-level: latent variable models versus multimode principal component analysis. *BMC Med Res Methodol* 2015; **15**: 88.

13 Hybels CF, Blazer DG, Pieper CF, Landerman LR, Steffens DC. Profiles of depressive symptoms in older adults diagnosed with major depression: latent cluster analysis. *Am J Geriatr Psychiatry Off J Am Assoc Geriatr Psychiatry* 2009; **17**: 387–396.

14 Lamers F, de Jonge P, Nolen WA, Smit JH, Zitman FG, Beekman ATF *et al.* Identifying depressive subtypes in a large cohort study: results from the Netherlands Study of Depression and Anxiety (NESDA). *J Clin Psychiatry* 2010; **71**: 1582–1589.

15 Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer: New York, NY, 2009.

16 Wierzchoń ST, Kłopotek MA. Algorithms of Combinatorial Cluster Analysis. In: *Modern Algorithms of Cluster Analysis*. Springer International Publishing: Cham, 2018, pp 67–161.

17 Kannan R, Vempala S, Vetta A. On clusterings: Good, bad and spectral. *J ACM* 2004; **51**: 497–515.

18 Jain AK, Dubes RC. *Algorithms for clustering data*. Prentice Hall: Englewood Cliffs, N.J, 1988.

19 Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering – A decade review. *Inf Syst* 2015; **53**: 16–38.

20 Oberg AL, Mahoney DW. Linear mixed effects models. *Methods Mol Biol Clifton NJ* 2007; **404**: 213–234.

21 Andrews JL, McNicholas PD. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Stat Comput* 2012; **22**: 1021–1029.

22 Larsen FB, Pedersen MH, Friis K, Glümer C, Lasgaard M. A Latent Class Analysis of Multimorbidity and the Relationship to Socio-Demographic Factors and Health-Related Quality of Life. A National Population-Based Study of 162,283 Danish Adults. *PLOS ONE* 2017; **12**: e0169426.

23 *Classification and Clustering*. Elsevier, 1977 doi:10.1016/C2013-0-11644-3.

24 Mechelli A, Price C, Friston K, Ashburner J. Voxel-Based Morphometry of the Human Brain: Methods and Applications. *Curr Med Imaging Rev* 2005; **1**: 105–113.

25 Ashburner J, Friston KJ. Unified segmentation. *NeuroImage* 2005; **26**: 839–851.

26 Ashburner J. A fast diffeomorphic image registration algorithm. *NeuroImage* 2007; **38**: 95–113.

27 Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* 2001; **14**: 21–36.

28 Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage* 2001; **14**: 1361–1369.

29 Yassa MA, Stark CEL. A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage* 2009; **44**: 319–327.

30 Whitwell JL. Voxel-Based Morphometry: An Automated Technique for Assessing Structural Changes in the Brain. *J Neurosci* 2009; **29**: 9661–9664.

31 Ashburner J, Friston KJ. Voxel-Based Morphometry—The Methods. *NeuroImage* 2000; **11**: 805–821.

32 Kennedy KM, Erickson KI, Rodrigue KM, Voss MW, Colcombe SJ, Kramer AF *et al.* Age-related differences in regional brain volumes: a comparison of optimized voxel-based morphometry to manual volumetry. *Neurobiol Aging* 2009; **30**: 1657–1676.

33 Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: A general linear approach. *Hum Brain Mapp* 1994; **2**: 189–210.

34 Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp* 1996; **4**: 58–73.

35 Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD. Neurodegenerative Diseases Target Large-Scale Human Brain Networks. *Neuron* 2009; **62**: 42–52.

36 Alexander-Bloch A, Giedd JN, Bullmore E. Imaging structural co-variance between human brain regions. *Nat Rev Neurosci* 2013; **14**: 322–336.

37 Tijms BM, Seriès P, Willshaw DJ, Lawrie SM. Similarity-based extraction of individual networks from gray matter MRI scans. *Cereb Cortex N Y N 1991* 2012; **22**: 1530–1541.

38 Li K, Luo X, Zeng Q, Huang P, Shen Z, Xu X *et al.* Gray matter structural covariance networks changes along the Alzheimer's disease continuum. *NeuroImage Clin* 2019; **23**: 101828.

39 Essen DCV. A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature* 1997; **385**: 313–318.

40 Gong G, He Y, Chen ZJ, Evans AC. Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *NeuroImage* 2012; **59**: 1239–1248.

41 Chen C-H, Fiecas M, Gutierrez ED, Panizzon MS, Eyler LT, Vuoksimaa E *et al.* Genetic topography of brain morphology. *Proc Natl Acad Sci* 2013; **110**: 17089–17094.

42 Schmitt JE, Lenroot RK, Ordaz SE, Wallace GL, Lerch JP, Evans AC *et al.* Variance decomposition of MRI-based covariance maps using genetically informative samples and structural equation modeling. *NeuroImage* 2009; **47**: 56–64.

43 Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933; **24**: 417–441.

44 Jolliffe IT. Principal Component Analysis. .

45 Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Netw* 2000; **13**: 411–430.

46 Bell AJ, Sejnowski TJ. An information-maximisation approach to blind separation and blind deconvolution. ; : 38.

47 Amari S, Cichocki A, Yang HH. A New Learning Algorithm for Blind Signal Separation. ; : 7.

48 Cardoso J-F. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process Lett* 1997; **4**: 112–114.

49 Wittchen HU, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jönsson B *et al.* The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol* 2011; **21**: 655–679.

50 Wittchen H-U, Hoyer J (eds.). *Klinische Psychologie & Psychotherapie*. 2., überarbeitete und erweiterte Auflage. Springer: Berlin Heidelberg, 2011.

51 World Health Organization. Depression. https://www.who.int/health-topics/depression#tab=tab_1.

52 Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. *Dialogues Clin Neurosci* 2008; **10**: 271–277.

53 Cleare A, Pariante C, Young A, Anderson I, Christmas D, Cowen P *et al.* Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. *J Psychopharmacol (Oxf)* 2015; **29**: 459–525.

54 Reddy MS. Depression: The Disorder and the Burden. *Indian J Psychol Med* 2010; **32**: 1–2.

55 Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L *et al.* Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry* 2006; **163**: 28–40.

56 Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA *et al.* Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatr Clin North Am* 2003; **26**: 457–494, x.

57 Probst JC, Laditka SB, Moore CG, Harun N, Powell MP, Baxley EG. Rural-urban differences in depression prevalence: implications for family medicine. *Fam Med* 2006; **38**: 653–660.

58 Rhoades H. The Hamilton Depression Scale: Factor Scoring and Profile Classification. *Psychopharmacol Bull* 1983; **VOL.19, No: 1**.

59 Hamilton M. A RATING SCALE FOR DEPRESSION. *J Neurol Neurosurg Psychiatry* 1960; **23**: 56–62.

60 Guillot-Valdés M, Guillén-Riquelme A, Buela-Casal G. Reliability and validity of the Basic Depression Questionnaire. *Int J Clin Health Psychol* 2019; **19**: 243–250.

61 Hennings JM, Owashi T, Binder EB, Horstmann S, Menke A, Kloiber S *et al.* Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - findings from the Munich Antidepressant Response Signature (MARS) project. *J Psychiatr Res* 2009; **43**: 215–29.

62 Binder EB, Salyakina D, Lichtner P, Wochnik GM, Ising M, Pütz B *et al.* Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment. *Nat Genet* 2004; **36**: 1319–1325.

63 Kloiber S, Ising M, Reppermund S, Horstmann S, Dose T, Majer M *et al.* Overweight and Obesity Affect Treatment Response in Major Depression. *Biol Psychiatry* 2007; **62**: 321–326.

64 Ising M. Bringing basic and clinical research together to an integrated understanding of psychiatric disorders. *J Psychiatr Res* 2007; **41**: 1–2.

65 Ising M, Horstmann S, Kloiber S, Lucae S, Binder EB, Kern N *et al.* Combined Dexamethasone/Corticotropin Releasing Hormone Test Predicts Treatment Response in Major Depression–A Potential Biomarker? *Biol Psychiatry* 2007; **62**: 47–54.

66 Frank E. Conceptualization and Rationale for Consensus Definitions of Terms in Major Depressive Disorder: Remission, Recovery, Relapse, and Recurrence. *Arch Gen Psychiatry* 1991; **48**: 851.

67 Rush AJ. The varied clinical presentations of major depressive disorder. *J Clin Psychiatry* 2007; **68 Suppl 8**: 4–10.

68 Souery D, Amsterdam J, de Montigny C, Lecrubier Y, Montgomery S, Lipp O *et al.* Treatment resistant depression: methodological overview and operational criteria. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol* 1999; **9**: 83–91.

69 Shallcross AJ, Gross JJ, Visvanathan PD, Kumar N, Palfrey A, Ford BQ *et al.* Relapse prevention in major depressive disorder: Mindfulness-based cognitive therapy versus an active control condition. *J Consult Clin Psychol* 2015; **83**: 964–975.

70 Olgiati P, Serretti A, Souery D, Dold M, Kasper S, Montgomery S *et al.* Early improvement and response to antidepressant medications in adults with major depressive disorder. Meta-analysis and study of a sample with treatment-resistant depression. *J Affect Disord* 2018; **227**: 777–786.

71 Williams JMG (ed.). *Cognitive psychology and emotional disorders*. 2nd ed. Wiley: Chichester ; New York, 1997.

72 Carrozzino D, Patierno C, Fava GA, Guidi J. The Hamilton Rating Scales for Depression: A Critical Review of Clinimetric Properties of Different Versions. *Psychother Psychosom* 2020; **89**: 133–150.

73 Paul R, Andlauer TF, Czamara D, Hoehn D, Lucae S, Pütz B *et al.* Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. *Transl Psychiatry* 2019; **9**: 1–15.

74 Trajković G, Starčević V, Latas M, Leštarević M, Ille T, Bukumirić Z *et al.* Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49years. *Psychiatry Res* 2011; **189**: 1–9.

75 McEwen BS. Neurobiological and Systemic Effects of Chronic Stress. *Chronic Stress Thousand Oaks Calif* 2017; **1**.

76 Park JH, Lee JJ, Lee SB, Huh Y, Choi EA, Youn JC *et al.* Prevalence of major depressive disorder and minor depressive disorder in an elderly Korean population: Results from the Korean Longitudinal Study on Health and Aging (KLoSHA). *J Affect Disord* 2010; **125**: 234–240.

77 Wardenaar KJ, Monden R, Conradi HJ, de Jonge P. Symptom-specific course trajectories and their determinants in primary care patients with Major Depressive Disorder: Evidence for two etiologically distinct prototypes. *J Affect Disord* 2015; **179**: 38–46.

78 Bühler J, Seemüller F, Läge D. The predictive power of subgroups: An empirical approach to identify depressive symptom patterns that predict response to treatment. *J Affect Disord* 2014; **163**: 81–87.

79 Fava M, Alpert JE, Carmin CN, Wisniewski SR, Trivedi MH, Biggs MM *et al.* Clinical correlates and symptom patterns of anxious depression among patients with major depressive disorder in STAR*D. *Psychol Med* 2004; **34**: 1299–1308.

80 Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH *et al.* Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 2016; **3**: 243–250.

81 Gili M, Roca M, Armengol S, Asensio D, Garcia-Campayo J, Parker G. Clinical Patterns and Treatment Outcome in Patients with Melancholic, Atypical and Non-Melancholic Depressions. *PLoS ONE* 2012; **7**: e48200.

82 Nie Z, Vairavan S, Narayan VA, Ye J, Li QS. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLOS ONE* 2018; **13**: e0197268.

83 Wardenaar KJ, van Loo HM, Cai T, Fava M, Gruber MJ, Li J *et al.* The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity. *Psychol Med* 2014; **44**: 3289–3302.

84 Habert J, Katzman MA, Oluboka OJ, McIntyre RS, McIntosh D, MacQueen GM *et al.* Functional Recovery in Major Depressive Disorder: Focus on Early Optimized Treatment. *Prim Care Companion CNS Disord* 2016. doi:10.4088/PCC.15r01926.

85 Szegedi A, Müller MJ, Anghelescu I, Klawe C, Kohnen R, Benkert O. Early improvement under mirtazapine and paroxetine predicts later stable response and remission with high sensitivity in patients with major depression. *J Clin Psychiatry* 2003; **64**: 413–420.

86 Nierenberg AA, Husain MM, Trivedi MH, Fava M, Warden D, Wisniewski SR *et al.* Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: a STAR*D report. *Psychol Med* 2010; **40**: 41.

87 Peciña M, Sikora M, Avery ET, Heffernan J, Peciña S, Mickey BJ *et al.* Striatal dopamine D2/3 receptor-mediated neurotransmission in major depression: Implications for anhedonia, anxiety and treatment response. *Eur Neuropsychopharmacol* 2017; **27**: 977–986.

88 Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. *Ann Data Sci* 2015; **2**: 165–193.

89 Maier W. Dimensions of the Hamilton-Depression-Scale (HAMD), A Factor Analytical Study. *Eur Arch Psychiatry Neurol Sci* 1985; **234**: 417–422.

90 Monden R, Wardenaar KJ, Stegeman A, Conradi HJ, de Jonge P. Simultaneous Decomposition of Depression Heterogeneity on the Person-, Symptom- and Time-Level: The Use of Three-Mode Principal Component Analysis. *PLOS ONE* 2015; **10**: e0132765.

91 Cotrena C, Damiani Branco L, Ponsoni A, Milman Shansis F, Paz Fonseca R. Neuropsychological Clustering in Bipolar and Major Depressive Disorder. *J Int Neuropsychol Soc* 2017; **23**: 584–593.

92 Zeng L-L, Shen H, Liu L, Hu D. Unsupervised classification of major depression using functional connectivity MRI: Unsupervised Classification of Depression. *Hum Brain Mapp* 2014; **35**: 1630–1641.

93 Kelley ME, Dunlop BW, Nemeroff CB, Lori A, Carrillo-Roa T, Binder EB *et al.* Response rate profiles for major depressive disorder: Characterizing early response and longitudinal nonresponse. *Depress Anxiety* 2018; **35**: 992–1000.

94 Uher R, Mors O, Rietschel M, Rajewska-Rager A, Petrovic A, Zobel A *et al.* Early and Delayed Onset of Response to Antidepressants in Individual Trajectories of Change During Treatment of Major Depression: A Secondary Analysis of Data From the Genome-Based Therapeutic Drugs for Depression (GENDEP) Study. *J Clin Psychiatry* 2011; **72**: 1478–1484.

95 Hartmann A, von Wietersheim J, Weiss H, Zeeck A. Patterns of symptom change in major depression: Classification and clustering of long term courses. *Psychiatry Res* 2018; **267**: 480–489.

96 Uher R, Perroud N, Ng MYM, Hauser J, Henigsberg N, Maier W *et al.* Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project. *Am J Psychiatry* 2010; **167**: 555–564.

97 Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A *et al.* Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychol Med* 2008; **38**. doi:10.1017/S0033291707001730.

98 Dunlop BW, Cole SP, Nemeroff CB, Mayberg HS, Craighead WE. Differential change on depressive symptom factors with antidepressant medication and cognitive behavior therapy for major depressive disorder. *J Affect Disord* 2018; **229**: 111–119.

99 Zimmerman M, Chelminski I, Posternak M. A review of studies of the Hamilton depression rating scale in healthy controls: implications for the definition of remission in treatment studies of depression. *J Nerv Ment Dis* 2004; **192**: 595–601.

100 Dilling H, Weltgesundheitsorganisation (eds.). *Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) ; klinisch-diagnostische Leitlinien*. 6., vollst. überarb. Aufl. unter Berücksichtigung der Änderungen entsprechend ICD-10-GM 2004/2008. Huber: Bern, 2008.

101 Wing JK, Sartorius N, Üstün TB. *Diagnosis and clinical measurement in psychiatry: a reference for SCAN*. Cambridge University Press: Cambridge, 2006.

102 Uher R, Tansey KE, Dew T, Maier W, Mors O, Hauser J *et al.* An Inflammatory Biomarker as a Differential Predictor of Outcome of Depression Treatment With Escitalopram and Nortriptyline. *Am J Psychiatry* 2014; **171**: 1278–1286.

103 Powell TR, Smith RG, Hackinger S, Schalkwyk LC, Uher R, McGuffin P *et al.* DNA methylation in interleukin-11 predicts clinical response to antidepressants in GENDEP. *Transl Psychiatry* 2013; **3**: e300–e300.

104 Leisch F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in *R*. *J Stat Softw* 2004; **11**. doi:10.18637/jss.v011.i08.

105 Grün B, Leisch F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J Stat Softw* 2008; **28**. doi:10.18637/jss.v028.i04.

106 Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell* 2000; **22**: 719–725.

107 Wright MN, Ziegler A. **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J Stat Softw* 2017; **77**. doi:10.18637/jss.v077.i01.

108 Derogatis LR. SCL-90-R, administration, scoring & procedures manual-I for the R(evised) version. Baltimore, MD: Johns Hopkins University, School of Medicine. Johns Hopkins University, School of Medicine, 1977.

109 Eysenck SBG, Eysenck HJ, Barrett P. A revised version of the psychoticism scale. *Personal Individ Differ* 1985; **6**: 21–29.

110 Cloninger CR, Przybeck TR, Svrakic DM. The Tridimensional Personality Questionnaire: U.S. Normative Data. *Psychol Rep* 1991; **69**: 1047–1057.

111 BREIMAN L. Random Forests. 2001; **45**: 5–32.

112 Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010; **26**: 1340–1347.

113 Kraus C, Kadriu B, Lanzenberger R, Zarate CA, Kasper S. Prognosis and Improved Outcomes in Major Depression: A Review. *FOCUS* 2020; **18**: 220–235.

114 Paul R, Andlauer TillFM, Czamara D, Hoehn D, Lucae S, Pütz B *et al.* Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. *Transl Psychiatry* 2019; **9**: 187.

115 Baudry J-P. Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electron J Stat* 2015; **9**: 1041–1077.

116 Williams JBW, Kobak KA. Development and reliability of a structured interview guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA). *Br J Psychiatry* 2008; **192**: 52–58.

117 Wright MN, Ziegler A. **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J Stat Softw* 2017; **77**. doi:10.18637/jss.v077.i01.

118 Kudlow PA, Cha DS, McLntyre RS. Predicting Treatment Response in Major Depressive Disorder: The Impact of Early Symptomatic Improvement. *Can J Psychiatry* 2012; **57**: 782–788.

119 McIntyre RS, Gorwood P, Thase ME, Liss C, Desai D, Chen J *et al.* Early Symptom Improvement as a Predictor of Response to Extended Release Quetiapine in Major Depressive Disorder: *J Clin Psychopharmacol* 2015; **35**: 706–710.

120 Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C *et al.* Does early improvement triggered by antidepressants predict response/remission? — Analysis of data from a naturalistic study on a large sample of inpatients with major depression. *J Affect Disord* 2009; **115**: 439–449.

121 Nie Z, Vairavan S, Narayan VA, Ye J, Li QS. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLOS ONE* 2018; **13**: e0197268.

122 Hung C-I, Liu C-Y, Yang C-H. Untreated duration predicted the severity of depression at the two-year follow-up point. *PLOS ONE* 2017; **12**: e0185119.

123 Gilmer WS, Gollan JK, Wisniewski SR, Howland RH, Trivedi MH, Miyahara S *et al.* Does the duration of index episode affect the treatment outcome of major depressive disorder? A STAR*D report. *J Clin Psychiatry* 2008; **69**: 1246–1256.

124 Sung SC, Haley CL, Wisniewski SR, Fava M, Nierenberg AA, Warden D *et al.* The Impact of Chronic Depression on Acute and Long-Term Outcomes in a Randomized Trial Comparing Selective Serotonin Reuptake Inhibitor Monotherapy Versus Each of 2 Different Antidepressant Medication Combinations. *J Clin Psychiatry* 2012; **73**: 967–976.

125 Otte C. Incomplete remission in depression: role of psychiatric and somatic comorbidity. *Dialogues Clin Neurosci* 2008; **10**: 453–460.

126 Inkster B, Rao AW, Ridler K, Nichols TE, Saemann PG, Auer DP *et al.* Structural Brain Changes in Patients with Recurrent Major Depressive Disorder Presenting with Anxiety Symptoms. *J Neuroimaging* 2011; **21**: 375–382.

127 Sämann PG, Höhn D, Chechko N, Kloiber S, Lucae S, Ising M *et al.* Prediction of antidepressant treatment response from gray matter volume across diagnostic categories. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol* 2013; **23**: 1503–15.

128 Quilty LC, Meusel L-AC, Bagby RM. Neuroticism as a mediator of treatment response to SSRIs in major depressive disorder. *J Affect Disord* 2008; **111**: 67–73.

129 Katon W, UnÃ¼tzer J, Russo J. Major depression: the importance of clinical characteristics and treatment response to prognosis. *Depress Anxiety* 2010; **27**: 19–26.

130 Uliaszek AA, Zinbarg RE, Mineka S, Craske MG, Sutton JM, Griffith JW *et al.* The role of neuroticism and extraversion in the stress–anxiety and stress–depression relationships. *Anxiety Stress Coping* 2010; **23**: 363–381.

131 Bulmash E, Harkness KL, Stewart JG, Bagby RM. Personality, stressful life events, and treatment response in major depression. *J Consult Clin Psychol* 2009; **77**: 1067–1077.

132 Mazure CM. Adverse Life Events and Cognitive-Personality Characteristics in the Prediction of Major Depression and Antidepressant Response. *Am J Psychiatry* 2000; **157**: 896–903.

133 van Calker D, Zobel I, Dykierek P, Deimel CM, Kech S, Lieb K *et al.* Time course of response to antidepressants: Predictive value of early improvement and effect of additional psychotherapy. *J Affect Disord* 2009; **114**: 243–253.

134 Joel I, Begley AE, Mulsant BH, Lenze EJ, Mazumdar S, Dew MA *et al.* Dynamic Prediction of Treatment Response in Late-Life Depression. *Am J Geriatr Psychiatry* 2014; **22**: 167–176.

135 Chandler GM, Iosifescu DV, Pollack MH, Targum SD, Fava M. RESEARCH: Validation of the Massachusetts General Hospital Antidepressant Treatment History Questionnaire (ATRQ): Validation of the MGH ATRQ. *CNS Neurosci Ther* 2010; **16**: 322–325.

136 Reynolds CF, Dew MA, Frank E, Begley AE, Miller MD, Cornes C *et al.* Effects of age at onset of first lifetime episode of recurrent major depression on treatment response and illness course in elderly patients. *Am J Psychiatry* 1998; **155**: 795–799.

137 Zisook S, Lesser I, Stewart JW, Wisniewski SR, Balasubramani GK, Fava M *et al.* Effect of age at onset on the course of major depressive disorder. *Am J Psychiatry* 2007; **164**: 1539–1546.

138 Park S-C, Hahn S-W, Hwang T-Y, Kim J-M, Jun T-Y, Lee M-S *et al.* Does age at onset of first major depressive episode indicate the subtype of major depressive disorder?: the clinical research center for depression study. *Yonsei Med J* 2014; **55**: 1712–1720.

139 Kloiber S, Domschke K, Ising M, Arolt V, Baune BT, Holsboer F *et al.* Clinical risk factors for weight gain during psychopharmacologic treatment of depression: results from 2 large German observational studies. *J Clin Psychiatry* 2015; **76**: e802-8.

140 Muhonen LH, Lönnqvist J, Lahti J, Alho H. Age at onset of first depressive episode as a predictor for escitalopram treatment of major depression comorbid with alcohol dependence. *Psychiatry Res* 2009; **167**: 115–122.

141 Schmaal L, Hibar DP, Sämann PG, Hall GB, Baune BT, Jahanshad N *et al.* Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry* 2017; **22**: 900–909.

142 Rentería ME, Schmaal L, Hibar DP, Couvy-Duchesne B, Strike LT, Mills NT *et al.* Subcortical brain structure and suicidal behaviour in major depressive disorder: a meta-analysis from the ENIGMA-MDD working group. *Transl Psychiatry* 2017; **7**: e1116.

143 Zobel AW, Nickel T, Sonntag A, Uhr M, Holsboer F, Ising M. Cortisol response in the combined dexamethasone/CRH test as predictor of relapse in patients with remitted depression. a prospective study. *J Psychiatr Res* 2001; **35**: 83–94.

144 Ashburner J, Friston KJ. Why Voxel-Based Morphometry Should Be Used. *NeuroImage* 2001; **14**: 1238–1243.

145 Beckmann CF, Smith SM. Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Trans Med Imaging* 2004; **23**: 137–152.

146 Smith SM, Vidaurre D, Beckmann CF, Glasser MF, Jenkinson M, Miller KL *et al.* Functional connectomics from resting-state fMRI. *Trends Cogn Sci* 2013; **17**: 666–682.

147 Tomasi D, Volkow ND. Association between Functional Connectivity Hubs and Brain Networks. *Cereb Cortex* 2011; **21**: 2003–2013.

148 Fox KCR, Nijeboer S, Dixon ML, Floman JL, Ellamil M, Rumak SP *et al.* Is meditation associated with altered brain structure? A systematic review and meta-analysis of morphometric neuroimaging in meditation practitioners. *Neurosci Biobehav Rev* 2014; **43**: 48–73.

149 Basser PJ, Pajevic S, Pierpaoli C, Duda J, Aldroubi A. In vivo fiber tractography using DT-MRI data. *Magn Reson Med* 2000; **44**: 625–632.

150 Ge R, Kot P, Liu X, Lang DJ, Wang JZ, Honer WG *et al.* Parcellation of the human hippocampus based on gray matter volume covariance: Replicable results on healthy young adults. *Hum Brain Mapp* 2019; : hbm.24628.

151 DuPre E, Spreng RN. Structural covariance networks across the life span, from 6 to 94 years of age. *Netw Neurosci* 2017; **1**: 302–323.

152 Keysers C, Gazzola V. Hebbian learning and predictive mirror neurons for actions, sensations and emotions. *Philos Trans R Soc B Biol Sci* 2014; **369**: 20130175.

153 Peelle JE, Cusack R, Henson RNA. Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging. *NeuroImage* 2012; **60**: 1503–1516.

154 Forest M, Iturria-Medina Y, Goldman JS, Kleinman CL, Lovato A, Oros Klein K *et al.* Gene networks show associations with seed region connectivity. *Hum Brain Mapp* 2017; **38**: 3126–3140.

155 Gupta CN, Turner JA, Calhoun VD. Source-based morphometry: a decade of covarying structural brain patterns. *Brain Struct Funct* 2019; **224**: 3031–3044.

156 Xu L, Groth KM, Pearlson G, Schretlen DJ, Calhoun VD. Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia. *Hum Brain Mapp* 2009; **30**: 711–724.

157 Guo X, Wang Y, Guo T, Chen K, Zhang J, Li K *et al.* Structural covariance networks across healthy young adults and their consistency: Structural Networks Across Healthy Adults. *J Magn Reson Imaging* 2015; **42**: 261–268.

158 Eickhoff SB, Yeo BTT, Genon S. Imaging-based parcellations of the human brain. *Nat Rev Neurosci* 2018; **19**: 672–686.

159 Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 2004; **22**: 1214–1222.

160 Endophenotype Potential of Nucleus Accumbens Functional Connectivity: Effects of Polygenic Risk for Schizophrenia Interacting with Childhood Adversity. *J Psychiatry Brain Sci* 2019. doi:10.20900/jpbs.20190011.

161 Brückl TM, Spoormaker VI, Sämann PG, Brem A-K, Henco L, Czamara D *et al.* The biological classification of mental disorders (BeCOME) study: a protocol for an observational deep-phenotyping study for the identification of biological subtypes. *BMC Psychiatry* 2020; **20**: 213.

162 Elbau IG, Brücklmeier B, Uhr M, Arloth J, Czamara D, Spoormaker VI *et al.* The brain's hemodynamic response function rapidly changes under acute psychosocial stress in association with genetic and endocrine stress response markers. *Proc Natl Acad Sci* 2018; **115**: E10206–E10215.

163 Ashburner J. *SPM12 Manual*. 2015http://web.mit.edu/spm_v12/manual.pdf.

164 F. Kurth. VBM8-Toolbox Manual. 2010.http://dbm.neuro.uni-jena.de/vbm8/VBM8-Manual.pdf.

165 Bendel P, Koivisto T, Äikiä M, Niskanen E, Könönen M, Hänninen T *et al.* Atrophic Enlargement of CSF Volume after Subarachnoid Hemorrhage: Correlation with Neuropsychological Outcome. *Am J Neuroradiol* 2010; **31**: 370–376.

166 O'Brien C, van Riper C, Myers DE. MAKING RELIABLE DECISIONS IN THE STUDY OF WILDLIFE DISEASES: USING HYPOTHESIS TESTS, STATISTICAL POWER, AND OBSERVED EFFECTS. *J Wildl Dis* 2009; **45**: 700–712.

167 Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* 2018; **167**: 104–120.

168 Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E *et al.* Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage* 2020; **208**: 116450.

169 Zhu T, Hu R, Qiu X, Taylor M, Tso Y, Yiannoutsos C *et al.* Quantification of accuracy and precision of multi-center DTI measurements: A diffusion phantom and human brain study. *NeuroImage* 2011; **56**: 1398–1411.

170 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**: 118–127.

171 Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 2017; **161**: 149–170.

172 Caprihan A, Abbott C, Yamamoto J, Pearlson G, Perrone-Bizzozero N, Sui J *et al.* Source-Based Morphometry Analysis of Group Differences in Fractional Anisotropy in Schizophrenia. *Brain Connect* 2011; **1**: 133–145.

173 Calhoun V, Rachakonda S. Group ICA/IVA of fMRI toolbox (GIFT) Manuel with SBM toolbox. 2004.

174 Sahonero-Alvarez G, Calderon H. A Comparison of SOBI, FastICA, JADE and Infomax Algorithms. 2017; : 6.

175 Langlois D, Chartier S, Gosselin D. An Introduction to Independent Component Analysis: InfoMax and FastICA algorithms. *Tutor Quant Methods Psychol* 2010; **6**: 31–38.

176 Li Y-O, Adali T, Calhoun VD. A Feature-Selective Independent Component Analysis Method for Functional MRI. *Int J Biomed Imaging* 2007; **2007**: 1–12.

177 Zhang Q, Hu G, Tian L, Ristaniemi T, Wang H, Chen H *et al.* Examining stability of independent component analysis based on coefficient and component matrices for voxel-based morphometry of structural magnetic resonance imaging. *Cogn Neurodyn* 2018; **12**: 461–470.

178 Institute of Electrical and Electronics Engineers (ed.). 2003 IEEE XIII Workshop on Neural Networks for Signal Processing--NNSP'03: Toulouse, France, September 17-19, 2003. IEEE: Piscataway, N.J, 2003.

179 Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 2004; **22**: 1214–1222.

180 Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* 2004; **22**: 1214–1222.

181 Hu G, Zhang Q, Waters AB, Li H, Zhang C, Wu J *et al.* Tensor clustering on outer-product of coefficient and component matrices of independent component analysis for reliable functional magnetic resonance imaging data decomposition. *J Neurosci Methods* 2019; **325**: 108359.

182 Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974; **19**: 716–723.

183 Schwarz G. Estimating the Dimension of a Model. *Ann Stat* 1978; **6**: 461–464.

184 Rissanen J. Modeling by shortest data description. *Automatica* 1978; **14**: 465–471.

185 Calhoun VD, Adali T, Pearlson GD, Pekar JJ. A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp* 2001; **14**: 140–151.

186 Wax M, Kailath T. Detection of signals by information theoretic criteria. *IEEE Trans Acoust Speech Signal Process* 1985; **33**: 387–392.

187 Jouan-Rimbaud Bouveresse D, Moya-González A, Ammari F, Rutledge DN. Two novel methods for the determination of the number of components in independent components analysis models. *Chemom Intell Lab Syst* 2012; **112**: 24–32.

188 Johnson SC. Hierarchical clustering schemes. *Psychometrika* 1967; **32**: 241–254.

189 Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 2014; **31**: 274–295.

190 Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* 1963; **58**: 236–244.

191 Thirion B, Varoquaux G, Dohmatob E, Poline J-B. Which fMRI clustering gives good brain parcellations? *Front Neurosci* 2014; **8**. doi:10.3389/fnins.2014.00167.

192 Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw* 1999; **10**: 626–634.

193 Zhang Q, Hu G, Tian L, Ristaniemi T, Wang H, Chen H *et al.* Examining stability of independent component analysis based on coefficient and component matrices for voxel-based morphometry of structural magnetic resonance imaging. *Cogn Neurodyn* 2018; **12**: 461–470.

194 Arslan S, Ktena SI, Makropoulos A, Robinson EC, Rueckert D, Parisot S. Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage* 2018; **170**: 5–30.

195 Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* 2016; **536**: 171–178.

196 Shen X, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage* 2013; **82**: 403–415.

197 Scrucca L. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* 2016; **8(1)**: 289–317.

198 Laparra V, Gutmann MU, Malo J, Hyvärinen A. Complex-Valued Independent Component Analysis of Natural Images. In: Honkela T, Duch W, Girolami M, Kaski S (eds). *Artificial Neural Networks and Machine Learning – ICANN 2011*. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011, pp 213–220.

199 Laparra V, Camps-Valls G, Malo J. Iterative Gaussianization: From ICA to Random Rotations. *IEEE Trans Neural Netw* 2011; **22**: 537–549.

200 Kato M, Asami Y, Wajsbrot DB, Wang X, Boucher M, Prieto R *et al.* Clustering patients by depression symptoms to predict venlafaxine ER antidepressant efficacy: Individual patient data analysis. *J Psychiatr Res* 2020; **129**: 160–167.

201 Kato TA, Katsuki R, Kubo H, Shimokawa N, Sato-Kasai M, Hayakawa K *et al.* Development and validation of the 22-item Tarumi's Modern-Type Depression Trait Scale: Avoidance of Social Roles, Complaint, and Low Self-Esteem (TACS-22). *Psychiatry Clin Neurosci* 2019; : pcn.12842.

202 Saraçli S, Doğan N, Doğan İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J Inequalities Appl* 2013; **2013**: 203.

203 Hebb DO. *The organization of behavior: a neuropsychological theory*. L. Erlbaum Associates: Mahwah, N.J, 2002.

Acknowledgements

This Ph.D. thesis is the product of many persons of whom I am immensely thankful for their contributions. First and foremost, I thank my supervisors Professor Bertram Müller-Myhsok, Professor Julien Gagneur, and Dr. Philipp G. Saemann. It has been a privilege to work with you all. I hope to be able to work with you again soon. Thanks for your responsiveness Bertram; you brought me to the Max Planck Institute of Psychiatry and inspired and guided me through this unpredictable Ph.D. journey. I have met so many Professors and teachers in my life, but no one like you. You have the magic power that allows a most ordinary student to dream big and achieve whatever he/she wants in their lives, and I was the most significant example of it. I have learned so much from you. Many thanks for always being there no matter how busy life was. My thanks for becoming a role model and a rock pillar for each student and me. You have taught me not only science and research but also to become a good human being and empathetic person. Julien, I am sincerely grateful for your presence and attention during these four years. Both made me go further professionally and personally.

Many thanks for your generous support when I needed it most. Philipp, thanks for your mentorship and tremendous support and guidance throughout my Ph.D. I have learned many things from you—starting from neuroimaging, data processing to research article writing. You had immense patience in handling my never-ending question. I can't express my gratitude to you for your contribution to get this thesis done on time. Your pieces of advice were of great help. I am also profoundly grateful for my co-authors' hard work and their substantial contribution to uplift the studies presented in this thesis—special thanks and gratitude for Dr. Benno Pütz. I have no words to describe how much you helped me become a better person in scientifically, analytical, and personal ways. You helped a lot throughout my Ph.D. process.

Many thanks to all of the other co-authors, mainly Marcus Ising, Till Andlauer, Darina Czamara, Cathryn M. Lewis, Benno Pütz, Rudolf Uher, David Hoehn, and Susanne Lucae for the immense help and the contribution to my first ever scientific journal. Without your help and kind guidance, my dream would not have come true. I am delighted and honored to have worked with you, and I look forward to working with you again. Thanks a lot to all the wonderful statgen colleagues and previous colleagues, specially Benno Pütz, Till Andlauer, Nazanin Mirza-Schreiber, Alessandro Gialluisi, Ilaria Bonavita, Dunja Kurtoic, Helena Pelin, Sylvain Moser, Lucas Miranda, who helped and provided me with a friendly and inspiring environment to work and have fun. I would also like to thank all the neuroimaging group members, especially Michael

Czisch and Christopher Eberle, for the immense help during my Ph.D. journey. I would also like to thank all the Binder group members, mainly Elisabeth Binder, for providing me the opportunity to work in the Max Planck Institute of Psychiatry. I am also grateful for having been supported by all the Max Planck Institute of Psychiatry members, especially all the IT and administrative agencies, to help me with IT and bureaucracy problems.

Finally, my deep and sincere gratitude to my family for their continuous and unparalleled love, help, and support. Many thanks to my father, Baidya Nath Paul, and my mother, Ruma Paul, for never giving up on me and becoming the pillar of my life. I am immensely grateful to my younger brother Soumya Paul for always being there for me as a friend and my moral supporter. I can't express my gratitude towards my brother. I can't thank enough my loving husband, Jayjit Dutta, for continuously supporting me in my Ph.D. Journey. I am blessed to have him in my life. I also have the privilege of supporting father-in-law Benoy Krishna Dutta and caring and loving mother in law Jayanti Dutta for the unconditional love and support throughout this challenging journey. I am forever indebted to my family for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them. Special thanks to all my friends; I thank you for all the incredible support. You light up my life and bring me so much happiness in my journey.

Riya Paul

Max Planck Institute of Psychiatry
31.12.2020

Curriculum vitae

PERSONAL INFORMATION

Riya Paul

[redacted]

[redacted]

riyaforms@gmail.com

https://www.linkedin.com/in/riya-paul-9b490b58/

AREAS OF EXPERTISE

- Medical image computing, neuroimaging and analysis applied in finding subtypes in psychiatric disorder.
- Multivariate modelling, Machine Learning, Deep Learning and statistical learning for feature selection, pattern recognition, classification.
- Transfer Learning Assisted Classification and Detection in psychiatric diseases.

WORK EXPERIENCES

1/10/2019–Present

Postdoctoral Researcher
Neurodiagnostic Applications in Psychiatry, Department of Psychiatry & Psychotherapy

Ludwig-Maximilian-University Munich (Germany)

Topic: Multi-scale, multimodal stratification and comorbidity analysis in psychiatric disorders

01/04/2016–30/09/2019

PhD Researcher

Research Group Statistical Genetics, Translational Research in Psychiatry

Max Planck Institute of Psychiatry, Munich (Germany)

Topic: Multivariate methods for analyzing Combined Multimodal Omics- and multimodal neuroimaging records

a Unsupervised clustering method using structural brain imaging data (Voxel-Based Morphometry clustering) to define volume based networks. Automated and purely data-driven grey matter parcellation for developing VBM (voxel-based morphometry) based atlas system

b Mixed model-based clustering to detect the subtypes in the depression data.

c An unsupervised clustering approach to detect subtypes in the depression using functional MRI datasets.

Scientific Researcher

01/11/2015–31/03/2016 Robert Bosch Krankenhaus, Stuttgart (Germany)

Worked as a scientific Researcher in Neuroimaging group in Robert Bosch Krankenhaus, Germany. I developed a framework for task fMRI image analysis based on stress study in sports.

01/03/2014–31/08/2015 Student Research Assistant

Max Planck Institute for Bio Cybernetics, Tübingen (Germany)

a Project 1: Worked as a student research assistant in the topic of "Statistical analysis of netfMRI data" by applying different machine learning and statistical analysis algorithm.

Algorithms: Generalized Linear Model and General Linear Model, Dictionary Learning and Sparsity – KSVD, K-means algorithms, Statistical analysis algorithms – t-test, sign-rank, sign-test and cosine similarity

b Project 2: Worked as a student assistant in 9.4 T MRI scan for developing proper parameter for image registration as well as image segmentation by using MIPAV, SPM, FSL software toolboxes. Learned the MIPAV and LIPSIA environment and tried various data to verify the registration and segmentation results.

01/06/2014–31/12/2014 Master Thesis Student

Max Planck Institute for Bio Cybernetics, Tübingen (Germany)

Worked as a master thesis student in the topic- the viability of EEG measurement of in8 degrees of freedom vehicle simulator by using EEGLAB, ERPLAB and MATLAB programming. My thesis work was related to analyzing mental workload by using signal processing techniques and classifying them using support vector machine classifier and also measuring signal to noise ratio in a motion simulator during different motion conditions and mental workload measuring tasks (auditory oddball task, auditory n-back task).

01/11/2013–31/05/2014 Student Research Intern

Max Planck Institute for Bio Cybernetics, Tübingen (Germany)

Worked as intern student in the topic- the viability of EEG measurement of in 8degrees of freedom vehicle simulator by using auditory oddball experimental task with EEGLAB and MATLAB programming.

01/05/2013–31/10/2013 Student Intern

Fraunhofer MeVis, Bremen (Germany)

Worked in quality control of Diffusion MRI images and image registration.

Student Research assistant

01/05/2013–31/10/2013	GEOMAR–Helmholtz Zentrum für Ozeanforschung, Kiel (Germany) Worked in Oceanographic research at GEOMAR, developing MATLAB codes for Oceanographic data (Image) analysis.
01/02/2012–01/02/2013	Student Research Assistant Christian Albrechts University zu Kiel (Technische Fakultät), Department of Digital Signal Processing and System Theory, Kiel (Germany) Worked in medical signal processing in developing programs for EEG, EMG, Parkinson and Tremor data analysis using EEGLAB, FastICA toolboxes and MATLAB programming, comparison of different independent component analysis methods.
01/03/2011–31/08/2011	Lecturer and Technical Assistant Electronics and Communication Engineering department, PAILAN Institute of Engineering & Technology, West Bengal University of Technology, Kolkata (India)

EDUCATION & TRAININGS

01/04/2016–Present	PhD Student Technische Universität München, School of Medicine, Department of Experimental Medicine Working as a Ph.D. student on Multivariate unsupervised clustering approach on high dimensional datasets (functional and structural brain MRI, genetics data as well as clinical items) to find the subtypes in the depression.
15/10/2011–31/03/2016	M.Sc. in Digital Communication Engineering Institute of Electrical Engineering and Information Technology (EE&IT), Christian-Albrechts-Universität zu Kiel, Kiel (Germany)
01/08/2005–31/05/2009	Bachelor of Engineering (1st Class) in Electronics and Instrumentation Engineering Netaji Subhash Engineering College (Techno India), West Bengal University of Technology, Kolkata, West Bengal, India, Kolkata (India)
Certifications	Neural Networks and Deep Learning – deeplearning.ai

PUBLICATIONS & SCIENTIFIC ACHIEVEMENTS

Awards	**Mifek-Kirschner Awardee 2019**

Publications

Journal:

- Paul, R., Andlauer, TF, Czamara, D., Hoehn, D., Lucae, S., Pütz, B., Lewis, CM, Uher, R., Müller-Myhsok, B., Ising, M. and Sämann, PG, 2019. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Translational psychiatry, 9 (1), pp.1-15.
- Popovic, D., Ruef, A., Dwyer, DB, Hedderich, D., Antonucci, LA, Kambeitz-Ilankovic, L., Öztürk, Ö.F., Dong, MS, Paul, R., Kambeitz, J. and Ruhrmann, S., 2020. O8. 5. Signs of adversity-a novel machine learning approach to childhood trauma, brain structure and clinical profiles. Schizophrenia Bulletin, 46 (Supplement_1), pp. S20-S20.
- Lucas Miranda, Riya Paul, Benno Pütz, Bertram Müller-Myhsok. Functional MRI applications for psychiatric disease subtyping: a review (https://arxiv.org/abs/2007.00126)

Posters:

- Structural covariance analysis to develop a data-driven voxel-wise grey matter volume atlas system, OHBM 2019, Rome (Poster presentation)
- Morphological ICA: Exploiting across subjects' covariance and independent component analysis to define volume-based networks, MAQC 2019, Riva Del Garda Italy, (Poster presented)
- Spatial structure of resting-state fMRI BOLD latency structure explored at the voxel level, OHBM (Human Brain Mapping) 2017, Vancouver Canada (Poster presented)

Conferences & Summer Schools

Attended Conferences and Summer Schools

Conferences:

- OHBM (Human Brain Mapping) 2019, Rome, Italy
- MAQC 2019, Riva Del Garda Italy
- OHBM (Human Brain Mapping) 2017, Vancouver Canada,
- OHBM (Human Brain Mapping) 2016, Geneva, Switzerland
- ECML-PKDD 2017
- European Society of Human Genetics (ESHG) 2016, Barcelona Spain.

Scientific talks:

MAQC 2019 (Riva del Garda) - Morphological ICA: Exploiting across subjects covariance and independent component analysis to define volume based networks

Summer Schools:

- Radboud Summer School 2016 – Computational Genetics
- DeepLearn 2018, Genoa, Italy – Deep Learning Summer School

PERSONAL SKILLS

Mother tongue(s)

Bengali

Foreign language(s)

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
German	B1	B1	B1	B1	B1
English	C1	C1	C1	C1	C1
Hindi	C1	C1	C1	C1	B1

Levels: A1 and A2: Basic user - B1 and B2: Independent user - C1 and C2: Proficient user (CEFR)
Common European Framework of Reference for Languages

Job-related skills

Programming languages:

MATLAB (native), R (native), Python (Intermediate), Shell Programming (Intermediate) and C++ (Basic)

Operating systems:

Mac, Linux, and Windows

Software:

FSL, SPM, MRICRO, MRICRON – used for image preprocessing and image visualization (DICOM, Nifti as well as Analyze images)

NILEARN – machine learning in structural MRI dataset (sparsity-based decomposition method for extracting spatial maps)

KERAS – classification model for clinical data.

MIPAV – image segmentation and image registration

EEGLAB – EEG signal processing and EEG data visualization.

Tools for Nifti and analyze image reader, different image format converter for DICOM, Nifti as well as analyze images.

Summer Schools:

- Radboud Summer School 2016 – Computational Genetics
- DeepLearn 2018, Genoa, Italy – Deep Learning Summer School

List of publications

1. Paul, R., Andlauer, TF, Czamara, D., Hoehn, D., Lucae, S., Pütz, B., Lewis, CM, Uher, R., Müller-Myhsok, B., Ising, M. and Sämann, PG, 2019. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Translational psychiatry, 9 (1), pp.1-15.
2. Popovic, D., Ruef, A., Dwyer, DB, Hedderich, D., Antonucci, LA, Kambeitz-Ilankovic, L., Öztürk, Ö.F., Dong, MS, Paul, R., Kambeitz, J. and Ruhrmann, S., 2020. O8. 5. Signs of adversity-a novel machine learning approach to childhood trauma, brain structure and clinical profiles. Schizophrenia Bulletin, 46 (Supplement_1), pp. S20-S20.
3. Lucas Miranda, Riya Paul, Benno Pütz, Bertram Müller-Myhsok. Functional MRI applications for psychiatric disease subtyping: a review (https://arxiv.org/abs/2007.00126)