Aus der Klinik für Psychiatrie und Psychotherapie
Klinik der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. med. Peter G. Falkai

Differentiation of Recent Onset Depression vs. Recent Onset Psychosis using Pattern Classification Methods on Neuropsychological Data: Diagnostic Performance and Generalizability

Dissertation
zum Erwerb des Doktorgrades der Medizin
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von
Yanis-Michael L. G. Köhler
aus Tunis

2019

Mit Genehmigung der Medizinischen Fakultät der Universität München

Berichterstatter:	Prof. Dr. med. Nikolaos Koutsouleris
Mitberichterstatter:	PD Dr. Alexander Brunnauer
	PD Dr. Leonhard Schilbach
	Prof. Dr. Andreas Dietmar Schuld
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	04.07.2019

to my loving parents and my wonderful sisters,
Maïnée and Sumiya

Contents

1 Summary

1.1 Deutsche Zusammenfassung

Hintergrund In der jüngeren Vergangenheit haben wissenschaftliche Studien gezeigt, dass es möglich ist, mittels MRT-basierten multivariaten Mustererkennung mit einer hohen Klassifizierungswahrscheinlichkeit zwischen neu erkrankt psychotischen (ROP) und neu erkrankt depressiven (ROD) Patienten zu unterscheiden [Koutsouleris et al., 2015]. Allerdings werden insbesondere depressive Patienten mit frühem Krankheitsbeginn öfter als ROP missklassifiziert, was zu der Vermutung führte, dass eine neuroanatomische Ähnlichkeit zwischen diesen beiden Gruppen bestehen könnte [Koutsouleris et al., 2015]. Des Weiteren deuten verschiedene Studien daraufhin, dass sowohl depressive, als auch psychotische Erkrankungsbilder bereits in frühen Stadien spezifische neurokognitive Beeinträchtigungsmuster zeigen [Lee et al., 2012, Bora and Murray, 2013]. In diesem Zusammenhang könnte die Anwendung multivariater Musteranalysen an neurokognitiven Daten die klinische Differentialdiagnostik in Zukunft unterstützen und erleichtern.

Ziele In dieser Studie wurden univariate und multivariate statistische Analysen durchgeführt um (i) mögliche Leistungsunterschiede in der neuropsychologischen Testbatterie zwischen ROD- und ROP-Patienten aufzudecken, (ii) Klassifizierungsmodelle mit Hilfe multivariater Musteranalysen zu erstellen, die anhand krankheitsspezifischer Muster in den neurokognitiven Daten verlässlich zwischen ROD- und ROP-Patienten unterscheiden können und (iii) diese Klassifizierungsmodelle unter Durchführung einer "Leave-center-out-Analyse" auf ihre Generalisierbarkeit und externe Validität hin zu prüfen.

Methodik Vorläufige Daten von 116 Studienteilnehmer (58 ROD (Alter: 25.5 ± 6.0 Jahre), 34 weiblich; 58 ROP (25.6 ± 5.2 Jahre), 16 weiblich) des PRONIA-Projektes wurden untersucht (vgl. https://www.pronia.eu/). Die Diagnosen der Studienteilnehmer wurden mittels dem Structured Clinical Interview for DSM-IV - Axis I (SCID-I) verifiziert. Gruppenbezogene Leistungsfähigkeit in der neurokognitiven Untersuchung wurde

mittels t-Tests an den Ergebnissen von 13 etablierten neurokognitiven Tests (u. A. Rey-Osterrieth Figure Complex Figure Test, Trail Making Test A and B, Rey Auditory Verbal Learning Test etc.) verglichen. Das Signifikanzniveau der t-Tests wurde entsprechend der Bonferroni-Methode korrigiert. Klassifizierungsmodelle wurden an den Daten der multidomänen, neuropsychologischen Testbatterie mittels Support Vector Machine (SVM)-basierter multivariater Mustererkennung trainiert. Dies wurde getrennt jeweils mit 23 und 214 neurokognitiven Variablen durchgeführt. Letztlich wurde die Generalisierbarkeit der Klassifikationsmodelle mittels einer "Leave-center-out-Analyse" überprüft.

Ergebnisse ROP-Patienten schnitten verglichen mit Teilnehmern der ROD-Gruppe in fast allen neuropsychologischen Tests schlechter ab. Statistisch signifikante Leistungsunterschiede konnten in 9 von 23 Variablen nachgewiesen werden. Folglich zeigten ROP-Patienten ausgeprägtere Defizite in den kognitiven Bereichen *Verarbeitungsgeschwindigkeit*, *Aufmerksamkeit*, *verbales Lernen*, *exekutive Funktionen* und *prämorbider IQ*.
In den multivariaten ROD vs. ROP Analysen wurden kreuzvalidierte Klassifikationsgenauigkeiten von 63.8% in der 23-Variablen-Analyse und 71.6% in der 214-Variablen-Analyse erreicht. Die durchschnittliche Klassifikationsgenauigkeit in der Leave-center-out-Analyse betrug 72.4% mit einer Sensitivität und einer Spezifizität von jeweils 72.4%.

Schlussfolgerung ROP-Patienten weisen insgesamt deutlich stärkere kognitive Defizite als ROD-Patienten auf. Durch die Anwendung von Machine-Learning-Algorhythmen auf neurokognitive Daten ist es möglich, Patienten mit kürzlich manifestierten depressiven und psychotischen Störungsbildern auf individueller Ebene und mit guter Genauigkeit zu identifizieren. Die Klassifikationsmodelle gewinnen dabei an Genauigkeit, wenn ihnen größere Datenbanken zur Verfügung gestellt werden. Unsere Ergebnisse weisen weiter daraufhin, dass die erstellten Klassifizierungsmodelle auch auf Patienten anderer Studienzentren anwendbar und somit generalisierbar sind. Zusammengefasst könnte die Anwendung neurokognitiver Daten und computerisierter Mustererkennungsmethoden die Differentialdiagnostik im klinischen Alltag in Zukunft weiter verbessern und unterstützen.

1.2 English Abstract

Background Recent findings have indicated that it is possible to differentiate between recent onset psychotic (ROP) and recent onset depressive (ROD) patients, using MRI-based markers in multivariate pattern classification with high accuracy [Koutsouleris et al., 2015]. However, especially those depressive patients with early onset of disease tend to be misclassified as ROP, leading to the assumption of a neuroanatomical likeness between these patient groups [Koutsouleris et al., 2015]. Furthermore, evidence suggests that major depressive as well as psychotic disorders present specific profiles of cognitive dysfunction early on in the course of the disease [Lee et al., 2012, Bora and Murray, 2013]. Therefore, using neurocognitive data in the framework of validated multivariate pattern classification could potentially facilitate clinical differentiation of recent onset psychotic (ROP) and recent onset depressive (ROD) patients.

Objective In this study, we conducted univariate and multivariate statistical analyses in order to (i) identify performance differences between ROD and ROP subjects on the PRONIA Neuropsychological Test Battery, (ii) establish classification models in the framework of multivariate pattern analysis that reliably differentiate between ROD and ROP subjects based on patterns in the neurocognitive data and (iii) examine these classification models in a leave-center-out analysis in order to assess their generalizability.

Methods Preliminary data of 116 subjects (58 ROD (mean age: 25.5 ± 6.0 yr), 34 female; 58 ROP (25.6 ± 5.2yr), 16 female) from the PRONIA-project was examined (cf. https://www.pronia.eu/). Subjects' diagnoses were assessed using the Structured Clinical Interview for DSM-IV - Axis I (SCID-I). Group performances on the neurocognitive assessment were compared by conducting Bonferroni-adjusted 2-sample t-tests on scores from 13 well established neurocognitive assessments (e.g. Rey-Osterrieth Figure Complex Figure Test, Trail Making Test A and B, Rey Auditory Verbal Learning Test etc). Classification models were trained on results from the multi-domain PRONIA Neuropsy-

chological Test Battery using Support Vector Machine (SVM)-based multivariate pattern analysis. This was conducted separately for 23 and 214 neurocognitive variables. Lastly, generalizability was estimated by conducting a leave-center-out multivariate analysis.

Results ROP patients performed relatively poorer than ROD subjects in almost all tests. Statistically significant performance differences were found in 9 out of 23 variables. Accordingly, ROP patients presented greater deficits in the cognitive domains *speed of processing*, *attention*, *verbal learning*, *executive functions* and *premorbid IQ*.
For the ROD vs. ROP analysis, cross-validated classification accuracies were 63.8% in the 23 variables analysis and 71.6% in the 214 variables analysis, respectively. The average classification accuracy in the leave-center-out analysis was 72.4%, with both sensitivity and specificity of 72.4%.

Conclusions ROP patients generally exhibit far greater cognitive deficits than ROD patients. By implementing machine learning algorithms trained on neurocognitive data to differentiate between diagnostic groups, it is possible to reliably identify patients with depressive and psychotic disorders on an individual level. These classifiers gain diagnostic accuracy when presented with larger data sets. Our findings further indicate that these classification models are generalizable to patients from other clinical centers. Hence, pattern recognition methods on neurocognitive data may improve machine based diagnostic differentiation and therefore enhance clinical diagnostic accuracy in the future.

2 Introduction

2.1 Depression

2.1.1 Definition and Epidemiology

Major depressive disorder (MDD) is one of the most prevalent mental disorders today, with patient numbers continuously increasing over the last years. In Germany, the lifetime prevalence of depression lies around 9.9% with women being affected about twice as often as men [Kessler and Bromet, 2013, Association et al., 2013].

Furthermore, the World Health Organization (WHO) expects MDD to rank second in leading diseases, measured in disability-adjusted life years, by the year 2020, surmounted only by heart disease [Chapman and Perry, 2008]. As a consequence, in 2004 the costs of MDD in Europe was already estimated to mount up to Euro 118 billion and therefore accounting for a major economic burden [Sobocki et al., 2006].

The typical age-of-onset of depression lies between the age 30 and 40 with an additional albeit smaller peak of incidence between the ages 50 and 60 [McGorry et al., 2011, Kessler and Bromet, 2013].

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM5), MDD is defined as the presence of two main depressive symptoms: depressed mood and anhedonia, at least one of which has to be present on most days for at least two consecutive weeks. In addition to that, MDD is accompanied by numerous symptoms such as loss of interest, change in eating, appetite or weight, sleep disorders, lethargic or overly active motor activity, fatigue and suicidal tendencies [Association et al., 2013]. For the exact diagnostic criteria, see chapter 2.1.3.

2.1.2 Aetiology

The exact cause of depression is still unknown and remains the focus of ongoing scientific investigation. However, it is generally believed that depression, like most other mood disorders, has a multifactorial aetiology where genetic, neurochemical, hormonal

and psychological factors have to be equally taken into account [Ehlers et al., 1988, Sullivan et al., 2000].

Concerning the pathophysiology, different theories addressing the monoaminergic system, autoimmune mechanisms and the circadian rhythm have been discussed in recent years. Theories about the monoaminergic circuit indicate a defective signal transmission regarding serotonin, dopamine and noradrenaline [Hamon and Blier, 2013], whereas other research gives reason to believe that immune system abnormalities have a propulsive effect in causing depressive symptoms [Köhler et al., 2014].

2.1.3 Diagnostic Criteria

In order to diagnose a patient with MDD, certain diagnostic criteria have to be fulfilled. Depending on the respective state or country, these criteria are defined by the Statistical Manual of Mental Disorders (DSM5) or the International Statistical Classification of Diseases and Related Health Problems (ICD-10) developed by the WHO. Below, I will only contemplate the criteria as described in the DSM-5 since diagnostic classification in this study was performed using the Structured Clinical Interview for DSM-5 (SCID).

Firstly, 5 or more of 9 symptoms have had to be present throughout two consecutive weeks. Furthermore, in order to fulfill the criteria at least one of the first two symptoms had to be present:

1) **depressed mood**

2) **loss of interest or pleasure**

3) significant weight loss or gain

4) insomnia or hypersomnia

5) psychomotor agitation or retardation

6) fatigue or loss of energy

7) feelings or worthlessness or guilt

8) loss of concentration and indecisiveness

9) recurrent thoughts of death or suicidal tendencies/attempts

Secondly, symptoms have to be severe enough to cause significant distress or impairment in social, occupational or other areas of functioning, while the depressive episode is not attributable to the effects of a substance or another medical condition.
Lastly, it has to be excluded that the depressive episode is not better explained by any mental disorder of the psychotic spectrum nor did a manic or hypomanic episode ever occur in the patient history [Association et al., 2013].

2.1.4 Therapy and Prognosis

The management and treatment of major depressive disorder consists of diverse therapeutic approaches such as antidepressant medication, psychotherapy, lifestyle adjustments, electroconvulsive therapy, transcranial magnetic stimulation and others. The specifics of the treatment may differ between individuals based on the severity, number of episodes or other peculiarities of the depressive episode.
Although about 50% of major depressive episodes resolve by themselves, whether they are treated or not, reoccurrence is much more likely when inappropriate or no treatment has been administered [Eaton et al., 2008, Geddes et al., 2003]. In general, about 80% of patients with a major depressive episode will experience at least one more episode in their lifetime [Fava et al., 2003].
Furthermore, people affected by MDD statistically exhibit a reduced life expectancy [Cassano and Fava, 2002]. This is for one due to the elevated number of suicides among this cohort, but also conditioned by an increased likelihood for an unhealthy conduct of life resulting in higher risk for heart disease and other medical conditions
[Leung et al., 2012, Alboni et al., 2008].

2.2 Psychosis

2.2.1 Definition and Epidemiology

Psychosis - or rather the psychotic episode - is generally understood as a temporary or permanent state of mind in which the affected person suffers from a detachment of reality. Thus, the patient may experience a diverse spectrum of symptoms including hallucinations, delusions, thought disorder, aberrant salience and catatonia. Psychosis however is not a psychiatric diagnosis by itself, rather than a condition that may occur in the context of a multitude of psychiatric entities such as schizophrenia, schizophreniform disorders, schizoaffective disorder, delusional disorder, or mood disorders like bipolar disorder and Major Depressive Disorder [Cardinal and Bullmore, 2011].

Likewise, psychotic episodes can be caused or induced by a vast variety of substances and medical conditions, making it essential in the diagnostic process to exclude any possible external causes before confirming a definite psychiatric diagnosis [Cummings, 1985, Cardinal and Bullmore, 2011].

Due to this polymorphism of psychosis, exact numbers for incidence and prevalence remain somewhat unclear. Targeting all forms of psychotic disorders, a study from 2007 conducted by researchers of the University of Helsinki described the all-over lifetime prevalence as 3.06%, in a general population [Perälä et al., 2007]. Thereof, Perälä *et al.* further state, the majority was made up by patients with schizophrenia, accounting for a prevalence of 0.87%. Accordingly, the prevalence of schizophrenia alone is generally estimated to lie between 0.3 and 0.8% in an average western population [McGrath et al., 2008].

For example, a systematic review of 2012 marked a pooled annual incidence of 32 cases per 100.000 people for psychotic disorders in general and 15 per 100.000 people for schizophrenia in specific [Kirkbride et al., 2012]. The average age-of-onset for psychotic disorders lies between 20 and 35 years, with a slightly later onset of episodes that have an underlying affective disorders. Moreover, it is a well-established fact that the average age-of-onset in women is approximately 5 years later than in men [Loranger, 1984].

2.2.2 Aetiology

Analog to MDD and most other mental health disorders, the exact causes of psychosis remain elusive. However, since the first medical reports of psychosis in the 20$^{\text{th}}$ century, numerous factors causing or contributing to the outbreak of psychosis have been identified and hypothesized.

On a neurochemical level, undoubtedly the most common theory is the dopamine hypothesis of schizophrenia and psychosis. This hypothesis attributes psychotic symptoms to a dysbalance of dopaminergic signal transductions, especially in the mesolimbic pathway [Kapur et al., 2005]. Originally, this theory arose due to the observation that drugs blocking D2 dopamine receptors in the brain seem to reduce psychotic symptoms, whereas drugs and substances leading to a higher dopamine output or concentration have shown to potentially induce or aggravate psychotic symptoms.

As to how these transmitter dysbalances are triggered remains unclear, with evidence suggesting hereditary, neurodegenerative, drug and stress related as well as social and environmental factors [Drake et al., 2000, Broome et al., 2005].

2.2.3 Diagnosis

When confronted with a patient that presents psychotic symptoms, such as delusions, hallucinations or any other form of bizarre thought disorder, it is of foremost importance to rule out any potential organic or drug related cause of the psychotic episode [Griswold et al., 2015]. To do so, health care professionals avail themselves with medical histories, brain scans, cerebrospinal fluid diagnostics, blood tests, neurological examinations and a multitude of further diagnostic instruments.

Only if the subsequent results are negative, can the physician rule out a secondary cause of the psychotic episode and proceed to examine the exact nature of the primary psychotic disorder at hand. Here, the intensity, frequency and duration of the respective symptoms, as well as the knowledge about any prior episodes of psychotic or affective disorders are decisive for the exact psychiatric diagnosis [Association et al., 2013].

2.2.4 Therapy and Prognosis

In order to treat acute psychotic episodes, healthcare professionals nowadays primarily rely on the use of antipsychotics [Stahl and Mignon, 2010]. Antipsychotics, traditionally divided into typical and atypical subtypes, are drugs targeting a range of dopamine receptors, but also histamine and serotonine receptors in the brain and have shown to effectively reduce psychotic symptoms in patients with primary psychosis.

However, in modern psychiatry, psychotherapeutic programs and occupational therapy play a substantial role in the treatment of psychotic disorders [Moritz et al., 2011]. By implementing psychotherapeutic, psychoeducational and psychosocial programs into the treatment, psychiatrists hope to achieve a better rehabilitation of people suffering from psychosis and therefore a higher outcome of compliance and remission in the long term [Segarra et al., 2012].

The prognosis for patients suffering from a psychotic episode is greatly dependent on the cause of the initial outbreak. While secondary psychoses often have a good prognosis when the underlying disorder is treated, the outcome of primary psychoses tends to vary from patient to patient. According to a systematic review on longitudinal outcome studies of first episode psychoses, about 42% of cases with first episode psychosis (FEP) will see a good outcome. 35% of the affected people will have an intermediate outcome with some remaining symptoms or impairments and about 27% of the patients will experience a poor outcome with a chronic progression of the disease [Menezes et al., 2006].

Additionally, according to Laursen *et al.*, patients with schizophrenia have a reduced life expectancy of 10 to 25 years, due to unhealthy lifestyles, the side effects of the antipsychotic treatment and an increased risk of committing suicide [Laursen et al., 2012].

2.3 Neuropsychology in Psychiatry

The scientific discipline of neuropsychology studies the processes and function of the brain in its diverse domains such as memory, intelligence, verbal and motoric function-

ality and many more. In that, neuropsychology links findings from the disciplines neurology, psychology, cognitive neuroscience and, with regards to artificial neural networks, even computer science. To do so, neuropsychologists utilize neuroimaging methods like magnetic resonance imaging (MRI), clinical neuropsychological assessments, electrophysiological monitoring of the brain like electroencephalography (EEG) and numerous other methods, to obtain a deeper understanding of the specific brain mechanisms leading to cognitive processes and behavioral patterns.

In psychiatry, clinical neuropsychology plays an increasingly important role by conducting comprehensive assessments with individual patients and therefore gathering valuable information contributing to an accurate diagnostic classification [Keefe, 1995].

Furthermore, neuropsychological assessments have become a popular instrument in psychiatric research for the quantification of therapeutic effects of new medications or psychotherapeutic treatments. On an individual level, results from neurocognitive assessments enable tailored cognitive training plans for patients suffering from cognitive deficits and thus enhancing disease remission and occupational rehabilitation [Keefe, 1995].

2.3.1 Neuropsychology and Depression

For many years, psychologists and psychiatrists accepted cognitive impairment in depressive patients as a mere side effect of depressed mood or understood poor performance in neuropsychological assessments as a result of lacking motivation and inattention. However, a meta-analysis of 2001, performed by Austin *et. al*, suggested a disease specific pattern of cognitive impairments in depressive patients. Interestingly, this pattern -mainly including deficits in episodic memory, learning, executive functions and attentional set-shifting- is not necessarily depending on the severity of the illness, nor on the age of the study participants [Austin et al., 2001]. Hence, Austin *et. al* concluded that impairment in memory and executive functioning are not simply attendant circumstances in depressive disorders, rather than a core feature of this diagnostic category [Austin et al., 2001]. Another study review, published by McIntyre *et al.* in 2013, confirmed cognitive deficits

in depression to mainly present in the domains executive functions, working memory, attention and processing speed [McIntyre et al., 2013].

In 2012, researchers around Rico S.C. Lee from the Macquarie University, NSW, Australia, conducted a meta-analysis exclusively focusing on cognitive deficits in patients with first episode MDD. Here, significant results were found for cognitive impairment in the domains psychomotor speed, attention, memory and executive functioning even in an early stage of the disease [Lee et al., 2012]. Furthermore, Lee and colleagues could not find any links between cognitive deficits and illness severity, indicating once again that the specific cognitive deficits may be a defining feature of MDD [Lee et al., 2012].

2.3.2 Neuropsychology and Psychosis

Neuropsychological findings concerning cognitive impairment in psychosis have been well described and extensively studied. Heinz and Zakzanis published a meta-analysis in 1998, indicating vast cognitive impairment in patients with schizophrenia compared to healthy controls, with an emphasis on the domains verbal memory, intelligence quotient (IQ), attention and verbal fluency [Heinrichs and Zakzanis, 1998]. A metaanalyis of 2009, conducted by Mesholam-Gately *et al.* confirmed these results and, while strictly focusing on first episode schizophrenia, found impairments to be maximal in verbal memory and processing speed [Mesholam-Gately et al., 2009].

Further research suggests that similar patterns of impairment can be found in other diagnostic groups of the psychosis spectrum, such as schizoaffective disorder or affective disorders with psychotic symptoms, albeit patients with schizophrenia still present greater measurable impairment in comparison [Reichenberg et al., 2008].

Nevertheless, cognitive deficits in the domains memory, executive functions, attention and processing speed seem to be a conjunctive pattern when examining different forms of psychosis [Reichenberg et al., 2008]. In addition, a 5 year follow-up study examining the neurocognitive development in first episode psychosis found the course of neurocognition to be more dependent on the number of recurring episodes than the exact diagnosis

[Barder et al., 2013].

It is also being discussed whether the described cognitive deficits may not simply be an epiphenomenon of full-blown psychosis rather than a reliable trait of the illness that often presents even before the prodromal stage of a psychotic episode becomes clinically apparent [Bora and Murray, 2013]. Hence, a number of researchers in psychiatry have suggested to include cognitive deficits into the diagnostic criteria of psychosis, especially since these deficits are objectively measureable as opposed to delusions or hallucinations [Lewis, 2004, Tsuang and Faraone, 2002].

2.4 Differential Diagnosis in Psychiatry

According to the handbook of the DSM-5, differential diagnosis in psychiatry happens in a number of subsequent steps [First, 2013]. Hence, as a first measure when confronted with a patient in psychiatry, it is elementary to rule out Malingering and Factitious Disorder. Secondly and thirdly, the clinician is expected to dismiss drugs and medications as well as any medical conditions as the aetiology of the symptomatology at hand. Only now can the psychiatric clinician begin to examine the exact nature of his or her patient's disorder by following a diagnostic tree that leads from symptom to syndrome to a final diagnosis. To do so, the clinician may make use of a vast number of diagnostic instruments such as the SCID-I and -II or psychometric scales like the Beck Depression Inventory - II (BDI) or the Positive and Negative Syndrome Scale (PANSS).

However, the process and therefore diagnostic accuracy depend on a multitude of diverse factors including the clinician's experience, the physician-patient relationship and the patient's ability to articulate his symptoms. Furthermore, in contrast to most other medical disciplines, physicians in the psychiatric field have few to no additional tools to verify or falsify a diagnosis, other than i.e. neurocognitive assessments and EEG. As a result, diagnosis in psychiatry is a fragile construct oftentimes susceptible to errors [Freedman et al., 2013].

2.5 Multivariate Analysis and Machine Learning Algorithms

Generally speaking, multivariate analysis techniques differ from univariate statistics by taking multiple variables into account when carrying out statistical calculations.
In the framework of supervised learning and computer based classification methods, machine learning algorithms that aim to recognize specific and generalizable patterns in the examined multivariate data have become increasingly popular in modern psychiatric research [Koutsouleris et al., 2014]. In this work's context, supervised learning means that I provide my classification algorithm (e.g. support vector machine) with the subject's label, i.e. diagnosis, when instructing it to establish classification models based on label-specific patterns in the data. Hereby, the classifier knows in which subgroups of the study subjects it has to search for patterns in the data presented to it.
In a nutshell, pattern recognition algorithms aim to identify label-specific patterns in comprehensive datasets, may it be voxel-based MR Imaging data, neurocognitive assessment results or even demographic variables, in order to facilitate inter-label classification.

2.5.1 Support Vector Machine (SVM)

SVMs are a kind of mathematical machine learning classification algorithm that function in a supervised machine learning environment by identifying patterns or differences in before defined samples (e.g. study groups). Subsequently, an SVM can use this acquired knowledge in order to classify new and unseen data instances to one of the known groups [Koutsouleris et al., 2009, Burges, 1998].
From a more technical prospective, the SVM operates as follows: In order to identify said label-specific patterns, each subject is projected in a hyperdimensional space, with the dimensionality corresponding to the number of variables the SVM is presented with. It is now the SVMs' task to create a separating hyperplane (SH) between the two label groups the classifier trains on. To do so, the SVM avails itself with so-called support vectors. These are vectors that indicate the minimal distance between the nearest data entries of the two opposite label groups and the separating hyperplane. In a linearly separable

training data set there are usually many competing hyperplanes. Therefore, in order to identify the optimal separating hyperplane (OSH), the SVM chooses the hyperplane with the longest support vectors. In other words: the best hyperplane is the one that always maintains the maximum distance to all data points of the two separable training groups [Mueller et al., 2015].

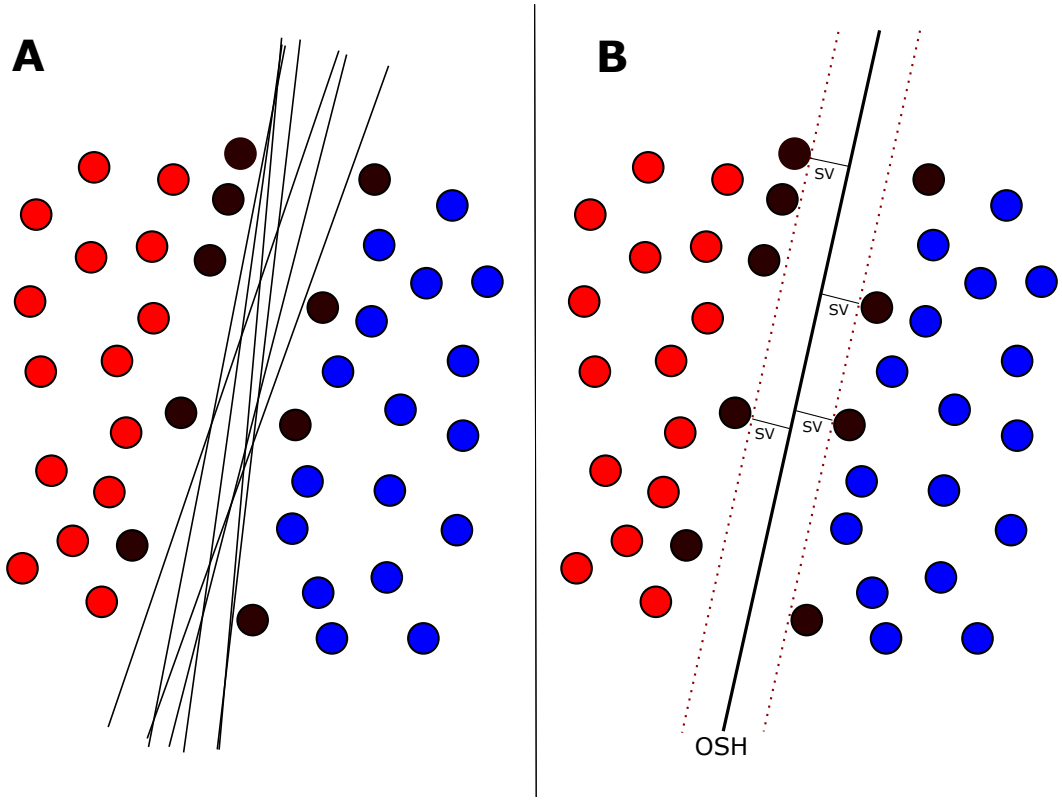

Figure 1: **Schematic representation of the identification of the OSH.** A, a vast number of hyperplanes separating the two training groups can be found. In B, the SVM identified the OSH by maximizing the distance between the nearest training subjects (black dots) of each training group and the OSH. SV: support vector; OSH: optimal separating hyperplane.

It is important to keep in mind that the OSH is a linear mathematical construct that can only function in a linear environment, whereas in reality almost all training data sets are nonlinear. Hence, in order to create such a linear environment, it is possible to transform the training data set into a high-dimensional feature map where linear separation becomes mathematically feasible. In reality however, this procedure would require huge amounts

of computational power. Therefore, it is more convenient to apply a so-called kernel trick. The kernel trick exploits the mathematical nature of the separation problem and makes non-linear seperation possible without projection to higher space. The most common kernel is the radial basis function kernel [Koutsouleris et al., 2009].

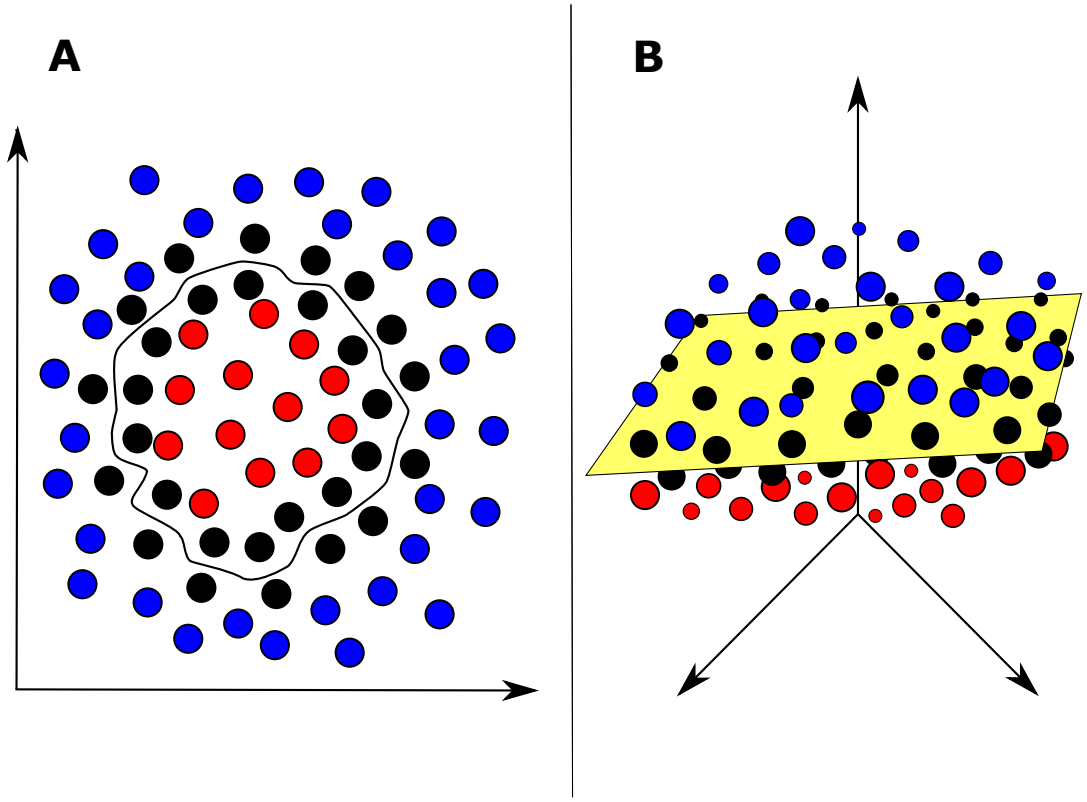

Figure 2: **Schematic representation of the transformation to high-dimensional feature map.** A, the space in non-linear and therefore not separable by a linear OSH. In B, by applying the radial basis functions kernel the data entries are transformed into a high-dimensional feature map, enabling the linear OSH to separate the two training groups. Note that this is only a theoretical representation since the kernel trick avoids actual projection to a higher-dimensional space. Blue dots: training group 1; red dots: training group 2; black dots: nearest subjects to OSH; yellow plane: OSH.

Lastly, since it may not always be possible to find a linear OSH that reliably separates all data entries of the two opposite data groups and also to mitigate the risk of overfitting, it is possible to determine so-called slack variables. These allow the SVM to knowingly misclassify data instances that lie within a certain margin around the OSH.

2.5.2 Cross-validation (CV)

When trying to establish a new diagnostic instrument, e.g. a self-rating questionnaire to measure the severity of a symptom or -as in this work- a classification model, it is always of primary interest to test this instrument for its accuracy, sensitivity and specificity in order to validate its performance. One way to do so is by performing a cross-validation. Here, the data is randomly divided into κ folds. Each fold has approximately the same size and composition. Now, the cross validation will perform κ validation runs with $\kappa - 1$ folds serving as the training data and the remaining fold serving as test data. Hence, each one of the κ folds will serve as test data exactly once. Now, using the $\kappa - 1$ folds as training data, the instrument -in this work the SVM- will generate a prediction model that is then applied to the test data in order to predict the properties of interest (in this work the group membership). Since these properties are known in beforehand, the accuracy of each of the κ models generated can be calculated. Finally, the parameters accuracy, sensitivity and specificity of the κ runs are averaged out and thus indicate the overall performance of the instrument or classifier at hand [Kohavi et al., 1995].

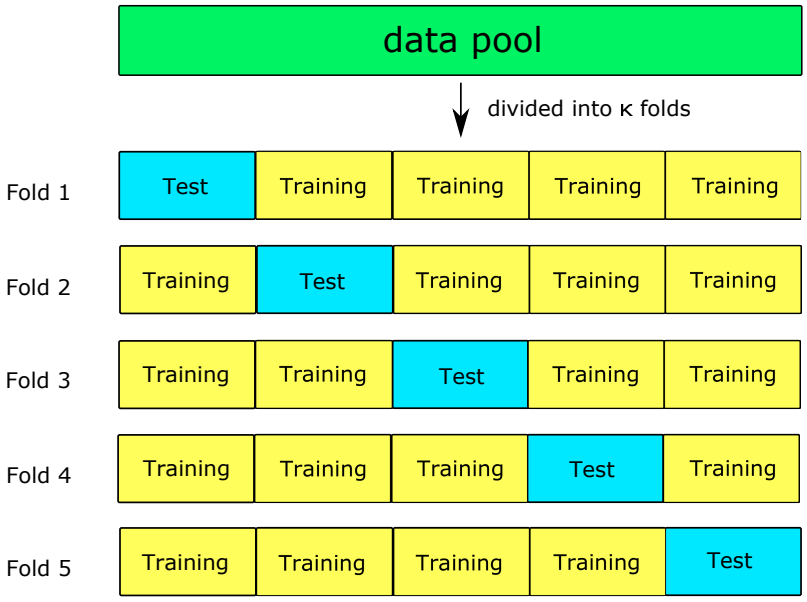

Figure 3: **κ-fold cross-validation in principle.**
For each fold accuracy, sensitivity and specificity of the model are calculated.

Repeated Double Cross-Validation (rdCV) CV is a suitable technique when aiming to estimate a prediction model's performance. However, with small sample sizes, simple κ-fold cross-validation is running the risk of eliding a possible overfitting effect [Faber, 2007]. In previous research, repeated double cross-validation has been proposed to be an eligible solution for this problem [Filzmoser et al., 2009].

Principally, rdCV functions the same way as CV with the only difference that it has an inner (CV1) and outer (CV2) loop. While the outer CV2 loop is comparable to the κ-fold cross-validation, the CV1 is a loop with again κ permutations generated from the data of each fold of the outer loop. This way, in the inner loop the generated models can tune its parameters for maximum performance and then test these models on the outer loop. By that, the number of tests is increased and therefore prediction performance and variability can be better estimated [Filzmoser et al., 2009].

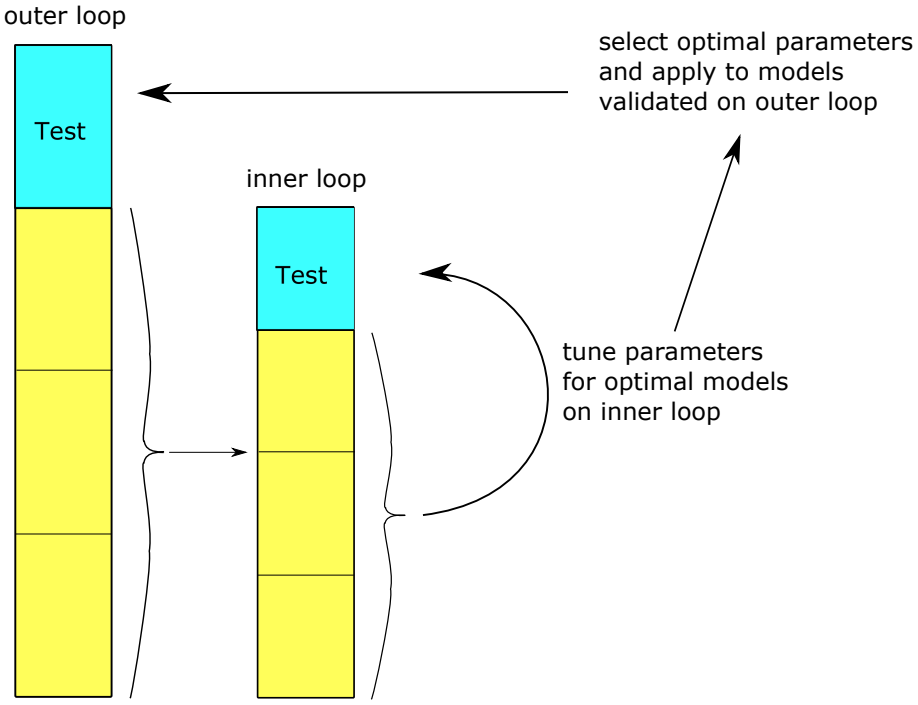

Figure 4: **Scheme of repeated double cross-validation.** Cross-validation is performed on the inner loop (= CV1). Best parameter and variable set-ups are applied to models validated on the outer loop (= CV2). Therefore 'double cross-validation'.

2.5.3 Pattern Recognition Analyses in Psychiatric Research

In recent years, machine learning algorithms have been increasingly implemented in the framework of psychiatric research, with scientists hoping to establish new classification models that will enhance diagnostic accuracy in psychiatry and that may even support clinicians in their therapeutic decision making.

Originally, in the early years of this millennium the application of machine learning algorithms to functional MRI data shaped the term multivariate pattern analysis (MVPA) [Haxby, 2012].

As a result, in the following years these methods have been frequently applied to MR Imaging data with the aim of revealing brain-morphologic, disease-specific patterns that remain elusive to mere visual inspection. For instance, in 2005, Davatzikos *et al.* used SVM algorithms on MRI data to differentiate between patients with schizophrenia and healthy controls, reaching an overall accuracy of 81.1% and thus providing additional evidence for the usefulness of MVPA [Davatzikos et al., 2005]. In a similar study of 2007 that performed a partial least squares analysis instead of operating an SVM, Kawasaki and colleagues reached an accuracy of 90% when classifying schizophrenic patients vs. healthy controls based on their brain MR images [Kawasaki et al., 2007].

In a meta-analysis of 2015, Kambeitz *et al.* assessed a total of 38 studies that implemented multivariate pattern recognition methods on brain scans in order to differentiate schizophrenic patients from healthy controls. As a result of their research, they found that differentiation based on functional and structural brain alterations was achieved with an overall sensitivity of 80.3% and a specificity of 80.3%, with studies focusing on resting state functional MRI reaching higher classification results than studies focusing on structural MRI data (sensitivity/specificity: 84.46%/76.9% and 76.4%/79.0%, respectively) [Kambeitz et al., 2015].

However, as a second metaanalysis of 2016 suggests, these patterns of brain alterations are not limited to psychotic disorders. Instead, when examining 33 studies that applied MVPA to magnetic resonance imaging in order to differentiate between healthy controls

and patients with major depressive disorder, classification results were almost equally as effective (sensitivity: 77% and specificity: 78%) [Kambeitz et al., 2016].

Further, a study of 2008 by Christos Davatzikos and his team of the University of Pennsylvania presented evidence that MVPA can detect patterns of brain alterations in prodromal Alzheimer's disease and therefore even before the breakout of the apparent illness [Davatzikos et al., 2008].

Based on these findings, the question has been raised whether multivariate pattern recognition may not only facilitate the differentiation of diseased vs. healthy subjects, but also present a serviceable instrument in differential diagnosis of diverse psychiatric disorders. Accordingly, a study of 2015 by Koutsouleris *et al.* conducted multivariate pattern analysis on structural MRI data in order to separate patients with schizophrenia and patients with MDD. As a result, the classifiers reached a balanced classification accuracy of 76%, with 79.8% of the schizophrenic and 72.2% of the depressive patients being correctly assigned to their respective cohort [Koutsouleris et al., 2015]. To test for generalizability, Koutsouleris *et al.* also applied the beforehand generated pattern recognition models to an independent patient cohort of 35 patients with bipolar disorder, 23 patients with first episode psychosis and 89 subjects with clinically defined at-risk mental states for psychosis (ARMS). Here, 74% of the bipolar patients were assigned to the MDD group, while 83% of the FEP patients and an average of 69% of the ARMS group were labelled as schizophrenic [Koutsouleris et al., 2015]. As a consequence to these results, Koutsouleris *et al.* conclude that neuroimaging pattern recognition may be a practical and valid instrument for differential diagnosis in psychiatry [Koutsouleris et al., 2015].

Furthermore, it has been discussed whether MVPA might also facilitate the early detection of mental disorders, as it may detect brain-morphologic alterations long before the outbreak of a psychiatric illness. Prior research has already shown promising results, with neuroanatomical biomarkers helping to identify ARMS subjects and their respective clinical outcome and therefore offering a reliable tool for early detection of psychosis [Koutsouleris et al., 2009].

In addition to that, evidence suggests that early detection and disease prediction is also feasible when focusing solely on disease-specific neurocognitive patterns, giving reason to believe that multi-modal pattern recognition methods may enhance diagnostic classification accuracy as well as early disease detection and outcome prediction in the future [Koutsouleris et al., 2011].

2.6 Aims of this Study

In this study, I performed both, univariate and multivariate statistics on neurocognitive data of patients with recent onset depression (ROD) and recent onset psychosis (ROP). The aims of the study were as follows:

i) identifying differences in the performance on the neurocognitive test battery between ROD and ROP subjects, by performing univariate statistical analysis,

ii) generating classification models based on neurocognitive pattern analysis that differentiate reliably between ROD vs. ROP patients and

iii) applying classification models that differentiate between ROD and ROP subjects to an independent study sample from an outside study center in order to test for generalizability (leave-center-out analysis).

In addition, I intend to compare the performance of multivariate classifiers that have been trained with variables reported in previous literature, with classifiers that have been provided with a larger set of variables of the neuropsychological test battery.
Lastly, I aim to compare those variables in our models that were most decisive for the assignment to a study group with those variables that indicate a cognitive deficit generally associated with the respective mental disorder. Hence, tackling the question whether our machine learning algorithms identify the same patterns of cognitive deficits that have been well established based on the findings of prior research relying solely on univariate statistics.

3 Materials and Methods

3.1 The PRONIA Study

All data analyzed in this study has been drawn from a preliminary PRONIA data set of 2015. The PRONIA Study (Personalised Prognostic Tools for Early Psychosis Management) is a European research project, under the coordination of Prof. Dr. Nikolaos Koutsouleris (https://www.pronia.eu/), with study centers at the Ludwig-Maximilians-University Munich (LMU), the University of Cologne (UKK), the University of Basel (UBS, Switzerland), the University of Birmingham (Uni BHAM, Great Britain), the University of Udine (Uni Udine, Italy), the University of Turku (Finland), the University of Milan (MilanNig, Italy) and the University of Melbourne (Australia). Additionally, the PRONIA consortium consists of four partners from industry and commerce (resp. Dynamic Evolution, General Electric Global Research, General Electric Healthcare and GABO:mi Gesellschaft für Ablauforganisation :milliarium).

The main goal of the PRONIA Study is to generate a software-based prognostic system that facilitates differential diagnosis as well as early risk prediction of different psychiatric disorders, using brain imaging and complementary data. In order to achieve this objective the PRONIA Study aims to recruit a total of 1700 participants. Each study subject is undergoing an extensive examination in which a variety of demographic, biometric and psychometric data is collected. After baseline examination, eligible study participants are separated into the study groups (i) Healthy Controls (HC), (ii) Recent onset Depression (ROD), (iii) Recent onset Psychosis (ROP) and (iiii) Clinical High Risk for Psychosis (CHR). For a detailed list of the psychometric and demographic instruments implemented at study baseline, see Table 1. Furthermore, all study participants receive between six and seven follow-up examinations, depending on the respective study group (Table 2). The PRONIA Study has been awarded a 6.000.000€ grant by the European Union within the 7th Framework Programme and runs from October 1st 2013 until October 2018. Grant agreement n°602152.

Observer rating instruments
1. Structured Interview for Prodromal Syndromes (P Items)
2. Schizophrenia Proneness Instrument - Adult Version
3. Comprehensive Assessment of At Risk Mental State
4. Global Assessment of Functioning
5. Structured Clinical Interview for DSM IV - 1 (Screening + Interview)
6. Demographic and Biographic Data
7. Premorbid Adjustment Scale
8. Positive and Negative Symptom Scale
9. Scale for the Assessment of Negative Symptoms
10. Functional Remission in General Schizophrenia
11. Global Functioning Scales (Social and Role)
12. Chart of Life Events
Self rating instruments
1. WHO Quality Of Life - Short Version
2. Multidimensional Scale of Perceived Social Support
3. Resilience Scale for Adults
4. Coping Inventory for Stressful Situations
5. Social Phobia Inventory
6. Beck Depression Inventory - II
7. Edinburgh Handedness Inventory - Short Version
8. Level of Expressed Emotion Scale
9. Wisconsin Schizotypy Scales - Short Form
10. The Everyday Discrimination Scale
11. Bullying Scale
12. Childhood Trauma Questionnaire
13. NEO Five-Factor Inventory
Neuropsychological instruments
1. Rey-Osterrieth Complex Figure Test
2. Diagnostic Analysis of Nonverbal Accuracy - 2nd Version
3. Forward and Backward Digit Span
4. Semantic and Phonemic Verbal Fluency Test
5. Rey Auditory Verbal Learning Test
6. Trail Making Test A + B
7. Continuous Performance Test - Identical Pairs
8. Self-Ordered Pointing Test
9. Digit Symbol Substitution Test
10. Wechsler Adult Intelligence Scale, Vocabulary + Matrices
11. Salience Attribution Test

Table 1: **Data acquisition instruments as used in the PRONIA Study baseline assessment (T0).**

Examination	HC	ROD	ROP	CHR
T0 - Baseline testing	X	X	X	X
IV3 - 3 months follow-up interview	-	X	X	X
IV6 - 6 months follow-up interview	-	X	X	X
T1 - 9 months follow-up testing	X	X	X	X
IV12 - 12 months follow-up interview	-	X	X	X
IV15 - 15 months follow-up interview	-	X	X	X
T2 - 18 months follow-up testing	-	X	X	X

Table 2: **Follow-up examinations and respective study groups.** Neuropsychological testing and MRI brain scans are only conducted at baseline, after 9 and 18 months (that is T0, T1 and T2).

3.2 Subjects

3.2.1 Study population

In the present study, I analyze baseline-data of 116 subjects drawn from a preliminary PRONIA dataset of 2015, originally consisting of 401 study subjects from the seven recruiting centers LMU (92 subjects), UKK (98 subjects), UBS (70 subjects), Uni BHAM (33 subjects), Uni Udine (51 subjects), MilanNig (15 subjects) and Turku (42 subjects). All data of said preliminary dataset has been collected between January 2014 and June 2015.

Of the 116 participants included in this study, 58 presented with recent onset psychosis (ROP) and 58 with recent onset depression (ROD). Subjects were recruited and tested at LMU, UBS, UKK, Uni BHAM and Uni Udine (Table 3).

Subjects recruited at the University of Turku were not included in this study, due to an incongruent Finnish version of the neuropsychological test protocol. Furthermore, none of the 15 subjects of the recruitment center MilanNig were included in this study, since no neuropsychological data was available from this center, at the time of data analysis. Another three subjects were missing more than 25% of data due to technical errors and were thereupon excluded prior to analysis (Table 4).

Study group	LMU	UBS	UKK	Uni BHAM	Uni Udine	TOTAL
ROD (n)	23	9	11	3	12	**58**
ROP (n)	18	13	17	5	5	**58**
TOTAL (n)	**41**	**22**	**28**	**8**	**17**	**116**

Table 3: **Distribution of subjects across centers.** Distribution of the 116 subjects drawn from the preliminary PRONIA dataset of 2015.

Subject	study center	study group	missing data (%)
#1	LMU	ROP	100.0
#2	UBS	ROP	95.8
#3	UBS	ROP	52.3

Table 4: **Subjects excluded due to missing data.** Information about the respective study centers, study groups and percentage of missing data.

3.2.2 Inclusion and Exclusion Criteria

Inclusion criteria In order to be considered as eligible for study inclusion, subjects had to be between 15 and 40 years old, needed sufficient language skills and had to be in possession of sufficient capacity to consent. Subjects under the age of 18 additionally had to produce the consent of their parents or legal guardians.

Subjects assigned to the ROD group had to fulfill the criteria for a Major Depressive Episode according to the DSM-IV-TR within the last three months prior to study inclusion (DSM-IV-TR, American Psychiatric Association, 2000). In addition, the duration of the first depressive episode had to be under 24 months, with the study screening visit as reference date.

For inclusion into the ROP group, subjects had to fulfill the criteria for DSM-IV-TR affective or non-affective Psychotic Episode within the past three months prior to the study screening visit. Altogether, the onset of the psychotic episode had to be within the past 24 months, again with the study screening visit as reference date.

Exclusion criteria Subjects were not considered for study inclusion if they fulfilled one or more of the following criteria:

i) Intelligence Quotient (IQ) below 70

ii) insufficient hearing for neuropsychological testing

iii) current or past head trauma with a loss of consciousness for more than 5 minutes

iv) current or past known neurological disorder of the brain

v) current or past known somatic disorder potentially affecting the structure or functioning of the brain

vi) current or past alcohol dependency

vii) current polytoxicomania or polytoxicomania (according to SCID-I) within the past six months

viii) infeasibility of MRI scanning due to medical reasons

A detailed list of all somatic disorders potentially affecting the structure or functioning of the brain that subsequently lead to study exclusion can be found in the appendix.
In addition to the general criteria of exclusion, potential ROD subjects were not considered for study inclusion if they (i) had had more than one Major Depressive Episode in their life time, (ii) had taken antipsychotic medication for more than 30 cumulative days in their life time or within three months prior to the study screening visit at or above minimum dosage of the '1st episode psychosis' recommendations of the DGPPN (Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde) S3 Guidelines.
Subjects assigned to the ROP group were not included if they had taken any antipsychotics for more than 90 cumulative days within the last 24 months with a daily dose at or above minimum dosage according to the '1st episode psychosis' recommendations of the DGPPN S3 Guidelines.

3.2.3 Demographic Data

Subjects included in this study were both inpatients as well as outpatients. Of the 116 subjects examined in this study, 43% were female. The mean age of the total sample was 25.5 ±5.6 with no significant age difference across study groups. A more detailed listing of the demographic and descriptive variables of the study population at hand can be found in Table 5.

Demographic and descriptive variables	**ROD**	**ROP**	**TOTAL**	**P**
n	58	58	116	
Sex (female) [n]	34	16	50	
Sex (male) [n]	24	42	66	
Mean age [yrs] (SD)	25.5 (6.0)	25.6 (5.2)	25.5 (5.6)	ns
Mean BDI (SD)	25.7 (13.1)	24.6 (12.3)	25.2 (12.6)	ns
Mean PANSS total (SD)	46.2 (9.9)	74.4 (19.5)	60.3 (20.9)	< 0.001
Mean PANSS positive (SD)	7.69 (2.2)	19.5 (6.0)	13.6 (7.3)	< 0.001
Mean PANSS negative (SD)	12.2 (4.2)	17.2 (7.3)	14.7 (6.5)	< 0.001
Mean PANSS general (SD)	26.5 (6.2)	37.7 (10.0)	32.1 (10.0)	< 0.001
Mean GAF-S (SD)	55.0 (12.8)	40.5 (13.3)	47.7 (14.9)	< 0.001
Mean GAF-DI (SD)	54.4 (14.6)	44.3 (12.8)	49.3 (14.5)	< 0.001
Mean WAIS-A	11.0 (2.3)	8.99 (2.6)	10.0 (2.7)	< 0.001

Table 5: **Demographic and descriptive measures.** Descriptive analyses were performed with t-tests. BDI: Beck Depression Inventory - II, PANSS: Positive and Negative Syndrome Scale, GAF: Global Assessment of Functioning, WAIS-A: Wechsler Adult Intelligence Scale - Average.

3.3 Psychometric Instruments

Beck Depression Inventory - II First presented by Beck *et al.* in 1961, the Beck Depression Inventory (BDI) is a self-report rating instrument measuring the severity of depressive attitudes and symptoms of depression [Beck et al., 1961]. In this study, we

utilize the revised second version of the BDI, as published by Beck and colleagues in 1996 [Beck et al., 1996]. The inventory contains 21 multiple-choice questions covering diverse symptoms of depression. Scores from 0 to 3 are assigned to the answer options of each item, thus resulting in possible sum scores of 0 to 63 points. Evaluation is as follows: scores above 29 indicate severe depression, scores between 20 and 28 indicate moderate depression, scores from 14 to 19 point towards mild depression and scores between 9 and 13 suggest minimal depression or depressive symptoms. Scores below 8 classify as no depression [Köllner and Schauenburg, 2012].

The BDI takes approximately 10 minutes for completion and requires basic reading abilities. When completing the BDI-II, subjects should only answer based on their state of health during the last two weeks.

Positive and Negative Syndrome Scale Divided into the three scales *positive*, *negative* and *general psychopathology*, the Positive and Negative Syndrome Scale (PANSS) is an interviewer-rated instrument, designed to quantify different symptom classes in schizophrenia. It was published by Stanley Kay and colleagues in 1987 and is a widely used instrument in both psychiatric research and clinical praxis [Kay et al., 1987]. The test interview takes 45 to 60 minutes and must be conducted by a psychological or medical professional, specifically trained for this instrument.

The PANSS was designed to detect the different qualities of symptoms in schizophrenia. Accordingly, the *positive scale* consists of 7 items that examine 'positive' symptoms such as delusions, hallucinations or hyperactivity, whilst the *negative scale* covers 7 symptoms like emotional withdrawal and blunted affect. The *general psychopathology scale* detects a variety of 16 other pathologies, such as anxiety, disorientation or poor impulse control. Each item is rated by the interviewer with 1 to 7 points depending on the severity of the respective symptom. Hence, the sum score for each scale varies as follows: *positive scale:* 7 to 49 points, *negative scale:* 7 to 49 points, *general psychopathology scale:* 16 to 112 points. When originally tested on patients with schizophrenia by Kayle and colleagues in

1987, the mean scores were 18.20 points on the *positive scale*, 21.01 points on the *negative scale* and 37.74 points on the *general psychopathology scale* [Kay et al., 1987].

Global Assessment of Functioning - Symptoms and Disability/Impairment The Global Assessment of Functioning (GAF) is a scale used to determine the level of functioning of the subject being examined. In this context, 'functioning' refers to the ability to cope with the diverse problems one encounters in one's social, occupational and psychological environments [Hall, 1995]. The GAF is rated by the clinician/interviewer, based on their evaluation of the patient's respective level of functioning. Scores range between 1 and 100, with 1 to 10 being the lowest and 91 to 100 being the highest score. In the PRONIA Study, the GAF is further divided into two assessments: GAF-Symptoms and GAF-Disability/Impairment. Here, the GAF-Symptoms focuses on the subject's actual symptoms, their severity and their impact on the subject's functioning. The GAF-Disability/Impairment takes into account the level of disability or impairment the subject suffers due to its mental health disorder.

3.4 The Neuropsychological Test Battery

All neuropsychological data examined in this study was collected using the standard PRONIA Neuropsychological Test Battery. This protocol consists of 15 neuropsychological assessments covering the domains *speed of processing, attention, working memory, verbal and visual learning, social cognition, executive functions*, *salience attribution* and *premorbid IQ* (Table 6).

Testing was conducted by trained psychologists and medical professionals using a tablet-computer running custom software programmed in PEBL (The Psychology Experiment Building Language; http://pebl.sourceforge.net/) set-up as well as pen and paper versions. Testing took about 120 minutes for completion with an optional 5 minutes break after 60 minutes. Participants were instructed not to consume any psychoactive or stimulating substances before or during testing (e.g. caffeine, alcohol).

Cognitive domain	Assessment	Format
Speed of processing	Digit Symbol Substitution Test	p&p
	Trail Making Test A	p&p
	Phonemic Verbal Fluency Task	p&p
Attention	Continuous Performance Test - Identical Pairs	PEBL
Working memory	Forward and Backward Digit Span	PEBL
	Self-Ordered Pointing Test	PEBL
Verbal learning	Rey Auditory Verbal Learning Test	PEBL
Visual learning	Rey-Osterrieth Complex Figure Test	p&p
Social cognition	Diagnostic Analysis of Nonverbal Accuracy - 2nd Version	PEBL
Executive functions	Trail Making Test B	p&p
	Semantic Verbal Fluency Task	p&p
Salience attribution	Salience Attribution Test	PEBL
Premorbid IQ	Wechsler Adult Intelligence Scale, Vocabulary	p&p
	Wechsler Adult Intelligence Scale, Matrices	p&p

Table 6: **Cognitive domains and neuropsychological assessments.** The Salience Attribution Test can also be assigned to the domain 'reward processing'. PEBL: tablet-based testing format; p&p: pen and paper version.

Rey-Osterrieth Complex Figure Test (ROCF) Originally invented in order to detect and quantify cognitive deficits in patients with traumatic brain injury, the Rey Complex Figure Test has been further developed and applied in different neuropsychological examinations, in order to tackle various problems [Rey, 1941]. In the PRONIA Neuropsychological Test Battery, we used the updated version by Paul-Alexandre Osterrieth, as it has been shown to detect specific cognitive deficits in both schizophrenia as

well as bipolar psychosis [Seidman et al., 2003].

When performing the ROCF, the examinee is asked in a first step to accurately copy a standardized geometrical figure shown on the tablet computer. In a subsequent second step, the participant has to produce another accurate copy of said image from his memory. After a 20 to 30 minutes delay, the participant is asked to again draw the image from his memory in a third and final step.

Altogether, the ROCF takes about 10 to 15 minutes to complete without taking the delay into account. Scoring was performed according to the guidelines developed by Taylor [Strauss et al., 2006]. Here the test figure is divided into 18 different subitems and which are then rated for completeness/accuracy and correct placement. In total we obtained 66 test features with *time of completion* for every one of the three steps being a PRONIA-specific feature originally not being surveyed in the ROCF. All examiners conducting and scoring the ROCF completed and passed a test for interrater reliability.

Diagnostic Analysis of Nonverbal Accuracy - 2nd Version (DANVA) The DANVA is a test examining the social cognition in which the examinee is presented with 24 pictures of adult faces to which he has to assign one of the basic emotions *happiness*, *sadness*, *anger* or *fear* [Nowicki and Duke, 2001]. Whilst the original DANVA also consists of three additional subtests (child faces, child paralanguage, adult paralanguage) the PRONIA Neuropsychological Test Battery limits itself to adult faces. The examination is conducted solely through the PEBL tablet application and the participants enter their answers by operating the tablet touch screen. Before testing, the participant performs one training run under the supervision of the examiner. For the time of testing the examiner must not speak to the participant nor influence him in any other way.

Forward and Backward Digit Span Task (FDS and BDS) The FDS/BDS are tasks examining the working memory and number processing skills of a subject. In these tasks, an audio file reads a series of numbers which the examinee has to immediately repeat to the examiner, who enters the subject's answers into the tablet computer (FDS).

With every two correctly repeated series of numbers the PEBL test software adds one additional number to the digit span. The test ends once the examinee failed two consecutive times to correctly repeat the series of numbers. In a second step, the subject is asked to repeat the numbers presented to him backwards (BDS). Depending on the subject's performance, this task takes 5 to 10 minutes for each subtest (FDS, BDS). We collected two kinds of measures in this test: (i) *maximum digit string length reminded at least once* and (ii) *number of correct trials*.

Semantic and Phonemic Verbal Fluency Task (PVF and SVF) Previous studies suggest that schizophrenic patients tend to present cognitive impairments in both semantic and phonemic verbal fluency [Aloia et al., 1996, Kremen et al., 2003]. Hence, in order to build on these findings the SVF and PVF were implemented in the PRONIA Neuropsychological Test Battery.

In the SVF - as in the PVF - the examinee is asked to produce as many words in 60 seconds with a certain characteristic as possible. In the SVF, these words have to be attributed to a specific category - in our case 'animals'. Furthermore, the subject is instructed not to repeat any words, nor to name groups of words that belong to the same semantic family. In the PVF, the participant is asked to list words that all start with the same letter - in our case the letter 'S'. Here, the semantic category of the words are of no importance.

Testing itself was conducted with the examiner writing down the words given by the participant while the tablet computer simultaneously voice recorded the subject's answers. At the beginning and the end of the 60 seconds time span, the tablet computer would make a beeping sound to indicate start and end of the exam. After testing, the examiner compared the written notes with the recorded audio files and scored the number of correct, as well as incorrect and repeated words. Collected measures were: *correct* , *error* and *repeated* words within the time spans 0 to 15 seconds, 16 to 30 seconds, 31 to 45 seconds, 46 to 60 seconds and 0 to 60 seconds in sum.

Rey Auditory Verbal Learning Test (RAVLT) In order to cover verbal learning as cognitive domain in our neuropsychological trials we implemented the RAVLT. This test is widely used in academia as well as clinical work as it provides a comprehensive understanding of the examinee's performance concerning short term auditory verbal learning memory, learning strategies and memory processing [Schmidt et al., 1996].

When conducting the RAVLT, the participant is given a list of 15 unrelated words (list A) presented to him by an audio file on the tablet computer. The participant is then asked to directly repeat as many of the 15 words as possible. The order in which the examinee repeats the words is of no concern. However, it is recorded if the participant repeats himself and names a certain word twice or more. In total, this part of the RAVLT repeats itself five times. After that the participant is presented with a different list of 15 unrelated words once (list B) and is again asked to repeat these words immediately after hearing the list out. Again, the examiner scores every word named by the subject. However, this time it is also noted if words from list A are being produced. In a next step, the participant is asked to again repeat all words from list A, without hearing list A again. Lastly, after a 20 to 30 minutes interval, the participant is again asked to reproduce as many words from list A as possible.

Altogether, the RAVLT takes about 15 to 20 minutes for completion, not taking into account the break interval.

In our analysis, due to a systematic error in the PEBL test set-up the variable 'Recognition words from list A' had to be excluded from all subsequent analyses.

Trail Making Test A + B (TMTA/TMTB) Testing the cognitive domains speed of processing (TMTA) and executive functions (TMTB), these neuropsychological tasks have been widely used and implemented in research, clinical work and even military admission tests [Tombaugh, 2004].

In the TMTA the participant is required to connect the numbers 1 to 25 spread randomly on a paper as quickly as possible. Furthermore, the examinee is instructed not to lift the

pen off the paper nor to turn or spin the exercise sheet once he has started the exercise. The exact task differs slightly in the TMTB, as the participant is now asked to connect numbers and letters in an alternating order (1-A-2-B-3-etc.). For both tasks the examiner records the time of completion, the number of violations (e.g. interrupting the exercise, turning the page etc.) and the number of errors (e.g. connecting the wrong dots). During the exercise, the examiner must oversee the participant's performance and - if needed - correct the examinee in case of errors or violations. In case of an error, the participant is to return to the last correct number and proceed from this point on.
Both tasks start off with a shortened example exercise to ensure the participant's understanding of the task's rules. For both, the TMTA and B the measures *Time of completion*, *Errors* and *Violations* were recorded.

Continuous Performance Test - Identical Pairs (CPT-IP) The CPT-IP has been widely used and investigated in the framework of schizophrenia and depression research and has shown to be a serviceable tool in detecting cognitive deficits in attention and working memory [Cornblatt et al., 1989, Cornblatt et al., 1988].
In this task we test for the examinee's ability to maintain attention and short-term memory capacity. After a short supervised test trial with 3 digits, the examinee is confronted with the actual task. Here, the participant is presented with rapidly flashing strings of 4 digits on the tablet screen. Whenever two identical strings of digits appear consecutively, the participant is to respond by clicking the left mouse button.
In total, there are 300 trials, 20% of which contain a target to which the examinee is expected to respond. Furthermore, there were 20% of catch-trials which are similar but not identical strings of digits. The PEBL program measures the number of correct responses (positive and negative response) as well as errors regarding distracting stimuli, filler stimuli and omissions. Additionally, the tablet computer records the participant's reaction times whenever there is a positive response (clicking of the mouse button).
For the univariate statistical analysis I further calculated the discriminability index *d'* in

order to better assess test performance. Precisely, d' is the ratio between hits at positive response and catch-trials and is calculated as $d' = \frac{\mu_p - \mu_c}{\sqrt{\frac{1}{2}(\sigma_p + \sigma_c)}}$, with μ_p and σ_p and μ_c and σ_c being the mean and standard deviation of correct positive response hits and false catch-trial hits, respectively [Macmillan and Creelman, 2004].
Consequently, I receive a statistical feature focusing especially on attentional processing [Mirzakhanian et al., 2013, Roitman et al., 1997].

Self-Ordered Pointing Test (SOPT) Originally developed by Petrides and Milner in order to test for working memory deficits in patients with frontal lobe brain lesions, the SOPT has become a popular tool in testing executive working memory in both children and adults [Petrides and Milner, 1982, Cragg and Nation, 2007]. In each trial of our version of the SOPT, the participant is presented with an array of 4, 6, 8 or 10 different symbols and is instructed to click each individual symbol exactly once. After every click the symbols will rearrange, forcing the examinee to exactly remember which symbols have already been clicked and which have not. The task starts off with the examiner supervising a test trial of 4 simple geometrical symbols. This test trial is conducted 3 times to ensure the participant's understanding of the task. In the actual task, the participant is confronted with first 4 then 6, 8 and 10 abstract symbols. Each trial is conducted three times. As variables, for each trial the tablet computer registers errors, perseveration errors and maximum correct responses before error.

Digit Symbol Substitution Test (DSST) Being a fast and simple testing instrument, the DSST has found its way into a multitude of neuropsychological test batteries (e.g. Brief assessment of cognition in schizophrenia [Keefe et al., 2004], Wechsler Adult Intelligence Scale [Wechsler, 2014]).
In this task, the participant is presented with 9 symbols, each accompanied by a corresponding digit (1 to 9). To complete the task, the participant is now given 90 seconds to add as many of the corresponding digits to a string of symbols as possible. In doing so, the participant has to 'translate' one symbol after another. After 90 seconds, the task

ends and the examiner records both, the number of correct answers and the errors.

Wechsler Adult Intelligence Scale, Vocabulary (WAIS-V) In order to assess the verbal IQ of the study participants, we implemented the Vocabulary test of the WAIS [Wechsler, 2014]. Here, the examinee is asked to explain a total of 33 words. The examiner scores the answers given using a standardized manual. Depending on the correctness of the response, the participant can achieve either 0, 1 or 2 points per answer, with two being the best result. Thus, the maximum obtainable raw score is 66. In a second step, the raw score is standardized to age according to the official WAIS scoring tables. Hereby, combined with the Wechsler Adult Intelligence Scale - Matrices, we expect to obtain a representative estimation of the premorbid IQ of each participant.

Wechsler Adult Intelligence Scale, Matrices (WAIS-M) Testing for non-verbal reasoning and IQ as well as visuo-spatial reasoning, this task in combination with the WAIS-V allows us to estimate the participant's premorbid IQ. The test itself is conducted using the official WAIS-Matrices charts. Here, the study participant is asked to complete the missing piece of a Matrix or row of symbols by choosing one of five given options. In total, there are 26 trials and a total of 26 achievable points maximum. The raw score is corrected for age using the official WAIS scoring tables. Both, the WAIS-V and WAIS-M are pen and paper versions. In this study, we used the official WAIS test materials. Both tests were conducted in German or English, depending on the study participant's preference and/or linguistic abilities.

Salience Attribution Test (SAT) In previous studies, this instrument has been shown to be a sensitive tool in revealing aberrant salience attribution in psychotic disorders such as schizophrenia [Roiser et al., 2009]. However, due to technical difficulties at the time of data analysis, this specific task did not find its way into the study at hand, since data needed further processing to be recoverable. Therefore, I refrain from going into the complex details of the structure and execution of the SAT.

3.5 Statistical Analysis

3.5.1 Univariate Analysis

For all univariate statistical analyses we used SPSS 15.0 for Windows. In order to examine group differences in ROD vs. ROP, univariate statistics two-sample t-tests were performed on a number of 23 variables. Significance was assumed for $P \leq 0.05$.

In order to avoid statistical error we corrected for multiple comparisons using the Bonferroni correction. Thus, the p-value was further divided by the number of comparisons conducted, giving us $P \leq 0.05/(23) = P \leq 0.00217$. For a detailed list of statistical variables analyzed, see *Table 7*.

Cognitive domain	**Assessment**	**Variables**
Speed of processing	DSST	Raw score correct
	TMTA	Time of completion
	PVF	Sum correct responses
Attention	CPT-IP	D prime (*d'*)
Working memory	FDS/BDS	Sum raw score correct
	SOPT	Mean error score trial 10 items
Verbal learning	RAVLT	Raw score trial 1 to 5
		Sum raw score 1 to 5
		Sum out of list words 1 to 5
		Raw score after interference list
		Raw score delayed repetition
Visual learning	ROCF	Raw score copy
		Raw score immediate memory
		Raw score delayed memory
Social cognition	DANVA	Raw score correct
Executive functions	TMTB	Time of completion
	SVF	Sum correct responses
Premorbid IQ	WAIS-V	Standardized Score
	WAIS-M	Standardized Score

Table 7: **List of univariate variables.** *Note*: Variables were chosen according to [Koutsouleris et al., 2011].

3.5.2 Multivariate Analysis (MVA)

All multivariate analyses were conducted using NeuroMiner©, a MATLAB-based fully automated machine learning software developed by Nikolaos Koutsouleris (cf. [Koutsouleris et al., 2014]).
In total we conducted 3 different analyses:

i) ROD vs. ROP using the 23 variables prior examined by univariate statistics,

ii) ROD vs. ROP using 214 variables collected by the PRONIA Neuropsychological Test Battery,

iii) Leave-Center-Out.

Each one of the analyses provides us with the measures classification accuracy, sensitivity and specificity. Additionally, NeuroMiner presents a list of variables most decisive to the classification process, permitting a further insight on the respective, illness-specific neurocognitive performance patterns.
In this study, we refrained from correcting the data for age, sex or level of education since the pool of Healthy Controls (HC) at the time of analysis was too sparse and did not allow for a valid statistical standardization.

Support Vector Machine (SVM) To create optimized models for diagnostic differentiation we utilized a L2-regularized support vector machine (SVM) from the LIBLINEAR toolbox implemented in NeuroMiner (see [Fan et al., 2008]). Methodologically, the SVM differentiates between the ROD and ROP groups by estimating a boundary between the two groups in which the support vectors mark the maximum distance between the most similar subjects of the opposite study populations [Koutsouleris et al., 2011]. See chapter 2.5.1 for a detailed explanation of the SVM's functionality.

Cross-validation (CV) Training and testing of the SVM was performed within a repeated nested cross-validation analysis (CV) [Filzmoser et al., 2009]. Here, our dataset is split into an inner (CV1) and outer (CV2) cross-validation. Each CV in itself is folded into 5 permutations x 10 folds of non-overlapping samples. In the CV1, these n folds function as training sets for the different parameter set-ups of the SVM with the $n-1^{th}$ fold being a test sample to validate each model's classification accuracy. For the purpose of detecting generalization errors, the best performing models generated in CV1 are then applied to the outer cross-validation. To assure the validity of said generalization analysis, all training and test samples were kept strictly separated (cf. [Zarogianni et al., 2013]).

Parameter Set-up Firstly, all data with zero variance was pruned from the dataset as these are redundant and hold no information for machine learning algorithms. Secondly, for subjects with less than 25% of missing data, missing values were imputed. The imputed data was calculated as the median value of the missing data, based on the values of the 7 nearest neighbours according to the Euclidian distance (cf. [Beretta and Santaniello, 2016]). Lastly, all data was scaled from -1 to 1 as this is demanded by LIBLINEAR SVM.

The SVM was fed with a range of slack variables. These slack variables were defined as $2^{[-5:2:11]}$. Consequently, the SVM tried different slack variable (or C parameter) set-ups (2^{-5}; 2^{-3}; 2^{-1};...; 2^{9}; 2^{11})on the CV1 and thereby defining different penalties for misclassifying subjects.

Feature selection To optimize model predictions in the CV1, it is possible to apply additional feature selection filters [Saeys et al., 2007]. In this work, I engaged a wrapper method with Greedy feature selection [John et al., 1994]. More specifically, I set up a forward stepping wrapper, up to 90% of the features in steps of 5%. The wrapper method was applied to the CV1 train and test partitions. Hereby, the SVM is presented with an iteratively increasing number of features and thus can detect the optimal conditions for

its classification models.

ROD vs. ROP Analysis - 23 variables As mentioned at the beginning of this chapter, I conducted 3 different multivariate analyses. In this first analysis, I aim to compare the multivariate classification performance between ROD and ROP subjects to our univariate statistics. Therefore, the analysis is restricted to the same 23 neurocognitive variables that are tested using the two-sampled t-tests. Classification is performed on all 116 available subjects.

ROD vs. ROP Analysis - 214 variables This second analysis investigates differential diagnostic performance between ROD and ROP in the framework of a much more comprehensive data-set (214 variables). Apart from the number of neurocognitive variables, all settings and subjects are identical to the 23-variables-analysis. A detailed list of the variables examined in this analysis can be found in the appendix.

Leave-Center-Out Analysis Finally, in order to test our classification models for multisite generalizability, I implemented a leave-center-out analysis. In this analysis, the SVM trains its models on data of only 4 of the 5 PRONIA study centers. The generated models are then applied to the data of the remaining study center in the CV2 test partitions. This is conducted for each of the centers as testing data, hence providing us with 5 different classification accuracies. For a measure of generalizability, these accuracies are then averaged.
In this analysis we use the same parameter setup and the same 214 variables as in the 'ROD vs. ROP - 214 variables' analysis to achieve maximum comparability.

4 Results

4.1 Demographic Data Analysis

To compare our 2 study groups in terms of illness-specific biometric scores, we performed 2-sample t-tests with ROD vs. ROP data on BDI, PANSS, GAF-S and GAF-DI scores (see *Table 5*). Significance was estimated for $P \leq 0.05$.

Participants with recent onset depression did not score significantly higher on the Beck Depression Inventory - II than subjects with recent onset psychosis (mean score ROD: 25.7 ± 13.1; mean score ROP: 24.6 ± 12.3). However, both groups classify as 'moderate' to 'severe' depression when comparing the mean scores to the benchmarks as described in previous literature (see chapter 3.3).

Concerning the Positive and Negative Syndrome Scale, ROP patients in average scored significantly higher than ROD patients in the PANSS total measure as well as all 3 subscales *positive*, *negative* and *general* symptoms. With mean scores of 19.5 ± 6.0 in the PANSS *positive*, 17.2 ± 7.3 in the PANSS *negative* and 37.7 ± 10.0 in the PANSS *general* subscale, the ROD group scores well in line the schizophrenic patients tested by Kayle and colleagues in 1987, with only the mean score in the PANSS *negative* lying below the correspondent score in literature (21.01 points).

Regarding the Global Assessment of Functioning, ROP subjects presented with a significantly lower level of functioning than ROD subjects in both scales, GAF-S and GAF-DI, indicating a generally lower level of functioning in the ROP group as compared to ROD subjects.

In order to attain an approximate measure for the premorbid IQ, we compared the WAIS-A (averaged score of WAIS-V and WAIS-M) of the two groups. Here, ROD participants again scored significantly higher than ROP patients (mean scores: 11.0 ± 2.3 and 8.99 ± 2.6, respectively).

4.2 Univariate Analysis Results

Examining the cognitive domain ***verbal learning***, we found significant results especially in comparing the sum scores of correctly remembered words throughout the first 5 trials of the task (RAVLT Sum Raw Score 1-5). Here, ROD patients scored significantly higher on average (58.3 ± 8.6) than their study counterparts with a mean score of 52.1 ± 12.1. On a single trial level, in particular trials 2 and 4 showed significant results (*Table 8*). In the variables for 'out of list words', raw score after the interference list and delayed repetition, no significant performance disparities could be detected.

In summary, it can be stated that in comparison ROP subjects perform worse in the cognitive domain *verbal learning* than subjects with recent onset depression.

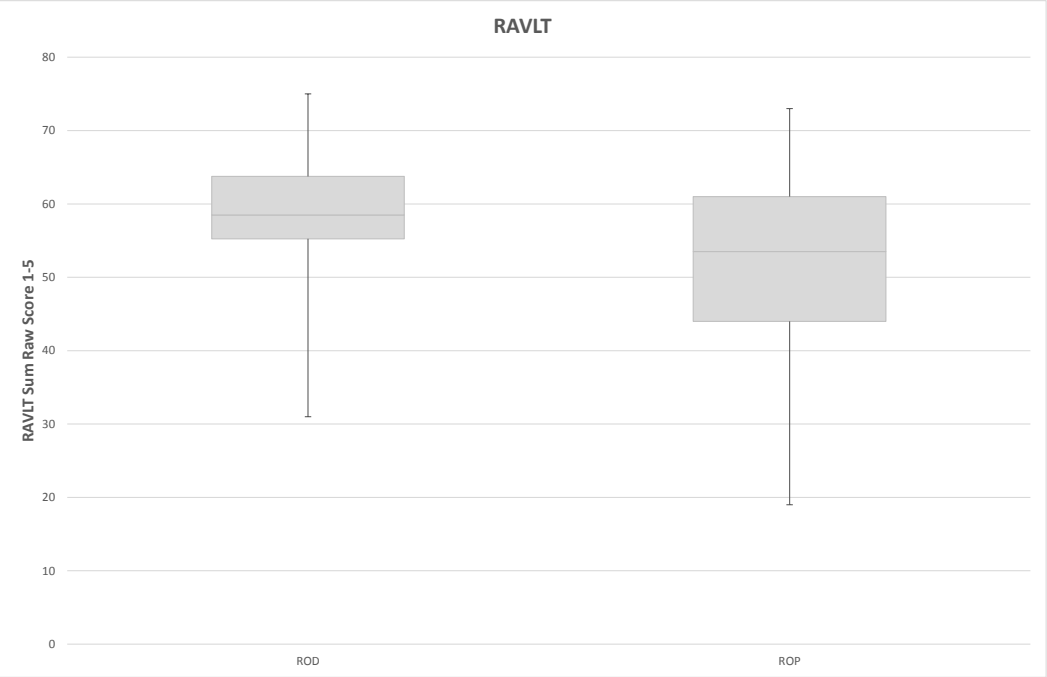

Figure 5: **Box plot RAVLT Sum Score trial 1 to 5.** This variable is a sum score of the variables RAVLT Raw Score trial 1 to 5. Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: Sum of Raw Score Trial 1 to 5; x-axis: ROD vs. ROP.

For sustained and onward ***attention***, we compared the mean performances on the discriminability index d'. Here the ROD group presented an average score of 2.49 ± 0.78, whereas ROP subjects reached a score of 1.96 ± 0.71. Again, this results is to be considered significant for $P < 0.00217$.

Accordingly, subjects suffering from depression exhibit a greater ability to identify stimuli to which they have to react, while dismissing false stimuli, than patients suffering from psychotic symptoms. Furthermore, depressive patients seem to be more capable in keeping their attention up over a longer time span.

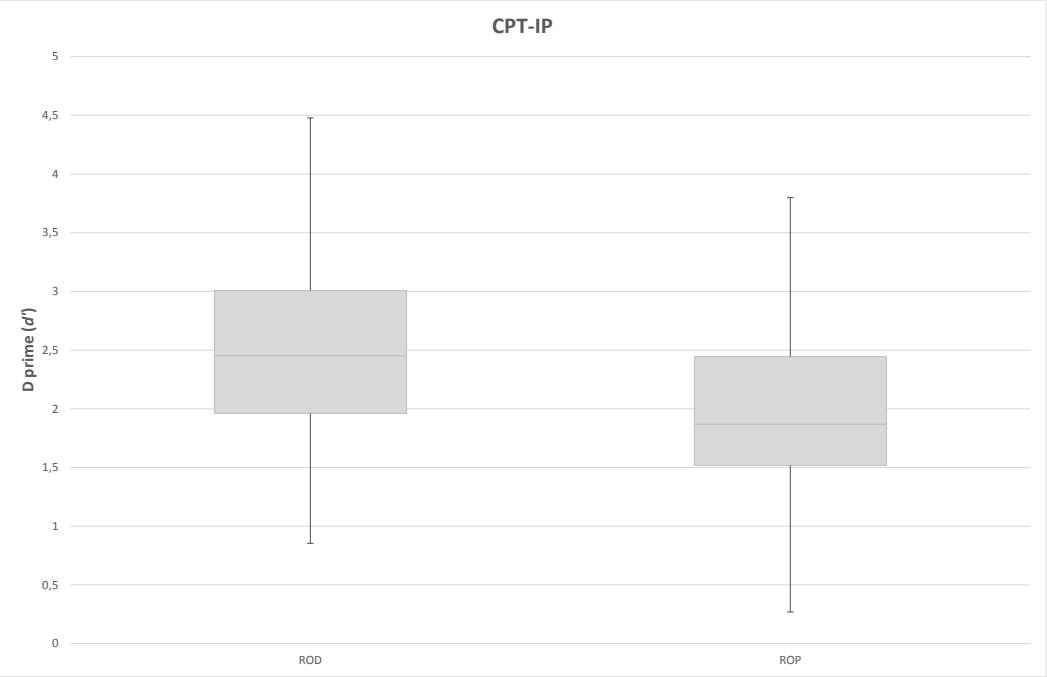

Figure 6: **Box plot CPT-IP D prime (d').** Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: D prime (d'); x-axis: ROD vs. ROP. See *Table 8* for exact measures.

By investigating the means of the 'FDS/BDS Sum Raw Score correct' and 'SOPT Mean error score 10 items' we can draw conclusions concerning potential group differences in ***working memory*** abilities. However, neither in the FDS/BDS sum score nor in the SOPT mean errors variable significant group differences could be revealed (11.3 ± 2.1 vs. 10.2 ± 2.0 and 1.63 ± 1.0 vs. 2.0 ± 0.88, respectively).

Social cognition was examined by comparing the mean 'DANVA Raw score correct' of each study group. ROD subjects on average reached a score of 19.4 ± 1.9 and ROP subject attained a mean score of 18.1 ± 2.9. While these values seem to present a notable difference, statistical significance could not be attested with $P = 0.003$.

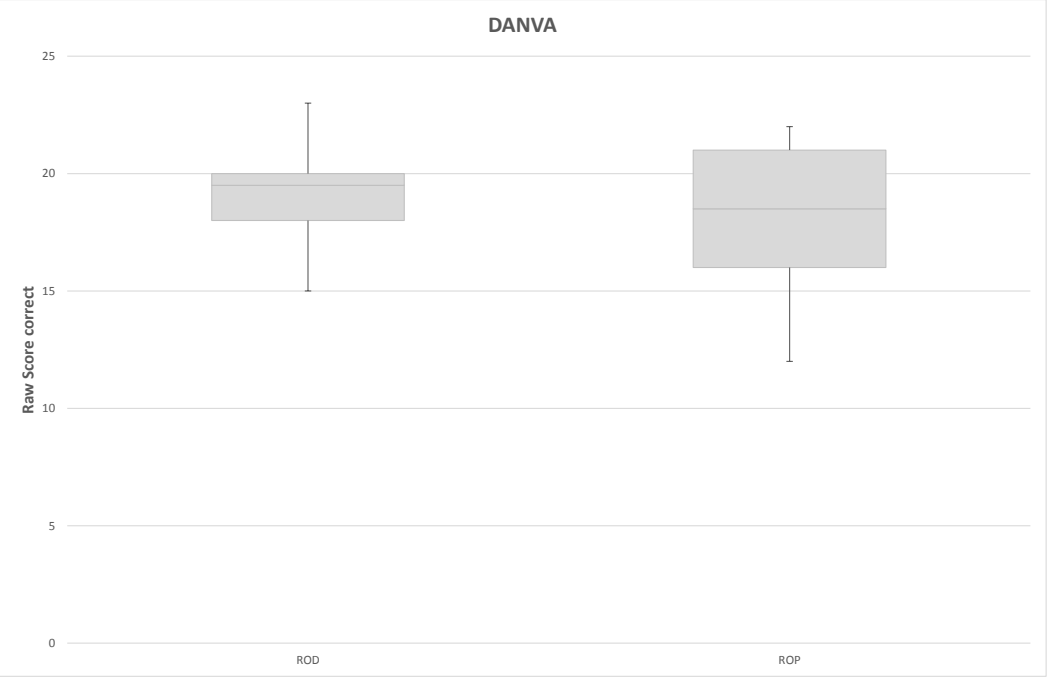

Figure 7: **Box plot DANVA Raw Score correct.** Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: DANVA Raw Score correct; x-axis: ROD vs. ROP. Mean differences were not statistically significant. See *Table 8* for exact measures.

In the domain ***speed of processing***, we analyzed variables from 3 different tasks: 'DSST Raw Score correct', 'TMTA Time of completion' and 'PVF Sum correct responses'. Of these 3 variables, only the Digit Symbol Substitution Test produces significant performance differences between the two study groups (ROD mean: 62.9 ± 11.0; ROP mean: 50.2 ± 12.6). With a respective P-value of 0.027 and 0.011 these results could not be confirmed when comparing the time of completion in the TMT A or the sum of correct responses in the PVF as additional measures for the *speed of processing*.

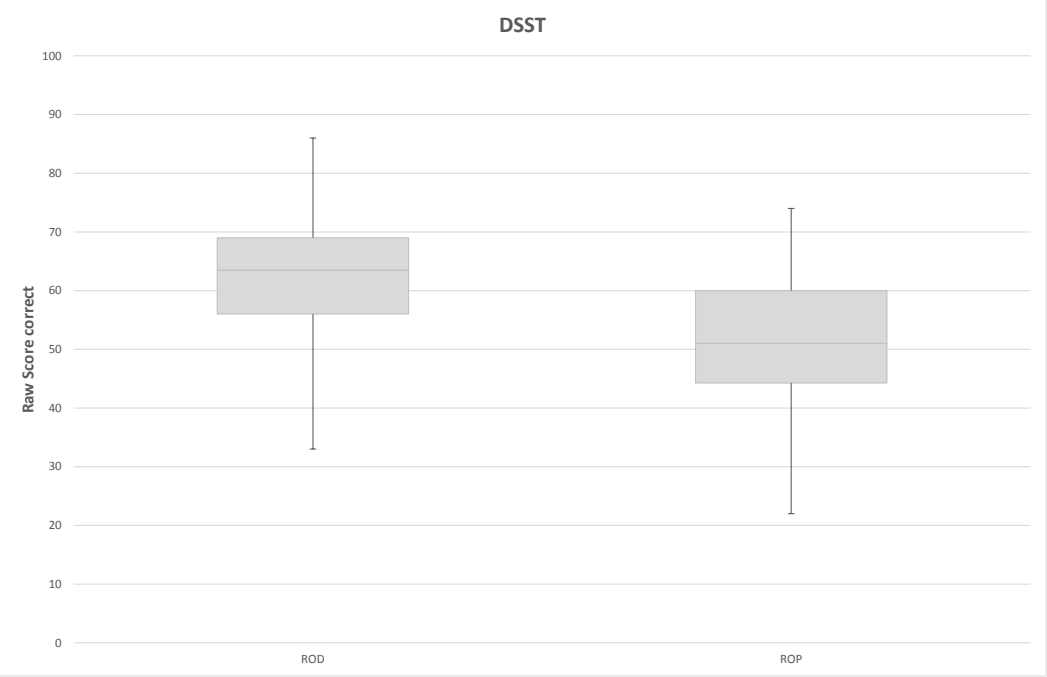

Figure 8: **Box plot DSST Raw Score.** Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: Raw Score of correctly translated symbols; x-axis: ROD vs. ROP. See *Table 8* for exact measures.

As mentioned before, the Rey-Osterrieth Complex Figure Test primarily covers the cognitive process of ***visual learning***. Accordingly, we investigated the participant's performance in the three trials direct copy, immediate recall and delayed recall. While all three trials show slightly higher scores for the ROD group (34.0 ± 3.0 vs. 32.7 ± 5.2, 23.7 ± 6.4 vs. 21.7 ± 9.5 and 23.4 ± 6.3 vs. 22.1 ± 9.2), none of these guarantee statistical significance.

Hence, neither depressive, nor psychotic patients seem to perform generally higher or lower than subjects from the respective opposite group.

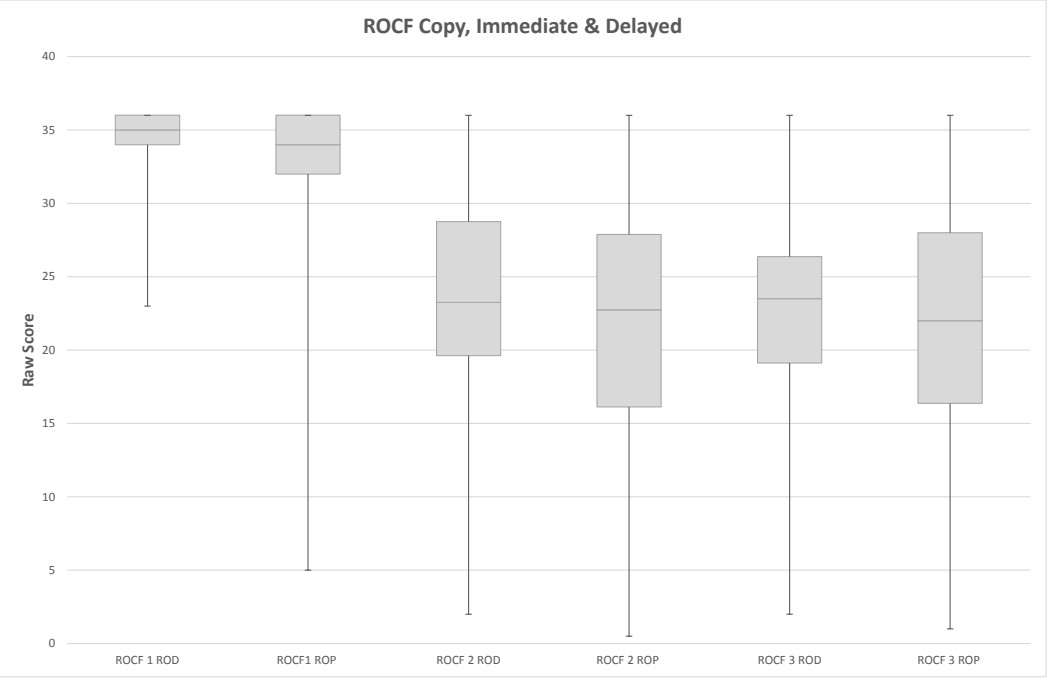

Figure 9: **Box plot ROCF copy, immediate and delayed.** Boxplots visualizing means, standard deviation, maximum and minimum scores. ROCF 1: Trial 1 copy; ROCF 2: Trial 2 immediate memory; ROCF 3: Trial 3 delayed memory. y-axis: ROCF Raw Score correct; x-axis: ROD vs. ROP. for each trial. See *Table 8* for exact measures.

Executive functionality disparities in the two study groups were highlighted by performing 2-sample t-tests on the 'TMT B Time of completion' variable and the sum of correct responses in the Semantic Verbal Fluency Task. Here, with average completion times of 60.9 ± 25.5 vs. 82.5 ± 42.1 seconds in the TMT B and 24.7 ± 6.7 vs. 20.9 ± 6.1 correct responses in the SVF we were able to identify performance differences between both groups.

Thus, these results suggest that patients with recent onset depression preserve a higher executive functionality than recent onset psychotic individuals.

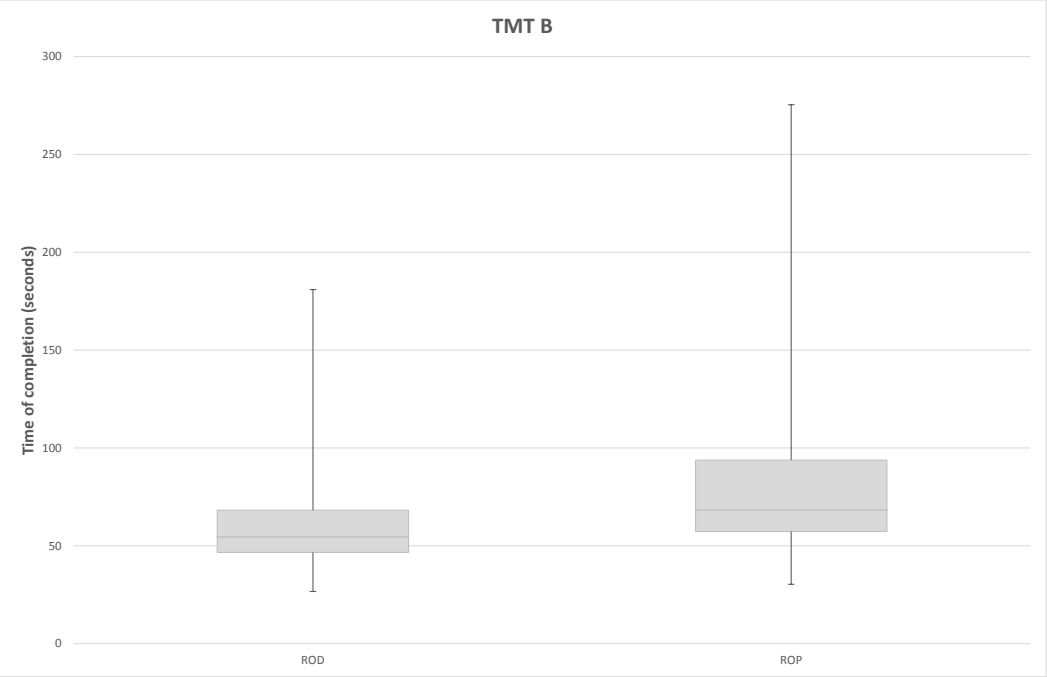

Figure 10: **Box plot TMT B Time of completion.** Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: Time of completion for the TMT B in seconds; x-axis: ROD vs. ROP. See *Table 8* for exact measures.

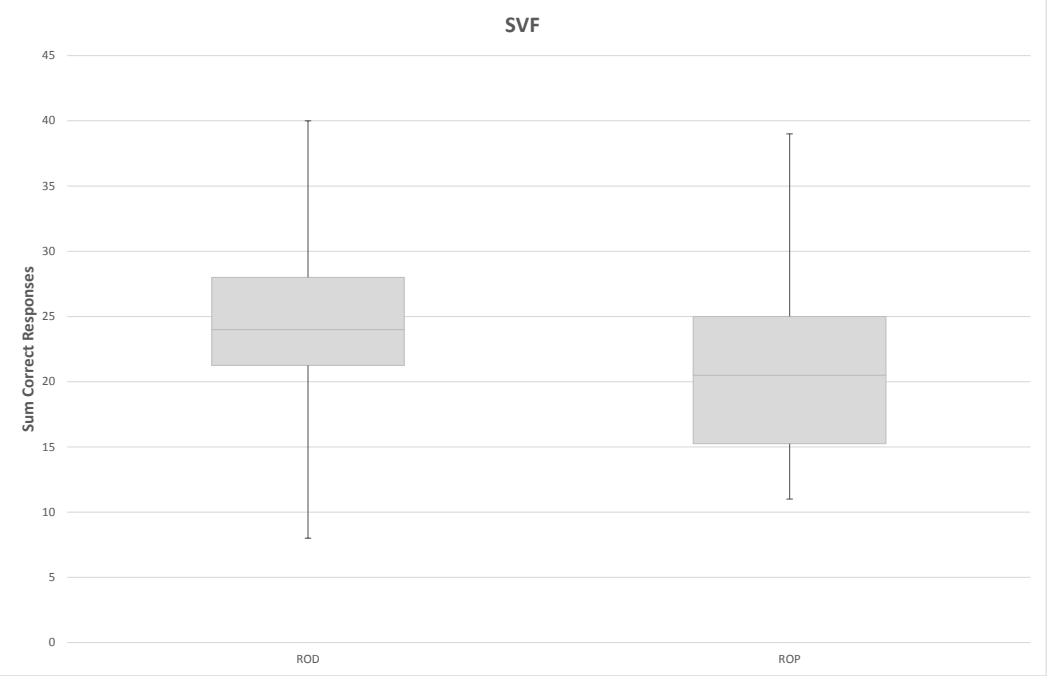

Figure 11: **Box plot SVF Sum correct responses.** Boxplots visualizing means, standard deviation, maximum and minimum scores. y-axis: Sum of correct responses; x-axis: ROD vs. ROP. See *Table 8* for exact measures.

We compared the overall performance of the two study populations on the WAIS-V and WAIS-M in order to receive an approximate measure for group differences concerning the ***premorbid IQ***. Results were significant for both variables. Participants with recent onset depression scored almost 2 points higher on average than participants with recent onset psychosis, with the respective means being ROD vs. ROP: 11.0 ± 2.8 vs. 9.1 ± 3.6 in the WAIS-V and 11.0 ± 2.4 vs. 8.9 ± 2.6 in the WAIS-M.

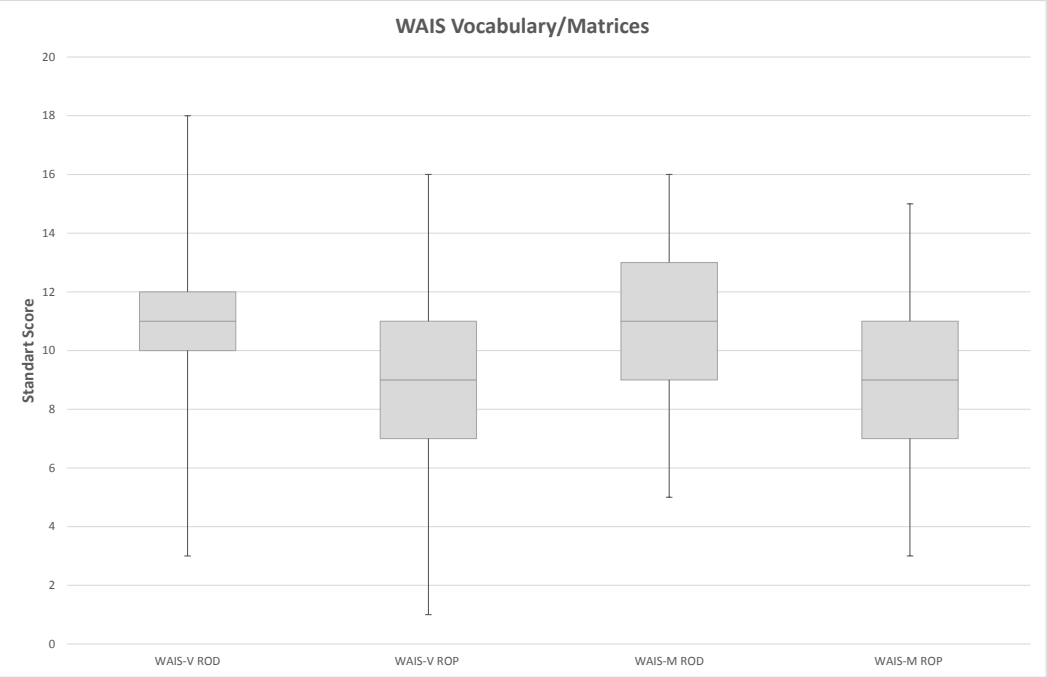

Figure 12: **Box plot WAIS Vocabulary/Matrices.** Means, standard deviation, maximum and minimum in the WAIS Vocabulary (WAIS-V) and Matrices (WAIS-M) for ROD and ROP group. See *Table 8* for exact measures.

Variables	**ROD mean (sd)**	**ROP mean (sd)**	**t**	**df**	***P***
RAVLT Raw Score trial 1	8.2 (2.3)	7.2 (2.7)	2.204	114	0.030
RAVLT Raw Score trial 2	11.3 (2.5)	9.7 (2.9)	3.184	114	**0.00***
RAVLT Raw Score trial 3	12.4 (2.2)	11.2 (2.7)	2.701	108.83	0.008
RAVLT Raw Score trial 4	13.0 (1.9)	11.7 (2.6)	3.229	104.42	**0.00***
RAVLT Raw Score trial 5	13.4 (1.8)	12.4 (2.6)	2.318	101.91	0.022
RAVLT Sum Raw Score 1-5	58.3 (8.6)	52.1 (12.1)	3.170	102.65	**0.00***
RAVLT Sum out of list words 1-5	0.95 (2.1)	1.29 (3.5)	0.649	114	0.518
RAVLT Raw Score interference list	12.5 (2.4)	11.2 (3.1)	2.408	114	0.018
RAVLT Raw Score delayed repet.	12.5 (2.3)	10.8 (3.5)	3.065	99.17	0.003
FDS/BDS Sum Raw Score correct	11.3 (2.1)	10.2 (2.0)	2.946	114	0.004
CPT-IP D prime (d')	2.49 (0.78)	1.96 (0.71)	3.874	114	**0.00***
DANVA Raw Score correct	19.4 (1.9)	18.1 (2.9)	3.049	114	0.003
DSST Raw Score correct	62.9 (11.0)	50.2 (12.6)	5.815	114	**0.00***
PVF Sum correct responses	14.6 (4.5)	12.2 (5.5)	2.582	109.89	0.011
SVF Sum correct responses	24.7 (6.7)	20.9 (6.1)	3.248	114	**0.00***
ROCF Raw Score copy	34.0 (3.0)	32.7 (5.2)	1.689	90.17	0.094
ROCF Raw Score immediate	23.7 (6.4)	21.7 (9.5)	1.331	99.93	0.186
ROCF Raw Score delayed	23.4 (6.3)	22.1 (9.2)	0.871	100.97	0.386
SOPT Mean error score 10 items	1.63 (1.0)	2.0 (0.88)	-2.359	114	0.020
TMTA Time of completion	29.1 (11.0)	34.7 (15.2)	-2.241	113	0.027
TMTB Time of completion	60.9 (25.5)	82.5 (42.7)	-3.309	93.16	**0.00***
WAIS-V Standardized Score	11.0 (2.8)	9.1 (3.6)	3.168	105.5	**0.00***
WAIS-M Standardized Score	11.0 (2.4)	8.9 (2.6)	4.528	114	**0.00***

Table 8: **Univariate Analysis Results.** sd = standard deviation; t = t-value; df = degrees of freedom; P = p-value. *Significant at $P < 0.00217$ according to Bonferroni correction for multiple comparisons.

4.3 Multivariate Analysis Results

4.3.1 ROD vs. ROP - 23 variables

In the ROD vs. ROP - 23 variables analysis we trained an SVM algorithm to classify subject groups based exclusively on data for the 23 neuropsychological variables that are examined in the prior univariate analysis (*Table 7*).

Here, of 58 patients with recent onset depression 41 have been rightly identified as such. Accordingly, 17 depressive patients have been misclassified as recent onset psychotic (ROP) subjects. Of the 58 ROP patients, 33 have been correctly assigned by the classifier to their respective category. However, a total of 25 ROP patients were wrongly categorized as subjects with recent onset depression.

As a result, we obtain an overall accuracy of **63.8%** with a sensitivity of **70.7%** and a specificity of **56.9%**. The positive predictive value reaches 62.1% and the negative predictive value 66.0%. A detailed list of results can be found in *Table 9*.

Variable	**Results**
ROD (n)	58
ROP (n)	58
True positive (TP)	41
True negative (TN)	33
False positive (FP)	25
False negative (FN)	17
Accuracy [%]	**63.8**
Sensitivity [%]	**70.7**
Specificity [%]	**56.9**
Balanced Accuracy [%]	63.8
Area under the Curve	0.72
Positive Predictive Value [%]	62.1
Negative Predictive Value [%]	66.0

Table 9: **Results ROD vs. ROP - 23 variables.**

In *Figure 9* the SVM's prediction scores for each subject individually are displayed.

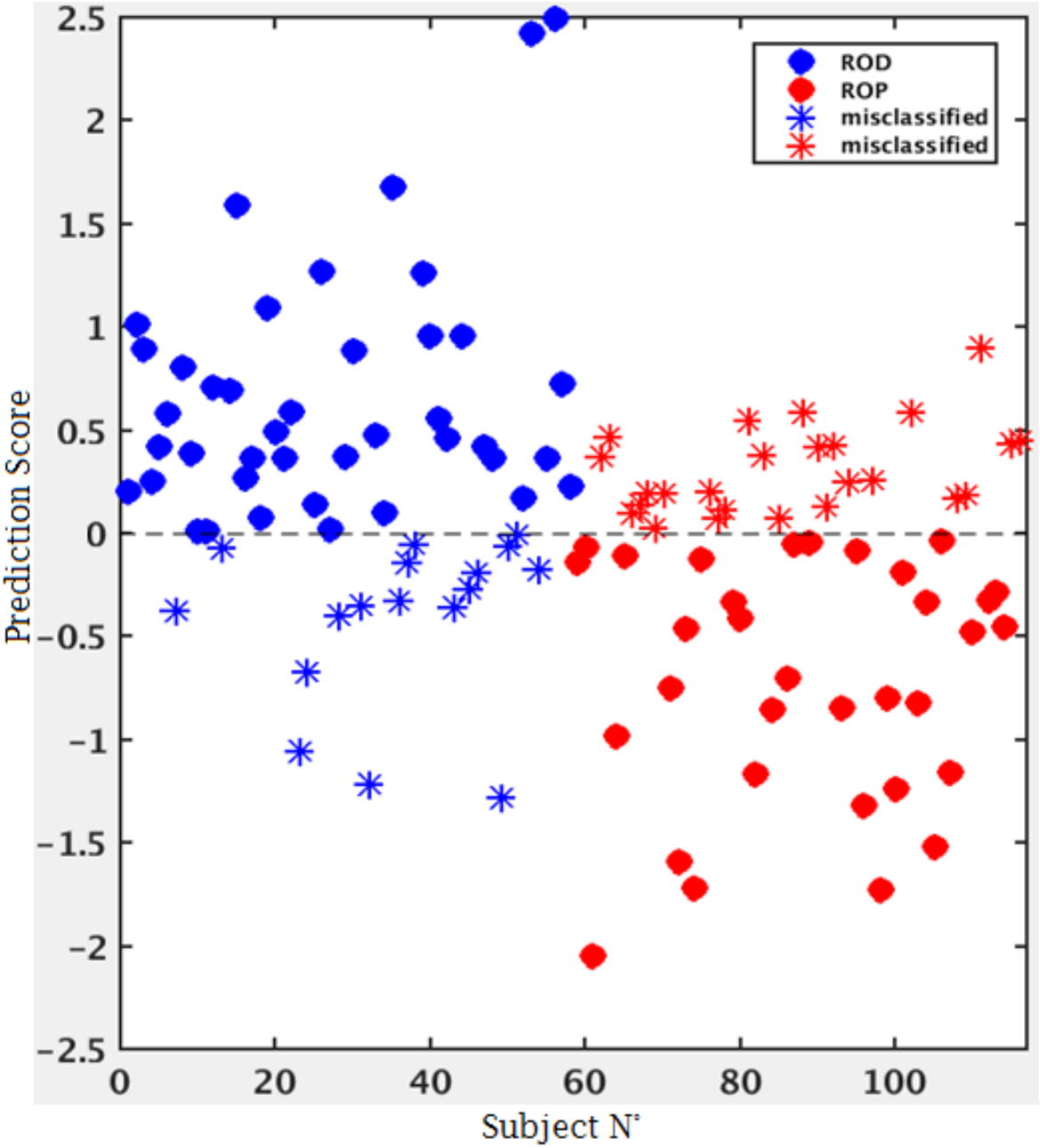

Figure 13: **Distribution plot ROD vs. ROP - 23 variables**. On the x-axis all 116 subjects are displayed according to their subject number. On the y-axis we see the 'Prediction Score' or likelihood-ratio that lead to classification. A positive Prediction Score indicates classification as ROD, a negative score as ROP. Blue circles: correctly classified ROD; Red circles: correctly classified ROP; Blue stars: misclassified ROD; Red stars misclassified ROP.

Most decisive variables for classification Another highly interesting aspect of the MVA is the possibility to identify which variables had the biggest decision value or feature weight in the SVM's decision-making.

Hence, variables with high feature weights in this specific analysis can be considered to contain relatively more illness-specific information than other variables analyzed by the SVM. A list of the 10 most decisive variables in the analysis at hand can be found in the *Table 10*:

Neuropsychological variable:
1. DSST Raw Score correct
2. WAIS-V Standardized Score
3. CPT-IP D prime (d')
4. FDS/BDS Sum Raw Score correct
5. ROCF Raw Score delayed
6. WAIS-M Standardized Score
7. RAVLT Raw Score trial 3
8. ROCF Raw Score immediate
9. RAVLT Raw Score interference list
10. RAVLT Sum out of list words 1-5

Table 10: **ROD vs. ROP - 23 variables analysis: 10 most decisive variables.** Variables are listed in a decreasing order of feature weight.

4.3.2 ROD vs. ROP - 214 variables

In this second neurocognitive pattern analysis we trained our SVM classifier on all 214 variables recorded throughout the PRONIA Neuropsychological Test Battery.

In this analysis, of 58 ROD subjects 42 were correctly classified as patients with depression (True positive = TP). 16 ROD subjects were falsely appointed to the recent onset psychosis group, giving us a False negative (FN) number of 16. Regarding the ROP group, 41 subjects were accurately classified as such, while 17 ROP subjects were incorrectly recognized as ROD subjects.

Consequently, in this analysis we achieved an overall classification accuracy of **71.6%**, with a sensitivity of **72.4%** and a specificity of **70.7%**. The positive and negative predictive values were set at 71.2% and 71.9%, respectively.

Variable	**Results**
ROD (n)	58
ROP (n)	58
True positive (TP)	42
True negative (TN)	41
False positive (FP)	17
False negative (FN)	16
Accuracy [%]	**71.6**
Sensitivity [%]	**72.4**
Specificity [%]	**70.7**
Balanced Accuracy [%]	71.6
Area under the Curve	0.76
Positive Predictive Value [%]	71.2
Negative Predictive Value [%]	71.9

Table 11: **Results ROD vs. ROP - 214 variables.**

In *Figure 10*, again the SVM algorithm's prediction score for each subject individually is displayed in a distribution plot.

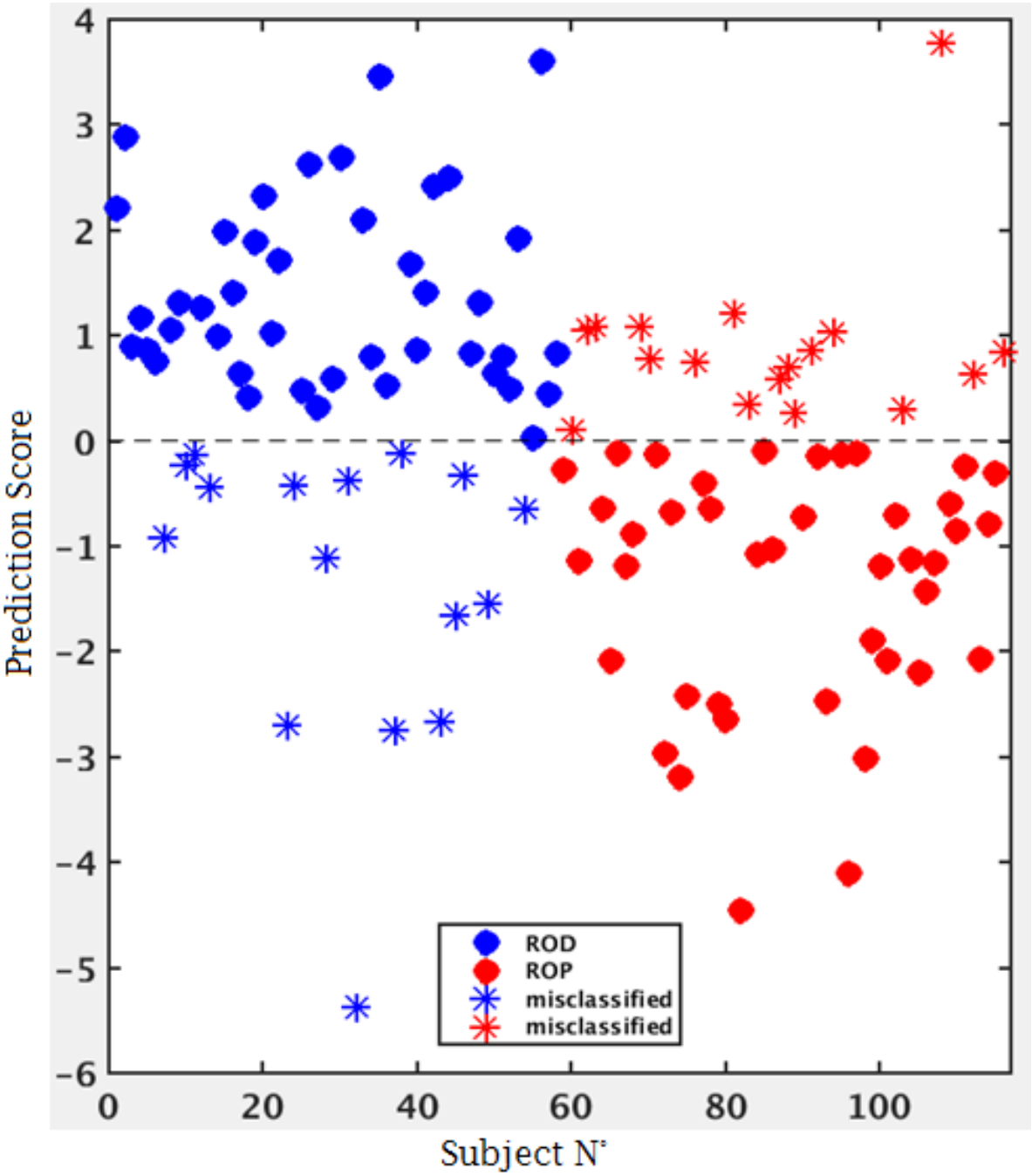

Figure 14: **Distribution plot ROD vs. ROP - 214 variables.** On the x-axis all 116 subjects are displayed according to their subject number. On the y-axis we see the 'Prediction Score' or likelihood-ratio that lead to classification. A positive Prediction Score indicates classification as ROD, a negative score as ROP. Blue circles: correctly classified ROD; Red circles: correctly classified ROP; Blue stars: misclassified ROD; Red stars misclassified ROP.

Most decisive variables for classification Since the data pool based on which the support vector machine trained its classification models is much more comprehensive in this analysis, it becomes even more interesting to highlight those neuropsychological variables with the greatest impact on the SVMs' models.

A list of the 15 most decisive variables can be found in *Table 12*:

Neuropsychological variable:
1. SOPT Errors 6 elements 02
2. WAIS-M Standardized Score
3. DANVA Raw Score correct
4. CPT-IP Reaction times correct 50 trials
5. CPT-IP Number correct 200 trials
6. WAIS-V Standardized Score
7. CPT-IP Number error filler stimuli 200 trials
8. SOPT Perseveration Errors 10 elements 03
9. DSST Correct number symbol matchings
10. PVF Correct 45 60 letter 1
11. DSST Raw score correct
12. CPT-IP Number correct 150 trials
13. TMTA Time of completion
14. RAVLT Immediate 5 repetition list A
15. SVF Error 15 30 category 1

Table 12: **ROD vs. ROP - 214 variables analysis: 15 most decisive variables.** Variables are listed in a decreasing order of feature weight. See chapter 7 'Appendix' for a list of all 214 variables.

4.3.3 Leave-center-out Analysis

As described in the paragraph 'Leave-center-out Analysis' in chapter 3.5.2, this analysis is primarily performed in order to quantify the classifiers' generalizability within the framework of a multicenter data pool. Hence, the question at hand is whether our classifier actually limits itself to differentiate between, in this case, disease-specific patterns or whether it also takes into account each subjects' center affiliation.

Subjects from one center are classified based on classification models that have been solely trained on data from the 4 remaining centers. Consequently, classification results in this analysis become a measure of generalizability.

Across all 5 sub-analyses, 42 of 58 subjects of both study groups were correctly classified and 16 of 58 subjects were falsely appointed to the opposite study group. Thus, we achieved an overall accuracy of **72.4%** with a sensitivity of **72.4%** and a corresponding specificity of **72.4%**.

Variable	**Results**
ROD (n)	58
ROP (n)	58
True positive (TP)	42
True negative (TN)	42
False positive (FP)	16
False negative (FN)	16
Accuracy [%]	**72.4**
Sensitivity [%]	**72.4**
Specificity [%]	**72.4**
Balanced Accuracy [%]	72.4
Area under the Curve	0.78
Positive Predictive Value [%]	72.4
Negative Predictive Value [%]	72.4

Table 13: **Results Leave-center-out Analysis**

When looking at the results of the Leave-center-out analysis individually, it becomes apparent that classification generalizability differs depending on the center the models are tested on.

Here, highest accuracies with **82.1%** are achieved when classifiers are trained on the centers LMU, UBS, Uni BHAM and Uni Udine and then tested on subjects from UKK. While results for Uni Udine and LMU (76.5% and 73.2%, respectively) as test data lie close to the overall accuracy of **72.4%**, subjects from Uni BHAM and especially UBS are much more likely to be misclassified with accuracies of 62.5% and 59.1%, respectively.

	LMU	**UBS**	**UKK**	**Uni BHAM**	**Uni Udine**	**TOTAL**
Accuracy [%]	**73.2**	**59.1**	**82.1**	**62.5**	**76.5**	**72.4**
Sensitivity [%]	75.0	50.0	80.0	50.0	83.3	**72.4**
Specificity [%]	70.5	66.7	83.3	66.7	60.0	**72.4**

Table 14: **Leave-center-out Analysis:** Results for each center independently.

5 Discussion

5.1 Summary of the Findings

The first aim of the study was to identify differences in the performance on the PRONIA Neuropsychological Test Battery between ROD and ROP subjects by comparing the respective test results conducting 2-sample t-tests. Significance in performance differences was assumed for $P < 0.00217$ since we implemented the Bonferroni correction to avoid statistical error due to multiple comparison.

Significant results were detected for the tasks RAVLT, CPT-IP, DSST, SVF, TMTB and both subtests of the WAIS (see *Table 8*). In all of these tasks, subjects suffering from recent onset psychosis performed significantly worse than their counterparts from the recent onset depression group. Hence, our findings suggest that patients in the early stage of psychosis exhibit greater deficits in the cognitive domains *speed of processing*, *attention*, *verbal learning*, *executive functions* and the estimate measures for *premorbid IQ* than patients that battle with their first depressive episode. In contrast, for the domains *working memory*, *visual learning* and *social cognition*, no statistically significant performance differences could be found.

The second aim of my study was to generate classification models with a support vector machine that differentiate between ROD and ROP subjects based on group-specific patterns in the neurocognitive data. Consequently, two analyses were conducted: the ROD vs. ROP - 23 variables Analysis which aims to identify said patterns within 23 neurocognitive features that are typically highlighted in univariate neuropsychological analyses, and the ROD vs. ROP - 214 variables Analysis that derives its information from all 214 neurocognitive features measured throughout the neurocognitive assessment of this study.

For the ROD vs. ROP - 23 variables Analysis, classification accuracy was set at 63.8%, with a respective sensitivity of 70.7% and a specificity of 56.9%. The ROD vs. ROP - 214 variables Analysis on the other side, being availed with a much more comprehensive

data set, reached a balanced accuracy of 71.6%, a sensitivity of 72.4% and a specificity of 70.7%, therefore presenting us with a considerably higher classification performance (cf. *Table 10* and *11*).

Lastly, I intended to test the generalizability of classification models generated through an SVM by testing a model trained with the data of 4 centers to classify the subjects of the remaining 5^{th} center (Leave-center-out Analysis). Accordingly, we obtained 5 different classification accuracies that were averaged out in order to gain an approximate measure of generalizability for this classification method.

In average, our models, when applied to an outside test sample, reached an overall accuracy of 72.4% with a likewise sensitivity and specificity of 72.4% (cf. *Table 14*).

5.1.1 Group Differences and Similarities in Demographic Data

In this study I examined a total of 116 subjects divided into two study groups. In total, 43% of the study participants were female. Each study group consisted of exactly 58 participants. However, study groups were not matched for sex, resulting in 27.6% (n= 16) of female subjects in the ROP group and a majority of 58.6% females in the ROD group. Concerning the age distribution, no significant differences between the two cohorts could be detected with a mean age of 25.5 ± 6.0 years in the ROD and 25.6 ± 5.2 years in the ROP group. This is somewhat surprising considering the typical age-of-onset of mood disorders being at the age of 30 (cf. [Kessler et al., 2005]) and therefore slightly later than non-affective psychosis [Kessler et al., 2007]. However, this may be partly explained by the strict inclusion criteria, only considering participants between the age of 15 to 40 and therefore eliding depressive patients that experienced their age-of-onset at the far end of the age distribution.

Regarding the psychometric instrument BDI, it is remarkable that patients with recent onset psychosis score almost just as high on depressive symptoms as patients with recent onset depression (24.6 ± 12.3 points vs. 25.7 ± 13.1 points, respectively). One possible explanation for this is the occurrence of negative symptoms such as loss of interest and

anhedonia in psychotic disorders that can reach a severity primarily observed in MDD [Velligan and Alphs, 2008, Andreasen, 1982].

However, this approach stands in contradiction with the fact that ROP subjects in this study scored lower on the PANSS *negative* subscale than schizophrenic patients in previous research [Kay et al., 1987], suggesting that in first episode psychosis, negative symptoms are less prominent or distinct than the positive and general symptomatology.

Eventually, this leaves us with two possible explanations: either there is a significant discrepancy between how subjects from the ROP group rate their symptoms themselves (BDI) and how the interviewer evaluates the negative symptoms of the subject (PANSS), or ROP subjects in our collective are generally more severely diseased than subjects from the ROD group.

The latter of these two explanations is additionally supported by the observation that ROP subjects score significantly lower on both versions, symptoms and impairment/ disability, of the Global Assessment of Functioning, when compared to ROD subjects (cf. *Table 5*). However, whether this theorized imbalance of illness severity between the ROD and ROP groups really exists, or whether this is a random constellation conditioned by relatively small sample sizes remains the focus of further investigation within the PRONIA study.

5.1.2 Performance Differences in the Neuropsychological Test Battery

As introduced in the chapters 2.3.1 and 2.3.2, both, major depressive as well as psychotic disorders exhibit diverse cognitive impairments, even from an early stage on [Lee et al., 2012, Mesholam-Gately et al., 2009]. Furthermore, the two disease spectra show certain similarities regarding the cognitive domains affected, with overlaps existing especially in the domains *executive functioning*, *(working) memory*, *attention* and *processing speed*.

In this study, I examine whether it is possible to reveal significant performance differences between ROD and ROP subjects, even in cognitive domains that are associated

with general impairments in both entities.

When focusing on the mean scores in the results, it quickly becomes apparent that subjects from the ROP group scored lower in almost every neurocognitive task examined (cf. *Table 8*). These differences in the mean scores where statistically significant for 9 out of 23 variables, therefore indicating significant performance differences in the cognitive domains *speed of processing*, *attention*, *verbal learning*, *executive functions* and *premorbid IQ*. Hence, ROP subjects also presented greater impairments in domains that have been proven to be generally impaired in depressive disorders as well (*speed of processing*, *attention*,*executive functions*, cf. [McIntyre et al., 2013] and [Lee et al., 2012]).

These findings are well in line with previous studies investigating similar questions. For instance, regarding the CPT-IP covering the domain *attention*, researchers have found deficits within psychotic patients to be much more comprehensive in comparison to depressive patients [Cornblatt et al., 1989, Nelson et al., 1998]. Further studies examining cognitive performance differences between psychotic depression and non-psychotic depression found patients with psychotic symptoms to perform generally worse, with cognitive impairments in psychotic depression resembling more the findings observed in schizophrenia than those in non-psychotic depression [Basso and Bornstein, 1999], [Schatzberg et al., 2000], [Jeste et al., 1996].

Another quite impressive result is the performance difference concerning the *premorbid IQ*. Here, on both our subscales, ROP subjects score significantly lower with results being up to one standard deviation below the ROD group's result. These results are supported by prior findings of decreased premorbid IQ in schizophrenic patients [Woodberry et al., 2008].

It must be taken into consideration that all results of the performance comparisons need to be critically examined, since in our study the study groups were not matched for sex nor was the data corrected for confounding variables such as years of education. Nevertheless, we find numerous disparities in the performance of ROD and ROP subjects that are concordant with previous research and the prevailing psychiatric doctrine.

5.1.3 The ROD vs. ROP -23 and -214 variables Analyses

To the best of my knowledge, this is the first study aiming to generate classification models to differentiate between recent onset depression and recent onset psychosis relying solely on neurocognitive data. However, previous studies have already shown that A) it is possible to differentiate between schizophrenia and mood disorders based on patterns in structural MRI data [Koutsouleris et al., 2015] and B) neurocognitive data bears valuable information that -when investigated in a multivariate machine learning environment- can facilitate the identification of ARMS subjects and furthermore even predict their further outcome [Koutsouleris et al., 2011].

To study the performance of neurocognitive pattern recognition models in the differentiation of psychotic and major depressive disorder, I performed two different analyses: The ROD vs. ROP -23 variables Analysis and the ROD vs. ROP -214 variables Analysis. Both approaches produced significant results with classification accuracies of 63.8% and 71.6%, thereby suggesting that pattern recognition models based on neuropsychological data hold a considerable diagnostic power that may facilitate diagnostic differentiation when applied to new unseen patients.

Interestingly, in this study the classification accuracy increases when the SVM is provided with greater, more comprehensive sets of variables: although the ROD vs. ROP -23 variables Analysis is provided with features that have already proven to hold significant information for the differentiation of ROD vs. ROP in the univariate statistical approach, the SVM reaches higher classification accuracies when provided with more variables. This finding underlines the importance of vast and comprehensive data sets in the framework of multivariate machine learning analysis.

Besides the mere classification accuracy, the two analyses also differed regarding the most decisive features for label assignment. Accordingly, in the ROD vs. ROP -23 variables Analysis, the most decisive features (cf. *Table 10*) mainly covered the cognitive domains *speed of processing*, *premorbid IQ*, *visual* and *verbal learning*. In previous research, especially the domains *speed of processing*, *premorbid IQ* and *verbal learning* have been

found to be significantly impaired in patients with psychotic disorders and therefore it is of no surprise that performance in these domains had a driving effect for classification [Heinrichs and Zakzanis, 1998, Mesholam-Gately et al., 2009, Reichenberg et al., 2008]. In the ROD vs. ROP -214 variables Analysis, features that had the greatest decision weights for classification mainly belonged to the cognitive domains *working memory*, *premorbid IQ*, *attention*, *speed of processing*, thus presenting a slightly different composition. Interestingly, for neither of the two approaches did the domains *executive functions* and *social cognition* show a notable decisive value for label assignment. This is somewhat surprising since in previous research deficits in executive functions and social cognition have been described as core features of cognitive impairment in psychotic patients [Lencz et al., 2006, Reichenberg et al., 2008, Green and Horan, 2010].

However, it must be kept in mind that the distribution of decision weights in multivariate analyses is a highly relative calculation whose results have to be handled with caution and may not allow direct conclusions. Besides, the PRONIA Neuropsychological Test Battery was originally assembled with the intention of detecting neurocognitive impairment in a broad spectrum of domains. Reversely, this means that our results may not always be comparable to those of studies extensively examining impairments in only one specific cognitive domain.

Nevertheless, the presumption that pattern recognition methods may identify new and different patterns of illness-specific impairments (than those detected using univariate statistical methods) is an exciting outlook that deserves and requires extensive further investigation.

5.1.4 Leave-Center-Out Analysis and Generalizability

The Leave-center-out Analysis was conducted in order to test the SVM classifiers for their generalizability. We wanted to examine whether the patterns of neurocognitive performance the SVM identified for both study groups actually apply to depressive and psychotic patients in general. In this context, generalizability is synonymous with the

external validity, since in this study subjects from the various centers were recruited in different countries and tested in different languages.

In detail, the idea of the Leave-center-out Analysis was that if classifiers, that were trained on data from 4 different centers and that are tested on the remaining 5^{th} center, reach similar classification accuracies as in the ROD vs. ROP -23 and -214 variables analyses, it can be concluded that the patterns based on which classification is conducted are actually illness-specific and do not randomly occur in the study demographic at hand.

Furthermore, we performed this Leave-center-out Analysis for each center as testing data, thereby gathering additional information regarding the heterogeneity between the data of the centers.

Accordingly, the mean classification accuracy of the Leave-Center-Out Analysis was 72.4% with a sensitivity and a specificity of likewise 72.4%, whereas the accuracies for each center individually ranged between 59.1 and 82.1%. These results, even on the lower end, are well above a random classification accuracy and therefore give reason to believe that our models are generalizable to a certain extend. Moreover, the mean accuracy of 72.4% is well in line with the results of the ROD vs. ROP -214 variables analysis (balanced accuracy: 71.6%).

However, the variety of accuracies between the different centers as test data remains rather conspicuous. These differences could be explained by an overfitting of the classification models, but also by other confounding variables like differences in recruiting and testing between the centers. Additionally, it has to be remembered that the subject cohorts of the 5 different centers have not been matched for age, sex or total number of subjects. Accordingly, the LMU study group consists of 41 subjects in total, whereas there are only 8 subjects in the Uni BHAM cohort (3 ROD, 5 ROP, cf. *Table 3*). Thus, when conducting an analysis that trains on the 108 subjects of the other 4 centers and tests its classification accuracy on the 8 subjects from Uni BHAM, a single misclassification will already lead to a decrease of 12.5% in balanced accuracy.

In conclusion, the results of the leave-center-out approach suggest that patterns of per-

formance and impairment in the neuropsychological assessment identified in recent onset depressive and recent onset psychotic patients do well generalize to outside subjects. Therefore, it is unlikely that these patterns are based on a statistical error or a strong overfitting effect of the SVM. However, to further evaluate the generalizability of neurocognitive patterns in the early stages of mental disorders it will be essential to test on larger study cohorts that have been matched to the training data for confounding variables such as sex, age and years of education.

5.2 Conclusions and Limitations

In a first step, we were able to identify statistically significant performance differences between ROD and ROP subjects on the PRONIA Neuropsychological Test Battery by conducting 2-sample t-tests. ROP participants scored poorer than ROD patients in almost every task. Significant differences were found in 9 out of 23 neurocognitive variables, covering all cognitive domains except *executive functions* and *social cognition*. Therefore, we can conclude that patients in the early stage of psychosis present greater and more profound cognitive deficits than people suffering from recent onset depression. However, the cognitive data examined in these analyses was not corrected for age, sex, years of education or symptom/illness severity. Therefore, it will be necessary to replicate these results with a greater data set, that has been standardized with the data of a comprehensive cohort of healthy controls. Furthermore, it would be of interest to compare the two study groups to a matched cohort of HCs in terms of performance differences.

Secondly, the results of the ROD vs. ROP analyses with 23 and 214 variables suggest that it is possible to identify patterns in the performance on neurocognitive assessments that are specific in depression and psychotic disorders and that further allow differential diagnosis with a considerable diagnostic accuracy. That being said, we further conclude that this differentiation gains diagnostic power with greater numbers of clinical variables that allow the SVM to recognize specific patterns within.

What is more, the machine learning algorithms seem to identify alternating patterns of

disease specific cognitive impairments to those that have been identified by conducting univariate statistics in previous studies. However, it is indispensable to reevaluate this finding by examining greater study samples with data that has been thoroughly controlled and standardized in order to avoid errors due to confounding variables.

Lastly, results of the Leave-center-out Analysis confirm our assumption that neurocognitive data holds valuable information for the diagnostic differentiation of early stages of psychosis and depression. With an overall accuracy of over 72%, our classification models in this analysis have proven to be generalizable to subjects from outside centers that have been recruited and tested by different clinicians, in a different environment and partly even in a different language and country. The results of this analysis are merely afflicted by the considerable differences in the sample sizes of the study centers. Due to the fact that this study had only access to a preliminary data set of the PRONIA study, this limitation could not be cleared at the time of analysis.

The results of this study give evidence to the assumption that neurocognitive data holds valuable information for differential diagnosis in psychiatry. As discussed above, these results will have to be validated by replication in the further process and evaluation of the PRONIA study.

5.2.1 Future Prospects

Besides the replication of the results of this study with larger and standardized data sets, it will be of great interest to examine the value of neurocognitive data in the framework of a multimodal multivariate approach. In previous studies, structural and functional brain imaging has proven its potential to enable accurate differential diagnosis when included and analyzed by pattern recognition algorithms like the SVM [Kambeitz et al., 2015, Kambeitz et al., 2016, Kambeitz et al., 2014]. In the future, the classification accuracy of machine learning algorithms could possibly be improved when provided with data of additional modalities such as psychometric instruments, biographic data, genetic data, blood results or -as examined in this study- neurocognitive performance.

Another promising application of pattern recognition algorithms concerns the domains early disease recognition and prediction. Recently, several studies have been published that do not only investigate multivariate pattern analysis in order to identify illness-specific alterations in diverse data modalities, but that also aim to establish classification models that allow accurate prediction concerning functional outcomes and transition probabilities from clinical high risk for psychosis to frank psychosis [Koutsouleris et al., 2009, Koutsouleris et al., 2011, Koutsouleris et al., 2014]. In this context, the introduction of multimodal classifiers constitutes an interesting and promising approach that may subsequently enhance diagnostic accuracy as well as patient outcome in clinical psychiatry in the future.

6 Acknowledgments

First and foremost I would like to thank Prof. Dr.med. Nikolaos Koutsouleris for introducing me into his team and the world of psychiatric research and giving me the opportunity, knowledge and resources to accomplish this work.

I am particularly grateful to Dr.med. Sebastian von Saldern who never grew tired of encouraging me, teaching me and guiding me along my way. Further, I would like to extend my thanks to Dr. Lana Kambeitz-Ilankovic and Dr.med. Joseph Kambeitz for their valuable insights and their kind hearts.

My grateful thanks are especially extended to Johanna Weiske who never grew tired of listening to my sorrows and complains and who always offered a helping hand.

Also I would like to thank my wonderful colleagues and former co-workers in the PRONIA group Maria Fernanda Urquijo, Dr. Dominic Dwyer, Shalaila Haas, Rachele Sanfelici and Carlos Cabral. I could not have done this work without your help.

Lastly, I want to express my eternal gratefulness to my parents, Ina and Michael, for always providing me with their love, wisdom, experience and kindness.

7 Appendix

7.1 List of neurological and somatic diseases leading to study exclusion

- **Somatic diseases:**
  - Hypertension (Grade II or higher)
  - Sarcoidosis (Boeck's disease)
  - Inflammatory vascular diseases
    - Systemic vasculitis
    - Polyarteritis nodosa
    - Giant-cell arteritis
  - Hepatic cirrhosis
  - Encephalopathy
    - Pancreatic encephalopathy
    - Wernicke's disease
  - Endocrinological diseases
    - Known History of Diabetes mellitus
    - Hyperthyroidism
    - Hypothyroidism; if untreated
    - Addison's disease
    - Cushing's syndrome
  - Hematologic disease
    - Leucaemia (all forms)
    - Polycythemia vera
  - Immunological diseases

    - Systemic lupus erythematosus
  - Cancer

- **Neurological diseases:**
  - Intrauterinely and perinatally acquired brain damages
    - Little's disease
    - Morbus haemolyticus neonatorum (Fetopathia serologica, Erythroblastosis fetalis)
    - Congenital rubella syndrome
    - Congenital toxoplasmosis
    - other congenital embryopathia (Lues, CMV, HIV, Mumps)
    - Fetal alcohol spectrum disorder
  - Malformations of the brain
    - Micro-/Macrocephaly
    - (Meningo-)encephalocele
    - Hydrocephalus
    - Neurofribromatosis
    - Tuberous sclerosis
    - Encephalotrigeminal angiomatosis
    - Von Hippel-Lindau disease
  - Intracranial tumors
    - Medulloblastoma
    - Astrocytoma
    - Oligodendroglioma
    - Glioblastoma
    - Vestibular schwannoma

    * Meningioma
    * (Pituitary) adenoma
    * Cerebral metastasis
- Dementia
    * Alzheimer's disease
    * Frontotemporal dementia
    * Dementia with Lewy bodies
- Dystonia
- Tourette syndrome
- Metabolic diseases
    * Lipoidosis
        · Leukodystrophy
        · Tay-Sachs disease
        · Refsum disease
        · Niemann-Pick disease
        · Gaucher's disease
    * Phenylketonuria
    * Maple syrup urine disease
    * Hartnup disease
    * Galactosemia
    * Wilson's disease
- Inflammatory neurological diseases
    * Multiple sclerosis (encephalomyelitis disseminata)
    * Meningitis
    * Encephalitis

    * Neurosyphilis
    * HIV-encephalitis
    * Behcet's disease
- Parkinson's disease
- Huntington's disease
- Epilepsy
- Autism
- Traumatically acquired brain damage
- Stroke
- Migraines with reoccurring episodes/symptoms within the last 3 months

7.2 List of variables for the ROD vs ROP - 214 Analysis

RAVLT:

Immediate 1 repetition list A
Immediate 2 repetition list A
Immediate 3 repetition list A
Immediate 4 repetition list A
Immediate 5 repetition list A
Immediate out of list words 1 repetition list A
Immediate out of list words 2 repetition list A
Immediate out of list words 3 repetition list A
Immediate out of list words 4 repetition list A
Immediate out of list words 5 repetition list A
Immediate repeated words 1 repetition list A
Immediate repeated words 2 repetition list A
Immediate repeated words 3 repetition list A
Immediate repeated words 4 repetition list A
Immediate repeated words 5 repetition list A
Interference Immediate repetition list B
Interference Immediate words from list A repetition list B
Interference Immediate out of list words repetition list B
Interference Immediate repeated words repetition list B
Interference Immediate 6 repetition list A
Interference Immediate out of list words 6 repetition list A
Interference Immediate repeated words 6 repetition list A
Interference Immediate words from list B 6 repetition list A
Delayed repetition list A
Delayed repetition out of list words list A
Delayed repetition repeated words list A
Delayed repetition words from list B list A

FDS/BDS:

BDS Number of correct trials
BDS Maximum digits string length correctly reminded at least once
FDS Number of correct trials
FDS Maximum digits string length correctly reminded at least once

DANVA:

Raw Score correct

CPT-IP:

Number correct 50 trials
Number error distracting stimuli 50 trials
Number error filler stimuli 50 trials
Number omissions 50 trials
Number correct 100 trials
Number error distracting stimuli 100 trials
Number error filler stimuli 100 trials
Number omissions 100 trials
Number correct 150 trials
Number error distracting stimuli 150 trials
Number error filler stimuli 150 trials
Number omissions 150 trials
Number correct 200 trials
Number error distracting stimuli 200 trials
Number error filler stimuli 200 trials
Number omissions 200 trials
Number correct 250 trials
Number error distracting stimuli 250 trials
Number error filler stimuli 250 trials
Number omissions 250 trials
Number correct 300 trials
Number error distracting stimuli 300 trials
Number error filler stimuli 300 trials
Number omissions 300 trials
Number correct whole test
Number error distracting stimuli whole test
Number error filler stimuli whole test
Number omissions whole test
Reaction times correct whole test
Reaction times error distracting whole test
Reaction times error filler whole test
Reaction times correct 50 trials
Reaction times correct 100 trials
Reaction times correct 150 trials
Reaction times correct 200 trials
Reaction times correct 250 trials
Reaction times correct 300 trials
Reaction times error distracting 50 trials
Reaction times error distracting 100 trials

DSST:

Correct number symbol matchings
Error number symbol matchings
Raw score correct

PVF:

PVF Correct 00 60 letter 1
PVF Correct 00 15 letter 1
PVF Correct 15 30 letter 1
PVF Correct 30 45 letter 1
PVF Correct 45 60 letter 1
PVF Error 00 60 letter 1
PVF Error 00 15 letter 1
PVF Error 15 30 letter 1
PVF Error 30 45 letter 1
PVF Error 45 60 letter 1
PVF Repetition 00 60 letter 1
PVF Repetition 00 15 letter 1
PVF Repetition 15 30 letter 1
PVF Repetition 30 45 letter 1
PVF Repetition 45 60 letter 1

SVF:

SVF Correct 00 60 category 1
SVF Correct 00 15 category 1
SVF Correct 15 30 category 1
SVF Correct 30 45 category 1
SVF Correct 45 60 category 1
SVF Error 00 60 category 1
SVF Error 00 15 category 1
SVF Error 15 30 category 1
SVF Error 30 45 category 1
SVF Error 45 60 category 1
SVF Repetition 00 60 category 1
SVF Repetition 00 15 category 1
SVF Repetition 15 30 category 1
SVF Repetition 30 45 category 1
SVF Repetition 45 60 category 1

SOPT:

Errors 4 elements 01
Errors 4 elements 02
Errors 4 elements 03
Perseveration Errors 4 elements 01
Perseveration Errors 4 elements 02
Perseveration Errors 4 elements 03
Maximum correct responses before error 4 elements 01
Maximum correct responses before error 4 elements 02
Maximum correct responses before error 4 elements 03
Errors 6 elements 01
Errors 6 elements 02
Errors 6 elements 03
Perseveration Errors 6 elements 01
Perseveration Errors 6 elements 02
Perseveration Errors 6 elements 03
Maximum correct responses before error 6 elements 01
Maximum correct responses before error 6 elements 02
Maximum correct responses before error 6 elements 03
Errors 8 elements 01
Errors 8 elements 02
Errors 8 elements 03
Perseveration Errors 8 elements 01
Perseveration Errors 8 elements 02
Perseveration Errors 8 elements 03
Maximum correct responses before error 8 elements 01
Maximum correct responses before error 8 elements 02
Maximum correct responses before error 8 elements 03
Errors 10 elements 01
Errors 10 elements 02
Errors 10 elements 03
Perseveration Errors 10 elements 01
Perseveration Errors 10 elements 02
Perseveration Errors 10 elements 03
Maximum correct responses before error 10 elements 01
Maximum correct responses before error 10 elements 02
Maximum correct responses before error 10 elements 03

ROCF - Copy:

Score element 01
Score element 02
Score element 03
Score element 04
Score element 05
Score element 06
Score element 07
Score element 08
Score element 09
Score element 10
Score element 11
Score element 12
Score element 13
Score element 14
Score element 15
Score element 16
Score element 17
Score element 18

ROCF - Immediate:

Score element 01 Immediate
Score element 02 Immediate
Score element 03 Immediate
Score element 04 Immediate
Score element 05 Immediate
Score element 06 Immediate
Score element 07 Immediate
Score element 08 Immediate
Score element 09 Immediate
Score element 10 Immediate
Score element 11 Immediate
Score element 12 Immediate
Score element 13 Immediate
Score element 14 Immediate
Score element 15 Immediate
Score element 16 Immediate
Score element 17 Immediate
Score element 18 Immediate

ROCF - Delayed:

Score element 01 Delayed
Score element 02 Delayed
Score element 03 Delayed
Score element 04 Delayed
Score element 05 Delayed
Score element 06 Delayed
Score element 07 Delayed
Score element 08 Delayed
Score element 09 Delayed
Score element 10 Delayed
Score element 11 Delayed
Score element 12 Delayed
Score element 13 Delayed
Score element 14 Delayed
Score element 15 Delayed
Score element 16 Delayed
Score element 17 Delayed
Score element 18 Delayed

ROCF - Whole:

Accuracy whole
Accuracy whole Immediate
Accuracy whole Delayed
Placement whole
Placement whole Immediate
Placement whole Delayed
Score whole
Score whole Immediate
Score whole Delayed
Time Copy
Time Immediate
Time Delayed

TMTA:

TMTA Time of completion
TMTA Errors
TMTA Violations

TMTB:

TMTB Time of completion
TMTB Errors
TMTB Violations

WAIS:

WAIS V Standard score
WAIS MR Standard score

7.3 List of Abbreviations

- **ARMS** - At-Risk Mental States for Psychosis
- **BDS** - Backward Digit Span Task
- **CPT-IP** - Continuous Performance Test - Identical Pairs
- **CV** - Cross-Validation
- **DANVA** - Diagnostic Analysis of Nonverbal Accuracy - 2nd Version
- **DGPPN** - Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde
- **DSM-5** - Diagnostic and Statistical Manual of Mental Disorders
- **DSST** - Digit Symbol Substitution Test
- **FDS** - Forward Digit Span Task
- **FEP** - First Episode Psychosis
- **IQ** - Intelligence Quotient
- **LMU** - Ludwig-Maximilians-University München
- **MDD** - Major Depressive Disorder

- **MilanNig** - University of Milan
- **MVA** - Multivariate Analysis
- **OSH** - Optimal Separating Hyperplane
- **PEBL** - The Psychology Experiment Building Language
- **PRONIA** - Personalised Prognostic Tools for Early Psychosis Management
- **PVF** - Phonemic Verbal Fluency Task
- **RAVLT** - Rey Auditory Verbal Learning Test
- **ROCF** - Rey-Osterrieth Complex Figure Test
- **ROD** - Recent onset Depression
- **ROP** - Recent onset Psychosis
- **SAT** - Salience Attribution Test
- **SCID** - Structured Clinical Interview for DSM-IV
- **SH** - Separating Hyperplane
- **SOPT** - Self-Ordered Pointing Test
- **SVF** - Semantic Verbal Fluency Task
- **SVM** - Support Vector Machine
- **TMTA** - Trail Making Test A
- **TMTB** - Trail Making Test B
- **UBS** - University of Basel
- **UKK** - University of Cologne

- **Uni BHAM** - University of Birmingham
- **Uni Udine** - University of Udine
- **WAIS-M** - Wechsler Adult Intelligence Scale: Matrices/Matrix Reasoning
- **WAIS-V** - Wechsler Adult Intelligence Scale: Vocabulary
- **WHO** - World Health Organization

List of Figures

List of Tables

References

[Alboni et al., 2008] Alboni, P., Favaron, E., Paparella, N., Sciammarella, M., and Pedaci, M. (2008). Is there an association between depression and cardiovascular mortality or sudden death? *Journal of Cardiovascular Medicine*, 9(4):356–362.

[Aloia et al., 1996] Aloia, M. S., Gourovitch, M. L., Weinberger, D. R., and Goldberg, T. E. (1996). An investigation of semantic space in patients with schizophrenia. *Journal of the International Neuropsychological Society*, 2(04):267–273.

[Andreasen, 1982] Andreasen, N. C. (1982). Negative symptoms in schizophrenia: definition and reliability. *Archives of general psychiatry*, 39(7):784–788.

[Association et al., 2013] Association, A. P. et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[Austin et al., 2001] Austin, M.-P., Mitchell, P., and Goodwin, G. M. (2001). Cognitive deficits in depression. *The British Journal of Psychiatry*, 178(3):200–206.

[Barder et al., 2013] Barder, H. E., Sundet, K., Rund, B. R., Evensen, J., Haahr, U., ten Velden Hegelstad, W., Joa, I., Johannessen, J. O., Langeveld, H., Larsen, T., et al. (2013). Neurocognitive development in first episode psychosis 5years follow-up: associations between illness severity and cognitive course. *Schizophrenia research*, 149(1):63–69.

[Basso and Bornstein, 1999] Basso, M. R. and Bornstein, R. A. (1999). Neuropsychological deficits in psychotic versus nonpsychotic unipolar depression. *NEUROPSYCHOLOGY-NEW YORK-*, 13:69–75.

[Beck et al., 1961] Beck, A., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 5:561–571.

[Beck et al., 1996] Beck, A. T., Steer, R. A., and Brown, G. K. (1996). Beck depression inventory-ii. *San Antonio, TX*, pages 78204–2498.

[Beretta and Santaniello, 2016] Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, 16(3):74.

[Bora and Murray, 2013] Bora, E. and Murray, R. M. (2013). Meta-analysis of cognitive deficits in ultra-high risk to psychosis and first-episode psychosis: do the cognitive deficits progress over, or after, the onset of psychosis? *Schizophrenia bulletin*, 40(4):744–755.

[Broome et al., 2005] Broome, M. R., Woolley, J. B., Tabraham, P., Johns, L. C., Bramon, E., Murray, G. K., Pariante, C., McGuire, P. K., and Murray, R. M. (2005). What causes the onset of psychosis? *Schizophrenia research*, 79(1):23–34.

[Burges, 1998] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

[Cardinal and Bullmore, 2011] Cardinal, R. N. and Bullmore, E. T. (2011). *The diagnosis of psychosis*. Cambridge University Press.

[Cassano and Fava, 2002] Cassano, P. and Fava, M. (2002). Depression and public health: an overview. *Journal of psychosomatic research*, 53(4):849–857.

[Chapman and Perry, 2008] Chapman, D. P. and Perry, G. S. (2008). Peer reviewed: depression as a major component of public health for older adults. *Preventing chronic disease*, 5(1).

[Cornblatt et al., 1989] Cornblatt, B. A., Lenzenweger, M. F., and Erlenmeyer-Kimling, L. (1989). The continuous performance test, identical pairs version: Ii. contrasting attentional profiles in schizophrenic and depressed patients. *Psychiatry research*, 29(1):65–85.

[Cornblatt et al., 1988] Cornblatt, B. A., Risch, N. J., Faris, G., Friedman, D., and Erlenmeyer-Kimling, L. (1988). The continuous performance test, identical pairs version (cpt-ip): I. new findings about sustained attention in normal families. *Psychiatry research*, 26(2):223–238.

[Cragg and Nation, 2007] Cragg, L. and Nation, K. (2007). Self-ordered pointing as a test of working memory in typically developing children. *Memory*, 15(5):526–535.

[Cummings, 1985] Cummings, J. L. (1985). Organic delusions: phenomenology, anatomical correlations, and review. *The British Journal of Psychiatry*, 146(2):184–197.

[Davatzikos et al., 2008] Davatzikos, C., Fan, Y., Wu, X., Shen, D., and Resnick, S. M. (2008). Detection of prodromal alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging*, 29(4):514–523.

[Davatzikos et al., 2005] Davatzikos, C., Shen, D., Gur, R. C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B. I., and Gur, R. E. (2005). Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Archives of general psychiatry*, 62(11):1218–1227.

[Drake et al., 2000] Drake, R. J., Haley, C. J., Akhtar, S., and Lewis, S. W. (2000). Causes and consequences of duration of untreated psychosis in schizophrenia. *The British Journal of Psychiatry*, 177(6):511–515.

[Eaton et al., 2008] Eaton, W. W., Shao, H., Nestadt, G., Lee, B. H., Bienvenu, O. J., and Zandi, P. (2008). Population-based study of first onset and chronicity in major depressive disorder. *Archives of general psychiatry*, 65(5):513–520.

[Ehlers et al., 1988] Ehlers, C. L., Frank, E., and Kupfer, D. J. (1988). Social zeitgebers and biological rhythms: a unified approach to understanding the etiology of depression. *Archives of general psychiatry*, 45(10):948–952.

[Faber, 2007] Faber, N. (2007). How to avoid over-fitting in multivariate calibrationthe conventional validation approach and an alternative. *Analytica Chimica Acta*, 595(1):98–106.

[Fan et al., 2008] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

[Fava et al., 2003] Fava, G. A., Ruini, C., and Sonino, N. (2003). Treatment of recurrent depression. *Cns Drugs*, 17(15):1109–1117.

[Filzmoser et al., 2009] Filzmoser, P., Liebmann, B., and Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics*, 23(4):160–171.

[First, 2013] First, M. B. (2013). *DSM-5 handbook of differential diagnosis*. American Psychiatric Pub.

[Freedman et al., 2013] Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., Gabbard, G. O., Gau, S. S.-F., Javitt, D. C., Oquendo, M. A., et al. (2013). The initial field trials of dsm-5: new blooms and old thorns.

[Geddes et al., 2003] Geddes, J. R., Carney, S. M., Davies, C., Furukawa, T. A., Kupfer, D. J., Frank, E., and Goodwin, G. M. (2003). Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *The Lancet*, 361(9358):653–661.

[Green and Horan, 2010] Green, M. F. and Horan, W. P. (2010). Social cognition in schizophrenia. *Current Directions in Psychological Science*, 19(4):243–248.

[Griswold et al., 2015] Griswold, K., Del Regno, P. A., and Berger, R. C. (2015). Recognition and differential diagnosis of psychosis in primary care. *Brain*, 100(11):18–37.

[Hall, 1995] Hall, R. C. (1995). Global assessment of functioning: a modified scale. *Psychosomatics*, 36(3):267–275.

[Hamon and Blier, 2013] Hamon, M. and Blier, P. (2013). Monoamine neurocircuitry in depression and strategies for new treatments. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 45:54–63.

[Haxby, 2012] Haxby, J. V. (2012). Multivariate pattern analysis of fmri: the early beginnings. *Neuroimage*, 62(2):852–855.

[Heinrichs and Zakzanis, 1998] Heinrichs, R. W. and Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology*, 12(3):426.

[Jeste et al., 1996] Jeste, D. V., Heaton, S. C., Paulsen, J. S., Ercoli, L., Harris, M. J., and Heaton, R. K. (1996). Clinical and neuropsychological comparison of psychotic depression with nonpsychotic depression and schizophrenia. *American Journal of Psychiatry*, 153(4):490–496.

[John et al., 1994] John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129.

[Kambeitz et al., 2014] Kambeitz, J., Abi-Dargham, A., Kapur, S., and Howes, O. D. (2014). Alterations in cortical and extrastriatal subcortical dopamine function in schizophrenia: systematic review and meta-analysis of imaging studies. *The British Journal of Psychiatry*, 204(6):420–429.

[Kambeitz et al., 2016] Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., Walter, M., Falkai, P., and Koutsouleris, N. (2016). Detecting neuroimaging biomarkers for depression: A meta-analysis of multivariate pattern recognition studies. *Biological psychiatry*.

[Kambeitz et al., 2015] Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., and Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, 40(7):1742.

[Kapur et al., 2005] Kapur, S., Mizrahi, R., and Li, M. (2005). From dopamine to salience to psychosislinking biology, pharmacology and phenomenology of psychosis. *Schizophrenia research*, 79(1):59–68.

[Kawasaki et al., 2007] Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.-Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., and Kurachi, M. (2007). Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *Neuroimage*, 34(1):235–242.

[Kay et al., 1987] Kay, S. R., Flszbein, A., and Opfer, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261.

[Keefe, 1995] Keefe, R. S. (1995). The contribution of neuropsychology to psychiatry. *American Journal of Psychiatry*, 152(1):6–15.

[Keefe et al., 2004] Keefe, R. S., Goldberg, T. E., Harvey, P. D., Gold, J. M., Poe, M. P., and Coughenour, L. (2004). The brief assessment of cognition in schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophrenia research*, 68(2):283–297.

[Kessler et al., 2007] Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., and Ustun, T. B. (2007). Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry*, 20(4):359.

[Kessler et al., 2005] Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., and Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602.

[Kessler and Bromet, 2013] Kessler, R. C. and Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual review of public health*, 34:119–138.

[Kirkbride et al., 2012] Kirkbride, J., Errazuriz, A., Croudace, T., Morgan, C., Jackson, D., McCrone, P., Murray, R., and Jones, P. (2012). Systematic review of the incidence and prevalence of schizophrenia and other psychoses in england, 1950-2009.

[Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA.

[Köhler et al., 2014] Köhler, O., Benros, M. E., Nordentoft, M., Farkouh, M. E., Iyengar, R. L., Mors, O., and Krogh, J. (2014). Effect of anti-inflammatory treatment on depression, depressive symptoms, and adverse effects: a systematic review and meta-analysis of randomized clinical trials. *JAMA psychiatry*, 71(12):1381–1391.

[Köllner and Schauenburg, 2012] Köllner, V. and Schauenburg, H. (2012). *Psychotherapie im Dialog-Diagnostik und Evaluation*. Georg Thieme Verlag.

[Koutsouleris et al., 2011] Koutsouleris, N., Davatzikos, C., Bottlender, R., Patschurek-Kliche, K., Scheuerecker, J., Decker, P., Gaser, C., Möller, H.-J., and Meisenzahl, E. M. (2011). Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophrenia bulletin*, page sbr037.

[Koutsouleris et al., 2015] Koutsouleris, N., Meisenzahl, E. M., Borgwardt, S., Riecher-Rössler, A., Frodl, T., Kambeitz, J., Köhler, Y., Falkai, P., Möller, H.-J., Reiser, M., et al. (2015). Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain*, 138(7):2059–2073.

[Koutsouleris et al., 2009] Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., et al. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of general psychiatry*, 66(7):700–712.

[Koutsouleris et al., 2014] Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., von Saldern, S., Cabral, C., Reiser, M., Falkai, P., et al. (2014). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia bulletin*, 41(2):471–482.

[Kremen et al., 2003] Kremen, W. S., Seidman, L. J., Faraone, S. V., and Tsuang, M. T. (2003). Is there disproportionate impairment in semantic or phonemic fluency in schizophrenia? *Journal of the International Neuropsychological Society*, 9(01):79–88.

[Laursen et al., 2012] Laursen, T. M., Munk-Olsen, T., and Vestergaard, M. (2012). Life expectancy and cardiovascular mortality in persons with schizophrenia. *Current opinion in psychiatry*, 25(2):83–88.

[Lee et al., 2012] Lee, R. S., Hermens, D. F., Porter, M. A., and Redoblado-Hodge, M. A. (2012). A meta-analysis of cognitive deficits in first-episode major depressive disorder. *Journal of affective disorders*, 140(2):113–124.

[Lencz et al., 2006] Lencz, T., Smith, C. W., McLaughlin, D., Auther, A., Nakayama, E., Hovey, L., and Cornblatt, B. A. (2006). Generalized and specific neurocognitive deficits in prodromal schizophrenia. *Biological psychiatry*, 59(9):863–871.

[Leung et al., 2012] Leung, Y. W., Flora, D. B., Gravely, S., Irvine, J., Carney, R. M., and Grace, S. L. (2012). The impact of pre-morbid and post-morbid depression onset on mortality and cardiac morbidity among coronary heart disease patients: A meta-analysis. *Psychosomatic medicine*, 74(8):786.

[Lewis, 2004] Lewis, R. (2004). Should cognitive deficit be a diagnostic criterion for schizophrenia? *Journal of Psychiatry and Neuroscience*, 29(2):102.

[Loranger, 1984] Loranger, A. W. (1984). Sex difference in age at onset of schizophrenia. *Archives of general psychiatry*, 41(2):157–161.

[Macmillan and Creelman, 2004] Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

[McGorry et al., 2011] McGorry, P. D., Purcell, R., Goldstone, S., and Amminger, G. P. (2011). Age of onset and timing of treatment for mental and substance use disorders: implications for preventive intervention strategies and models of care. *Current opinion in psychiatry*, 24(4):301–306.

[McGrath et al., 2008] McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews*, 30(1):67–76.

[McIntyre et al., 2013] McIntyre, R. S., Cha, D. S., Soczynska, J. K., Woldeyohannes, H. O., Gallaugher, L. A., Kudlow, P., Alsuwaidan, M., and Baskaran, A. (2013). Cognitive deficits and functional outcomes in major depressive disorder: determinants, substrates, and treatment interventions. *Depression and anxiety*, 30(6):515–527.

[Menezes et al., 2006] Menezes, N., Arenovich, T., and Zipursky, R. (2006). A systematic review of longitudinal outcome studies of first-episode psychosis. *Psychological medicine*, 36(10):1349–1362.

[Mesholam-Gately et al., 2009] Mesholam-Gately, R. I., Giuliano, A. J., Goff, K. P., Faraone, S. V., and Seidman, L. J. (2009). Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology*, 23(3):315.

[Mirzakhanian et al., 2013] Mirzakhanian, H., Singh, F., Seeber, K., Shafer, K. M., and Cadenhead, K. S. (2013). A developmental look at the attentional system in the at risk and first episode of psychosis: age related changes in attention along the psychosis spectrum. *Cognitive neuropsychiatry*, 18(1-2):26–43.

[Moritz et al., 2011] Moritz, S., Veckenstedt, R., Randjbar, S., Vitzthum, F., and Woodward, T. (2011). Antipsychotic treatment beyond antipsychotics: metacognitive in-

tervention for schizophrenia patients improves delusional symptoms. *Psychological medicine*, 41(9):1823–1832.

[Mueller et al., 2015] Mueller, T., Kusne, A. G., and Ramprasad, R. (2015). Machine learning in materials science: Recent progress and emerging applications. *Rev. Comput. Chem.*

[Nelson et al., 1998] Nelson, E. B., Sax, K. W., and Strakowski, S. M. (1998). Attentional performance in patients with psychotic and nonpsychotic major depression and schizophrenia. *American Journal of Psychiatry*, 155(1):137–139.

[Nowicki and Duke, 2001] Nowicki, S. and Duke, M. P. (2001). Nonverbal receptivity: The diagnostic analysis of nonverbal accuracy (danva).

[Perälä et al., 2007] Perälä, J., Suvisaari, J., Saarni, S. I., Kuoppasalmi, K., Isometsä, E., Pirkola, S., Partonen, T., Tuulio-Henriksson, A., Hintikka, J., Kieseppä, T., et al. (2007). Lifetime prevalence of psychotic and bipolar i disorders in a general population. *Archives of general psychiatry*, 64(1):19–28.

[Petrides and Milner, 1982] Petrides, M. and Milner, B. (1982). Deficits on subject-ordered tasks after frontal-and temporal-lobe lesions in man. *Neuropsychologia*, 20(3):249–262.

[Reichenberg et al., 2008] Reichenberg, A., Harvey, P. D., Bowie, C. R., Mojtabai, R., Rabinowitz, J., Heaton, R. K., and Bromet, E. (2008). Neuropsychological function and dysfunction in schizophrenia and psychotic affective disorders. *Schizophrenia bulletin*, 35(5):1022–1029.

[Rey, 1941] Rey, A. (1941). L'examen psychologique dans les cas d'encphalopathie traumatique. (les problems.). /the psychological examination in cases of traumatic encepholopathy. problems. *Archives de Psychologie*, 28:215–285.

[Roiser et al., 2009] Roiser, J., Stephan, K., Den Ouden, H., Barnes, T., Friston, K., and Joyce, E. (2009). Do patients with schizophrenia exhibit aberrant salience? *Psychological medicine*, 39(02):199–209.

[Roitman et al., 1997] Roitman, S. E. L., Keefe, R. S., Harvey, P. D., Siever, L. J., and Mohs, R. C. (1997). Attentional and eye tracking deficits correlate with negative symptoms in schizophrenia. *Schizophrenia research*, 26(2):139–146.

[Saeys et al., 2007] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

[Schatzberg et al., 2000] Schatzberg, A. F., Posener, J. A., DeBattista, C., Kalehzan, B. M., Rothschild, A. J., and Shear, P. K. (2000). Neuropsychological deficits in psychotic versus nonpsychotic major depression and no mental illness. *American Journal of Psychiatry*, 157(7):1095–1100.

[Schmidt et al., 1996] Schmidt, M. et al. (1996). *Rey auditory verbal learning test: A handbook*. Western Psychological Services Los Angeles, CA.

[Segarra et al., 2012] Segarra, R., Ojeda, N., Pena, J., Garcia, J., Rodriguez-Morales, A., Ruiz, I., Hidalgo, R., Buron, J., Eguiluz, J., and Gutierrez, M. (2012). Longitudinal changes of insight in first episode psychosis and its relation to clinical symptoms, treatment adherence and global functioning: one-year follow-up from the eiffel study. *European Psychiatry*, 27(1):43–49.

[Seidman et al., 2003] Seidman, L. J., Lanca, M., Kremen, W. S., Faraone, S. V., and Tsuang, M. T. (2003). Organizational and visual memory deficits in schizophrenia and bipolar psychoses using the rey-osterrieth complex figure: effects of duration of illness. *Journal of Clinical and Experimental Neuropsychology*, 25(7):949–964.

[Sobocki et al., 2006] Sobocki, P., Jönsson, B., Angst, J., and Rehnberg, C. (2006). Cost of depression in europe. *The journal of mental health policy and economics*, 9(2):87–98.

[Stahl and Mignon, 2010] Stahl, S. M. and Mignon, L. (2010). *Stahl's Illustrated Antipsychotics: Treating Psychosis, Mania and Depression*. Cambridge University Press.

[Strauss et al., 2006] Strauss, E., Sherman, E. M., and Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society.

[Sullivan et al., 2000] Sullivan, P. F., Neale, M. C., and Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, 157(10):1552–1562.

[Tombaugh, 2004] Tombaugh, T. N. (2004). Trail making test a and b: normative data stratified by age and education. *Archives of clinical neuropsychology*, 19(2):203–214.

[Tsuang and Faraone, 2002] Tsuang, M. T. and Faraone, S. V. (2002). Diagnostic concepts and the prevention of schizophrenia.

[Velligan and Alphs, 2008] Velligan, D. I. and Alphs, L. D. (2008). Negative symptoms in schizophrenia: the importance of identification and treatment. *Psychiatric Times*, 25(3):12.

[Wechsler, 2014] Wechsler, D. (2014). Wechsler adult intelligence scale–fourth edition (wais–iv).

[Woodberry et al., 2008] Woodberry, K. A., Giuliano, A. J., and Seidman, L. J. (2008). Premorbid iq in schizophrenia: a meta-analytic review. *American Journal of Psychiatry*, 165(5):579–587.

[Zarogianni et al., 2013] Zarogianni, E., Moorhead, T. W., and Lawrie, S. M. (2013). Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. *NeuroImage: Clinical*, 3:279–289.

Eidesstattliche Versicherung

Köhler, Yanis-Michael L. G.

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Titel

Differentiation of Recent Onset Depression vs. Recent Onset Psychosis using Pattern Classification Methods on Neuropsychological Data: Diagnostic Performance and Generalizability

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Hamburg, 15.07.2019 Yanis-Michael L. G. Köhler

_____

Unterschrift Doktorandin bzw. Doktorand