The studies described in this thesis were performed at the Rudolf Magnus Institute of Neuroscience, Department of Psychiatry, University Medical Center Utrecht, the Netherlands.

ISBN: 978-90-393-6497-0

Cover design: Sinds1961 Grafisch Ontwerp
Design: Mireille Nieuwenhuis
Printed by: Printservice Ede, the Netherlands

The Ghost in the Machine

Machine learning models of the brain and genome in patients with schizophrenia and bipolar disorder

De ziel in een machine

Geautomatiseerde modellen van het brein en het genoom in patienten met schizofrenie en bipolaire stoornis

(met een samenvatting in het nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht op gezag van de
rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor
promoties in het openbaar te verdedigen
op donderdag 11 februari 2016 des ochtends te 10.30 uur

door

Mireille Nieuwenhuis
geboren op 15 februari 1985 te Utrecht

Promotor:	Prof dr. R.S. Kahn
Copromotor:	Dr. H.G. Schnack

Contents

Chapter 1

Introduction

Introduction

The title of this thesis “The Ghost in the Machine” refers to the title of a 1967 book about philosophical psychology, quoting Gilbert Ryle who disputes the Cartesian dualist account of the body-mind relationship (Descartes argues that the body and mind are two distinct substances). Next to that, in computer sciences “The Ghost in the Machine” refers to intelligent machines responding in an inexplicable human-like manner. We cannot program consciousness, but can this “ghost” arise in complex enough systems.

In this thesis we explore the mental health-body relationship, trying to capture the ghost in the human machine. It is well known that brains of schizophrenia and bipolar patients show structural abnormalities (S. V Haijma et al., 2013; Kempton et al., 2008). Building on this information we investigated if these brain anomalies can be detected by machine learning models, and to predict outcome or illness course in patients with schizophrenia or bipolar disorder. To broaden our search for an objective model attempting to differentiate between schizophrenia patients and healthy controls, we also built several models on common genetic variants.

According to the World Health Organization mental illness is in the top three of disease burden worldwide (WHO, 2008). Currently, the diagnosis of these disorders such as schizophrenia and bipolar disorder is based predominantly on their clinical manifestations. These illnesses can worsen if left undiagnosed and untreated. More objective measures would help psychiatrists in the process of diagnosis and increase its reliability, which could in turn lead to a healthier and higher quality of live of patients.
In the studies in this thesis I attempt to apply machine learning to assist in individualized diagnosis and prognosis of patients with mental disorders.

Machine learning

Machine learning originated from the field of computer sciences. The concept of machine learning entails that a machine learns from already gathered data to make predictions in a new data set. It finds regularities in the data, which are similar, but need not be exactly the same, thus taking into account statistical variation.

General

In this paragraph I will explain why we chose the support vector machine as machine learning method in our research. There are several categories of machine learning methods; supervised vs. unsupervised; black box vs. interpretable. In supervised learning the model is based on information including a label or category, whereas in unsupervised learning, the machine has to find a pattern plus label for the data. The data we obtained is already categorized by illness status, or illness course, so called labeled data, thus we applied a supervised machine learning algorithm.

In a black box methodology a prediction is made for future subjects, but what exactly made the machine come to this decision is not determinable. However, knowing where in the brain or genome the abnormalities underlying the decision are is important from a scientific point of view. These patterns could in the future possibly help determine cause, origin and/or treatment of the illness at hand. Therefore, an interpretable method was elected.

Another advantage of the technique we chose is that it is able to handle data sets that have a low number of cases (participants, in our case) compared to the number of features describing that dataset.

Support Vector Machine

A support vector machine (SVM) is a high-dimensional supervised learning algorithm (Vapnik, 1999). It learns a function to divide individuals to one of two classes from presented data, which can be used to predict the class based on data from new individuals. Figure 1 shows a two-dimensional illustrative example of an SVM model. A SVM model is a function $y(x_i)$ that divides the space into two parts, labeling all new data in one part to belong to the same class, i.e., a subject is classified according to the sign of $y(x_i)$.

Function: $y(x_i) = w^{\mathsf{T}} \cdot x_i - b$

The w is the weight vector, b is an offset, and xi represents a subject. In the training phase each subject has a label, t_i. During the training phase the function is optimized by requiring $y(x_i)$ < 0 if t_i = -1, and $y(x_i)$ > 0 if t_i = +1. The weight vector contains information on feature importance, and also on whether the particular feature shows an increase or decrease in the pattern when comparing one class to the other. Theoretically, there can be several surfaces that exactly separate the classes. To determine which one to use, the SVM chooses the optimal separating hyperplane (OSH) such that the space between the two classes, which is called the margin, is made as large as possible. The size of the margin is 2/$\| w \|$, so minimizing $\| w \|$ will maximize the margin. Because the problem is not per se linearly separable, an error measure ε is brought into the equation. If the subject is classified correctly ξ = 0; otherwise it is the distance from the OSH to the subject. There is a free parameter in linear SVM to influence the narrowness of the margin, a penalty C is multiplied by the error per subject. The OSH is now dependent on both the margin and the error. On the one hand the margin is maximized and on the other hand the error times C is minimized leading to a minimization of:

$$C\sum_{n=1}^{N}\xi_n + \frac{2}{\| w \|}$$

This means that if C is larger, then the penalty of misclassified subjects will be higher and the margin will most likely be smaller. Tuning C can increase the model's performance (Franke et al., 2010; Nieuwenhuis et al., 2012).

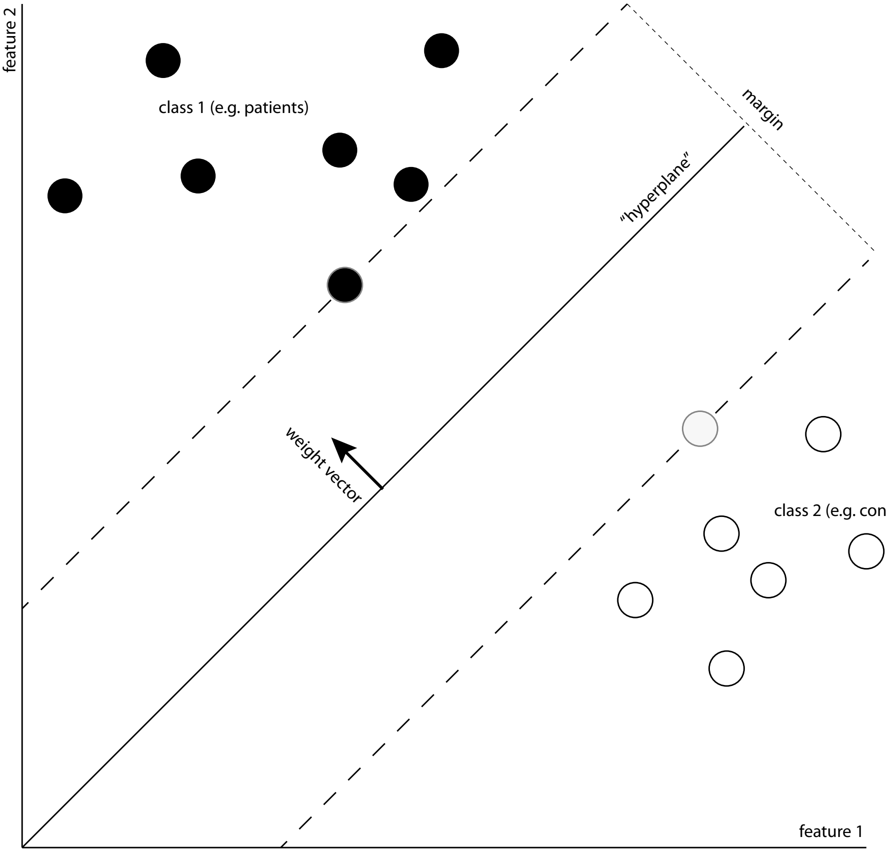

Figure 1 | An illustrative example of a two-dimensional SVM-model. The circles represent the training data; the black circles class 1 and the white circles class 2. The hyperplane is the solid line separating the two classes. In a two-dimensional example this is a line.

In a support vector machine model each individual subject is represented by a set of features. In this thesis, we either use local gray matter volumes of the brain in voxels, explained in the next Magnetic Resonance Imaging section, or common genetic variations, which is explained in the Genome section.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a non-invasive technique that can be used to study the brain in vivo. It is a well-established technique to study brain abnormalities in mental disorders. MRI uses radio waves and magnetic fields to determine structure and tissue of the object that is being studied. The technique uses the fact that the nuclei of many atoms (e.g., hydrogen (H)) rotate rapidly around their axes; this motion is called spin. Due to the magnetic field inside an MRI scanner, Hydrogen proton spins are all aligned in the same direction. When a radio transmitter generates an electromagnetic pulse, the protons are triggered to precess around the orientation of the scanner's magnetic field. These protons than return to their aligned state; different types of tissue lead to different characteristic return times, or relaxation times. The signals from different tissues at different locations can be combined to form a three-dimensional image.

Image acquisition

There are several parameter settings that can be set to acquire different types of MR images. These settings are referred to as the MRI protocol. For example, the flip angle is the angle over which the proton spins are rotated by the radio pulse, influencing the signal strength. Different types of images emphasize different contrasts between different tissue types; the most commonly used images in structural brain studies are T1 and T2 weighted images. In the studies in this thesis T1 weighted images are used. Tissues that show up with different intensities in these images are fluids such as Cerebral Spinal Fluid (CSF), muscle and fat, and more importantly gray matter (GM) and white matter (WM).

The field strength of an MRI scanner is measured in tesla (T). Most of the scanners that are used for clinical purposes operate on 1.5 T or 3.0 T. In recent years, more and more hospitals acquire scanners with a field as strong as 7.0 T. The stronger the magnetic field the higher the possible quality of the scans (signal to noise ratio; resolution; or acquisition speed) of the images.

Image processing

A three-dimensional image is built up out of voxels, which is the equivalent of a pixel in a 2D image. Each voxel has a gray scale value that is used to determine the tissue of which the voxel is comprised. In a T1-weighted image CSF appears almost black, GM is gray and WM is white. The process of assigning a tissue type to each voxel is called segmentation.

To be able to compare different brains, all brains need to have the same orientation and are thus transformed into the same coordinate system. One commonly used technique to do this is voxel based morphometry (VBM)(Ashburner and Friston, 2001), where all images are nonlinearly registered to a template brain image, so that the brain tissue can be compared voxel by voxel through the whole brain.

Genome

The genome is the complete set of genes present in an organism. Genes consist of DNA, which is divided into separate pieces called chromosomes. Humans have 23 chromosome pairs, these chromosomes are made up out of nucleotide base pairs: cytosine (C) and guanine (G) or adenine (A) and thymine (T). Our genome consists of about 3.2 billion base pairs. Most of the human DNA is identical, between individuals, however some of our DNA is not. When individuals have different base pairs at the same location, it is called a single nucleotide polymorphism (SNP).

To increase our understanding on the human genome and psychiatric illnesses the Psychiatric Genomics Consortium (PGC) was founded in 2007 (PGC 2015). Currently, the consortium has obtained genetic data of about 40,000 schizophrenia patients. The consortium performs genome-wide association studies (GWAS), initially examining common genetic variants, focusing on associations between single nucleotide polymorphisms (SNPs) and psychiatric illnesses.

Mental disorders

Someone who has a mental disorder suffers from a psychological syndrome or behavioral pattern that causes this person to function poorly in daily life. Assessment if someone suffers from a mental disorder is done by observation and questioning. There are two widely used guidelines for diagnosis, i.e., the Diagnostic Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2013) and the International Classification of Diseases (ICD) (WHO, 2010).

According to the latest Diagnostic Statistical Manual of Mental Disorders (DSM-V) a mental disorder is defined as "a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning.”

Schizophrenia

Schizophrenia is a severe mental disorder. Patients affected by schizophrenia can display any of a broad array of symptoms, including psychosis, delusions, paranoia, hallucinations, disorganized speech, but also lack of motivation, lack of interest, apathy, lack of speech, and flatness. Even though psychosis or loss of contact with reality is a symptom of schizophrenia, only one third of the patients that undergo a first psychotic episode evolve to schizophrenia (Harrison et al., 2001).

Etiology and Risk factors

What causes schizophrenia is not known. There is consensus that the combination of a genetic vulnerability combined with environmental factors increase the risk to develop the illness. Several alleles have been found to be associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), however not one allele in particular is believed to cause the disease (Harrison and Weinberger, 2005). Onset of schizophrenia typically occurs between the age of 16 and 20 years old (Sham et al., 1994). Lifetime prevalence for schizophrenia is 0.5% (Simeone et al., 2015).

Brain abnormalities

Schizophrenia is associated with smaller gray matter volume. Predominant gray matter volume losses are found in the prefrontal cortex, but also in superior and medial frontal and temporal gyri, insula and thalamus (Fornito et al., 2009; S. Haijma et al., 2013; Honea et al., 2000; Shepherd et al., 2012). Even at the first psychotic episode, patients show thalamic, insular and hippocampal volume reductions and larger ventricular volume compared to healthy controls (Levitt et al., 2010; Rosa et al., 2010; Schaufelberger et al., 2007; Steen et al., 2006). Although these statistical findings are scientifically interesting and could help us fathom aspects such as illness development, illness course and its effect on the brain, they are of no avail to diagnose individuals.

Bipolar disorder

Bipolar disorder is more commonly known as manic-depressive illness. It is characterized by alternating periods of emotional highs and lows. When patients experience an emotional high or mania, symptoms include: loss of reality, persistent elevated or irritated mood, increased activity or energy, talkativeness, increased sexual activity, decreased need for sleep, racing thoughts, increased distractibility, and unrestrained buying sprees. When patients experience an emotional low or depression, symptoms include depressed mood, disturbed sleep patterns resulting in insomnia or excessive sleeping, feelings of worthlessness, lack of joy, indecisiveness, and recurrent thoughts of death. Most patients spend more time in a depressive state than in a manic state. Even though the disease is treatable, episodes of mania and depression typically recurrent over time.

Etiology and Risk factors

There is no single known cause for bipolar disorder. There are alleles that have been found associated with bipolar disorder (Frey et al., 2013). While children or siblings of patients have an increased risk to develop the illness, its cause is not purely genetic.

The average age of onset of bipolar disorder is during the early 20s, although there have been reports of the disorder beginning as early as elementary school. Lifetime prevalence of bipolar disorder is 1% to 3% worldwide (Merikangas et al., 2011).

Brain abnormalities

Patients with bipolar disorder show larger volumes of the lateral and third ventricles and a smaller area of the corpus callosum. Gray matter volume is found to be larger in patients using lithium as compared to those not on lithium (Kempton et al., 2008). Locally, the prefrontal cortex in adults with bipolar disorder tends to be smaller than in healthy controls (Soares et al., 2005). Again, these findings are group differences and they provide no information about an individual.

Thesis overview

The overall aim of this work is to integrate machine learning techniques with available unbiased case-control data into clinically predictive models.

When I started my PhD project in 2010, several smaller sized studies used machine learning to predict schizophrenia based on magnetic resonance images of the brain. In Chapter 2, we build such a MRI based prediction model and study the generalizability and possibility of clinical application of such a model. We use two large independent samples, one to create a model and another independent sample to test its accuracy.

A psychiatrist can easily determine if someone is ill or not, i.e., simply using an MRI scan to diagnose is not adding additional information. A more challenging problem is to differentiate between different psychiatric disorders, using MRI based information. In Chapter 3, we aim to separate not only healthy controls from schizophrenia patients, but also to separate bipolar patients from schizophrenia patients.

In Chapter 4, we attempted to replicate earlier studies that predicted future illness course from a baseline structural brain scan. Being able to predict future illness course could improve treatment and direct necessary care more adequately. Five longitudinal first episode patient samples from three different continents were included. All patients underwent a baseline scan soon after their first psychotic episode and they were followed for several years to obtain information on illness outcome. In this chapter we also explore the possibility to combine data from multiple centers into one model.

In Chapter 5, we explore the possibility to use genotype data to model schizophrenia. We select different varieties of Single Nucleotide Polymorphisms (SNPs) and create machine learning models for individualized prediction of schizophrenia.

Finally, in Chapter 6 we provide a brief summary and the future implications of the above-mentioned studies.

References

American Psychiatric Association, 2013. *Diagnostic and Statistical Manual of Mental Disorders,* Arlington. doi:10.1176/appi.books.9780890425596.744053

Ashburner, J., Friston, K.J., 2001. Why voxel-based morphometry should be used. *Neuroimage* 14, 1238–43. doi:10.1006/nimg.2001.0961

Fornito, A., Yücel, M., Patti, J., Wood, S.J., Pantelis, C., 2009. Mapping grey matter reductions in schizophrenia: An anatomical likelihood estimation analysis of voxel-based morphometry studies. *Schizophr. Res.* 108, 104–113. doi:10.1016/j.schres.2008.12.011

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *Neuroimage* 50, 883–92. doi:10.1016/j.neuroimage.2010.01.005

Frey, B.N., Andreazza, A.C., Houenou, J., Jamain, S., Goldstein, B.I., Frye, M. a, Leboyer, M., Berk, M., Malhi, G.S., Lopez-Jaramillo, C., Taylor, V.H., Dodd, S., Frangou, S., Hall, G.B., Fernandes, B.S., Kauer-Sant'Anna, M., Yatham, L.N., Kapczinski, F., Young, L.T., 2013. Biomarkers in bipolar disorder: a positional paper from the International Society for Bipolar Disorders Biomarkers Task Force. *Aust. N. Z. J. Psychiatry* 47, 321–32. doi:10.1177/0004867413478217
Haijma, S., Van Haren, N., Cahn, W., Koolschijn, P.C.M.P., Hulshoff Pol, H.E., Kahn, R.S., 2013. Brain volumes in schizophrenia: A meta-analysis in over 18 000 subjects. *Schizophr. Bull.* 39, 1129–1138. doi:10.1093/schbul/sbs118

Haijma, S. V, Van Haren, N., Cahn, W., Koolschijn, P.C.M.P., Hulshoff Pol, H.E., Kahn, R.S., 2013. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. Schizophr. Bull. 39, 1129–38. doi:10.1093/schbul/sbs118

Harrison, G., Hopper, K., Craig, T., Laska, E., Siegel, C., Wanderling, J., Dube, K.C., Ganev, K., Giel, R., an der Heiden, W., Holmberg, S.K., Janca, a, Lee, P.W., León, C. a, Malhotra, S., Marsella, a J., Nakane, Y., Sartorius, N., Shen, Y., Skoda, C., Thara, R., Tsirkin, S.J., Varma, V.K., Walsh, D., Wiersma, D., 2001. Recovery from psychotic illness: a 15- and 25-year international follow-up study. *Br. J. Psychiatry* 178, 506–17.

Harrison, P.J., Weinberger, D.R., 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol. Psychiatry* 10, 40–68; image 5. doi:10.1038/sj.mp.4001686
Honea, R., Sc, B., Crow, T.J., Ph, D., Passingham, D., Mackay, C.E., 2000. Reviews and Overviews Regional Deficits in Brain Volume in Schizophrenia : A Meta-Analysis of Voxel-Based Morphometry Studies i, 2233–2245.

Honea, R., Sc, B., Crow, T.J., Ph, D., Passingham, D., Mackay, C.E., 2000. Reviews and Overviews Regional Deficits in Brain Volume in Schizophrenia : A Meta-Analysis of Voxel-Based Morphometry Studies, 2233–2245.

Kempton, M.J., Geddes, J.R., Ettinger, U., Williams, S.C.R., Grasby, P.M., 2008. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Arch. Gen. Psychiatry* 65, 1017–1032. doi:10.1001/archpsyc.65.9.1017

Levitt, J., Bobrow, L., Lucia, D., Srinivasan, P., 2010. A selective review of volumetric and morphometric imaging in schizophrenia. *Curr Top Behav Neurosci.* doi:10.1007/7854

Merikangas, K.R., Jin, R., He, J.-P., Kessler, R.C., Lee, S., Sampson, N.A., Viana, M.C., Andrade, L.H., Hu, C., Karam, E.G., Ladea, M., Medina-Mora, M.E., Ono, Y., Posada-Villa, J., Sagar, R., Wells, J.E., Zarkov, Z., 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Arch. Gen. Psychiatry* 68, 241–251. doi:10.1001/archgenpsychiatry.2011.12

Nieuwenhuis, M., van Haren, N.E.M., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 61, 606–12. doi:10.1016/j.neuroimage.2012.03.079

Olde Loohuis, L., Vorstman, J. a. S., Ori, A.P., Staats, K. a., Wang, T., Richards, A.L., Leonenko, G., Walters, J.T., DeYoung, J., Kahn, R.S., Linszen, D., Os, J. Van, Wiersma, D., Bruggeman, R., Cahn, W., Haan, L. De, Krabbendam, L., Myin-Germeys, I., Cantor, R.M., Ophoff, R. a., 2015. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nat. Commun.* 6, 7501. doi:10.1038/ncomms8501

PGC 2015 www.med.unc.edu/pgc

Rosa, P.G.P., Schaufelberger, M.S., Uchida, R.R., Duran, F.L.S., Lappin, J.M., Menezes, P.R., Scazufca, M., McGuire, P.K., Murray, R.M., Busatto, G.F., 2010. Lateral ventricle differences between first-episode schizophrenia and first-episode psychotic bipolar disorder: A population-based morphometric MRI study. *World J. Biol. Psychiatry* 11, 873–87. doi:10.3109/15622975.2010.486042

Schaufelberger, M.S., Duran, F.L.S., Lappin, J.M., Scazufca, M., Amaro, E., Leite, C.C., de Castro, C.C., Murray, R.M., McGuire, P.K., Menezes, P.R., Busatto, G.F., 2007. Grey matter abnormalities in Brazilians with first-episode psychosis. *Br. J. Psychiatry. Suppl.* 51, s117–s122. doi:10.1192/bjp.191.51.s117

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–7. doi:10.1038/nature13595

Sham, P.C., MacLean, C.J., Kendler, K.S., 1994. A typological model of schizophrenia based on age at onset, sex and familial morbidity. *Acta Psychiatr. Scand.* 89, 135–141. doi:10.1111/j.1600-0447.1994.tb01501.x

Shepherd, A.M., Laurens, K.R., Matheson, S.L., Carr, V.J., Green, M.J., 2012. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neurosci. Biobehav. Rev.* 36, 1342–56. doi:10.1016/j.neubiorev.2011.12.015

Simeone, J.C., Ward, A.J., Rotella, P., Collins, J., Windisch, R., 2015. An evaluation of variation in published estimates of schizophrenia prevalence from 1990–2013: a systematic literature review. *BMC Psychiatry* 15, 193. doi:10.1186/s12888-015-0578-7

Soares, J.C., Kochunov, P., Monkul, E.S., Nicoletti, M.A., Brambilla, P., Sassi, R.B., Mallinger, A.G., Frank, E., Kupfer, D.J., Lancaster, J., Fox, P., 2005. Structural brain changes in bipolar disorder using deformation field morphometry., *Neuroreport.* doi:10.1097/00001756-200504250-00004

Steen, R.G., Mull, C., McClure, R., Hamer, R.M., Lieberman, J.A., 2006. Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. *Br. J. Psychiatry* 188, 510–518. doi:10.1192/bjp.188.6.510

WHO, 2010. International Statistical Classification of Diseases and Related Health Problems (International Classification of Diseases)(ICD) 10th Revision - Version:2010, Occupational Health.

WHO, 2008. The Global Burden of Disease: 2004 doi:10.1038/npp.2011.85

Chapter 2

Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples

Mireille Nieuwenhuis | Neeltje E.M. van Haren |

Hilleke E. Hulshoff Pol | Wiepke Cahn | René S. Kahn | Hugo G. Schnack

NeuroImage 61 (2012) 606–612

Abstract

The purpose of this study is to create a model that can classify schizophrenia patients and healthy controls based on whole brain gray matter densities (voxel-based morphometry, VBM) from structural magnetic resonance imaging (MRI) scans. In addition, we investigated the stability of the accuracy of the models, when built with different sample sizes.

Using a support vector machine, we built a model from 239 subjects (128 patients and 111 healthy controls) and classified 71.4% correct (leave-one-out). We replicated and validated this result by testing the unaltered model on a completely independent sample of 277 subjects (155 patients and 122 healthy controls), scanned with a different scanner. The classification rate of the validation sample was 70.4%. The model's discriminative pattern showed, amongst other differences, gray matter density decreases in frontal and superior temporal lobes and hippocampus in schizophrenia patients with respect to healthy controls and increases in gray matter density in basal ganglia and left occipital lobe and. Larger training samples gave more reliable models: Models based on sample sizes smaller than N=130 should be considered unstable and can even score below chance.

Introduction

Currently, the diagnosis of schizophrenia is based purely on clinical manifestations. The availability of a more objective measure could help psychiatrists in the process of diagnosis and increase its reliability. In addition, an objective measure would serve as a basis for diagnosis at an earlier stage, which in turn could lead to better treatment. Throughout the years, magnetic resonance imaging (MRI) has proven to be an effective technique to detect structural brain abnormalities in schizophrenia patients (Honea et al., 2005; Olabi et al., 2011; Wright et al., 2000). These observations are usually based on statistical analyses, comparing groups of patients to groups of healthy controls. Unfortunately, statistical group differences do not imply the possibility to discover deviations from normal in single individuals and therefore do not suffice to aid in diagnosis.

A considerable amount of work has been done to establish possible detectable patterns in the brain that distinguish between individual schizophrenia patients and healthy controls. Usually, the underlying methodology is machine learning classification by means of pattern recognition. These discriminating patterns are generated by means of input features; in structural MRI most common features are so-called brain tissue densities (obtained from voxel based morphometry). Frequently used methods to create these classification models are: support vectors machine (SVM) (Davatzikos et al., 2005; Fan et al., 2005, 2007, 2008; Ingalhalikar et al., 2010; Koutsouleris et al., 2009; Pohl and Sabuncu, 2009); Discriminant Function Analysis (Karageorgiou et al., 2011; Kasparek et al., 2011; Leonard et al., 1999; Liu et al., 2004; Nakamura et al., 2004; Takayanagi et al., 2010, 2011); and some other methods (Caprihan et al., 2008; Kawasaki et al., 2007; Sun et al., 2009). Although considerable accuracies have been achieved ranging from 70.5% to 91.8%, these were often obtained from relatively small data sets and without testing the model in validation samples. To our knowledge there is only one study that used a separate, though small, cohort (16 patients and 16 controls) to validate their initial results (Kawasaki et al., 2007).

Most classification studies included around 30 subjects per class with sample sizes ranging from 10 to 69 patients. Since subjects have to be divided into a subset from which the model is built and a set, which is subsequently used to test the model's predictive value, samples of this size may be too small for robust model building and testing. Moreover, in prior studies the predictive capacity of models was not based on using a separate validation sample, but on using a cross validation method, such as leave one out, providing an estimation of the percentage correctly classified subjects using virtually all data to create the models. A more robust method is using a completely independent sample to validate the model. Therefore, we used an independent sample to validate the results we found with our discovery sample. The goal of this study is twofold: 1) Test whether a large sample is necessary to build a stable classification model; and 2) Investigate whether the classification results obtained with such a model can be validated, using the same model, in an independent sample. As input features we started with all gray matter densities in the brain, which enables us to compare the model's classification patterns to brain abnormalities found in group-level statistical analyses of schizophrenia brain images. Next to this full model, we used two forms of feature reduction. First, we excluded the striatum, since this structure is known to be affected by (typical) antipsychotic medication (Smieskova et al., 2009), and we wish to separate patients from controls, rather than medication-users from non-users. We further reduced the number of features by ranking them and keeping only the 10% features that had the most influence on the model. In doing so, we reduced the risk of overfitting the model to our training set and thus made it potentially more general.

Materials and methods

Subjects

In both samples, the presence or absence of psychopathological abnormality was established using the Comprehensive Assessment of Symptoms and History (Andreasen et al., 1992) and Schedule for Affective Disorders and Schizophrenia Lifetime Version (Endicott and Spitzer, 1978) assessed by at least one independent rater who was trained to assess this interview. All healthy comparison subjects met Research Diagnostic Criteria (Spitzer et al., 1978) of "never [being] mentally ill." All patients met DSM-IV criteria for a nonaffective psychotic disorder, diagnosis of schizophrenia, schizophreniform disorder or schizoaffective disorder. Written informed consent was obtained from all subjects. Subjects were matched for age, sex and socioeconomic status of their parents which is expressed as the highest completed level of education by one of their parents.

Discovery sample

The discovery sample was selected from a sample that has been described before (Hulshoff Pol et al., 2001). For the current study, subjects older than 50 years of age were excluded, to match the validation sample's age range. Furthermore, an upgraded version of our image processing pipeline was used (Brouwer et al., 2010), which discarded eight scans as too noisy for reliable segmentation. The sample included 128 patients (93 males and 35 females) and 111 matched healthy controls (79 males and 32 females). All patients had received antipsychotic medication in the past and all but four patientsreceived antipsychotic medications at the time of the MRI scan. Medication included typical (49% of the patients received haloperidol) and atypical (46% of the patients received clozapine, risperidone, olanzapine, or sertindole) antipsychotic agents.

Validation sample

An independent validation sample was used to test the model. The sample consisted of 155 patients (125 males and 30 females). In addition, 122 matched healthy controls (61 males and 61 females) were included. The age range was between 17 and 50 years. The sample is part of an ongoing longitudinal study in the Netherlands (Genetic Risk and Outcome of Psychosis; GROUP) and has been described before (Boos et al., 2011). The majority of patients (81%) were taking atypical antipsychotic medication at the time of scan, with olanzapine and risperidone being most often prescribed, 8% of the patients were on typical antipsychotic medication. 10% of the patients did not use medication at the time of the scan.

All scans were acquired on a 1.5 T Philips scanner (discovery sample: NT; validation sample: Achieva) using the identical acquisition protocol. Three-dimensional T1-weighted, fast field echo scans with 160 to 180 contiguous coronal slices (echo time [TE], 4.6 ms; repetition time, 30 ms; flip angle, 30°; field of view, 256 mm; 1×1×1.2 mm^3 voxels) were made of all subjects. The samples were acquired 10 years apart from one another (discovery sample: 1995–1998; validation sample: 2005–2007). More information on the samples can be found in **Table 1**.

Image processing

All scans were processed on the computer network of the Department of Psychiatry at the University Medical Center Utrecht. The features we used were extracted from the processed T1-weighted images. The images were transformed into Talairach orientation (no scaling), after which they were corrected for scanner RF-field nonuniformity with the N3 algorithm (Sled et al., 1998). Using a partial volume segmentation technique (Brouwer et al., 2010) the brain was segmented into gray matter, white matter and cerebrospinal fluid. To compare voxels between subjects we used voxel-based morphometry (VBM) (Ashburner and Friston, 2000). The gray matter segments were blurred using a three-dimensional Gaussian kernel (full-width half-maximum (FWHM)=8 mm). The voxel values of these blurred segments reflect the local presence, or concentration, of gray matter and will be referred to as gray matter 'densities' (GMDs). In order to compare GMDs at the same anatomical location between all subjects, the GMD images were transformed into a standardized coordinate system using a two step process. First, the T1-weighted images were linearly transformed to a model brain (Hulshoff Pol et al., 2001). In this linear step a joint entropy mutual information metric was optimized (Maes et al., 1997).

Table 1 | Characteristics of the subjects for both the discovery and the validation sample.

Sample:	**Discovery: N = 239**	**Validation: N = 277**
Subjects: patients/healthy controls	128/111	155/122
Age in years: mean (SD)	30.87 (9.52)	27.18 (6.87)
Sex: male/female	172/67	186/91
Handedness: right/left	203/36	251/26
PANNS-positive symptoms score mean (SD) [range]	16.85 (5.62) [7–30][1]	15.34 (5.7) [7–35][2]
PANNS-negative symptoms score mean/(SD) [range]	18.60 (5.49) [9–32][1]	15.41 (5.5) [6–31][2]
PANNS-total symptoms score mean/(SD) [range]	71.56 (17.01) [40–117][3]	62.22 (17.17) [30–133][2]
Illness duration at scan time in years mean/(SD) [range]	10.3 (5.0) [0–36][5] [7]	5.0 (4.0) [0–16][6] [7]
Patients on typical medication	46%[4]	81%[5]
Patients on typical medication	49%[4]	8%[5]
Patients not medicated	0.7%[4]	10%[5]
Scanner type	1.5 T Philips NT	1.5 T Achieva

[1] Information is missing in 18 patients; [2] Information is missing in eight patients; [3] Information is missing in 23 patients; [4] Information is missing in five patients; [5] Information is missing in one patient; [6] Information is missing in 16 patient; [7] p<0.0001.

In the second step nonlinear (elastic) transformations were calculated to register the linearly transformed images to the model brain up to a scale of 4 mm (FWHM), thus removing global shape differences between the brains, but retaining local differences (ANIMAL; Collins et al., 1995). The GMD maps were now transformed to the model space by applying the concatenated linear and nonlinear transformations. Since the density maps have been blurred to an effective resolution of 8 mm, it is not necessary to keep this information at the 1-mm level. Therefore, the maps were resampled to voxels of size 2 x 2 x 2.4 mm3, i.e., doubling the original voxel sizes. For all voxels, GMD was regressed on age, sex, and handedness.

The resulting b-maps were used to correct the GMD maps for the effects of these factors and calculate GMD residuals:

$$\text{GMD residual}_i = \text{GMD}_i - \sum_r a_{i,r} \times b_r$$

where index r refers to the regressor, for instance age, b_r is the outcome of the regression (b-map) and $a_{i,r}$ is subject i's value for this regressor. These GMD residuals are used as features for the support vector machine model.

Preprocessing validation sample

The scans of the validation sample followed the same preprocessing steps. However, no regression analyses were carried out: the GMDs of the validation sample were corrected for age, sex, and handedness using the b-maps from the discovery sample. In this manner, no information of the validation sample is included in the calculation process and a proper simulation is created of what would happen if a new person would be scanned and tested with this model.

Linear support vector machine

A support vector machine (SVM) is a high-dimensional, pattern recognition, supervised learning algorithm (Vapnik, 1999) used to solve classification problems. In our case this problem consists of separating schizophrenia patients from healthy controls. The SVM classification process includes two phases (**Figure 1**); the first phase is the creation of a model by means of training; and the second phase is model validation. We integrated LIBSVM (Chang, 2011) with our software to carry out the classification.

Subjects are represented by features congregated into a vector x_i per subject. These vectors exist in a high dimensional feature space, in which a flat decision surface is constructed to separate the subjects from different classes. This is accomplished by the introduction of a decision function $y(x_i)$:

$$y(x_i) = w^{\top} \cdot x_i - b$$

that vanishes at the decision surface. The weight vector w is a normal vector to this surface; b is an offset. In the training phase each subject has a label t_i (schizophrenia patients 1; healthy control −1), and the function is optimized by requiring $y(x_i)$ < 0 if t_i = -1, and $y(x_i)$ > 0 if t_i = +1. In the test phase this decision function is used to classify the test subjects according to the sign of $y(x_i)$. The weight-vector not only contains information on feature importance, but also on whether the particular feature shows an increase or decrease in the pattern comparing patients to healthy controls.

There can be several surfaces that exactly separate the classes. The SVM chooses the so called optimal separating hyperplane (OSH) so that the space between the two classes, which is called the margin, is made as large as possible. The size of the margin is 2/$\| w \|$so minimizing $\| w \|$ will maximize the margin. Because the problem is not per se linearly separable, an error measure ξ is brought into the equation. If the subject is classified correctly ξ = 0; otherwise it is the distance from the OSH to the subject.

There is a free parameter in linear SVM to influence the narrowness of the margin, a penalty C is multiplied by the error per subject. The OSH is now dependent on boththe margin and the error. On the one hand the margin is maximized and on the other hand the error times C is minimized leading to a minimization of:

$$C\sum_{n=1}^{N}\xi_n + \frac{2}{\| w \|}$$

This means that if C is larger, then the penalty of misclassified subjects will be higher

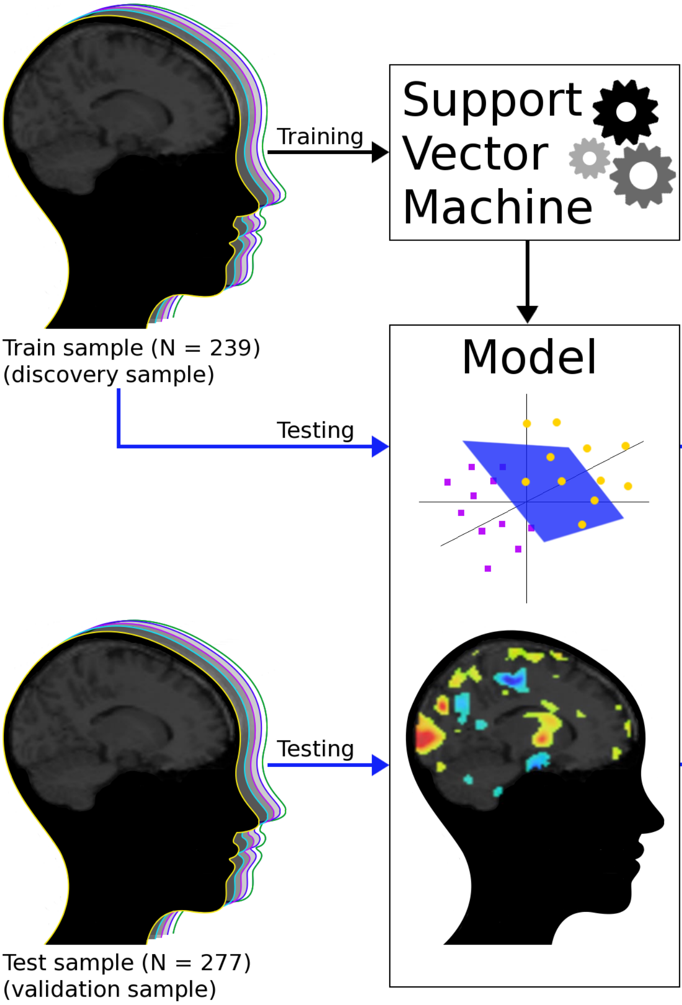

Figure 1 | Classification train and test procedure.

In the top left corner the sample is depicted that is used to train the support vector machine, to create the model. An abstract 3-dimensional example of such a model is shown in the center. The optimal separating hyperplane (OSH) is shown in blue, the pink squares represent one class and the yellow circles the other class. Below this abstract model, the model is visualized on the brain (weight vector w). In the bottom left corner the validation sample is depicted; this sample is used to test the model created with the discovery sample.

and the margin will most likely be smaller. It was shown earlier (Franke et al., 2010) that tuning C can increase the model's performance. We optimized C for our discovery sample (see Appendix A).

Feature selection

A whole brain analysis includes 157,256 features; to reduce influence of noise and runtime significantly, feature reduction can be invoked. First a complete model is built from which a selection of the top 10% ranked features is taken. Feature ranking is based on the absolute values of the elements of the weight-vector, representative of their influence in the model. This selection of features is then used to build a new model.

Another selection method we used is knowledge based. The size of the striatum is known to change if a subject is on medication (Smieskova et al., 2009). To exclude this possible confounding effect, we created a model where the striatum was masked out. The striatum was segmented manually from the model brain image and, using mathematical morphology operations, enlarged, to ensure that all spots possibly affected by medication were excluded for all subjects. For comparison, we created a top 10% model of both the whole brain analysis and the model where the striatum was excluded.

Quality measures

The quality of a model is assessed by three quantities:

Sensitivity = TP / (TP + FP), where TP is the number of true positives (correctly classified patients), and FP is it the number of false positives.
Specificity = TN / (TN + FN), where TN is the number of true negatives, and FN is the number of false negatives.
Average accuracy = (sensitivity + specificity) / 2.

Next to the replication using the independent validation sample, we also tested the accuracy of the model on the discovery sample itself. This is done by leave-one-out (LOO) cross validation. LOO gives an estimate of how well the model will generalize to a new data set. First a model is trained on all subjects but one that is then used to test this model. This is done until all subjects are left out once. In our case the model is trained 239 times.

To test the statistical significance of the accuracy obtained with the validation sample, we randomly permuted the labels of the validation group and applied the modelto these data. We repeated this process 10,000 times to determine a null-distribution of accuracies and calculate the p-value of the accuracy found from our full model.

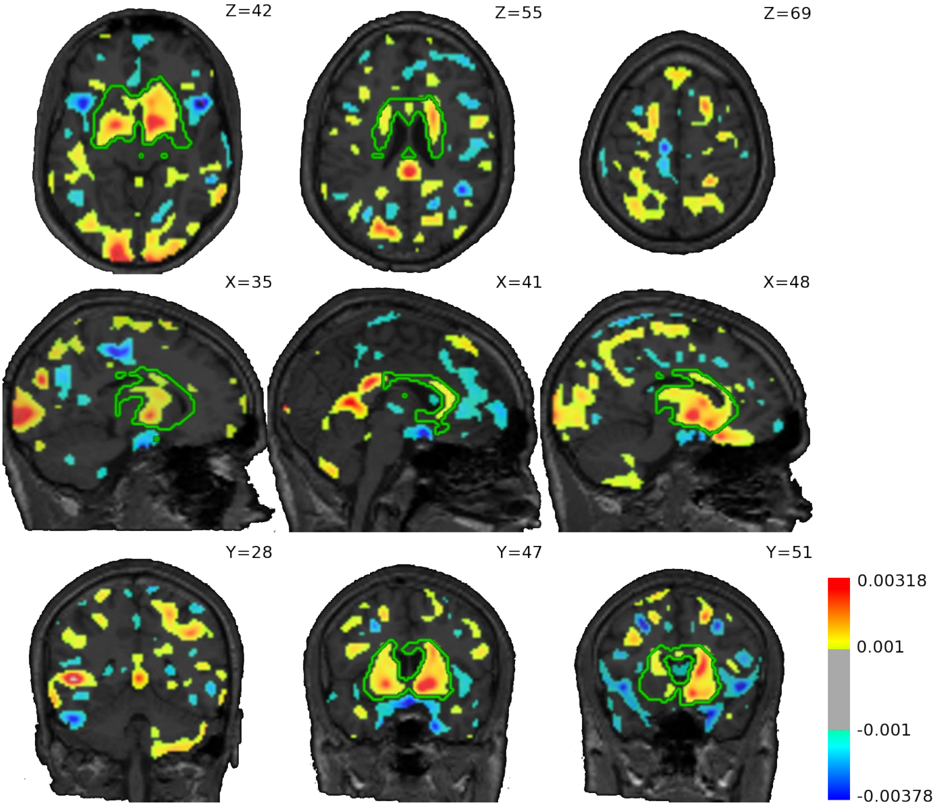

Figure 2 | Weight-vector (w-map) of the full model, i.e. the pattern differentiating between schizophrenia patients and healthy controls in axial (top row), coronal (middle row) and sagittal (bottom row) slices of the brain. The w-map is thresholded at 0.001 and −0.001 to show only the more relevant features. Warm colors indicate increases in gray matter densities in patients compared to controls and cool colors indicate decreases. The green line demarcates the border of the enlarged striatum that was excluded in the feature reduction models.

Stability

A model should be trained on a set of subjects that represent all the variability within their class. This would result in a model of which the accuracy and the separating pattern do not change much for different selections of subjects. To investigate what amount ofsubjects is necessary to have such a representative set, we conducted 100 bootstraps (repetitions) with different selections of subjects. Every bootstrap starts with a training set of ten subjects (five patients and five control subjects) randomly drawn from the discovery sample. A model is built from this set and tested on the validation sample, after which ten subjects are added to the training set for the next train/test step, until the maximum training set size of N=220 is reached. The change in accuracy between two steps is taken as a measure of stability. For every training set size N_t the mean absolute change over all bootstraps is calculated to reflect the average accuracy of a model built from N_t subjects.

Results

Figure 2 shows the weight-vector w mapped onto the brain. Warm colors indicate increases in GM-densities and cool colors indicate decreases in GM-densities when examining patients as compared to controls. Substantial contributions to the full model's discriminative pattern were found for the basal ganglia and left occipital lobe (relatively large GMD in patients) and for the frontal and superior temporal lobes and hippocampus (relatively small GMD in patients). The LOO accuracy reached on the full model was 71.4%. Replication in the validation sample (277 subjects) yielded a classification accuracy of 70.4% ($p < 0.0001$).

Exclusion of the striatum resulted in a model with a very similar weight-vector, with exception of the voxels that were excluded. The model produced a slightly decreased accuracy in the discovery sample (67.5%), compared to the full model, but approximately the same accuracy in the validation sample (70.6%). The sensitivity in the validation sample improved, leading to 74.8% of the patients being correctly classified. The reduction model, containing only the top 10% ranked features, yielded 86.8% correctly classified subjects in the discovery sample. The validation sample's result (69.1%) for this model was in the same range as the validation results for the more extensive models (see **Table 2**). Reduction of the full model, thus including the striatum, by keeping only the top 10% of its features, led to a sensitivity of 92.2% and a specificity of 84.7% in the discovery sample and 72.3% and 72.1%, respectively, in the validation sample.

Figure 3 shows the results of the stability test for the full model. From N=130 subjects onward in the train group, all bootstrap test accuracies were above chance. The mean absolute change in accuracy when ten subjects are added to the model decreases for higher N down to 1.4%. The mean accuracy keeps rising until all 220 subjects are included in the creation of the model and appears not to have reached its maximum height yet. Moreover, the mean absolute change is still diminishing; suggesting that including even more subjects would increase the model's robustness.

Table 2 | The LOO and test set accuracies of the three different three different models: full model; model excluding the striatum; 10% reduction (excluding the striatum).

	Full model	**Model excluding the striatum**	**10% reduction model (excluding the striatum)**
Discovery sample			
Sensitivity (LOO)	73.4%	71.1%	89.8%
Specificity (LOO)	69.4%	64.0%	83.8%
Average accuracy (LOO)	71.4%	67.5%	86.8%
Validation sample			
Sensitivity	67.1%	74.8%	74.2%
Specificity	73.8%	66.4%	63.9%
Average accuracy	70.4%	70.6%	69.1%

Discussion

The purpose of this study was to create a model that classifies schizophrenia patients and healthy controls based on structural MRI scans. First, we used a support vector machine (SVM) to create a model from a large sample of patients and control subjects (discovery sample, N=239), which we then applied to a large independent sample (validation sample, N=277). We demonstrated that it is possible to attain approximately the same classification accuracy (70.4%) in a completely independent set of subjects, as the accuracy achieved in the sample from which the model was built (71.4%). In view of these results it is likely that any new individual, being healthy or schizophrenia patient, will be classified equally well, provided the individual is scanned with a 1.5 T scanner, using a comparable acquisition protocol and processing steps.

One other study validated its classification model in an independent sample (Kawasaki et al., 2007). The discovery sample's (N=60) accuracy was 75%, while their validation sample (N=32) led to 80% correctly classified subjects. This unexpected increase of accuracy in the validation sample may be attributable to the small sample size: To indicate the reliability of this estimate we calculated the 95%- confidence interval. The percentages in samples with size N=32 drawn from a population in which 75% is ‘correct’ will vary from 60% to 90%.

Apart from influencing the accuracy of a replication study, sample size also determines the reliability of the classification model itself. Our experiments showed that an SVM requires a large dataset (at least larger than about 130 subjects) for building a stable model that can differentiate between schizophrenia patients and healthy controls.

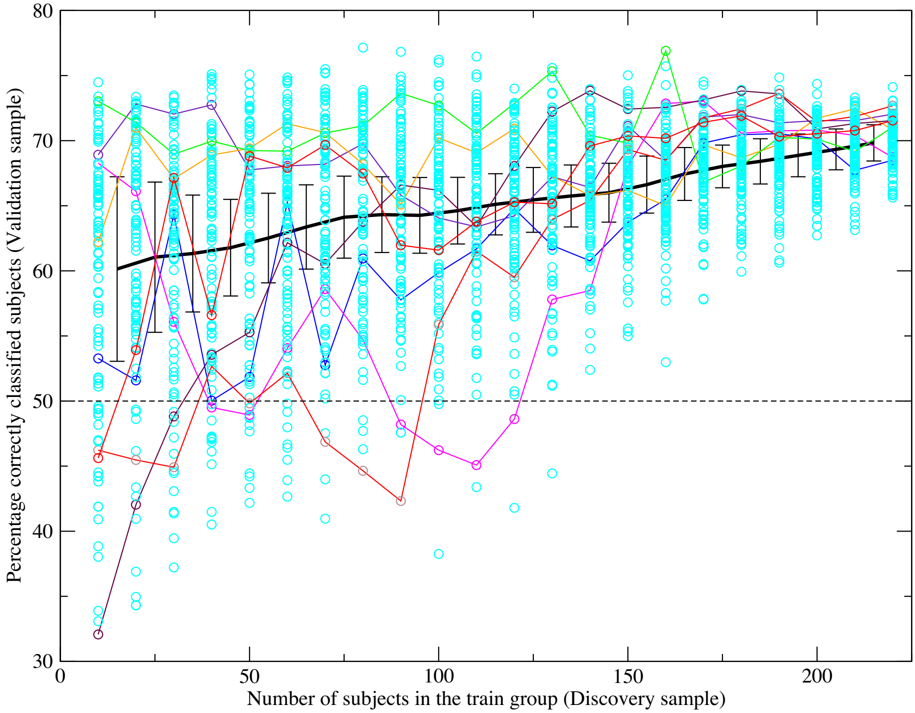

Figure 3 | Stability test results, demonstrating the relationship between accuracy obtained in the validation sample and the size of the training sample (subsets of the discovery sample) of the full model. The colored lines and circles show the trajectories of eight complete bootstraps starting with ten subjects and increasing by steps of ten subjects up to 220 subjects. The light blue circles represent the results for all 100 bootstraps at all sample sizes. The black line shows the average accuracy for each sample size and the mean absolute change in accuracy when ten subjects are added to the model (error bars). The dashed line indicates the 50% (chance) line.

Models based on smaller samples led to large fluctuations in classification accuracies and sometimes resulted in accuracies lower than chance level. Even with 140 subjects the accuracies fluctuated between 52% and 74%, indicating that these models still depend on the selection of subjects. An additional advantage of a larger training sample is, of course, that it leads to higher classification accuracy.

The 87% classification accuracy (LOO) in the discovery sample, which we obtained after reduction of the amount of features to 10%, is comparable to previous classification studies (Caprihan et al., 2008; Davatzikos et al., 2005; Fan et al., 2005, 2007, 2008;Koutsouleris et al., 2009; Leonard et al., 1999; Liu et al., 2004). Moreover, the discriminative patterns of the model built on all gray matter densities of the discovery sample appear to be consistent with the reported structural brain abnormalities in schizophrenia patients, e.g., decreases in frontal and superior temporal gray matter volumes (Honea et al., 2005; Hulshoff Pol et al., 2001; Wright et al., 2000). Moreover, in contrast to most statistical analyses, the SVM detects interactions between the voxels, thus providing a pattern indicating that patients have density decreases in certain areas and simultaneous increases in other areas as compared to healthy controls. To rule out the possibility that the SVM patterns rely heavily on medication effects in the brain, resulting in separation of medicated from non-medicated subjects rather than patients from controls, we masked out the striatum, a structure known to be affected by (typical) antipsychotic medication. When included, the striatum contributed to the separation of controls from patients on typical medication, due to gray matter increases in the latter group, at the cost of correctly classifying patients on atypical medication. While the classification accuracy in discovery sample dropped by 4%, it did not change in the validation sample. Apparently the full model had partly been based on medication effects, but a 'pure disease' model without these effects turned out to classify the new subjects equally well. We can however not exclude an effect of medication entirely. It has been argued that antipsychotic medication also affects whole brain gray matter density (Ho et al., 2011). Since the reported effects of medication on cortical gray matter density are inconsistent (Shepherd et al., 2012), it is not clear which brain regions should be left out of the model. Since typical, but not atypical, antipsychotics have been reported to lead to increased gray matter volume of the basal ganglia (Smieskova et al., 2009), the different reactions of discovery sample and validation sample to removing this structure from the model can be explained by differences in medication use between the two sets. Due to change in clinical practice throughout the years, in our discovery sample (used to train the model) half of the patients were on typical medication while in the validation sample this was only 8%. The general application of a model is thus increased by the removal of medication-sensitive structures from it. Another difference between the two samples is the inclusion of more chronic patients in the discovery sample than in the validation sample. It may be possible that inclusion of chronic patients, with more marked brain changes, enabled the model to classify younger patients with less severe brain alterations.

Although the present study has shown that we can classify schizophrenia patients and controls with 71% accuracy through pattern recognition, this does not mean that we can classify patients with other psychiatric disorders as being ill, and, more importantly, if classified as ill, that we can separate them from those with schizophrenia. To provetheir clinical utility it will be inevitable to create classification models on multiple psychiatric disorders. (Dis) similarities between the brain pattern found for schizophrenia and brain patterns of disorders such as bipolar disorder, depression or borderline personality disorder are unknown. However, recent findings of genetic factors influencing brain structure that are unique for schizophrenia and bipolar disorder suggest that it may be possible to find those patterns (Hulshoff Pol et al., 2012). Next to classification of other disorders, prediction of disease outcome is of clinical interest (Mourao-Miranda et al., 2011), as well as early detection of diseases. All of these goals require an output from the models that is more refined than the simple binary yes/no presented here. For these models we need large data sets, probably only reachable in a multicenter setting. This latter approach will also be of use for further tests of generalizability. While recent work demonstrated the possibility to combine multicenter VBM data for statistical analyses (Schnack et al., 2010), this replication study using different scanners is a first promising step in the direction of cross-site classification of individuals.

In conclusion, we have shown that a large set (N>130) of structural MRI images is required to build a classification model that is able to separate new, unrelated, subjects into healthy controls and schizophrenia patients. The current model reached 71% accuracy. Further investigations will determine whether this result can be improved towards prediction models that are clinically useful.

Appendix A

As explained in the Materials and methods section, the OSH is dependent on two terms; the margin is maximized while the error times C is minimized leading to a minimization of:

$$C\sum_{n=1}^{N}\xi_n + \frac{2}{\| w \|}$$

C is the penalty parameter that controls the tradeoff between training errors and the narrowness of the margin. Increasing its value narrows the margin and forces better classification of the subjects in the training set. The goal was to identify the optimal C that would create a model that could predict classes as accurately as possible from VBM data. To find this value of C a parameter search was carried out. Starting from C = 0.000001 and multiplying it by 2 for each next step, C was raised to 16.78 in 25 steps. For each value of C a model was created from the training data and tested on an independent validation set, yielding a predicted accuracy for this C.

From the complete set of 294 subjects, we randomly selected a training set (N=210) and a validation set (N=50). We repeated this procedure one hundred times giving usaverage prediction accuracies as a function of C. Since the extreme values of C always produce suboptimal accuracies, there must be a C-value between the extremes that produces a model with the highest average accuracy. This value is taken as the optimal C-value for the current amount of features (about 160,000) and a large amount of subjects (N=210) in the model.

To investigate the dependency of the optimal C on the amount of features in the model, different numbers of features were selected by applying checkerboard-like masks to the images. Masks with increasing spaces between the selected voxels resulted in selections of about 80,000; 40,000; 20,000; 6,000; 2500; 1300; 700; and 450 features. For each amount of features F_i, the parameter search was carried out, yielding an optimal C (F_i). The (F_i, C(F_i)) data appeared to obey an inverse relationship: A continuous C(F) function was therefore obtained by a linear fit to (1/F_i, C(F_i)).

Apart from the empirically determined dependency on the amount of features, the optimal C depends on the number of subjects. Since C is multiplied by the summation of errors of all N subjects, C~1/N seems to be a reasonable scaling. This resulted in:

$$C = 131,48 \times 210/NF = 27610,885/NF$$

We adopt this optimal C formula for all our models using VBM features from schizophrenia patients and healthy controls.

References

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. *Arch. Gen. Psychiatry* 49, 615–623.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821.

Boos, H.B., Cahn, W., van Haren, N.E., Derks, E.M., Brouwer, R.M., Schnack, H.G., Hulshoff Pol, H.E., Kahn, R.S., 2011. Focal and global brain measurements in siblings of patients with schizophrenia. *Schizophr. Bull.*(Electronic publication ahead of print).

Brouwer, R.M., Hulshoff Pol, H.E., Schnack, H.G., 2010. Segmentation of MRI brain scans using non-uniform partial volume densities. *Neuroimage* 49, 467–477.

Caprihan, A., Pearlson, G.D., Calhoun, V.D., 2008. Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements. *Neuroimage* 42, 675–682.

Chang, C.-C., 2011. A library for support vector machines, In: Lin, C.-J. (Ed.), ACM Transactions on Intelligent Systems and Technology, 2nd ed. pp. 27:1–27:27. Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-d model-based neuroanatomical segmentation. *Hum. Brain Mapp*. 3, 190–208.

Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-d model-based neuroanatomical segmentation. *Hum. Brain Mapp.* 3, 190–208.

Davatzikos, C., Shen, D., Gur, R.C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B.I., Gur, R.E., 2005. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch. Gen. Psychiatry* 62, 1218–1227.

Endicott, J., Spitzer, R.L., 1978. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch. Gen. Psychiatry* 35, 837–844.

Fan, Y., Shen, D., Davatzikos, C., 2005. Classification of structural images via highdimensional image warping, robust feature extraction, and SVM. *Med. Image Comput. Comput. Assist. Interv.* 8, 1–8.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* 26, 93–105.

Fan, Y., Gur, R.E., Gur, R.C., Wu, X., Shen, D., Calkins, M.E., Davatzikos, C., 2008. Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study. *Biol. Psychiatry* 63, 118–124.

Franke, K., Ziegler, G., Kloppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *Neuroimage* 50, 883–892.

Ho, B.C., Andreasen, N.C., Ziebell, S., Pierson, R., Magnotta, V., 2011. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Arch. Gen. Psychiatry* 68, 128–137.

Honea, R., Crow, T.J., Passingham, D., Mackay, C.E., 2005. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *Am. J. Psychiatry* 162, 2233–2245.

Hulshoff Pol, H.E., Schnack, H.G., Mandl, R.C., van Haren, N.E., Koning, H., Collins, D.L., Evans, A.C., Kahn, R.S., 2001. Focal gray matter density changes in schizophrenia. *Arch. Gen. Psychiatry* 58, 1118–1125.

Hulshoff Pol, H.E., van Baal, G.C., Schnack, H.G., Brans, R.G., van der Schot, A.C., Brouwer, R.M., van Haren, N.E., Lepage, C., Collins, D.L., Evans, A.C., Boomsma, D.I., Nolen, W., Kahn, R.S., 2012. Overlapping and segregating structural brain abnormalities in twins with schizophrenia or bipolar disorder. *Arch. Gen. Psychiatry* 69 (4), 349–359.

Ingalhalikar, M., Kanterakis, S., Gur, R., Roberts, T.P., Verma, R., 2010. DTI based diagnostic prediction of a disease via pattern classification. *Med. Image Comput. Comput. Assist. Interv.* 13, 558–565.

Karageorgiou, E., Schulz, S.C.,Gollub, R.L., Andreasen,N.C.,Ho, B.C., Lauriello, J., Calhoun, V.D., Bockholt, H.J., Sponheim, S.R., Georgopoulos, A.P., 2011. Neuropsychological Testing and Structural Magnetic Resonance Imaging as Diagnostic Biomarkers Early in the Course of Schizophrenia and Related Psychoses. *Neuroinformatics* 9 (4), 321–333.

Kasparek, T., Thomaz, C.E., Sato, J.R., Schwarz, D., Janousova, E., Marecek, R., Prikryl, R., Vanicek, J., Fujita, A., Ceskova, E., 2011. Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Res.* 191, 174–181.

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

Koutsouleris, N., Meisenzahl, E.M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Moller, H.J., Gaser, C., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch. Gen. Psychiatry* 66, 700–712.

Leonard, C.M., Kuldau, J.M., Breier, J.I., Zuffante, P.A., Gautier, E.R., Heron, D.C., Lavery, E.M., Packing, J., Williams, S.A., DeBose, C.A., 1999. Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. *Biol. Psychiatry* 46, 374–382.

Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C.S., Stenger, V.A., Davis, S., Aizenstein, H., Becker, J.T., Lopez, O.L., Meltzer, C.C., 2004. Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. *Lect. Notes Comput. Sci.* 3216, 393–401.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* 16, 187–198.

Mourao-Miranda, J., Reinders, A.A., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., Morgan, K.D., Fearon, P., Jones, P.B., Doody, G.A., Murray, R.M., Kapur, S., Dazzan, P., 2011. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychol. Med.* 1–11.

Nakamura, K., Kawasaki, Y., Suzuki, M., Hagino, H., Kurokawa, K., Takahashi, T., Niu, L., Matsui, M., Seto, H., Kurachi, M., 2004. Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. *Schizophr. Bull.* 30, 393–404.

Olabi, B., Ellison-Wright, I., McIntosh, A.M., Wood, S.J., Bullmore, E., Lawrie, S.M., 2011. Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biol. Psychiatry* 70, 88–96.

Pohl, K.M., Sabuncu, M.R., 2009. A unified framework for MR based disease classification. *Inf. Process. Med. Imaging* 21, 300–313.

Schnack, H.G., van Haren, N.E., Brouwer, R.M., van Baal, G.C., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T.D., Huttunen, M., Lepage, C., Collins, D.L., Evans, A., Murray, R.M., Kahn, R.S., Hulshoff Pol, H.E., 2010. Mapping reliability in multicenter MRI: voxelbased morphometry and cortical thickness. *Hum. Brain Mapp.* 31, 1967–1982.

Shepherd, A.M., Laurens, K.R., Matheson, S.L., Carr, J.V., Green, M.J., 2012. Systematic Meta-review and Quality Assessment of the Structural Brain Alterations in Schizophrenia. *Neuroscience and Biobehavioral Reviews*, 36(4), 1342–56.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans.Med. Imaging* 17, 87–97.

Smieskova, R., Fusar-Poli, P., Allen, P., Bendfeldt, K., Stieglitz, R.D., Drewe, J., Radue, E.W., McGuire, P.K., Riecher-Rossler, A., Borgwardt, S.J., 2009. The effects of antipsychotics on the brain: what have we learnt from structural imaging of schizophrenia?—a systematic review. *Curr. Pharm. Des.* 15, 2535–2549.

Spitzer, R.L., Endicott, J., Robins, E., 1978. Research diagnostic criteria: rationale and reliability. *Arch. Gen. Psychiatry* 35, 773–782.

Sun, D., van Erp, T.G., Thompson, P.M., Bearden, C.E., Daley, M., Kushan, L., Hardt, M.E., Nuechterlein, K.H., Toga, A.W., Cannon, T.D., 2009. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biol. Psychiatry* 66, 1055–1060.

Takayanagi, Y., Kawasaki, Y., Nakamura, K., Takahashi, T., Orikabe, L., Toyoda, E., Mozue, Y., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., Matsushita, M., Suzuki, M., 2010. Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 34, 10–17.

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., Suzuki, M., 2011. Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. *PLoS One* 6, e21047.

Vapnik, V.N., 1999. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* 10, 988–999.

Wright, I.C., Rabe-Hesketh, S.,Woodruff, P.W.,David, A.S.,Murray, R.M., Bullmore, E.T., 2000. Meta-analysis of regional brain volumes in schizophrenia. *Am. J. Psychiatry* 157, 16–25.

Chapter 3

Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolardisorder and healthy subjects

Mireille Nieuwenhuis | Hugo G. Schnack | Neeltje E.M. van Haren | Lucija Abramovic | Thomas W. Scheewe |Rachel M. Brouwer | Hilleke E. Hulshoff Pol | René S. Kahn

NeuroImage 84 (2014) 299–306

Abstract

Although structural magnetic resonance imaging (MRI) has revealed partly non-overlapping brain abnormalities in schizophrenia and bipolar disorder, it is unknown whether structural MRI scans can be used to separate individuals with schizophrenia from those with bipolar disorder. An algorithm capable of discriminating between these two disorders could become a diagnostic aid for psychiatrists.

Here, we scanned 66 schizophrenia patients, 66 patientswith bipolar disorder and 66 healthy subjects on a 1.5 TMRI scanner. Three support vector machines were trained to separate patients with schizophrenia from healthy subjects, patients with schizophrenia from those with bipolar disorder, and patients with bipolar disorder from healthy subjects, respectively, based on their gray matter density images. The predictive power of the models was tested using cross-validation and in an independent validation set of 46 schizophrenia patients, 47 patients with bipolar disorder and 43 healthy subjects scanned on a 3 TMRI scanner.

Schizophrenia patients could be separated from healthy subjects with an average accuracy of 90%. Additionally, schizophrenia patients and patients with bipolar disorder could be distinguished with an average accuracy of 88%. The model delineating bipolar patients from healthy subjects was less accurate, correctly classifying 67% of the healthy subjects and only 53% of the patients with bipolar disorder. In the latter group, lithium and antipsychotics use had no influence on the classification results. Application of the 1.5 T models on the 3 T validation set yielded average classification accuracies of 76% (healthy vs schizophrenia), 66% (bipolar vs schizophrenia) and 61% (healthy vs bipolar).

In conclusion, the accurate separation of schizophrenia from bipolar patients on the basis of structural MRI scans, as demonstrated here, could be of added value in the differential diagnosis of these two disorders. The results also suggest that gray matter pathology in schizophrenia and bipolar disorder differs to such an extent that they can be reliably differentiated using machine learning paradigms.

Introduction

Currently, the diagnosis of psychiatric disorders such as schizophrenia and bipolar disorder is based predominantly on their clinical manifestations. While psychiatrists can establish the presence of illness (as distinct from its absence) with relative ease, discrimination between several possible diagnoses is far more complicated, especially in the early phase of schizophrenia and bipolar disorder. The availability of additional (objective) measures would assist psychiatrists in the process of diagnosis, with obvious benefits to efficiency of treatment and improved outcome. Magnetic resonance imaging (MRI) has proven to be an effective technique to detect structural brain abnormalities at group-level in schizophrenia patients (meta-analyses: Haijma et al., 2012; Olabi et al., 2011) and those with bipolar disorder (meta-analyses: Kempton et al., 2008; McDonald et al., 2004). Unfortunately, statistical group differences do not translate to discovering deviations from normal on an individual basis and therefore are not sufficient as a diagnostic aid.

Using machine learning techniques, promising results have been obtained for the classification of schizophrenia patients and healthy subjects based on MRI scans. Pioneering work was done by Davatzikos et al. (2005), followed by numerous other investigations. The support vector machine (SVM; Fan et al., 2008; Ingalhalikar et al., 2010; Koutsouleris et al., 2009; Pohl and Sabuncu, 2009; Vapnik, 1999) and the Discriminant Function Analysis (Karageorgiou et al., 2011; Kasparek et al., 2011; Leonard et al., 1999; Liu et al., 2004; Nakamura et al., 2004; Takayanagi et al., 2011) are the most frequently used methods (for an overview of schizophrenia classification studies using structural MRI, see Nieuwenhuis et al. (2012)). We recently demonstrated in two large independent samples that a classification model built from one data set can be used to classify new subjects as schizophrenia patients or healthy subjects with 71% accuracy. To the best of our knowledge, no studies have been published investigating the use ofMRI to separate bipolar patients from healthy subjects or schizophrenia patients [although one study combined structural MRI brain measures and neuropsychological test scores for this purpose (Pardo et al., 2006)]. Given the brain abnormalities found in bipolar disorder and schizophrenia and the differences between these abnormalities (Arnone et al., 2009; Ellison-Wright and Bullmore, 2010; Hulshoff Pol et al., 2012; Koo et al., 2008; McDonald et al., 2005; Qiu et al., 2008; Rimol et al., 2010, 2012), it may be fruitful to apply these classification models to help separate these two disorders. We train three SVM models to separate patients with schizophrenia from those with bipolar disorder, healthy subjects from patients with schizophrenia, and healthy subjects from patients with bipolar disorder. Although it could be of theoretical interest to build a three-group classifier that separates the three groups in a single step, our approach addresses the clinical relevant issue of separating the two disorders using MRI. Furthermore, it provides brain patterns that discriminate between the respective groups, which can be analyzed to indicate which features are unique to the discrimination between schizophrenia and bipolar disorder. We test the predictive power of the models both in the dataset they were built on and in an independent dataset.

Materials and methods

General

In this study we used two datasets. The first set, called discovery sample, was used to build classification models for the separation of healthy subjects and patients with schizophrenia and bipolar disorder. The models were tested on this set too. On the second set, called validation sample, no models were built; this independent sample was used to test the generalizability of the models built on the first set.

Subjects - discovery sample

Schizophrenia patients (SZ), patients with bipolar disorder (BP) and healthy subjects (HC) were selected from our database. Since the quality of machine learning models strongly benefits from large and balanced training data sets, we extracted the largest possible groups of subjects that were same-sized and matched on gender (exactly) and age. This resulted in three groups of each 66 subjects (24 males), aged 37 ± 11 years. The subjects overlap to a large extent with the sample described in Hulshoff Pol et al. (2012); additional SZ patients and HC subjects are part of the study described by Hulshoff Pol et al. (2001). The sample included singletons and twins. To ensure independency between the three groups, only one twin from discordant twin pairs was included. The presence or absence of psychopathological abnormality was established using the Comprehensive Assessment of Symptoms and History (Andreasen et al., 1992) and Schedule for Affective Disorders and Schizophrenia Lifetime Version (Endicott and Spitzer, 1978) assessed by at least one independent rater who was trained to assess this interview. All healthy subjects met Research Diagnostic Criteria (Spitzer et al., 1978) of “never [being] mentally ill.” Patients in the schizophrenia group met DSM-IV criteria for schizophrenia. Patients in the bipolar group met DSM-IV criteria for bipolar I (N = 50), II (N = 14) or NOS (N = 2). Subjects were matched for age, sex and socioeconomic status of their parents expressed as the highest completed level of education by one of their parents.

All SZ patients had received antipsychotic medication in the past and all but one patient received antipsychotic medication at the time of the MRI scan. Medication included typical (N = 36) and atypical (N = 23) antipsychotic agents. Forty-five BP patients were using lithium at the time of the scan and 13 patients were using antipsychotics (of whom 5 were using both lithium and antipsychotics). See **Table 1** for demographic information.

Table 1 | Demographics

	Discovery sample			Validation sample		
	SZ	BP	HC	SZ	BP	HC
N	66[1]	66[2]	66[3]	46	47	43
Male/female	24/42	24/42	24/42	33/13	22/25	21/22
Age (year) mean (SD)	36.5 (11.0)	37.7 (11.0)	38.2 (10.8)	31.0 (7.5)	41.6 (10.0)	33.8 (9.4)
Range	18-57	18-60	18-62	19-48	22-60	19-60
Duration of illness mean (SD) (year)	15.4[4] (11.0)	12.8[5] (9.7)	-	7.2 (6.5)	20.5[6] (7.4)	-
Medication:						
Antipsychotic (yes/no)	59/1[7]	12/54	-	46/0	36/11	-
Lithium (yes/no)	0/66	45/21	-	0/46	31/15[8]	-

[1] 24 twins (no complete pairs); [2] 32 twins and 13 complete concordant twin pairs; [3] 23 twins and 18 complete pairs. Information missing in; [4] 17, [5] 8, [6] 5, [7] 6, [8] 1 patients.

Subjects - validation sample

Forty-six SZ patients, 47 BP patients and 43 HC subjects were drawn from our 3 Tesla MRI database. The SZ patients and part of the HC subjects were part of an earlier study (Scheewe et al., 2012). When composing the validation data set, we had to make a trade-off between the size of the sample and the matching on sex and age with the discovery set. A fair test of the validity of the models requires distributions of subjects with respect to variables such as age and sex that match those of the discovery sample as close as possible, but not at the cost of excluding too many subjects, since this would reduce the power of the validation test. The resulting sample included about equal numbers of subjects per group, and matched the discovery sample on age, but not on gender (significantly more males). Patients in the SZ group met DSM-IV criteria for schizophrenia (N = 35) or schizoaffective disorder (N = 11). Patients in the bipolar group all met DSM-IV criteria for bipolar I. Medication of SZ patients included typical (N = 4) and atypical (N = 38) antipsychotic agents. Thirty-one BP patients were using lithium at the time of the scan and 36 patients were using antipsychotics (of whom 22 were using both). See **Table 1** for demographic information. All participants gave written informed consent to participate in the study. The study was approved by the Medical Ethical Research Committee for human research (METC) from the University Medical Center Utrecht and was carried out under the directives of the Declaration of Helsinki (Amendment South Africa 2000).

Imaging and preprocessing

All scans from the discovery sample were acquired on a 1.5 Tesla Philips NT scanner (Philips, Best, The Netherlands). Three-dimensional T1-weighted, fast field echo scans with 160 to 180 contiguous coronal slices (echo time [TE], 4.6 ms; repetition time [TR], 30 ms; flip angle, 30°; field of view [FOV], 256 mm; 1×1×1.2 mm^3 voxels) were made of all subjects. All scans from the validation sample were acquired on a 3 Tesla Philips Achieva scanner. Three-dimensional T1-weighted, fast field echo scans with 180 contiguous sagittal slices (TE, 4.6 ms; TR, 10 ms; flip angle, 90°; FOV, 240 mm; 0.75×0.75×0.80 mm^3 voxels) were made of all subjects.

The scans were processed with our standard image processing pipeline (Brouwer et al., 2010; Hulshoff Pol et al., 2001) on the computer network of the Department of Psychiatry at the University Medical Center Utrecht. The features we used were extracted from the processed T1-weighted images. The images were transformed into Talairach orientation (no scaling), after which they were corrected for scanner RF-field nonuniformity. Using a partial volume segmentation technique (Brouwer et al., 2010) the brain was segmented into gray matter, white matter and cerebrospinal fluid. The gray matter segments were blurred using a three-dimensional Gaussian kernel (full-width half-maximum (FWHM) = 8 mm). The voxel values of these blurred segments reflect the local presence, or concentration, of gray matter and will be referred to as gray matter 'densities' (GMDs). In order to compare GMDs at the same anatomical location between all subjects, the GMD images were transformed into a standardized coordinate system using a two-step process. First, the T1-weighted images were linearly transformed to a model brain (Hulshoff Pol et al., 2001). In this linear step joint entropy mutual information metric was optimized. In the second step nonlinear (elastic) transformations were calculated to register the linearly transformed images to the model brain up to a scale of 4 mm (FWHM), thus removing global shape differences between the brains, but retaining local differences (ANIMAL; Collins et al., 1995). The GMD maps were then transformed to the model space by applying the concatenated linear and nonlinear transformations. The GMD maps were not modulated, thus representing relative amounts of gray matter. We made this choice (i) to keep the preprocessing steps and interpretation the same as for our VBM studies on these data and (ii) because of recent evidence that unmodulated analyses outperform modulated analyses, both in VBM (Radua et al., in press) and analyses aiming to discriminate between groups of patients and controls (Dashjamts et al., 2012). Since the density maps have been blurred to an effective resolution of 8 mm, it is not necessary to keep this information at the 1-mm level. Therefore, the maps were resampled to voxels of size 2 × 2 × 2.4 mm³, i.e., doubling the original voxel sizes. The use of smoothed, resampled, GMD maps removes noise originating from imperfect segmentation and warping (due to image noise and different brain topology). This is especially important when combining 1.5 T and 3 T images, since the latter are much more detailed. For all voxels, GMD was regressed on age and sex for all subjects in the sample together. The resulting b-maps were used to correct the GMD maps for the effects of these factors and calculate GMD residuals, which were used as features for the support vector machine model.

Support vector machine models

The support vector machine (SVM) is a high-dimensional, pattern recognition, supervised learning algorithm (Vapnik, 1999) used to solve classification problems. In our case this problem consists of separating three groups of subjects. Therefore, we build three models: M(sz-hc) to separate SZ from HC; M(hc-bp) to separate HC from BP; and M(bp-sz) to separate BP from SZ (**Figure 1**). The SVM model is trained to classify subjects based on their features, in our case gray matter densities. We integrated LIBSVM (Chang, 2011) with our software to carry out the classification. Subjects are represented by features congregated into a vector x_i per subject. These vectors exist in a high dimensional feature space, in which a flat decision surface is constructed to separate the subjects from different classes (shown schematically in the center of **Figure 1**). This is accomplished by the introduction of a decision function **Equation 1** $y(x_i)$:
that vanishes at the decision surface. The weight vector w is a normal vector to this surface; b is an offset. In the training phase each subject has a label t_i (e.g., patients 1;

$$y(x_i) = w^\top \cdot x_i - b$$

healthy control -1), and the function is optimized by requiring $y(x_i) < 0$ if $\mathrm{t}_i = -1$, and $y(x_i) > 0$ if $\mathrm{t}_i = +1$. When applying the model this decision function is used to classify the test subjects according to the sign of $y(x_i)$. The weight-vector not only contains information on feature importance, but also on whether it is either an increase or decrease of a particular feature's value that contributes to being classified as a patient. There can be several surfaces that exactly separate the classes. The SVM chooses the so called optimal separating hyperplane (OSH) such that the space between the two classes, which is called the margin, is made as large as possible. This is a necessary condition for generalization of the model to new subjects. There is a free parameter C in SVM that influences the narrowness of the margin. It was shown earlier (Franke et al., 2010) that tuning C can increase the model's performance. We used C as optimized by Nieuwenhuis et al. (2012).

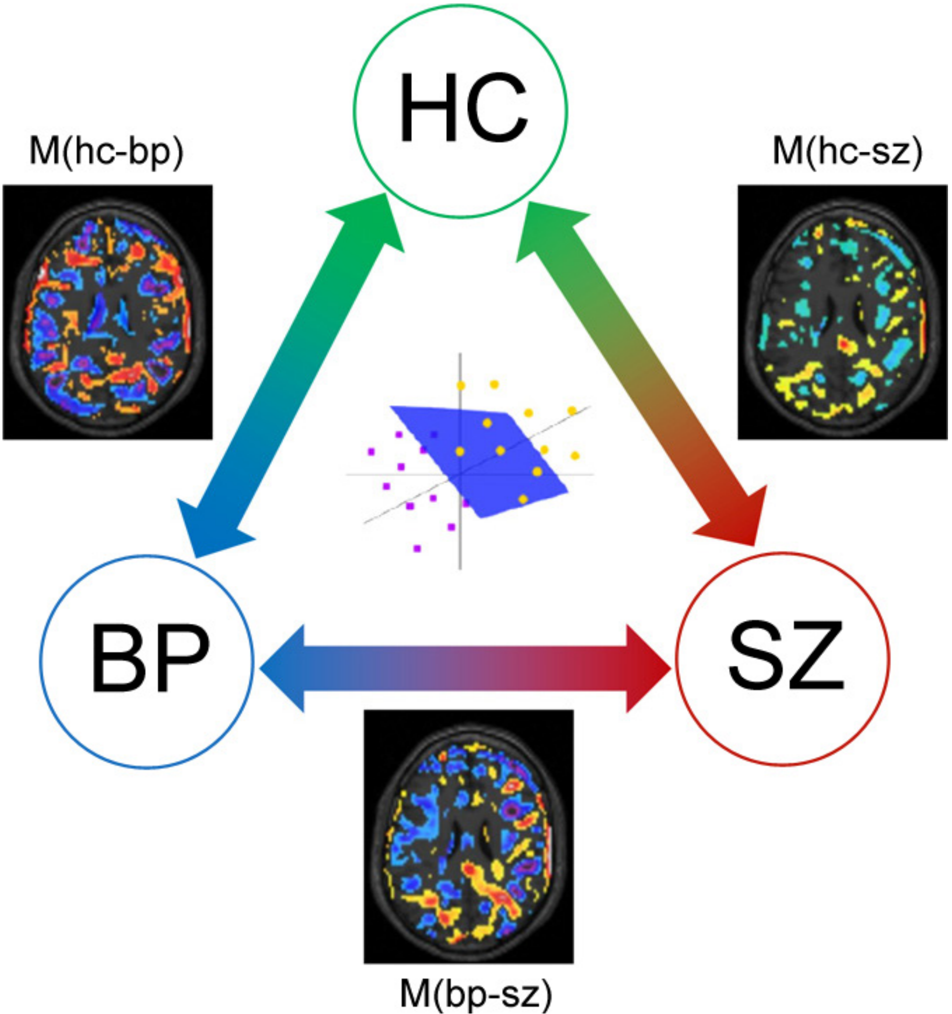

Figure 1 | Classification scheme. The three groups: healthy subjects (HC); patients with bipolar disorder (BP) and schizophrenia patients (SZ), are depicted by circles. The three models that are trained to perform pair-wise separations of the groups are indicated by arrows, labeled with the model's name (M) and a symbolic picture of its discriminative brain pattern (w-map). In the center a schematic picture of the support vectormachine (SVM): an optimal separation plane (OSH; blue) separates the two classes of subjects based on their positions in a high-dimensional feature space (yellow and purple dots).

Limiting and avoiding the influences of medication

The size of the striatum is known to be affected by (typical) antipsychotic medication (Smieskova et al., 2009). To exclude this possible confounding effect, we created a model where the striatum was masked out. The striatum was segmented manually from the model brain image and, using mathematical morphology operations, enlarged, to ensure the entire striatum was excluded for all subjects. We applied two-sample t-tests to test for possible differences in model performance between BP patients on antipsychotic medication and thosewhowere not. Lithium has been shown to influence gray matter (density) (Kempton et al., 2008). Two-sample t-tests were used to test for differences in model performance (ie, accuracy) between BP patients using/not using lithium. To avoid separating lithium users fromnon-users, instead of BP patients from other subjects, we also built SVM models including only subjects not on lithium (these models are referred to as MŁ). Since only 25 of the BP patients did not use lithium, models using these subjects would include only 50 subjects in total. To solve this problem of small training sets, we built 100 of such models, each time using random selections (but accounting for twin dependency — see 'Quality measures' section) of SZ and HC subjects. For comparison, we also built 100 models MŁ(hc-sz) including random selections of 25 SZ patients and 25 healthy subjects.

Quality measures I: tests on the discovery sample

The quality of a model M(g1–g2) is assessed by the percentage correctly classified subjects belonging to group g1 and the percentage correctly classified subjects belonging to group g2. We tested the accuracy of the models by using cross validation,which gives an estimate of how well the model will generalize to a new data set. In a leave-k-out cross validation setup, each time k subjects are left out of the training set, which are subsequently used to test the model. Often k = 1 is chosen, but to keep equal numbers of subjects in the two groups, k = 2 would be more appropriate here. We chose k = 4, which gave us the opportunity to leave complete twin pairs out of the training set simultaneously. This avoids a bias in the prediction of a twin's class if his/her cotwin (in the same class) is in the training set. The procedure is thus as follows: First a model is trained on all subjects but four, which are then used to test this model. This is repeated with four different subjects left out, until all subjects have been left out once. In our case each model M is thus trained 33 times. For the MŁ models, twin bias was avoided by not including the cotwin when a twin from a complete pair was included in the selection. This gave us the opportunity to apply leave-2-out cross-validation, leaving the training set as large as possible. The significance of a model M(g1–g2) is tested by means of bootstrapping, testing the null-hypothesis that the observed separation of the two groups is by chance and not due to true group differences. All subject-labels (g1, g2) are randomly permuted prior to the model's training and testing phase. This process is repeated a thousand times to estimate the null-distribution of separation accuracy percentages and calculate the probability of finding by chance an accuracy at least as extreme as observed (p-value). To test whether being a twin biases a subject's chance to be correctly classified, a Fisher's exact test was applied to the frequencies of correctly and wrongly classified singletons and twins. In addition, a multi-class SVM [max-wins-voting SVM, MWV_SVM (Knerr et al., 1990)] model was trained and tested using the same cross validation set-up.

Quality measures II: tests on the validation sample

The generalizability of the models built on the discovery sample was tested by applying them to the data of the validation sample. Large systematic differences in scan quality between the two samples (1.5 T vs 3 T),which will be reflected in the GMD values and thus in the feature vectors x, were expected to lead to shifts in the output values y of the classifiers (see Eq. (1)), and thus to a misbalance between false positives and false negatives in the classification of subjects scanned at 3 T. These possible shiftswere accounted for in two steps. Based on a 1.5 T – 3 T calibration study (Brouwer et al., unpublished; see Appendix A) a reliability mask was applied to the features, thus only including voxels with GMD values comparable between 1.5 T and 3 T into the model. The 1.5 T models were thus rebuilt on a restricted feature set (76% of the voxels). A receiver operating characteristic (ROC) curve analysis was used to remove the remaining misbalance between false positives and false negatives. In a post hoc analysis the accuracy percentages on the validation setwere recalculated,weighing themale and female percentages according to the discovery sample's gender distribution. Within the SZ group, a Fisher's exact test was used to test whether diagnosis (schizophrenia vs schizoaffective) biases a subject's chance to be correctly classified. To test the significance of the results obtained in the validation sample the same 1000 label permutations of the discovery setwere used. The models built fromthese 1000 sets (without leaving out any subjects this time) were applied to the validation set's subjects to estimate the nulldistribution of separation accuracy percentages.The validation accuracies of the truemodelswere comparedwith these null-distributions to calculate the probability that these accuracies were found by chance.

Analysis of the discriminative patterns

If the different models are built from the same set of features (as is the case in our approach), their weight vectors lie in the same space and can be directly compared. The groups is by chance and not due to true group differences. All subject-labels (g1, g2) are randomly permuted prior to the model's training and testing phase. This process is repeated a thousand times to estimate the null-distribution of separation accuracy percentages and calculate the probability of finding by chance an accuracy at least as extreme as observed (p-value). To testwhether being a twin biases a subject's chance to be correctly classified, a Fisher's exact testwas applied to the frequencies of correctly and wrongly classified singletons and twins. In addition, a multi-class SVM [max-wins-voting SVM, MWV_SVM (Knerr et al., 1990)] model was trained and tested using the same cross validation set-up.

Quality measures II: tests on the validation sample

The generalizability of the models built on the discovery sample was tested by applying them to the data of the validation sample. Large systematic differences in scan quality between the two samples (1.5 T vs 3 T),which will be reflected in the GMD values and thus in the feature vectors x, were expected to lead to shifts in the output values y of the classifiers (see Eq. (1)), and thus to a misbalance between false positives and false negatives in the classification of subjects scanned at 3 T. These possible shifts were accounted for in two steps. Based on a 1.5 T – 3 T calibration study (Brouwer et al., unpublished; see Appendix A) a reliability mask was applied to the features, thus only including voxels with GMD values comparable between 1.5 T and 3 T into the model. The 1.5 T models were thus rebuilt on a restricted feature set (76% of the voxels). A receiver operating characteristic (ROC) curve analysis was used to remove the remaining misbalance between false positives and false negatives. In a post hoc analysis the accuracy percentages on the validation setwere recalculated,weighing the male and female percentages according to the discovery sample's gender distribution. Within the SZ group, a Fisher's exact test was used to test whether diagnosis (schizophrenia vs schizoaffective) biases a subject's chance to be correctly classified. To test the significance of the results obtained in the validation sample the same 1000 label permutations of the discovery setwere used. The models built from these 1000 sets (without leaving out any subjects this time) were applied to the validation set's subjects to estimate the null distribution of separation accuracy percentages.The validation accuracies of the true models were compared with these null-distributions to calculate the probability that these accuracies were found by chance.

Analysis of the discriminative patterns

If the different models are built from the same set of features (as is the case in our approach), their weight vectors lie in the same space and can be directly compared. The groups is by chance and not due to true group differences. All subject-labels (g1, g2) are randomly permuted prior to the model's training and testing phase. This process is repeated a thousand times to estimate the null-distribution of separation accuracy percentages and calculate the probability of finding by chance an accuracy at least as extreme as observed (p-value). To test whether being a twin biases a subject's chance to be correctly classified, a Fisher's exact test was applied to the frequencies of correctly and wrongly classified singletons and twins. In addition, a multi-class SVM [max-wins-voting SVM, MWV_SVM (Knerr et al., 1990)] model was trained and tested using the same cross validation set-up.

Quality measures II: tests on the validation sample

The generalizability of the models built on the discovery sample was tested by applying them to the data of the validation sample. Large systematic differences in scan quality between the two samples (1.5 T vs 3 T),which will be reflected in the GMD values and thus in the feature vectors x, were expected to lead to shifts in the output values y of the classifiers (see Eq. (1)), and thus to a misbalance between false positives and false negatives in the classification of subjects scanned at 3 T. These possible shiftswere accounted for in two steps. Based on a 1.5 T – 3 T calibration study (Brouwer et al., unpublished; see Appendix A) a reliability mask was applied to the features, thus only including voxels with GMD values comparable between 1.5 T and 3 T into the model. The 1.5 T models were thus rebuilt on a restricted feature set (76% of the voxels). A receiver operating characteristic (ROC) curve analysis was used to remove the remaining misbalance between false positives and false negatives. In a post hoc analysis the accuracy percentages on the validation setwere recalculated,weighing themale and female percentages according to the discovery sample's gender distribution. Within the SZ group, a Fisher's exact test was used to test whether diagnosis (schizophrenia vs schizoaffective) biases a subject's chance to be correctly classified. To test the significance of the results obtained in the validation sample the same 1000 label permutations of the discovery setwere used. The models built fromthese 1000 sets (without leaving out any subjects this time) were applied to the validation set's subjects to estimate the nulldistribution of separation accuracy percentages. The validation accuracies of the truemodelswere comparedwith these null-distributions to calculate the probability that these accuracies were found by chance.

Analysis of the discriminative patterns

If the different models are built from the same set of features (as is the case in our approach), their weight vectors lie in the same space and can be directly compared. The gray matter pattern that discriminates between schizophrenia and bipolar disorder may share part of it with the pattern that discriminates between schizophrenia and health. Mathematically, the vectors w in Eq. (1) of the different models may not be orthogonal. By projection, we can decompose $w\,(bp - sz)$ into a part that coincides with $w\,(hc - sz)$

and a part perpendicular to it (**Equation 2**):

$$w\left(bp-sz\right)=w_{/\!/hc-sz}\left(bp-sz\right)+w_{\perp hc-sz}\left(bp-sz\right)$$

which can be written as (**Equation 3**):

$$w\left(bp-sz\right)=\cos\left(\varphi\right)w\left(hc-sz\right)+w_{\perp hc-sz}\left(bp-sz\right)$$

with: $\cos\left(\varphi\right)=w^{\top}\left(bp-sz\right)\cdot w\left(hc-sz\right)/\left(\parallel w\left(bp-sz\right)\parallel\parallel w\left(hc-sz\right)\parallel\right)$ ie, the cosine of the angle φ between the weight vectors, determining the size of the parallel, or shared part. The perpendicular part is obtained by subtraction: $w_{\perp}=w-w_{/\!/}$ The offset b in Eq. (1) is divided accordingly:$b=b_{/\!/}+b_{\perp}$ The separation of sz and bp is thus built from a part that employs the same pattern as is used for the $hc-sz$ separation, and a part that is unique for the discrimination between sz and bp. The latter tells us which brain pattern drives the discrimination between the two disorders. We will refer to the vectors was w-maps. The two components of w-map $(bp-sz)$ will also be applied to the data.

Results

The classification results of the models are given in **Table 2**. In the discovery sample, the models involving SZ patients performed best, with classification accuracies of 86% or higher for themodels including all subjects (M). SZ patients can thus be separated well from HC subjects and from BP patients (p b 0.001). The separation of BP patients and HC subjects, on the other hand, turned out to be less accurate: 67% of the HC subjects were correctly classified (p = 0.02) and only 53% of the BP patients (p = 0.4). There was no significant difference in classification accuracy between BP patients on antipsychotic medication and those who were not (M($sz-bp$): 83% and 87%, respectively, t(14.7) = 0.32, p = 0.63; M($hc-bp$): 50% and 54%, t(15.7) = 0.23, p = 0.41). There was no significant difference in classification accuracy between BP lithium users and non-users (M($sz-bp$): 90% and 80%, t(39.6) = 1.11, p = 0.14; M($hc-bp$): 51% and 56%, t(50.6) = 0.38, p = 0.71). The models including only non-users (MŁ) produced comparable outcomes. These accuracies were on average a few percentage points lower. There was no significant difference in classification accuracy between twins and singletons. Contingency **Table 3** presents themulti-class results,which indicate a comparable classification of SZ subjects (86% correct) and less accurate separation of HC and BP subjects, although with accuracy percentages of 59 and 50 well above chance (33%). Application of the discovery sample's models to the validation sample yielded unbalanced specificity/sensitivity percentages: 91%/54% (HC/SZ); 79%/48% (BP/SZ); 70%/49% (HC/BP). Adjustment of the classification thresholds after ROC curve analyses led to accuracies of 77%/74% (HC/SZ), 66%/65% (BP/SZ), 63%/55% (HC/BP). The performance of the HC/SZ and BP/SZ models was significantly better than chance (p < 0.005 and p < 0.05); the HC/BP model did not perform significantly better (p = 0.2). Reweighting the male/female accuracy percentages to match the discovery sample's gender distribution lead to specificity/sensitivity percentages of 79%/79% (HC/SZ), 65%/71% (BP/SZ), 63%/55% (HC/BP). There was no significant difference in classification accuracy between the diagnoses schizophrenia and schizoaffective disorder. The results are presented in **Table 4**, together with those from the discovery sample, for comparison.

Table 3 | Contingency table (multi-class SVM)

	Classification:		
	Healthy control	**Schizophrenia patient**	**Bipolar patient**
Group:			
Healthy control	59	11	30
Schizophrenia patient	6	86	6
Bipolar patient	41	8	50

Class predictions for subjects from the three groups in %. In the SZ and BP groups one subject could not be assigned to a class because there was a tie.

Table 4 | Classification accuracies in %. Raw: application of discovery-model to validation set; ROC:application results after ROC curve analysis; ROC-reweighted: the ROC results reweighted to match gender distribution of discovery sample; calib-masked: model performance on the discovery sample (using L4O) using the reduced feature set, for direct comparison with the validation set results; the 'original' results are taken from **Table 2**, for comparison.

	Validation set			**Discovery set**	
Model	Raw	ROC[a]	ROC-reweighted	Calib-masked	Original
BP/SZ	78.7/47.8	66.0/65.2	64.6/71.0	86.4/89.4	86.4/89.4
HC/BP	69.8/48.9	62.8/55.3	63.0/55.5	66.7/54.5	66.7/53.0
HC/SZ	90.7/54.3	76.7/73.9	79.2/79.2	86.4/90.9	87.9/92.4

[a] Corresponding shifts in b (see Eq. (1)): −0.16 (BP/SZ), −0.08 (HC/BP), and −0.20 (HC/SZ).

Figure 2 shows the w-maps overlaid on the brain and the decomposition of w-map$(bp - sz)$ into the part shared with w-map $(hc - sz)$, ie, $w_{/\!/ hc-sz}(bp - sz)$and the part perpendicular to it, $w_{\perp hc-sz}(bp - sz)$. The cosine of the angle between the two weight vectors was $\cos(\varphi) = 0.65$ (Eq. (3)). Schizophrenia patients are separated from other subjects by a pattern that includes decreases in (pre- and orbito) frontal and (superior) temporal gray matter. The pattern that drives the discrimination between patients with bipolar disorder and schizophrenia patients includes decreases (in SZ relative to BP) in superior frontal and parietal gray matter; deeper in the frontal, parietal and occipital cortices, paired regions of increases and decreases are found (**Figure 2**, third column). **Figure 3** displays the y-values (Eq. (1)) after application of the different w-maps to the individual brain data, ie, the scattering around the optimal separating hyperplanes (OSHs), here depicted as separation lines. The figure illustrates the good separability of SZ patients from both BP patients and HC subjects. $M_{/\!/ hc-sz}(bp-sz)$, the part shared by HC and BP, already produces a good separation of SZ patients and the other subjects; $w_{\perp hc-sz}(bp-sz)$, the part unique for BP-SZ separation, further increases the distance of the BP subjects to the OSH (all subjects' y < 0 in the training phase), while it does not displace the HC subjects from the OSH (subjects' mean y close to 0). In the test phase (leave-4-out), the discriminating effect of: $w_{\perp hc-sz}(bp-sz)$ was smaller, but significant (mean difference in y between the two groups = 0.11, t = 2.06, p < 0.041).

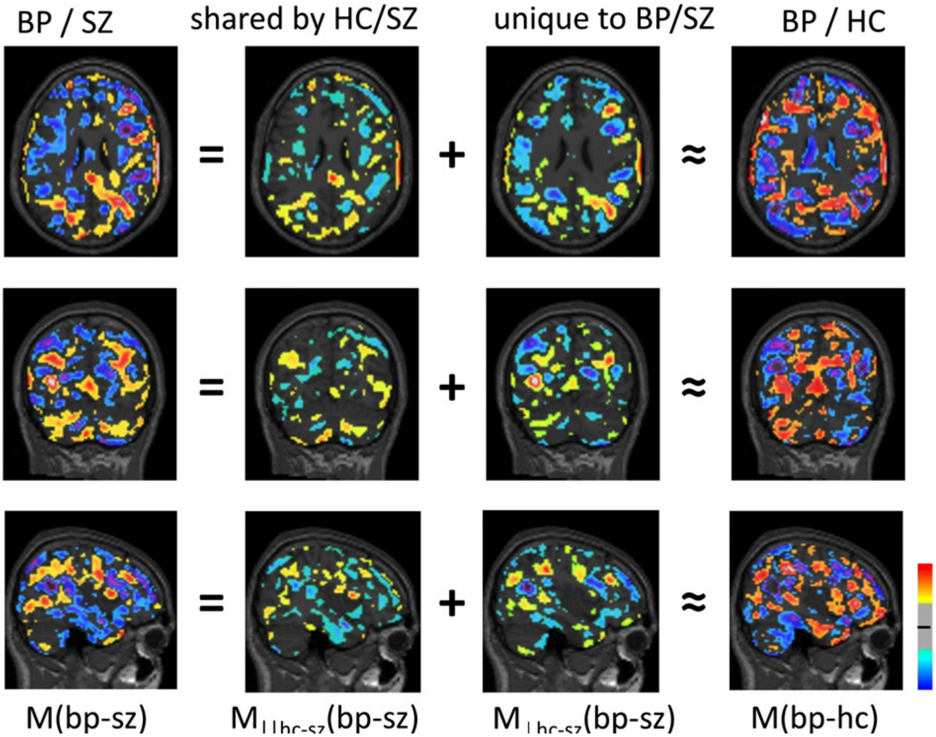

Figure 2 | Axial, coronal and sagittal slices (fromtop to bottom) of the template brain with w-maps projected on it. Separation pattern (w-map) of BP and SZ (M$(bp-sz)$; first column), which has been decomposed into the part shared with the separation of HC and SZ $M_{/\!/ hc-sz}(bp-sz) = 0.65 * M(hc-sz);$ second column) and the remaining part, orthogonal to the former part, being unique to the separation of BP and SZ ($M_{\perp hc-sz}(bp-sz)$; third column). The latter pattern can be compared to M$(\mathrm{bp}-\mathrm{hc})$ ($=-$ M$(\mathrm{hc}-\mathrm{bp})$; fourth column). The pattern in the third column is characterized by neighboring pairs of regions with hot and cool colors. For all models M(g1–g2), hot colors (clamped between 0.0005 and 0.0025) refer to relative increase of GM density in g2 subjects with respect to g1 subjects, and vice versa for cool colors (clamped between −0.0005 and −0.0025).

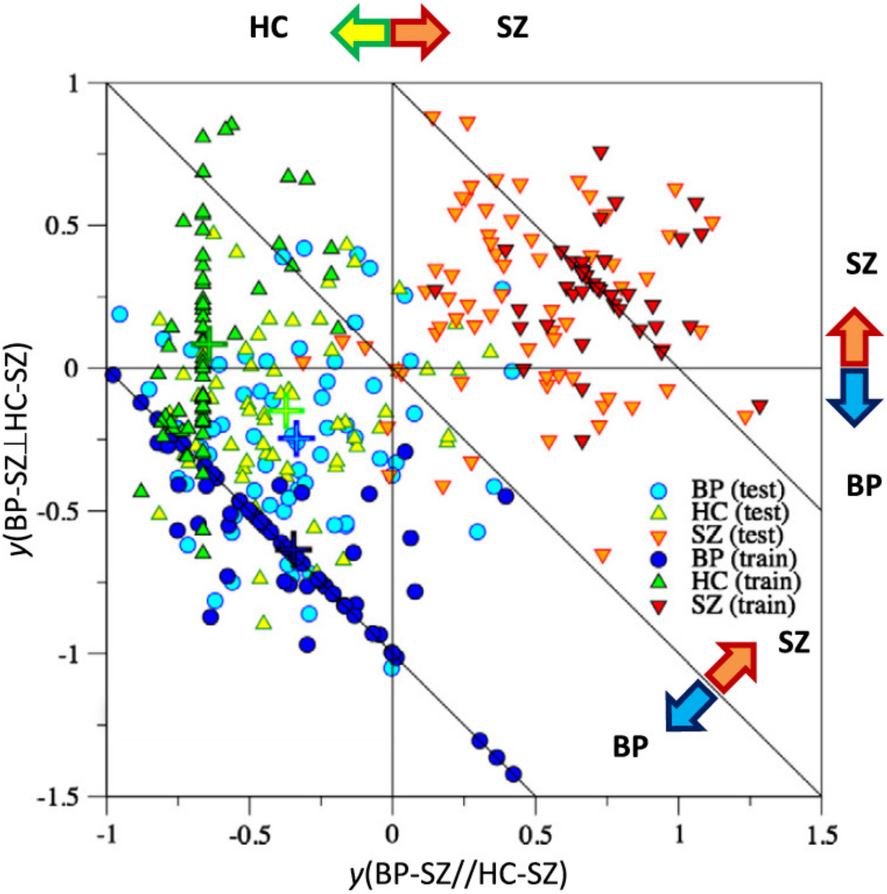

Figure 3 | Application of M$(bp - sz)$ to all subjects. The horizontal axis shows the effect $y-$ value, Eq. (1)) of the part shared with M$(hc - sz)$, ie, $0.65(w^{\top}(hc - sz) \cdot x_i - b)$ with $\cos(\varphi)$=0.65 (Eq. (3)). A y-value smaller than 0 means not-SZ, larger than 0 means SZ. The vertical axis shows the effect of the remaining part of the model (unique for BP-SZ), ie, the part perpendicular to M$(hc - sz)$. A y-value smaller than 0 means classification as BP, larger than 0 means classification as SZ. The sum of the two parts gives the y-value of the M$(bp - sz)$ model itself, with subjects on the lower-left of the diagonal being classified as BP (or not-SZ), and on the upper-right of the diagonal as SZ. Darker symbols (red, blue, green) refer to the effects using the models trained on all subjects; brighter symbols (orange, cyan, yellow) to effects on subjects that were left out of the model (test). The majority of training subjects is placed at distance ±1 (or ± $\cos(\varphi)$) by application of the model they are part of: this is reflected by the lines of blue (BP) and green (HC) symbols; since SZ is part of all models, the majority of SZ subjects is placed at ($\cos(\varphi)$, 1 – $\cos(\varphi) = (0.65,\ 0.35)$. The plus symbols are placed at the means of the distributions.

Discussion

The purpose of this study was to classify patients with schizophrenia, bipolar disorder, and healthy controls on the basis of their structural MRI scans. We used a support vector machine (SVM) to create three models from gray matter density images, each separating two of the three groups. We confirmed that it is possible to separate schizophrenia patients from healthy subjects with an average accuracy of 90%. Additionally, we demonstrated that schizophrenia patients can be separated from those with bipolar disorder with a nearly identical degree of precision (88%). The delineation of patients with bipolar disorder and healthy subjects reached much lower accuracy: 67% of the healthy subjects and only 53% of the patients with bipolar disorder were classified correctly. In an independent data set we replicated these results, yielding average classification accuracies of 75% and 66% for the separation of schizophrenia patients from healthy subjects and patients with bipolar disorder, and 59% for the separation of bipolar patients from healthy subjects. The accuracy of the HC-SZ and BP-SZ models in this set is lower than in the discovery set, but remains significantly better than chance.

Since lithium is known to affect gray matter density (Kempton et al., 2008; Moore et al., 2000; van der Schot et al., 2010), patients with bipolar disease were split into those who were on lithium and those who were not; separate models were created including only subjects not on lithium. Lithium turned out not to influence classification accuracy. Models including only non-users performed with slightly lower precision, but this was a systematic effect also present in the SZ-HC model with reduced number of subjects, and thus likely to be a decrease in robustness due to fewer training subjects (Nieuwenhuis et al., 2012). It is well known that (typical) antipsychotic medication affects striatal volume, which is why we excluded features from the striatum in the models. Although some studies also suggest that antipsychotic medication affects cortical gray matter (Ho et al., 2011), the reported effects are inconsistent (Shepherd et al., 2012) and so it is not clear which brain regions should be left out of the model. However, when we compared the patients with bipolar disorder who were on antipsychotic medication with those who were not, we did not find an effect of antipsychotic medication on the classification of these patients suggesting that, for classification purposes at least, the effect of antipsychotics on cortical gray matter is limited. Nevertheless, effects of medication on classification can only be ruled out by studies in medication naive patients.

Since the discriminative brain pattern is a description of the cumulative contributions of all features, the interpretation of the effects of single brain regions on the separation of the groups is complicated. Some marked contributions are, however, present. Schizophrenia patients are separated from other subjects by a pattern that includes decreases in (pre- and orbito)frontal and (superior)temporal gray matter, consistent with reported structural brain abnormalities in schizophrenia (Fornito et al., 2009; Haijma et al., 2012). The pattern that separates bipolar patients from healthy subjects shows widespread contributions, lacking any marked regions, "consistent" with recent literature that concludes that many brain abnormalities in BP may be obscured by (clinical) heterogeneity (Kempton et al., 2008; Selvaraj et al., 2012). The pattern that drives the discrimination between patients with bipolar disorder and schizophrenia is diffuse, but includes decreases (in SZ relative to BP) in superior frontal and parietal gray matter. Deeper in the frontal, parietal and occipital cortices, neighboring regions with opposite contributions to the discrimination are found: decreases are flanked by increases (**Figure 2**; third column). These paired regions may reflect local shape changes in which some structure is larger in one disorder relative to the other, at the cost of a neighboring structure. The large extent of these changes over the (right) hemisphere may reflect a brain network problem in SZ (van den Heuvel et al., 2013), reflecting symptoms that could be attributable to problemswith the proper integration of information.

The classification accuracies in the validation sample are all themore remarkable given the differences in scanner field strength and demographics between discovery and validation samples. While a 1.5 T–3 T calibration allowed us to exclude the least reliable features, the 24% reduction in number of features to build the model from could have reduced the model's generalizability. Correcting for the difference in gender distribution improved the performance in the validation sample. Other factors that may have affected the performance reproducibility are the differences in duration of illness and the number of SZ patients on typical antipsychotic medication, although we tried to reduce the impact of the latter as much as possible by excluding the striatum from the model (Nieuwenhuis et al., 2012). The classification accuracy of patients with schizoaffective disorder was not significantly different from those with schizophrenia. The relatively low performance of the model separating BP from HC may be attributable to the less marked brain abnormalities found in BP (meta-analyses: Kempton et al., 2008; McDonald et al., 2004) as compared to those found in SZ (meta-analyses: Haijma et al., 2012; Olabi et al., 2011).

This study has two major implications. First, the accurate separation of individuals suffering from schizophrenia and bipolar disorder on the basis of their structural MRI scans suggests that gray matter pathology in schizophrenia and bipolar disorder differs to such an extent that they can be reliably differentiated using machine learning paradigms. The second, and more practically useful, implication is that structural MRI could aid at separating schizophrenia from bipolar disorder in the differential diagnostic process.

Although the current MRI classification results are promising, they are limited by the small number of patients with bipolar disorder not using lithium (N = 25). Indeed, larger datasets and independent replications are needed before thismodel can be put to clinical use. Also, the model was built, and tested on, chronically ill patients,while its clinical use will be mostly in the more recently ill subjects. Thus, longitudinal studies including recent onset psychotic patients with diagnostic follow-up (of many years) are needed to test the clinical usefulness of this model. Although the psychiatrist's diagnosis serves as gold standard to train our SV machine, his/her reliability is not 100%; in fact, for SCID-I, inter-rater reliabilities for bipolar disorder are in the range of 80% and for schizophrenia are 94% (Skre et al., 1991), while agreement between clinical diagnoses and SCID-I is poor (Steiner et al., 1995). Machines cannot be expected to approach 100% accuracy, and, consequently, an MRI-SVM can only become a tool aiding the diagnostic process. However, this tool comeswith two virtues: it is based on information not available to the psychiatrist, and it is objective.

In conclusion,we have shown that structural MRI can be used to separate schizophrenia patients fromboth healthy subjects and bipolar patients with good accuracy, despite using discovery and validation samples acquired on scannerswith different field strengths and varying in gender distribution. This suggests not only that gray matter pathology is different in schizophrenia and bipolar disorder, but that, eventually, MRI may prove to be a useful instrument in the diagnostic process in psychiatry.

Acknowledgments

We thank Dorret Boomsma for providing us with healthy twin subjects from the Netherlands Twin Registry.We thank and Astrid van der Schot for assistance with data acquisition and processing.

Appendix A. 1.5 Tesla–3 Tesla calibration

Fifteen healthy volunteers (12 female;mean (SD) age 26.1 (8.1) year) were scanned on both our 1.5 T and 3 T MRI scannerwith the acquisition protocols as described in the Methods section. Twelve of them were scanned twice on the same day and the others less than 4 months apart. Scans were preprocessed using our standard image processing pipeline (including putting scans in Talairach orientation, nonuniformity correction, and partial volume segmentation of gray/white matter and CSF, as described in the Methods section). Like the images from this study, the calibration scans were non-linearly transformed to a model brain. These transformations were then applied to the partial volume gray matter maps. The transformed segments were blurred (full-width-half-max, 8 mm) and then resampled to a 2 × 2 × 2.4 mm3 resolution to reduce noise and increase statistical power as would be done in a voxel-based analysis. To assess systematic offsets and reliability between the scanners, mean (SD) gray matter density per scanner and intraclass correlation (Shrout and Fleiss, 1979) between scanners was computed in each voxel. A between-scanner reliabilitymask was created by selecting voxels with ICC > 0.7 or SDs < 0.05.

References

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. *Arch. Gen. Psychiatry* 49, 615–623.

Arnone, D., Cavanagh, J., Gerber, D., Lawrie, S.M., Ebmeier, K.P.,McIntosh, A.M., 2009. Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta analysis. *Br. J. Psychiatry* 195, 194–201.

Brouwer, R.M., Hulshoff Pol, H.E., Schnack, H.G., 2010. Segmentation of MRI brain scans using non-uniform partial volume densities. *NeuroImage* 49, 467–477.

Chang, C.-C., 2011. A library for support vector machines, In: Lin, C.-J. (Ed.), *ACM Transactions on Intelligent Systems and Technology, 2nd ed.* pp. 1–27.

Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-d model-based neuroanatomical segmentation. *Hum. Brain Mapp.* 3, 190–208.

Dashjamts, T., Yoshiura, T., Hiwatashi, A., Togao, O., Yamashita, K., Ohyagi, Y.,Monji, A., Kamano, H., Kawashima, T., Kira, J., Honda, H., 2012. Alzheimer's disease: diagnosis by different methods of voxel-based morphometry. *Fukuoka Igaku Zasshi* 103, 59–69.

Davatzikos, C., Shen, D., Gur, R.C.,Wu, X., Liu, D., Fan, Y., et al., 2005.Whole-brainmorphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch. Gen. Psychiatry* 62, 1218–1227.

Ellison-Wright, I., Bullmore, E., 2010. Anatomy of bipolar disorder and schizophrenia: a meta-analysis. *Schizophr. Res.* 117, 1–12.

Endicott, J., Spitzer, R.L., 1978. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch. Gen. Psychiatry* 35, 837–844.

Fan, Y., Gur, R.E., Gur, R.C.,Wu, X., Shen, D., Calkins, M.E., Davatzikos, C., 2008. Unaffected family members and schizophrenia patients share brain structure patterns: a highdimensional pattern classification study. *Biol. Psychiatry* 63, 118–124.

Fornito, A., Yücel, M., Patti, J., Wood, S.J., Pantelis, C., 2009. Mapping grey matter reductions in schizophrenia: an anatomical likelihood estimation analysis of voxel-based morphometry studies. *Schizophr. Res.* 108, 104–113.

Franke, K., Ziegler, G., Kloppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage* 50, 883–892.

Haijma, S.V., Van Haren, N., Cahn,W., Koolschijn, P.C., Hulshoff Pol, H.E., Kahn, R.S., 2012. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophr. Bull.* 39, 1129–1138.

Ho, B.C., Andreasen, N.C., Ziebell, S., Pierson, R., Magnotta, V., 2011. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Arch. Gen.*

Psychiatry 68, 128–137.

Hulshoff Pol, H.E., Schnack, H.G.,Mandl, R.C., van Haren, N.E., Koning, H., Collins, D.L., et al., 2001. Focal gray matter density changes in schizophrenia. *Arch. Gen. Psychiatry* 58, 1118–1125.

Hulshoff Pol, H.E., van Baal, C.M., Schnack, H.G., Brans, R., van der Schot, A.C., Brouwer, R.M., et al., 2012. Overlapping and segregating structural brain abnormalities in twins with schizophrenia or bipolar disorder. *Arch. Gen. Psychiatry* 69, 349–359.

Ingalhalikar,M., Kanterakis, S., Gur, R., Roberts, T.P., Verma, R., 2010. DTI based diagnostic prediction of a disease via pattern classification. *Med. Image Comput. Comput. Assist. Interv.* 13, 558–565.

Karageorgiou, E., Schulz, S.C., Gollub, R.L., Andreasen, N.C., Ho, B.C., Lauriello, J., et al., 2011. Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. *Neuroinformatics* 9, 321–333.

Kasparek, T., Thomaz, C.E., Sato, J.R., Schwarz, D., Janousova, E., Marecek, R., et al., 2011. Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Res.* 191, 174–181.

Kempton, M.J., Geddes, J.R., Ettinger, U., Williams, S.C., Grasby, P.M., 2008. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Arch. Gen. Psychiatry* 65, 1017–1032.

Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Fogelman-Soulie, Herault (Ed.),*Neurocomputing:Algorithms, Architectures and Applications, NATO ASI.* Springer.

Koo, M.S., Levitt, J.J., Salisbury, D.F., Nakamura, M., Shenton, M.E., McCarley, R.W., 2008. A cross-sectional and longitudinal magnetic resonance imaging study of cingulate gyrus gray matter volume abnormalities in first-episode schizophrenia and first episode affective psychosis. *Arch. Gen. Psychiatry* 65, 746–760.

Koutsouleris, N., Meisenzahl, E.M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., et al., 2009. Use of neuroanatomical pattern classification to identify subjects in at risk mental states of psychosis and predict disease transition. *Arch. Gen. Psychiatry* 66, 700–712.

Leonard, C.M., Kuldau, J.M., Breier, J.I., Zuffante, P.A., Gautier, E.R., Heron, D.C., et al., 1999. Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. *Biol. Psychiatry* 46, 374–382.

Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton,M., Carter, C.S., et al., 2004. Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. *Lect. Notes Comput. Sci.* 3216, 393–401.

McDonald, C., Zanelli, J., Rabe-Hesketh, S., Ellison-Wright, I., Sham, P., Kalidindi, S., et al., 2004. Meta-analysis of magnetic resonance imaging brain morphometry studies in bipolar disorder. *Biol. Psychiatry* 56, 411–417.

McDonald, C., Bullmore, E., Sham, P.,

Chitnis, X., Suckling, J., MacCabe, J., et al., 2005. Regional volume deviations of brain structure in schizophrenia and psychotic bipolar disorder: computational morphometry study. *Br. J. Psychiatry* 186, 369–377.

Moore, G.J., Bebchuk, J.M.,Wilds, I.B., Chen, G.,Manji, H.K., 2000. Lithium-induced increase in human brain grey matter. *Lancet* 356, 1241–1242.

Nakamura, K., Kawasaki, Y., Suzuki,M., Hagino, H., Kurokawa, K., Takahashi, T., et al., 2004. Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. *Schizophr. Bull.* 30, 393–404.

Nieuwenhuis, M., van Haren, N.E., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage* 61, 606–612.

Olabi, B., Ellison-Wright, I.,McIntosh, A.M.,Wood, S.J., Bullmore, E., Lawrie, S.M., 2011. Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biol. Psychiatry* 70, 88–96.

Pardo, P.J., Georgopoulos, A.P., Kenny, J.T., Stuve, T.A., Findling, R.L., Schulz, S.C., 2006. Classification of adolescent psychotic disorders using linear discriminant analysis. *Schizophr. Res.* 87, 297–306.

Pohl, K.M., Sabuncu, M.R., 2009. A unified framework for MR based disease classification. *Inf. Process. Med. Imaging* 21, 300–313.

Qiu, A., Vaillant, M., Barta, P., Ratnanather, J.T.,Miller, M.I., 2008. Region-of-interest-based analysis with application of cortical thickness variation of left planum temporale in schizophrenia and psychotic bipolar disorder. *Hum. Brain Mapp.* 29, 973–985.z

Radua, J., Canales-Rodríguez, E.J., Pomarol-Clotet, E., Salvador, R., 2013. Validity of modulation and optimal settings for advanced voxel-based morphometry. *NeuroImage.* http://dx.doi.org/10.1016/j.neuroimage.2013.07.084 (in press).

Rimol, L.M., Hartberg, C.B., Nesvåg, R., Fennema-Notestine, C., Hagler Jr., D.J., Pung, C.J., et al., 2010. Cortical thickness and subcortical volumes in schizophrenia and bipolar disorder. *Biol. Psychiatry* 68, 41–50.

Rimol, L.M., Nesvåg, R., Hagler Jr., D.J., Bergmann, O., Fennema-Notestine, C., Hartberg, C.B., et al., 2012. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biol. Psychiatry* 71, 552–560.

Scheewe, T.W., van Haren, N.E., Sarkisyan, G., Schnack, H.G., Brouwer, R.M., de Glint, M., et al., 2012. Exercise therapy, cardiorespiratory fitness and their effect on brain volumes: a randomised controlled trial in patients with schizophrenia and healthy controls. *Eur. Neuropsychopharmacol.* 23, 675–685.

Selvaraj, S., Arnone, D., Job, D., Stanfield, A., Farrow, T.F.,Nugent, A.C., Scherk, H., Gruber, O., Chen, X., Sachdev, P.S., Dickstein, D.P., Malhi, G.S., Ha, T.H., Ha, K., Phillips, M.L., McIntosh, A.M., 2012. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. *Bipolar Disord.* 14, 135–145.

Shepherd, A.M., Laurens, K.R.,Matheson, S.L., Carr, J.V., Green,M.J., 2012. Systematic meta review and quality assessment of the structural brain alterations in schizophrenia. *Neurosci. Biobehav. Rev.* 36, 1342–1356.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 2, 420–428.

Skre, I., Onstad, S., Torgersen, S., Kringlen, E., 1991. High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). *Acta Psychiatr. Scand.* 84, 167–173.

Smieskova, R., Fusar-Poli, P., Allen, P., Bendfeldt, K., Stieglitz, R.D., Drewe, J., et al., 2009. The effects of antipsychotics on the brain: what have we learnt from structural imaging of schizophrenia? — a systematic review. *Curr. Pharm.* Des. 15, 2535–2549.

Spitzer, R.L., Endicott, J., Robins, E., 1978. Research diagnostic criteria: rationale and reliability. *Arch. Gen. Psychiatry* 35, 773–782.

Steiner, J.L., Tebes, J.K., Sledge, W.H., Walker, M.L., 1995. A comparison of the structured clinical interview for DSM-III-R and clinical diagnoses. *J. Nerv. Ment. Dis.*183, 365–369.

Takayanagi, Y., Takahashi, T., Orikabe, L.,Mozue, Y., Kawasaki, Y.,Nakamura, K., et al., 2011. Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. *PLoS One* 6, e21047.

van den Heuvel, M.P., Sporns, O., Collin, G., Scheewe, T., Mandl, R.C., Cahn, W., Goñi, J., Hulshoff Pol, H.E., Kahn, R.S., 2013. Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry* 70, 783–792.

van der Schot, A.C., Vonk, R., Brouwer, R.M., van Baal, G.C., Brans, R.G., van Haren, N.E., et al., 2010. Genetic and environmental influences on focal brain density in bipolar disorder. *Brain* 133, 3080–3092.

Vapnik, V.N., 1999. An overviewof statistical learning theory. IEEE Trans. *Neural Netw.* 10, 988–999.

Chapter 4

Multi-center MRI prediction models: predicting gender and illness course in first episode psychosis patients.

Mireille Nieuwenhuis| Hugo G. Schnack | Neeltje E. van Haren | Julia Lappin |

Craig Morgan | Antje A. Reinders | Diana Gutierrez-Tordesillas |

Roberto Roiz-Santiañez | Maristela S. Schaufelberger | Pedro G. Rosa |

Marcus V. Zanetti |Geraldo F. Busatto | Benedicto Crespo-Facorro |

Patrick D. McGorry | Dennis Velakoulis | Christos Pantelis | Stephen J. Wood |

René S. Kahn | Janaina Mourao-Miranda | Paola Dazzan

Under review (NeuroImage)

Abstract

Structural Magnetic Resonance Imaging (MRI) studies have attempted to use brain measures to predict outcome in first-episode schizophrenia patients, but results have been inconsistent. This has lead to one of the biggest challenges in the translation of neuroimaging-based findings into clinical practice: the need to validate predictive measures across large independent samples and across data obtained from different MRI scanners and centers.

This study had three main aims: 1) to investigate whether structural MRI data from multiple centers can be combined to create a machine-learning model able to predict gender, a strong biological variable; 2) to replicate previous evidence that illness course in first episode psychosis (FEP) patients can be predicted at the individual level; and finally 3) to investigate whether MRI scans can be used to predict illness course in multi-center models. Finally, in an additional exploratory analysis the study investigated whether reducing sample heterogeneity would improve models.

The complete multi-center sample included brain structural MRI scans from 256 males and 133 females patients with first episode psychosis, acquired across five centers: University Medical Center Utrecht (n=67)(The Netherlands); Institute of Psychiatry, Psychology and Neuroscience (n=97)(London, United Kingdom); University of São Paulo (n=64)(Brazil); the University of Cantabria (n=107)(Santander, Spain) and the University of Melbourne (n=54)(Australia). All images were acquired on 1.5-Tesla scanners and all studies provided information on illness course during a follow-up period ranging 3 to 7 years. We only included in the analyses patients for whom illness course was categorized as “continuous” (n=94 patients) or “remitting” (118 patients).

Using structural brain scans from all centers, gender was predicted with a significant accuracy (89%; p<0.001). In the single- or multi-center models illness course could not be predicted with significant accuracy. However, when reducing heterogeneity by restricting the analyses to male patients only, classification accuracy improved in some samples.

In conclusion, this study provides proof of concept that combining multi-center MRI data to create a well performing classification model is possible. However, to create complex multi-center models that perform accurately, each center should contribute a sample large or homogeneous enough to first allow accurate classification within the single-center.

Introduction

Population-based studies indicate a lifetime prevalence of psychoses above 3% (Perälä and Suvisaari, 2007). While one third of affected individuals experience psychotic symptoms only for a short period of time, others are affected throughout their entire lives (Harrison et al., 2001). Unfortunately, there is no marker to predict, early on, which individuals will develop this incapacitating course. The possibility to identify such individuals at illness onset is important as it could help provide better care to those more at risk.

Schizophrenia, and to a lesser extent other psychotic disorders, are associated with smaller gray matter volume, predominantly in the prefrontal cortex, but also in superior and medial frontal and temporal gyri, insula and thalamus (Fornito et al., 2009; Haijma et al., 2013; Honea et al., 2000; Shepherd et al., 2012). Even following the first psychotic episode, patients show thalamic, insular and hippocampal volume reductions and larger ventricular volume compared to healthy controls (Levitt et al., 2010; Rosa et al., 2010; Schaufelberger et al., 2007; Steen et al., 2006).

Several structural Magnetic Resonance Imaging (MRI) studies have tried to use these brain measures to predict outcome in first-episode schizophrenia patients, but they have varied in sample size, brain regions studied, outcome measures used, and follow-up interval. Most studies did not find significant differences between poor and good outcome patients even when they had adequate power (Molina et al., 2010; van Haren et al., 2003). Others had contradictory findings: a 1-year follow-up outcome study reported a smaller area of the anterior limb of the internal capsule in first-episode schizophrenia (FE-SZ) patients with clinical deterioration compared to those with stable psychopathology, but no differences in either its volume or in any of the other 32 regions of interest (ROI) studied (Wobrock et al., 2009). Another short-term study found that volume of the left dorsolateral prefrontal cortex was predictive of outcome at 1-year, but not at 2-years (Prasad et al., 2005). Studies with a longer follow-up time – 5 to 6 years – more consistently found that smaller initial gray matter volume was predictive of poorer outcome at follow-up (Cahn et al., 2006; Lieberman et al., 2001; Milev et al., 2003). This evidence suggests that studies conducted over a longer follow-up time could potentially achieve better accuracy in the prediction of outcome.

The research summarized above applied a univariate approach to identify brain abnormalities related to subsequent outcome. While this statistical approach allows for inference about regional effects, it does not enable predictions at the level of the individual subject. More recently, machine-learning approaches have shown potential for clinical translation (Fu and Costafreda, 2013), and multivariate pattern recognition techniques have been applied to MRI data for the individualized prediction of clinical characteristics. Pattern recognition is a field within the area of machine learning concerned with automatic discovery of regularities in data through the use of computer algorithms, and using these regularities to take actions such as classifying data into different categories (Bishop, 2006). When applied to data such as structural MRI, brain scans are treated as spatial patterns, and pattern recognition models are used to identify statistical properties of the data, which in turn enable discrimination between groups of subjects, for example patients from healthy subjects (Kambeitz et al., 2015; Klöppel et al., 2012; Nieuwenhuis et al., 2012; Orrù et al., 2012). The models created to date have shown promising results. For example, even in first episode samples, multivariable models have been used to predict diagnoses with accuracies ranging between 79.3% and 91.5% (Karageorgiou et al., 2011; Pohl and Sabuncu, 2009; Sun et al., 2009; Takayanagi et al., 2011, 2010). Furthermore, our own previous work has shown that MRI obtained at time of the first psychotic episode (in a sample of 56 patients) could be used to predict illness course 6 years later with an accuracy of 70% (Mourao-Miranda et al., 2012). This accuracy is substantially higher than that reported in an earlier one-year follow-up study that used pattern recognition to separate remitting (n=15) and not remitting (n=21) patients, which only achieved an accuracy of 58% (Zanetti et al., 2013). Although these results are modest, they support the potential clinical utility of biological markers for the prediction of outcome in schizophrenia.

One of the biggest challenges in the translation of neuroimaging-based findings into clinical practice is the need to validate these models across large independent samples and across data obtained from different MRI scanners (Schnack et al., 2010). This is important to demonstrate robustness with respect to the variability introduced by factors such as scanner type, acquisition protocols and clinical evaluation. In addition, combining multiple samples increases the overall sample size, overcoming a common limitation of existing neuroimaging studies. Recently, several studies on Alzheimer's disease (Dukart et al., 2013; Dyrba et al., 2013; Li et al., 2014) and major depression (Mwangi et al., 2012) have in fact used multi-center data with high classification accuracies, ranging from ~80% to ~90%.

In this study we combined five independent structural MRI datasets from leading international centers for the study of psychosis, to test whether using machine learning on the MRI data acquired at the time of first episode of psychosis (FEP) we could predict illness course 3-7 years later.

This study had three main aims: first, to provide proof of concept for multi-center classification. To achieve this, we tested whether a strong biological outcome like gender could be accurately predicted from data acquired across multiple centers and scanners.

Second, we aimed to replicate previous findings that an MRI scan obtained at first episode can be used to predict illness course 5 years later (Mourao-Miranda et al., 2012; Zanetti et al., 2013). In addition to our original sample from the Institute of Psychiatry in London, this replication was conducted in four single center samples from different research centers (University Medical Center Utrecht; University of São Paulo; University of Cantabria; University of Melbourne). All samples had a similar long duration of follow-up (3-7 years), and the sample size we achieved makes this the largest replication sample yet. Third and last, we aimed to investigate whether the scans obtained at first episode could be used to predict illness course in multicenter-models.

We predicted that combining data and thus creating larger samples would increase the classification accuracy of the models, and that we could replicate earlier illness-course prediction findings in the four additional samples. Moreover, we expected that combining samples from the four different scanners would make it easier for the classifier to learn the classification task at hand and for scanner or site effects to be considered as noise, thus resulting in a more robust model of illness-course and also resulting in higher classification accuracy.

Method

Samples

The overall sample comprised five datasets of patients (total n=389) who had an MRI scan at the time of their first episode of any psychotic illness (including DSM-IV diagnosis of schizophrenia, schizophreniform disorder, schizotypal disorder, schizoaffective disorder, depression with psychotic symptoms, bipolar affective disorder, psychosis not otherwise specified), and who were followed up over a period ranging between three and seven years, when clinical outcome was evaluated. The samples included: n=67 patients from the University Medical Center Utrecht (UMCU The Netherlands) (Cahn et al., 2002); n=97 patients from the Institute of Psychiatry, Psychology and Neuroscience (IoPPN, London, United Kingdom) (Mourao-Miranda et al., 2012); n=64 patients from the University of São Paulo (São Paulo, Brazil) (Schaufelberger et al., 2007); n=107 patients from the University of Cantabria (Santander, Spain) (Crespo-Facorro et al., 2009); and n=54 patients from the University of Melbourne (Melbourne, Australia) (Velakoulis et al., 2006). The samples were derived from well-established studies, the main findings of which have been extensively published elsewhere. All subjects were scanned in a 1.5T scanner (protocol details provided below). All participants gave written informed consent and local ethics committees approved the studies.

Gender groups

Although classifying gender is a complex task, the classification problem is relatively easy, as gender can be unequivocally determined and brain sexual dimorphisms is well established, particularly in the Heschl's gyrus, the planum temporale and the hypocampal formation (Good et al., 2001). The dataset used in this analysis included 133 females and 256 males. To create a model based on a balanced dataset, all 133 females were included, together with a subset of 133 males randomly selected from the overall sample, matched for site. To compare multi-center models to single-center models, five single-center models were created with all the females and a random selection of males of the same size. To reduce chances of selection bias, we built one hundred models with these random selections. Cross-validation (see details below section on pattern recognition) was used to estimate average prediction percentages.

Outcome groups

All studies provided information on the number of episodes that patients had experienced during the follow-up period and on whether they had achieved remission. Centers differed in the instruments used to evaluate illness course during follow-up (Table 2). These included the World Health Organization Life Chart (World Health organization, 1992), Schedules for Clinical Assessment in Neuropsychiatry (SCAN (Wing et al., 1990)), Positive and Negative Syndrome Scale, (PANNS (Kay et al., 1989)), Comprehensive Assessment of Symptoms and History, (CASH (Andreasen et al., 1992)) and Scale for the Assessment of Negative Symptoms (SANS, (Andreasen, 1984)). Illness course was therefore evaluated using a conservative approach to identify two patient groups with an "extreme" type of outcome: one group with a "continuous" illness course (no remission of symptoms of greater than 6 months); and another with a "remitting" illness course (one or more periods of remission of at least 6 months, and no episode lasting longer than 6 months), based on retrospectively available data on the course of illness. All patients who had experienced only a single psychotic episode (lasting no longer than 6 months) were included in the "remitting" group. Three centers (London, Utrecht, São Paulo) provided additional information on duration of the psychotic episodes. When data were available, patients were excluded from the remitting group if the first episode lasted longer than six months. Patients who were neither in the continuous nor in the remitting group (i.e., had a remission and an episode lasting longer than 6 months) were excluded from further analyses on illness course. Within the London, Melbourne and Utrecht samples, 59%, 56% and 58% of all patients were included. In the samples from Santander and São Paulo these percentages were smaller, being 53% and 45% of the entire samples respectively. In total, 94 continuous patients and 118 remitting patients were included in the analyses. Most patients had a diagnosis of schizophrenia (n=114), followed by schizophreniform disorder and schizoaffective disorder (n=39). Other diagnoses included bipolar affective disorder (n=13), brief psychotic disorder (n=5) and depression with psychotic symptoms (n=9). The mean duration of follow-up in years was 6.3 (SD=2.2), 4.9 (SD=0.8), 7.0 (SD=1.4), 3.0 (SD=0.0) and 3.7(SD=0.6) respectively for London, Utrecht, Melbourne, Santander, and São Paulo.

After the exclusion of patients who could not be classified as either continuous or remitting, the sample included 141 male patients (66 remitting and 75 continuous) and 71 females (52 episodic and only 19 continuous). When modeling less heterogeneous samples we only included male patients, since the number of females per illness-course group per center was very small, and the samples size would have been too small to create a female-only model and thus this was not tested (see **Table 1** for details).

MRI protocols and processing

All images were acquired on 1.5-Tesla scanners (see Table 3: information on scanners and protocols per center.). The T1-weighted images were pre-processed using the same processing protocol. All scans were manually oriented into MNI space, after which a non-uniformity correction (Sled et al., 1998) was applied to remove radio frequency (RF) field inhomogeneity. Spatially normalized gray matter probabilities were obtained by running “segment” in SPM8 (Ashburner and Friston, 2005). This method segments, spatially normalizes (modulated normalized) and bias-corrects (10 full width half maximum (FWHM) 150mm) all brains into the same space and dimension (dimension: 91x109x91; voxel size: 2x2x2mm). To reduce noise, all scans were smoothed employing a 4-mm FWHM Gaussian kernel.

To ensure that only actual gray matter was used in the analysis, all voxels with gray matter probabilities below 0.03 for any given subject were excluded from the analysis. This resulted in 170,000 voxels or features per subject being included in the analyses.

Pattern Recognition Analyses

All models were created using a linear Support Vector Machine (SVM) (Vapnik, 1999), which is a supervised machine learning method commonly applied to binary classification problems in neuroimaging (Klöppel et al., 2012; Orrù et al., 2012). In supervised learning approaches a predictive function is “learned” from labeled training data, which is a data set consisting of examples (e.g. gray matter patterns) and labels (e.g. patients or healthy controls). The binary problem in this case consists of classification of the two previously defined groups, for example, female vs. male or remitting vs. continuous. Every subject is represented by its gray matter probability map, which defines a high

Table 1 | Demographic information on the samples

	Kings College London		University medical Center Utrecht		The University of Melbourne		University of Cantabria, Santander		University of São Paulo	
Remitting Continuous	R	C	R	C	R	C	R	C	R	C
Patients included (males)	27 (13)	30 (22)	15 (14)	24 (21)	16 (9)	14 (12)	41 (21)	16 (12)	19 (9)	10 (8)
Age in years (SD)	28.1 (6.3)	29.2 (9.7)	22.2 (3.9)	24.6 (5.3)	21.3 (3.6)	21.6 (3.1)	31.4 (9.0)	29.8 (9.5)	26.2 (8.8)	31.7 (9.2)
DUP in days (SD)	245 (825)	579 (1052)	144 (190)	243 (425)			373 (577)	229 (281)	59 (82)	122 (199)
Schizophrenia diagnosis*	9	22	12	23	0	10	22	10	1	5
Entire sample Males - Females	61	36	58	9	37	17	62	45	38	26

* The number of patients diagnosed with schizophrenia at follow-up per sample per group.

Table 2 | Questionnaires used in each center

	DSM-IV / DSM-V	ICD-10	PANSS	WHO Life chart	SCAN	CASH	GAF	SAPS	SANS
London	x	x		x	x				
Utrecht	x		x			x	x		
Melbourne	x			x					
Santander	x	x						x	x
São Paulo	x		x				x		

dimensional feature vector (in which each voxel in the map corresponds to a feature in the feature-vector). In order to build the binary classifier, the labeled data are used to create a model or decision boundary based on training examples (such as gray matter probability maps from remitting and continuous course patients). In the linear case, this decision boundary corresponds to a hyperplane in the voxel space. The SVM finds the hyperplane that has the largest margin or separation between the two groups, also know as optimal hyperplane (Vapnik, 1999). The advantages of SVM compared to other classification techniques are its scalability and computational efficiency in higher dimensional problems.

A pre-existing implementation (Chang and Lin, 2009) of LIBSVM in Matlab (version 2009b) was used to compute models with a linear kernel. The parameter C was determined through nested cross validation. Nested cross validation has one more loop than normal cross validation (explained below). The inner loop is used for optimizing model parameters and the outer loop is used to estimate model performance based on the test subjects, which were not used during the parameter optimization process.

Cross validation is a technique for estimating model performance using partitions of the sample for training and testing. One part of the data is used for model estimation or training and the other part for model testing. We elected to use a leave-two-out cross-validation framework (L2o), which allows for one subject of each class to be left out for testing and for the remaining subjects from both classes to be used for model creation. In multi-center models pairs that were left out were always from the same center. We bootstrapped one hundred models: each time, a balanced group (as large as possible with an equal amount of subjects per class) was selected randomly, after which a complete leave-two-out cross-validation was performed.

Performance measures

In the gender classification model the percentage of correctly classified females and males was estimated as the correctly classified females divided by all the females, and the correctly classified males divided by all the males.

The performance of the disease course models is reflected in a value per class: the positive and negative predictive accuracies (PPA and NPA). These are the ratios of correctly classified subjects in a specific class true positives (TP) or true negatives (TN), divided by all the subjects classified by the model as belonging to that class TP + false positives (FP) or TN + false negatives (FN):

Positive predictive accuracy = TP / (TP + FP)
Negative predictive accuracy = TN / (TN + FN)

Significance of the models was determined by permutation test. During the permutation test the labels were permuted one thousand times before re-training the models. The occurrence of accuracies equal to or higher than the accuracy of the model that is being tested were counted and then divided by the number of permutations, resulting in their p-value.

Results

Gender classification models

The five single-center model accuracies can be found on the right hand side of **Table 4 (a)**. The centers with the larger samples, London and Santander, performed 5-10% better than the other, smaller sample centers. The left part of the table contains the results of the multi-center-SVM-model, in which data from all centers were combined to train the models. The average accuracy in separating females and males was 89% (range 81% to 94%) (males=88%, females=90% p=0.001). This multi-center model performed as well as the single-center model of the two larger samples (London 89% and Santander 88% vs. multi-center 89%). Interestingly, in the multi-center model, the accuracies of the centers with smaller samples (Utrecht and Melbourne) improved considerably, especially in the percentage of correctly classified females. Overall, the accuracy improved by 2% to 9% when compared to the single-center models. Moreover, the difference in accuracy between correctly predicted males and correctly predicted females was much smaller, indicating a more balanced model, implying that the model was as likely to correctly classify a male as it was to correctly classify a female. The large differences we had seen in the smaller single-center models (Utrecht (20%), Melbourne (9%) and São Paulo (10%)) were reduced to only 6%, 5% and 0.3% respectively in the combined model.

Illness course classification in individual centers

To investigate whether gray matter density at first episode could predict illness course, each dataset was first analyzed individually **(Table 4, b).** When both male and female patients were included, the classification into continuous and remitting course was significant and above chance only in the sample from London at 68% and 70% ($p < 0.02$ and $p < 0.007$ respectively).

To investigate if a less heterogeneous sample would lead to better models, we built new models including only male patients **(Table 4, c)**. This increased the accuracy in the Santander and São Paulo datasets by 11% and 14% respectively. However, it negatively affected the classification accuracy in London (from 69% to 64%), while it did not change the classification for the Utrecht and Melbourne datasets. Only the accuracy in the

Table 3 | Scanner-protocols and scanner-type per center.

							Voxel dimension (mm)**		
	Field strength	**System**	**Sequence**	**Flip angle**	**Repetition time ms**	**Echo time (TE) ms**	**x**	**y**	**z**
University of Cantabria Santander	1.5T	General Electric SIGNA System	SPGR*	45°	24	5	1.02	1.02	1.50
University Medical Center Utrecht	1.5T	Philips	Fast field echo	30°	30	4.6	1.00	1.20	1.00
The University of Melbourne	1.5T	General Electric SIGNA System	SPGR*	30°	14.3	3.3	0.94	0.94	1.50
Kings College London	1.5T	General Electric SIGNA System	SPGR*	20°	13.8	2.8	0.94	1.50	0.94
University of São Paulo	1.5T	General Electric SIGNA System	SPGR*	20°	21.7	5.2	0.86	0.86	1.50

*Spoiled Gradient Recalled Acquisition in Steady State. **All the scans had a coronal acquisition orientation.

sample from São Paulo reached significance (p=0.005), and the episodic patients from London were classified with a significance of p=0.077. The lack of significance in the other samples is probably due to the small sample sizes. The lack if improvement in accuracy after reducing the data heterogeneity suggests that heterogeneity was not the only factor leading to the poor performance of these models. Reducing heterogeneity also led to smaller sample sizes, which could have negatively affected the results.

Illness course classification with multi-center models

In the third and last set of analyses we combined all data into multi-center models to classify illness course. Combining data from all centers into one model did not improve the results obtained from single-center models. The classification accuracy remained at chance level and did not reach significance **(Table 4, b)**.

The less heterogeneous multi-center model, including only male patients, showed that illness course in subjects from London and São Paulo was correctly classified with an average of 62% and 74% respectively (depicted in green and light blue in Figure 1). Unfortunately, combining all five centers did not increase the accuracies in the others centers.

Table 4 | Results of the classification models. The right hand side shows accuracies of the single-center models and the left hand side shows the accuracies of the multi-center models. Part (a) of the table shows the percentage of correctly classified males and females in the gender classification model; (b) shows the negative and positive predictive accuracies of the multi-center and single-center models on illness course classification; (c) shows the less heterogeneous illness course models, including only males.

(a)	Gender classification, males vs. females					
			Combined into one model		Five Seperate models	
	N males	N females	male	female	male	female
London	61	36	93%*	85%*	90%*	88%*
Santander	62	45	91%*	89%*	87%*	90%*
Utrecht	58	9	81%*	87%*	89%*	69%*
Melbourne	37	17	94%*	89%*	87%*	78%
Sao Paulo	38	26	87%*	86%*	89%*	79%
Overall	256	133	90%*	87%*	88%*	81%

(b)	Illness course classification, continuous vs. remitted patients					
Entire sample			Combined into one model		Five Seperate models	
	N continuous	N remitted	PPV	NPV	PPV	NPV
London	30	27	55%	55%	68%	70%
Santander	16	41	44%	45%	44%	42%
Utrecht	24	15	49%	48%	48%	48%
Melbourne	14	16	52%	51%	54%	53%
Sao Paulo	10	19	56%	62%	61%	62%
Overall	94	118	52%	52%	55%	55%

(c)	Illness course classification, continuous vs. remitted patients			
Males only			Combined into one model	
	N continuous	N remitted	PPV	NPV
London	22	13	62	62
Santander	12	21	46	45
Utrecht	21	14	45	45
Melbourne	12	9	55	54
Sao Paulo	8	9	68	80
Overall	75	66	54	54

*significant models with p-value < 0.001

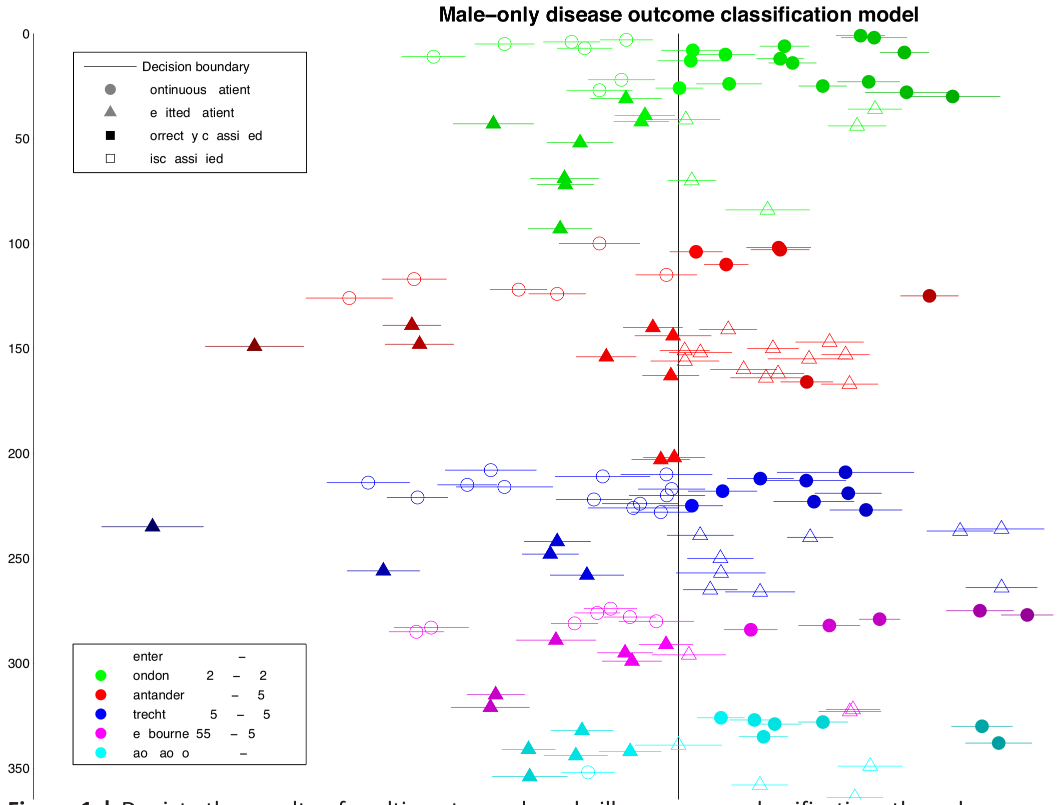

Figure 1 | Depicts the results of multi-center male-only illness course-classification; the colors represent subjects from the different centers. The vertical black line represents the decision boundary. Ideally, all continuous patients (circles) would appear right of the line and the remitting patients (triangles) left of the line.

Discussion

To the best of our knowledge, this is the first study that has investigated whether a multi-center model using neuroimaging data obtained at the first psychotic episode can be reliably combined and used to predict subsequent illness course. Our main finding is that combining multi-center neuroimaging data can lead to a single classifying model that performs well (90% accurate) when classifying strong, unequivocal biological outcomes such as gender. In contrast, our second main finding shows that when classifying clinical characteristics, such as illness course, classification accuracies are only modest, and significant only in centers with most similar definitions of outcome. However, the results also show that multi-center models can be used to increase the performance of smaller and heterogeneous samples. Taken together, these findings suggest that with larger samples, standardized clinical information and clear-cut outcome groups, multi-center-models have the potential to yield generalizable, clinically useful predictions.

Gender classification models

Can we pool data from different centers to increase classification accuracy and create better predictive models? The models based on larger samples (London and Santander) correctly classified up to 90% of individuals from both genders, while those based on smaller samples (Melbourne and São Paulo's) classified males with approximately 88% accuracy, but females only with 10% lower accuracy (female mean 79%). This could be due to the smaller number of females in the models: consistently with previous studies, models based on small samples show larger fluctuations in classification accuracies (Nieuwenhuis et al., 2012). By combining data from all centers (133 females and 133 males) the accuracy of the model increased by approximately 8%, also in those centers with small groups (for example, the Utrecht sample only included 9 females). This confirms our hypothesis that combining datasets, even if acquired on different scanners and from different centers, can improve predictive models. Thus, the limitations of small sample sizes can be overcome by combining multi-center-data, and unbalanced datasets (with fewer subjects of one group and more of the other group) can benefit from a more balanced distribution deriving from the merge with other datasets.

Illness course classification

Our second main finding was that when creating single-center models to predict future outcome, only one center (London) achieved an accuracy significantly above chance (Zanetti et al., 2013). In contrast with the gender classification problem, this clinical question is much more challenging. The low accuracies observed in Utrecht, Melbourne, Santander and São Paulo are probably due to the small sample size in these centers (ranging between 10 – 16 subjects). In addition, illness course is much more heterogeneous than captured by a simple continuous vs. remitting classification, and clinical differences between these two course types could be extremely subtle and difficult to capture. This highlights the importance of using strong, valid and standardized instruments for the classification of outcome (Fu and Costafreda, 2013; Mayberg, 2014).

Multi-center illness course classification

The same modest results were seen when the five centers were combined into one model to predict illness course. Compared to single-center models the variability in multi-center-illness prediction models increases due to the use of multiple scanners and protocols, and to differences across clinical samples. Still, although classifying outcome is more complex than gender, we expected that combining data would overcome outcome variability by increasing sample size.

The difficulty could lie in the single-center classifiers, as the accuracy of the single-center-illness-prediction models was poor. This could indicate that for small number of subjects differences between classes are not large enough to overcome sample variability (i.e. the signal-to-noise ratio was very low). Including these "low signal-to-noise centers" might have increased the noise by adding another center and scanner. One cannot exclude the possibility that if samples had been larger and single-center models had a better performance, the multi-center model would have also been more accurate. Similar findings are also reported in Schnack et al (Schnack et al., 2010), where the influence of noise and sample size in multicenter MRI studies were also examined.

Male-only multi-center illness course classification

To investigate whether making the patient population more homogeneous could improve the models, we repeated the illness course classification analyses including only male patients. Combining all males into a multi-center-illness-prediction model led to an average accuracy of 54% overall, compared to the multi-center-illness-prediction model containing both genders, the model performed slightly better, albeit not significantly. The sample from São Paulo showed a significant improvement in accuracy of 12% of PPA and 17% of NPA compared to the multi-center mixed-gender models ($p < 0.089$ and $p < 0.02$ respectively). São Paulo's continuous group contained mainly male subjects it could thus have benefited most by the exclusion of female data. While a decrease in sample size typically leads to a decrease in accuracy, this may have been compensated by an increase in signal-to-noise ratio in the more homogeneous sample. This increased homogeneity when considering only male patients could be due to a combination of factors. For example, there is evidence of gender-specific brain abnormalities associated with schizophrenia, and of gender-specific differences in illness characteristics, with males being likely to have an earlier age of onset, to experience more severe symptoms and to relapse more frequently (Aleman et al., 2003; Bryant et al., 1999; Goldstein, 2002; Leung and Chue, 2000). Single-gender models could therefore more easily find a "typical" predictive pattern in one gender group during the training process, which would become more difficult in the presence of sexual dimorphism.

Limitations

Together with the heterogeneity introduced by different neuroimaging acquisition parameters, the one introduced by clinical factors may have played a major role in the accuracies we achieved. To this extent, the scales used to assess illness course were not the same across all studies and this may have resulted in heterogeneous and overlapping patient groups. Furthermore, the length of follow up, although long in all groups, ranged from 3 to 7 years and one cannot exclude that illness course may become more established with time. Furthermore, some centers had more clinical information available (for example on duration of psychosis) which could have potentially allowed for more accurate decisions about illness course. These factors highlight the importance of gathering precise and standardized information on the duration and characteristics of each psychotic episode when trying to classify outcome, which could eventually result in better models.

Another limitation of this study is the sample size per center. Even though we had data from 5 centers, we decided to restrict our analyses only to those subjects with the two most extreme illness course types (remitting and continuous), which resulted in some centers having only a small sample size, with a very small number of female subjects. Finally, we only used one brain scan, and it is possible that a better prediction of illness course could be achieved by measuring change over time or studying trajectories of change (Cropley and Pantelis, 2014). This would also eliminate inter-individual baseline differences while also limiting between-scanner differences.

Several clinical baseline markers were studied for their predictive values on outcome in first episode patients. Most yielded no predictive value, but some did; a diagnosis of schizophrenia, shorter education and poor premorbid social adjustment have been shown to predict 3-year outcome with 56% accuracy (Ayesa-Arriola et al., 2013). One study used logistic regression and showed that global functioning (GAF) in the year before admission, total score of the Strauss-Carpenter Prognostic Scale and the PANSS negative sub-score at admission were predictive of symptom remission. The regression model showed a predictive value of about 70% (Jäger et al., 2009). This would suggest that at present, by using MRI only we would be unlikely to predict better than clinicians, and thus highlights the need to improve our approach before MRI becomes clinically useful. Future studies should certainly be based on more detailed and standardized clinical data.

Conclusion

In summary, we provide proof of concept that combining multi-center MRI data to create a single well performing model is possible. Theoretically, multi-center models could lead to more robust accuracy when using brain structure to predict illness course or diagnosis, which would be more generalizable to new patient samples than models based on just a single center. We expect that multi-center models will treat the scanner and acquisition protocols as noise and find effects common to all centers, reducing the risk of misclassification. However, the effect within each single center needs to be strong enough to contribute to the multi-center model. This effectively means that a center has to contribute a sample that is large or homogeneous enough to individually classify with significant accuracy, in order to be of use in generating robust multi-center models.

Acknowledgements

The Melbourne research was supported by grants from the National Health and Medical Research Council of Australia (Program Grants, IDs: 350241; 566529). Christos Pantelis and Patrick McGorry were supported by NHMRC Senior Principal Research Fellowships (ID: 628386; 1060996). The London research was supported by UK Medical Research Council (Ref: G0500817) and the Department of Health via the National Institute for Health Research (NIHR) Specialist Biomedical Research Center for Mental Health award to South London and Maudsley NHS Foundation Trust (SLaM) and the Institute of Psychiatry at King's College London. London

Diana Tordesillas-Gutiérrez is funded by a contract from the Carlos III Health Institute (CA12/00312). A.A.T.S. Reinders was supported by the Netherlands Organization for Scientific Research (www.nwo.nl), NWO-VENI grant no. 451-07-009.

Janaina Mourao-Miranda was supported by the Wellcome Trust under grants no. WT086565/Z/08/Z and no. WT102845/Z/13/Z.

References

Aleman, A., Kahn, R.S., Selten, J.P., 2003. Sex differences in the risk of schizophrenia: evidence from meta-analysis. Arch Gen Psychiatry 60, 565–571. doi:10.1001/archpsyc.60.6.565

Andreasen, N.C., 1984. SCALE FOR THE ASSESSMENT OF (SANS).

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49, 615–623. doi:10.1001/archpsyc.1992.01820080023004

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839¬¬851. doi:10.1016/j.neuroimage.2005.02.018

Ayesa-Arriola, R., Manuel Rodríguez-Sánchez, J., Pérez-Iglesias, R., González-Blanch, C., Pardo-García, G., Tabares-Seisdedos, R., Vázquez-Barquero, J.L., Crespo-Facorro, B., 2013. The relevance of cognitive, clinical and premorbid variables in predicting functional outcome for individuals with first-episode psychosis: A 3 year longitudinal study. Psychiatry Res. 209, 302–308. doi:10.1016/j.psychres.2013.01.024

Bishop, C.M.C.C.M., 2006. Pattern recognition and machine learning, Pattern Recognition. doi:10.1117/1.2819119

Bryant, N.L., Buchanan, R.W., Vladar, K., Breier, A., Rothman, M., 1999. Gender Differences in Temporal Lobe Structures of Patients With Schizophrenia : A Volumetric MRI Study 603–609.

Cahn, W., Pol, H.E.H., Lems, E.B.T.E., van Haren, N.E.M., Schnack, H.G., van der Linden, J.A., Schothorst, P.F., van Engeland, H., Kahn, R.S., 2002. Brain Volume Changes in First-Episode Schizophrenia. Arch. Gen. Psychiatry 59, 1002. doi:10.1001/archpsyc.59.11.1002

Cahn, W., van Haren, N.E.M., Hulshoff Pol, H.E., Schnack, H.G., Caspers, E., Laponder, D. a J., Kahn, R.S., 2006. Brain volume changes in the first year of illness and 5-year outcome of schizophrenia. Br. J. Psychiatry 189, 381–2. doi:10.1192/bjp.bp.105.015701

Chang, C., Lin, C., 2009. LIBSVM : a Library for Support Vector Machines 1–30.

Crespo-Facorro, B., Roiz-Santiáñez, R., Pérez-Iglesias, R., Tordesillas-Gutiérrez, D., Mata, I., Rodríguez-Sánchez, J.M., de Lucas, E.M., Vázquez-Barquero, J.L., 2009. Specific brain structural abnormalities in first-episode schizophrenia. A comparative study with patients with schizophreniform disorder, non-schizophrenic non-affective psychoses and healthy volunteers. Schizophr. Res. 115, 191–201. doi:10.1016/j.schres.2009.09.007

Cropley, V.L., Pantelis, C., 2014. Using longitudinal imaging to map the “relapse signature” of schizophrenia and other psychoses. Epidemiol. Psychiatr. Sci. 23, 219–225. doi:10.1017/S2045796014000341

Dukart, J., Mueller, K., Barthel, H., Villringer, A., Sabri, O., Schroeter, M.L., 2013. Meta-analysis based SVM classification enables accurate detection of Alzheimer’s disease across different clinical centers using FDG-PET and MRI. Psychiatry Res. - Neuroimaging 212, 230–236. doi:10.1016/j.pscychresns.2012.04.007

Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., Meindl, T., Pievani, M., Bokde, A.L.W., Fellgiebel, A., Filippi, M., Hampel, H., Klöppel, S., Hauenstein, K., Kirste, T., Teipel, S.J., 2013. Robust Automated Detection of Microstructural White Matter Degeneration in Alzheimer's Disease Using Machine Learning Classification of Multicenter DTI Data. PLoS One 8. doi:10.1371/journal.pone.0064925

Fornito, A., Yücel, M., Patti, J., Wood, S.J., Pantelis, C., 2009. Mapping grey matter reductions in schizophrenia: An anatomical likelihood estimation analysis of voxel-based morphometry studies. Schizophr. Res. 108, 104–113. doi:10.1016/j.schres.2008.12.011

Fu, C.H.Y., Costafreda, S.G., 2013. Neuroimaging-based biomarkers in psychiatry: Clinical opportunities of a paradigm shift. Can. J. Psychiatry.

Goldstein, J., 2002. Impact of normal sexual dimorphisms on sex differences in structural brain abnormalities in schizophrenia assessed by magnetic resonance imaging. Arch. … 59.

Good, C.D., Johnsrude, I., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. Neuroimage 14, 685–700. doi:10.1006/nimg.2001.0857

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P.C.M.P., Hulshoff Pol, H.E., Kahn, R.S., 2013. Brain volumes in schizophrenia: A meta-analysis in over 18 000 subjects. Schizophr. Bull. 39, 1129–1138. doi:10.1093/schbul/sbs118

Harrison, G., Hopper, K., Craig, T., Laska, E., Siegel, C., Wanderling, J., Dube, K.C., Ganev, K., Giel, R., an der Heiden, W., Holmberg, S.K., Janca, a, Lee, P.W., León, C. a, Malhotra, S., Marsella, a J., Nakane, Y., Sartorius, N., Shen, Y., Skoda, C., Thara, R., Tsirkin, S.J., Varma, V.K., Walsh, D., Wiersma, D., 2001. Recovery from psychotic illness: a 15- and 25-year international follow-up study. Br. J. Psychiatry 178, 506–17.

Honea, R., Sc, B., Crow, T.J., Ph, D., Passingham, D., Mackay, C.E., 2000. Reviews and Overviews Regional Deficits in Brain Volume in Schizophrenia : A Meta-Analysis of Voxel-Based Morphometry Studies i, 2233–2245.

Jäger, M., Riedel, M., Schmauss, M., Laux, G., Pfeiffer, H., Naber, D., Schmidt, L.G., Gaebel, W., Klosterkötter, J., Heuser, I., Kühn, K.-U., Lemke, M.R., Rüther, E., Buchkremer, G., Gastpar, M., Bottlender, R., Strauss, A., Möller, H.-J., 2009. Prediction of symptom remission in schizophrenia during inpatient treatment. World J. Biol. Psychiatry 10, 426–434. doi:781956961 [pii]\r10.1080/15622970701541054

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., Koutsouleris, N., 2015. Detecting Neuroimaging Biomarkers for Schizophrenia: A Meta-Analysis of Multivariate Pattern Recognition Studies. Neuropsychopharmacology 40, 1742–1751. doi:10.1038/npp.2015.22

Karageorgiou, E., Schulz, S.C., Gollub, R.L., Andreasen, N.C., Ho, B.-C., Lauriello, J., Calhoun, V.D., Bockholt, H.J., Sponheim, S.R., Georgopoulos, A.P., 2011. Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. Neuroinformatics 9, 321–33. doi:10.1007/s12021-010-9094-6

Kay, S.R., Opler, L.A., Lindenmayer, J.P., 1989. The Positive and Negative Syndrome Scale (PANSS): Rationale and standardisation. Br. J. Psychiatry. doi:10.1093/schbul/13.2.261

Klöppel, S., Abdulkadir, a, Jack, C., Koutsouleris, N., Mourão-Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61, 457–463. doi:10.1016/j.neuroimage.2011.11.002

Leung, A., Chue, P., 2000. Sex differences in schizophrenia, a review of the literature. Acta Psychiatr. Scand. Suppl. 401, 3–38. doi:10.1111/j.0065-1591.2000.0ap25.x

Levitt, J., Bobrow, L., Lucia, D., Srinivasan, P., 2010. A selective review of volumetric and morphometric imaging in schizophrenia. Curr Top Behav Neurosci. doi:10.1007/7854

Li, M., Oishi, K., He, X., Qin, Y., Gao, F., Mori, S., 2014. An Efficient Approach for Differentiating Alzheimer’s Disease from Normal Elderly Based on Multicenter MRI Using Gray-Level Invariant Features. PLoS One 9, e105563. doi:10.1371/journal.pone.0105563

Lieberman, J., Chakos, M., Wu, H., Alvir, J., Hoffman, E., Robinson, D., Bilder, R., 2001. Longitudinal study of brain morphology in first episode schizophrenia. Biol. Psychiatry 49, 487–499.

Mayberg, H.S., 2014. Neuroimaging and Psychiatry: The Long Road from Bench to Bedside. Hastings Cent. Rep. 44. doi:10.1002/hast.296

Milev, P., Ho, B.-C., Arndt, S., Nopoulos, P., Andreasen, N.C., 2003. Initial magnetic resonance imaging volumetric brain measurements and outcome in schizophrenia: a prospective longitudinal study with 5-year follow-up. Biol. Psychiatry 54, 608–615. doi:10.1016/S0006-3223(03)00293-2

Molina, V., Sanz, J., Villa, R., Pérez, J., González, D., Sarramea, F., Ballesteros, A., Galindo, G., Hernández, J.A., 2010. Voxel-based morphometry comparison between first episodes of psychosis with and without evolution to schizophrenia. Psychiatry Res. 181, 204–10. doi:10.1016/j.pscychresns.2009.09.003

Mourao-Miranda, J., Reinders, a a T.S., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., Morgan, K.D., Fearon, P., Jones, P.B., Doody, G. a, Murray, R.M., Kapur, S., Dazzan, P., 2012. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. Psychol. Med. 42, 1037–47. doi:10.1017/S0033291711002005

Mwangi, B., Ebmeier, K.P., Matthews, K., Douglas Steele, J., 2012. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 135, 1508–1521. doi:10.1093/brain/aws084

Nieuwenhuis, M., van Haren, N.E.M., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage 61, 606–12. doi:10.1016/j.neuroimage.2012.03.079

Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36, 1140–52. doi:10.1016/j.neubiorev.2012.01.004

Perälä, J., Suvisaari, J., 2007. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch. … 64.

Pohl, K.M., Sabuncu, M.R., 2009. A unified framework for MR based disease classification. Inf. Process. Med. Imaging 21, 300–13.

Prasad, K.M.R., Sahni, S.D., Rohm, B.R., Keshavan, M.S., 2005. Dorsolateral prefrontal cortex morphology and short-term outcome in first-episode schizophrenia. Psychiatry Res. 140, 147–55. doi:10.1016/j.pscychresns.2004.05.009

Rosa, P.G.P., Schaufelberger, M.S., Uchida, R.R., Duran, F.L.S., Lappin, J.M., Menezes, P.R., Scazufca, M., McGuire, P.K., Murray, R.M., Busatto, G.F., 2010. Lateral ventricle differences between first-episode schizophrenia and first-episode psychotic bipolar disorder: A population-based morphometric MRI study. World J. Biol. Psychiatry 11, 873–87. doi:10.3109/15622975.2010.486042

Schaufelberger, M.S., Duran, F.L.S., Lappin, J.M., Scazufca, M., Amaro, E., Leite, C.C., de Castro, C.C., Murray, R.M., McGuire, P.K., Menezes, P.R., Busatto, G.F., 2007. Grey matter abnormalities in Brazilians with first-episode psychosis. Br. J. Psychiatry. Suppl. 51, s117–s122. doi:10.1192/bjp.191.51.s117

Schnack, H.G., van Haren, N.E.M., Brouwer, R.M., van Baal, G.C.M., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T.D., Huttunen, M., Lepage, C., Collins, D.L., Evans, A., Murray, R.M., Kahn, R.S., Hulshoff Pol, H.E., 2010. Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. Hum. Brain Mapp. 31, 1967–82. doi:10.1002/hbm.20991

Shepherd, A.M., Laurens, K.R., Matheson, S.L., Carr, V.J., Green, M.J., 2012. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. Neurosci. Biobehav. Rev. 36, 1342–56. doi:10.1016/j.neubiorev.2011.12.015

Sled, J.G., Zijdenbos, a P., Evans, a C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97. doi:10.1109/42.668698

Steen, R.G., Mull, C., McClure, R., Hamer, R.M., Lieberman, J.A., 2006. Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. Br. J. Psychiatry 188, 510–518. doi:10.1192/bjp.188.6.510

Sun, D., van Erp, T.G.M., Thompson, P.M., Bearden, C.E., Daley, M., Kushan, L., Hardt, M.E., Nuechterlein, K.H., Toga, A.W., Cannon, T.D., 2009. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. Biol. Psychiatry 66, 1055–60. doi:10.1016/j.biopsych.2009.07.019

Takayanagi, Y., Kawasaki, Y., Nakamura, K., Takahashi, T., Orikabe, L., Toyoda, E., Mozue, Y., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., Matsushita, M., Suzuki, M., 2010. Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. Prog. Neuropsychopharmacol. Biol. Psychiatry 34, 10–7. doi:10.1016/j.pnpbp.2009.09.004

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., Suzuki, M., 2011. Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. PLoS One 6, e21047. doi:10.1371/journal.pone.0021047

Van Haren, N.E.M., Cahn, W., Hulshoff Pol, H.E., Schnack, H.G., Caspers, E., Lemstra, A., Sitskoorn, M.M., Wiersma, D., van den Bosch, R.J., Dingemans, P.M., Schene, A.H., Kahn, R.S., 2003. Brain volumes as predictor of outcome in recent-onset schizophrenia: a multi-center MRI study. Schizophr. Res. 64, 41–52. doi:10.1016/S0920-9964(03)00018-5

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 988–999. doi:10.1109/72.788640

Velakoulis, D., Wood, S.J., Wong, M.T., McGorry, P.D., Yung, A., Phillips, L., Smith, D., Brewer, W., Proffitt, T., Desmond, P., Pantelis, C., 2006. Hippocampal and amygdala volumes according to psychosis stage and diagnosis. Arch. Gen. Psychiatry 63, 139–149.

Wing, J.K., Babor, T., Brugha, T., Burke, J., Cooper, J.E., Giel, R., Jablenski, A., Regier, D., Sartorius, N., 1990. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch. Gen. Psychiatry 47, 589–593. doi:10.1001/archpsyc.1990.01810180089012

Wobrock, T., Gruber, O., Schneider-Axmann, T., Wölwer, W., Gaebel, W., Riesbeck, M., Maier, W., Klosterkötter, J., Schneider, F., Buchkremer, G., Möller, H.J., Schmitt, A., Bender, S., Schlösser, R., Falkai, P., 2009. Internal capsule size associated with outcome in first-episode schizophrenia. Eur. Arch. Psychiatry Clin. Neurosci. 259, 278–283. doi:10.1007/s00406-008-0867-y

World Health organization, 1992. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. World Heal. Organ. 1–267.

Zanetti, M. V., Schaufelberger, M.S., Doshi, J., Ou, Y., Ferreira, L.K., Menezes, P.R., Scazufca, M., Davatzikos, C., Busatto, G.F., 2013. Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. Prog. Neuro-Psychopharmacology Biol. Psychiatry 43, 116–125. doi:10.1016/j.pnpbp.2012.12.005

Chapter 5

Can machine-learning models based on genetic data be used to classify individuals with schizophrenia and healthy controls?

M. Nieuwenhuis | H.G. Schnack | L. Olde Loohuis | K.R. Van Eijk |

N.E.M. Van Haren | R. Ophoff | R. Kahn

Abstract

Schizophrenia is a highly heritable brain disease. Recent studies have identified a large number of alleles with small effect that increase the risk to develop schizophrenia. Machine learning methods could possibly detect patterns in the genome that distinguish between individual schizophrenia patients and healthy controls. The goal of this study is to investigate if genotype data can be used for individualized prediction of schizophrenia.

Three single nucleotide polymorphisms (SNP)-based support vector machine (SVM) models were constructed on data from 705 cases and 637 controls. First, an unbiased model-driven approach was used, including all SNPs that were available (N=7,164,809 SNPs per individual). To investigate if machine learning models can find interaction between SNPs and compare that to current additive models, the second SVM-model was based on SNPs that showed nominally significant association ($P < 0.05$) in the most recent GWAS study including 36,989 cases and 113,075 controls (N=728,869 SNPs per individual). The third SVM-model was based on a selection of SNPs of the 108 independent known loci associated with schizophrenia (N=74 SNPs per individual). We compare our results against the polygenic risk score (PGRS), which is an additive score based on strength of association and frequency.

When all SNPs were included in the analysis, schizophrenia diagnosis was predicted with an accuracy of 54%. The accuracy of the GWAS-SVM model based on SNPs nominally significantly associated with schizophrenia was 60%. The GWAS-significant SVM-model based on the 74 independent SNPs from the 108 loci found to be genome-wide significantly associated with schizophrenia reached an accuracy of 56%.

The model based on the risk score SNPs had an accuracy that was similar to the predictive value (63% accurate) of the PGRS value in this sample. This indicates that machine learning could in principle be used to detect differences between schizophrenia patients and healthy controls. However, to be clinically beneficial these results are too modest.

Introduction

Schizophrenia is a severe psychiatric disorder with a lifetime prevalence of 0.5% (Simeone et al., 2015). It is known that early and accurate diagnosis of schizophrenia improves illness course and treatment considerably (Perkins et al., 2005). Diagnoses are currently performed through the International Classification of Diseases (ICD) and Diagnostic Statistical Manual of Mental Disorders (DSM). Studies investigating the reliability of diagnosing schizophrenia between psychiatrists find congruence coefficients ranging between 0.56-0.80, while the congruence between DSM-IV and ICD-10 is found to be 0.61 (Cheniaux et al., 2009; Richieri et al., 2011). Compared to congruence rates of other psychiatric disorders such as bipolar disorder those for schizophrenia are low. Therefore, a more objective measure to diagnose schizophrenia would be beneficial to clinicians.

There is consistent evidence that schizophrenia is highly heritable, i.e., approximately 80% (Singh et al., 2014). However, so far, the susceptibility alleles or loci found to be related to an increase risk of schizophrenia cannot account for this high heritability (Manolio et al., 2009; Zuk et al., 2012). The Psychiatric Genomics Consortium (PGC) was founded in 2007 with the goal to increase our understanding on the genetic basis of the illness and it currently includes data of about 40,000 schizophrenia patients. This consortium performs genome-wide association studies (GWAS) initially examining common genetic variants, focusing on associations between single-nucleotide polymorphisms (SNPs). Thousands of SNPs have been found to be weakly associated with schizophrenia, i.e. genotypic relative risk < 1.05 (Purcell et al., 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). However, to this day 108 loci reached genome-wide significance (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). A recent study estimated that SNPs capture about 23% of the variation in liability to schizophrenia, they imply that more individual SNP associations will be detected for this disease as sample size increases (Lee et al. 2012).

A well-established topic in genetics is the polygenic risk score (PGRS). This is a cumulative score constructed from alleles with a low to modest association with schizophrenia. The association between PGRS and relevant phenotypes, such as IQ or brain measures, has been investigated (Derks et al., 2012; French et al., 2015; Power et al., 2015; Terwisscha Van Scheltinga et al., 2013). The PGRS has only found to explain 7% of the variation in liability to schizophrenia (Ripke et al., 2014). This low sensitivity and specificity does not support its use as a predictive test (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

The findings so far are based on statistical analyses that compare groups of patients to groups of healthy controls. Unfortunately, statistical group differences do not imply the possibility to discover deviations from normal in single individuals and therefore do not suffice to aid in diagnosis. Machine-learning methods could possibly detect patterns in the genome that distinguish between individual schizophrenia patients and healthy controls. These discriminating patterns are generated by means of input features; the most commonly used features from genotype data are SNPs. In 2010 the first machine-learning methods were used to try and classify schizophrenia based on about 40-50 SNPs (in the HTR2A and DRD3 genes). A sample of 260 patients and 354 controls reached 66.6% classification accuracy, adding data of simulated controls (N=3070) resulted in an increased accuracy of 93.8% (Aguiar-Pulido et al., 2010). This suggests that large sample sizes are needed in order to classify with high accuracy. One other study used 384 SNPS from 222 genes in 20 patients and 20 controls and classified schizophrenia patients from healthy controls with an accuracy of 74% (Yang et al., 2010).

The objective of this study was to investigate if machine learning can be applied on large numbers of single nucleotide polymorphisms to classify schizophrenia patients from healthy controls in a modest sample size compared to genetic studies, but relatively homogeneous population in The Netherlands. Since this sample is relatively homogenous possible population stratification is minimal. Hidden population structure is unlikely and therefore this sample is suitable for the machine learning approach without further correction. Machine learning can detect SNP-patterns that discriminate between the groups. These patterns not only contain information on SNP importance, but also on whether the particular feature shows a higher value or lower value, in this study a higher or lower minor allele count, in the pattern comparing patients to healthy controls. Consequently, this allows us to make an individual prediction about the presence of the disease. Here, we investigate three SNP-based models on data from 705 cases and 637 controls.

To this day machine-learning studies have only investigated a selection of SNPs, based on a-priori knowledge of association with schizophrenia. These SNP-sets were selected based on univariate methodologies. To fully utilize the multivariate nature of machine learning our first model is a model-driven unbiased approach including all available SNPs after quality control (N=7,164,809 SNPs per individual). The large amount of SNPs compared to the amount of subjects in this study could cause concern for over-fitting. However, because of the nature of the SVM selecting an optimal separating plane i.e. with the largest margin between two classes, and through careful selection of the regularization parameter C (Pattern Recognition Analyses), SVM-models tends to be very resistant to over-fitting. Comparing this model to the GWAS results could give insight into if a machine-learning approach on this relatively small sample will detect similar SNPs as the statistical approach on very large samples. To investigate if a more complex SNP-pattern could predict schizophrenia better than the PGRS the second model was based on over seven hundred thousand SNPs, selected based on that these had the best predictive value in the GWAS polygenic risk score analysis (N=728,869 SNPs per individual). Since there is a pre-selection of features known to be associated with schizophrenia we expect this model to perform better than the first model. The third and final model was based on a selection of 74 independent SNPs from the 108 loci found to be genome-wide significantly associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Since these SNPs are significantly associated with schizophrenia and our sample is much larger than that of previous machine-learning studies, and the predictive power of the PGRS in this sample is better compared to that of other datasets in the PGC we expect this model to perform better than previously published studies. Overall, we expected that these machine-learning based models could surpass the predictive value of the PGRS because of their multivariate nature and the homogeneity of this sample. To our knowledge, this is the largest sample yet to investigate this. Moreover, this is the first machine-learning study that includes all available SNPs to classify schizophrenia patients and healthy controls.

Method

Samples

All individuals are of Dutch ancestry based on the fact that three of their four grandparents were born in The Netherlands, which makes this a homogeneous and therefor advantageous sample. The sample contains 705 cases (males 530 and females 175) and 637 controls (313 males and 324 females). All subjects gave written informed consent. Patients were recruited from a variety of hospitals and institutions in the Netherlands partly by the Genetic Risk and Outcome of Psychosis (GROUP)-investigators (Korver et al., 2012). All cases were diagnosed for subtypes of schizophrenia according to the DSM-IV. Controls were all volunteers and showed no psychopathological abnormalities according to Comprehensive Assessment of Symptoms and History (CASH) (Andreasen et al., 1992). Several studies are already published on these subjects or a selection of these subjects (Buizer-Voskamp et al., 2011; Olde Loohuis et al., 2015; Ripke et al., 2013, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

SNP Data Collection and Preprocessing

The subjects were genotyped at the UCLA Neurosciences Genomics Core (UNGC). The genotypes were imputed and quality control was performed by the PGC. Common quality control parameters were applied, which included selection of all SNPs with an imputation Information content metric score (INFO- score) ≥ 0.1, minor allele frequency (MAF) ≥0.005. A filter for missingness < 0.02 (call with p > 0.8) was applied, which means that no more than 2% of the subjects can have a missing value or imputed value with a probability larger than 0.8. From the dose files genotypes with high probability were inferred. We acquired the resulting imputed data i.e. single nucleotide polymorphisms (SNPs) of chromosomes 1 through 22. The X chromosome was excluded to prevent males and females to cluster differently.

Using Plink, all genotypes were recoded additively, which reflects the number of minor alleles. The minor alleles are the least common alleles in a given population. All ATCG values were recoded to 0, 1 or 2, with 0 indicating that both alleles are major alleles 1 indicates that one of the alleles is a minor allele and 2 indicates that both alleles are minor alleles. All missing values were defined as 0. In the last preprocessing step, we excluded all insertions and deletions to compare only similar features per individual. The number of SNPs after these preprocessing steps was 7,164,809 per individual.

SNP selection based on polygenic risk score (PGRS)

The GWAS-SVM model is based on the SNPs that were found nominally significantly associated with schizophrenia at a p-value cutoff < 0.05 (in the GWAS). This was the threshold with the highest predictive value according to (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). After selecting all the SNPs in our sample according to this threshold each individual was represented by 728,869 SNPs

SNP selection based on significant GWAS SNPs

The most significant SNPs of the 108 independent known loci associated with schizophrenia are a total of 128 linkage-disequilibrium-independent SNPs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Sixteen of these SNPs are insertions and deletions, which were removed in our sample; three are located at the X chromosome, which was excluded and 35 of these SNPs did not pass the filter with missing < 0.02 (call with P > 0.8). To find if our sample included linkage disequilibrium proxy SNPs of these excluded SNPs, a thorough search with SNAP (SNP annotation and proxy search) was performed. The search threshold was set to r2<0.8 and the distance limit was set to maximum (500), no SNPs were found. This resulted in a set of 74 SNPs per individual after quality control.

Pattern Recognition Analyses

All models were created using a linear Support Vector Machine (SVM) (Vapnik, 1999), which is a supervised machine learning method commonly applied to binary classification problems. In supervised learning approaches a predictive function is "learned" from labeled training data. The problem in this case consists of classification of the two previously defined groups, or classes, schizophrenia patient vs. healthy control. Every subject is represented by single nucleotide polymorphisms (SNPs), which defines a high-dimensional feature-vector (in which each SNP corresponds to a feature in the feature-vector). In order to build the binary classifier, the labeled data are used to create a model or decision boundary based on training examples. In the linear case, this decision boundary corresponds to a hyperplane in the feature space. The SVM finds the hyperplane that has the largest margin or separation between the two groups, also know as optimal separating hyperplane (Vapnik, 1999). This hyperplane can be expressed by an intercept term b and a normal vector w that is commonly referred to as the weight-vector. The advantages of SVM compared to other classification techniques are its scalability and computational efficiency in higher dimensional problems. We integrated LIBSVM (Chang and Lin, 2011) with our own software to carry out the classification.

The parameter C, which penalizes classification error, was determined through nested cross validation. Nested cross validation has one more loop than normal cross validation (explained below). The inner loop is used for optimizing model parameters and the outer loop is used to estimate model performance based on the test subjects, which were not used during the parameter optimization process.

Cross validation is a technique for estimating model performance using partitions of the sample for training and testing. One part of the data is used for model estimation or training and the other part for model testing. We elected to use a leave-ten-out cross-validation framework (L10o), which allows for five subjects of each class to be left out for testing and for the remaining subjects from both classes to be used for model creation. We bootstrapped several models: each time, a balanced group (as large as possible with an equal amount of subjects per class) was selected randomly, after which a complete leave-ten-out cross-validation was performed.

Outcome measures

The accuracy of the different models is assessed by three statistical measures, sensitivity (true positive rate), specificity (true negative rate) and the average accuracy:

Sensitivity = TP / (TP + FP), where TP is the number of true positives (correctly classified patients), and FP is it the number of false positives.

Specificity = TN / (TN + FN), where TN is the number of true negatives and FN is the number of false negatives.
Average accuracy = (sensitivity + specificity) / 2.

Predictive value of the PGRS in this sample

We obtained the polygenic risk score of 1307 subjects (700 cases and 607 controls) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and used this as bench mark for our SVM prediction models. To do so, we calculated the individualized predictive value of the PGRS. For this we implemented a leave-one-out cross validation procedure. All subjects but one were used to determine the best cut-off parameter to divide the group in cases and controls, the left-out subject was then classified according to this cut-off value and as such used to calculate the predictive accuracy. This process was repeated until all subjects were left out once.

Results

The full SNP SVM-model classified 54% subjects accurately (Table 1). Figure 1 shows the weight-vector of this model sorted by weight and it shows the positions of the GWAS-significant SNPs. Forty-five percent of the GWAS-significant SNPs (n=33) was present in the top 15% of the weight-vector. Seventy percent of the SNPs (n=52) were present in the top 30% of the weight-vector. The polygenic Risk (PGRS) GWAS-SVM model had a sensitivity of 66% and a specificity of 57%. The PGRS-value by itself led to an accuracy of 63%.

The GWAS-significant SVM-model had a sensitivity and specificity of 60%.

Table 1 | the leave-one-out results of the three support-vector-machine models. The top row shows the results based on all SNPs available after quality control. The second row shows the results based on the SNPs that were used in the GWAS analysis to calculate the PGRS and the third row shows the results of the selection of 74 SNPs that reached genome wide significant association with schizophrenia.

Models	Sensitivity	Specificity	Average
All SNPs	54%	55%	54%
GWAS-SVM 728869	60%	60%	60%
GWAS-significant SVM 74	58%	54%	56%

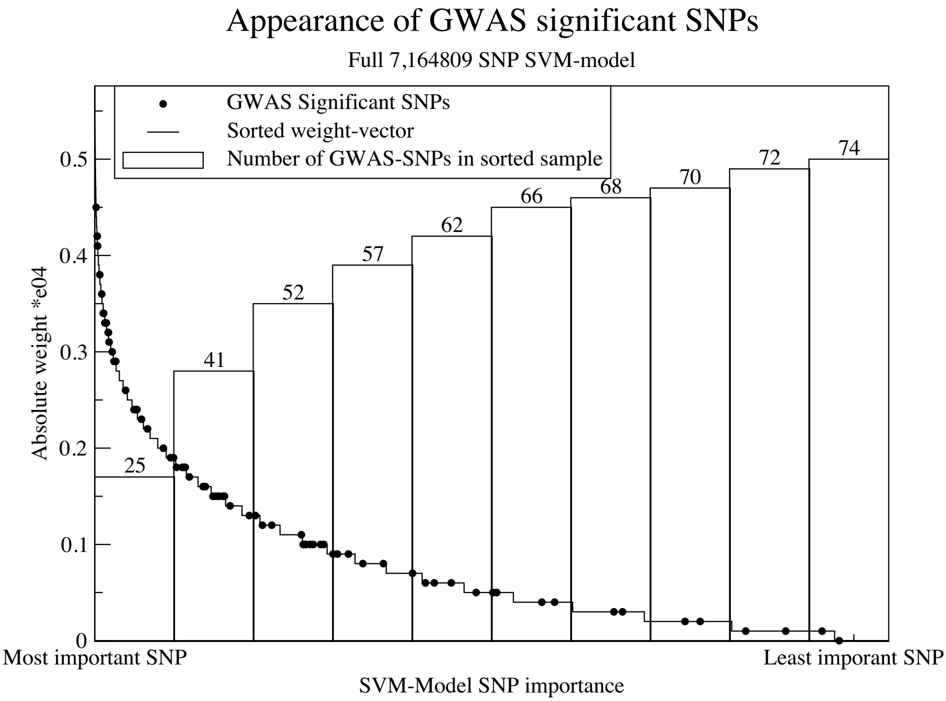

Figure 1 | Appearance of the GWAS significant SNPs in the full SVM-SNP model. Depicted on the y-axis is the absolute weight of the SNPs in the SVM-model. On the x-axis are the SNPs sorted by importance (absolute weight). The SNPs on the left hand side are the most important and on the right hand side are the least important. The blue stars are the 74 GWAS significant SNPs. The bars depict how many of the 74 SNPs are in the sample at that particular location of the sorted w-vector. The bars have a step size of 10% of all the features.

Discussion

To the best of our knowledge, this is the first study to apply machine learning on more than seven million common genetic variants (single nucleotide polymorphisms; SNPs) to predict the presence of illness in a large sample of schizophrenia patients and healthy controls (n=1342). In addition to the full SNP-model, we created two models based on a selection of SNPs known to be associated with schizophrenia. We found accuracies ranging between 56-60%. Our main findings indicate that machine learning can predict illness status to the same extent as the polygenic risk score (PGRS) alone (63% accurate). The full SNP-model did not perform well enough to be clinically valuable.

To this day, no substantial evidence has been found in favor of non-additive effects in schizophrenia genetics (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). However, one of the possible causes of missing heritability is that estimates of total heritability implicitly assume the trait involves no genetic interactions (Zuk et al., 2012). Investigating non-additive relationships in a more homogeneous population like this sample might yield noteworthy results.

We used a support vector machine (SVM) to create three models based on a sample of patients and control subjects. The first model was based on all available SNPs per individual after quality control. Sensitivity and specificity of this model was 54% and 55% respectively. That yields an average accuracy of only 54%, which is not nearly accurate enough to be beneficial for clinical diagnoses. A SVM model generates information about if a certain SNP has a decreased or increased number of minor alleles when comparing patients to controls, this is however beyond the scope of this study. Additionally each SNP has a particular weight in the model, this weight gives an indication of how much influence a SNP has. Even though classification accuracy was low, when investigating where in this model the genome-wide association study (GWAS) significant SNPs were located, we found that more than 70% of these SNPs were present in the top 30% of the weight-vector. This indicates that a SVM may be sensitive to detect the majority of important discriminating SNPs. However, to verify this similar analyses in larger samples are needed. A direct comparison of these models might not be completely fair due to the large differences in ratio of subjects and SNPs. Increasing the number of subjects and thus increasing the ratio could increase the accuracy significantly (Aguiar-Pulido et al., 2013).

The second model was based on 728.869 SNPs that had the best predictive value in the GWAS-PGRS analysis. This GWAS-SVM model yielded a sensitivity and specificity of 60%. This result is comparable to the predictive value (63% accurate) of the stand-alone PGRS in this sample. These comparable results are not surprising if you take into account that no epistasis has been observed before (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

The final model was based on a selection of 74 independent SNPs that were at the level of genome-wide significance related to schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Sensitivity and specificity of this model was 58% and 54% respectively. Even though only 74 SNPs were included, the model still classified the subjects with 56% accuracy, which is just above chance level. Since these SNPs are known to explain approximately half of the genetic variation currently explained by SNPs (Ripke et al., 2014) these results are remarkably high. However, if more independent SNPs are found that reach GWAS-significance and these were added to this model, these results could well improve and potentially lead to a future diagnostic tool. We estimate that the best performing SNP SVM-classifier could be one that includes a number of SNPs somewhere between the numbers in the PGRS-SVM model and in this model. Another interesting research area would be combining phenotypic data such as MRI with genetic data; this could possibly reveal interesting interactions. However, for this to be advanced further the genetic models have to be improved first.

Limitations

Schizophrenia has a high heritability, but studies on only common genetic variants have had no success in explaining this heritability. Several potential sources have been suggested that may explain this ‘missing’ heritability, such as gene-environment interactions, epigenetic variation, and rare genetic variation (Manolio et al., 2009; Zuk et al., 2012). To investigate how much ‘missing ’ heritability these rare variants possibly explain, they should be included in the SVM-models. This was beyond the scope of this study.

In this paper the insertions and deletions were left out to keep all features of similar form. Coding SNPs in a dichotomous fashion instead of counting the number of minor alleles could be a method to keep similar features and include insertions and deletions. This, however, would result in a three-fold increase of the data because than three features instead of one are needed to represent one SNP.

Only 74 independent SNPs from the 108 loci known to be associated with schizophrenia were present in our sample. Including more SNPs might improve the results considerably.

The current sample is part of the Psychiatric Genetics Consortium (PGC) and is thus included in the GWAS polygenic risk score analysis (Ripke et al., 2014). The SNPs that were GWAS significant were calculated on a sample including our sample. Hence, we may overestimate the predictive value of the GWAS-SVM model and the GWAS-significant SVM model. Even though we expect that excluding our 705 patients from the analyses in 30,000 patients will not change the results, for future research on samples that are in the PGC sample we would suggest that the 108 loci and the GWAS significant SNPs are calculated excluding the sample that is used in that particular study. This way the results used to select SNPs would be independent of the sample.

Due to the modest results of the full model we have not investigated what SNPs drove the classification or if SNPs showed an increase or decrease in minor allele count in schizophrenia patients relative to controls. This information is available with support vector machines and could hold interesting information for future research.

The strength of our approach lies in the large sample (relative to the samples in earlier SVM studies on genes) and the large number of SNPs included in the model. However, the number of subjects is still small relative to the number of included SNPs, therefore, there is a need to replicate these findings in larger samples.

Conclusion

We show that machine learning is a technique that can be used to classify schizophrenia patients and healthy controls from genotype data with significant accuracies between 56%

and 60%. However, sensitivity and specificity are currently not adequate to be of diagnostic value. The use of larger samples to create a full SNP-SVM model may reveal interesting differentiating SNP-patterns, which could be used to develop an objective diagnostic tool for schizophrenia. Moreover, in contrast to brain-imaging models, SNP models can be applied to subjects at a very young age and might thus be used to predict if an individual is at risk to develop schizophrenia very early in life.

References

Aguiar-Pulido, V., Gestal, M., Fernandez-Lozano, C., Rivero, D., Munteanu, C.R., 2013. Applied computational techniques on schizophrenia using genetic mutations. Curr. Top. Med. Chem. 13, 675–84. doi:10.2174/1568026611313050010

Aguiar-Pulido, V., Seoane, J. a, Rabuñal, J.R., Dorado, J., Pazos, A., Munteanu, C.R., 2010. Machine learning techniques for single nucleotide polymorphism--disease classification models in schizophrenia. Molecules 15, 4875–89. doi:10.3390/molecules15074875

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH). An instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49, 615–623. doi:10.1001/archpsyc.1992.01820080023004

Buizer-Voskamp, J.E., Muntjewerff, J.-W., Strengman, E., Sabatti, C., Stefansson, H., Vorstman, J.A.S., Ophoff, R.A., 2011. Genome-Wide Analysis Shows Increased Frequency of Copy Number Variation Deletions in Dutch Schizophrenia Patients. Biol. Psychiatry. doi:10.1016/j.biopsych.2011.02.015

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27. doi:10.1145/1961189.1961199

Cheniaux, E., Landeira-Fernandez, J., Versiani, M., 2009. The diagnoses of schizophrenia, schizoaffective disorder, bipolar disorder and unipolar depression: Interrater reliability and congruence between DSM-IV and ICD-10. Psychopathology 42, 293–298. doi:10.1159/000228838

Derks, E.M., Vorstman, J. a S., Ripke, S., Kahn, R.S., Ophoff, R. a, 2012. Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis. PLoS One 7, e37852. doi:10.1371/journal.pone.0037852

French, L., Gray, C., Leonard, G., Perron, M., Pike, G.B., Richer, L., Séguin, J.R., Veillette, S., Evans, C.J., Artiges, E., Banaschewski, T., Bokde, A.W.L., Bromberg, U., Bruehl, R., Buchel, C., Cattrell, A., Conrod, P.J., Flor, H., Frouin, V., Gallinat, J., Garavan, H., Gowland, P., Heinz, A., Lemaitre, H., Martinot, J.-L., Nees, F., Orfanos, D.P., Pangelinan, M.M., Poustka, L., Rietschel, M., Smolka, M.N., Walter, H., Whelan, R., Timpson, N.J., Schumann, G., Smith, G.D., Pausova, Z., Paus, T., 2015. Early Cannabis Use, Polygenic Risk Score for Schizophrenia and Brain Maturation in Adolescence. JAMA Psychiatry 1–10. doi:10.1001/jamapsychiatry.2015.1131

Korver, N., Quee, P.J., Boos, H.B., Simons, C.J., de Haan, L., 2012. Genetic Risk and Outcome of Psychosis (GROUP), a multi-site longitudinal cohort study focused on gene-environment interaction: objectives, sample characteristics, recruitment and assessment methods. Int J Methods Psychiatr Res 21, 205–221. doi:10.1002/mpr.1352

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

Olde Loohuis, L., Vorstman, J. a. S., Ori, A.P., Staats, K. a., Wang, T., Richards, A.L., Leonenko, G., Walters, J.T., DeYoung, J., Kahn, R.S., Linszen, D., Os, J. Van, Wiersma, D., Bruggeman, R., Cahn, W., Haan, L. De, Krabbendam, L., Myin-Germeys, I., Cantor, R.M., Ophoff, R. a., 2015. Genome-wide burden of deleterious coding variants increased in schizophrenia. Nat. Commun. 6, 7501. doi:10.1038/ncomms8501

Perkins, D.O., Gu, H., Boteva, K., Lieberman, J.A., 2005. Relationship between duration of untreated psychosis and outcome in first-episode schizophrenia: A critical review and meta-analysis. Am. J. Psychiatry. doi:10.1176/appi.ajp.162.10.1785

Power, R.A., Steinberg, S., Bjornsdottir, G., Rietveld, C.A., Abdellaoui, A., Nivard, M.M., Johannesson, M., Galesloot, T.E., Hottenga, J.J., Willemsen, G., Cesarini, D., Benjamin, D.J., Magnusson, P.K.E., Ullén, F., Tiemeier, H., Hofman, A., van Rooij, F.J.A., Walters, G.B., Sigurdsson, E., Thorgeirsson, T.E., Ingason, A., Helgason, A., Kong, A., Kiemeney, L.A., Koellinger, P., Boomsma, D.I., Gudbjartsson, D., Stefansson, H., Stefansson, K., 2015. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. Nat. Neurosci. 18, 953–955. doi:10.1038/nn.4040

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., 2009. xxCommon polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–52. doi:10.1038/nature08185

Richieri, R. (Assistance P.-H. de M.-P.P.H.S.M.-13009 M., Boyer, L., Lancon, C., 2011. [Analysis of the reliability of diagnostic criteria and classifications in psychiatry]. Sante Publique (Paris). 23, S31–S38.

Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P. a., Lee, P., Bulik-Sullivan, B., Collier, D. a., Huang, H., Pers, T.H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. a., Begemann, M., Belliveau Jr, R. a., Bene, J., Bergen, S.E., Bevilacqua, E., Bigdeli, T.B., Black, D.W., Bruggeman, R., Buccola, N.G., Buckner, R.L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R.M., Carr, V.J., Carrera, N., Catts, S. V., Chambert, K.D., Chan, R.C.K., Chen, R.Y.L., Chen, E.Y.H., Cheng, W., Cheung, E.F.C., Ann Chong, S., Robert Cloninger, C., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J.J., Curtis, D., Davidson, M., Davis, K.L., Degenhardt, F., Del Favero, J., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A.H., Farrell, M.S., Frank, J., Franke, L., Freedman, R., Freimer, N.B., Friedl, M., Friedman, J.I., Fromer, M., Genovese, G., Georgieva, L., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J.I., Golimbet, V., Gopal, S., Gratten, J., de Haan, L., Hammer, C., Hamshere, M.L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A.M., Henskens, F. a., Herms, S., Hirschhorn, J.N., Hoffmann, P., Hofman, A., Hollegaard, M. V., Hougaard, D.M., Ikeda, M., Joa, I., Julià, A., Kahn, R.S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M.C., Kennedy, J.L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J. a., Konte, B., Kucinskas, V., Ausrele Kucinskiene, Z., Kuzelova-Ptackova, H., Kähler, A.K., Laurent, C., Lee Chee Keong, J., Hong Lee, S., Legge, S.E., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C.M., Lubinski, J., Lönnqvist, J., Macek Jr, M., Magnusson, P.K.E., Maher, B.S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R.W., McDonald, C., McIntosh, A.M., Meier, S., Meijer, C.J., Melegh, B., Melle, I., Mesholam-Gately, R.I., Metspalu, A., Michie, P.T., Milani, L., Milanova, V., Mokrab, Y., Morris, D.W., Mors, O., Murphy, K.C., Murray, R.M., Myin-Germeys, I., Müller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D. a., Nestadt, G., Nicodemus, K.K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F.A., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Endophenotypes International Consortium, P., Pantelis, C., Papadimitriou, G.N., Papiol, S., Parkhomenko, E., Pato, M.T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D.O., Pietiläinen, O., Pimm, J., Pocklington, A.J., Powell, J., Price, A., Pulver, A.E., Purcell, S.M., Quested, D., Rasmussen, H.B., Reichenberg, A., Reimers, M. a., Richards, A.L., Roffman, J.L., Roussos, P., Ruderfer, D.M., Salomaa, V., Sanders, A.R., Schall, U., Schubert, C.R., Schulze, T.G., Schwab, S.G., Scolnick, E.M., Scott, R.J., Seidman, L.J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J.M., Sim, K., Slominsky, P., Smoller, J.W., So, H.-C., Spencer, C. a., Stahl, E. a., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R.E., Strengman, E., Strohmaier, J., Scott Stroup, T., Subramaniam, M., Suvisaari, J., Svrakic, D.M., Szatkiewicz, J.P., Söderman, E., Thirumalai, S., Toncheva, D., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B.T., Weiser, M., Wildenauer, D.B., Williams, N.M., Williams, S., Witt, S.H., Wolen, A.R., Wong, E.H.M., Wormley, B.K., Simon Xi, H., Zai, C.C., Zheng, X., Zimprich, F., Wray, N.R., Stefansson, K., Visscher, P.M., Trust Case-Control Consortium, W., Adolfsson, R., Andreassen, O. a., Blackwood, D.H.R., Bramon, E., Buxbaum, J.D., Børglum, A.D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C.M., Iwata, N., Jablensky, A. V., Jönsson, E.G., Kendler, K.S., Kirov, G., Knight, J., Lencz, T., Levinson, D.F., Li, Q.S., Liu, J., Malhotra, A.K., McCarroll, S. a., McQuillin, A., Moran, J.L., Mortensen, P.B., Mowry, B.J., Nöthen, M.M., Ophoff, R. a., Owen, M.J., Palotie, A., Pato, C.N., Petryshen, T.L., Posthuma, D., Rietschel, M., Riley, B.P., Rujescu, D., Sham, P.C., Sklar, P., St Clair, D., Weinberger, D.R., Wendland, J.R., Werge, T., Daly, M.J., Sullivan, P.F., O'Donovan, M.C., 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature. doi:10.1038/nature13595

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., Kim, Y., Lee, S.H., Magnusson, P.K.E., Sanchez, N., Stahl, E. a, Williams, S., Wray, N.R., Xia, K., Bettella, F., Borglum, A.D., Bulik-Sullivan, B.K., Cormican, P., Craddock, N., de Leeuw, C., Durmishi, N., Gill, M., Golimbet, V., Hamshere, M.L., Holmans, P., Hougaard, D.M., Kendler, K.S., Lin, K., Morris, D.W., Mors, O., Mortensen, P.B., Neale, B.M., O'Neill, F. a, Owen, M.J., Milovancevic, M.P., Posthuma, D., Powell, J., Richards, A.L., Riley, B.P., Ruderfer, D., Rujescu, D., Sigurdsson, E., Silagadze, T., Smit, A.B., Stefansson, H., Steinberg, S., Suvisaari, J., Tosato, S., Verhage, M., Walters, J.T., Levinson, D.F., Gejman, P. V, Laurent, C., Mowry, B.J., O'Donovan, M.C., Pulver, A.E., Schwab, S.G., Wildenauer, D.B., Dudbridge, F., Shi, J., Albus, M., Alexander, M., Campion, D., Cohen, D., Dikeos, D., Duan, J., Eichhammer, P., Godard, S., Hansen, M., Lerer, F.B., Liang, K.-Y., Maier, W., Mallet, J., Nertney, D. a, Nestadt, G., Norton, N., Papadimitriou, G.N., Ribble, R., Sanders, A.R., Silverman, J.M., Walsh, D., Williams, N.M., Wormley, B., Arranz, M.J., Bakker, S., Bender, S., Bramon, E., Collier, D., Crespo-Facorro, B., Hall, J., Iyegbe, C., Jablensky, A., Kahn, R.S., Kalaydjieva, L., Lawrie, S., Lewis, C.M., Linszen, D.H., Mata, I., McIntosh, A., Murray, R.M., Ophoff, R. a, Van Os, J., Walshe, M., Weisbrod, M., Wiersma, D., Donnelly, P., Barroso, I., Blackwell, J.M., Brown, M. a, Casas, J.P., Corvin, A.P., Deloukas, P., Duncanson, A., Jankowski, J., Markus, H.S., Mathew, C.G., Palmer, C.N. a, Plomin, R., Rautanen, A., Sawcer, S.J., Trembath, R.C., Viswanathan, A.C., Wood, N.W., Spencer, C.C. a, Band, G., Bellenguez, C., Freeman, C., Hellenthal, G., Giannoulatou, E., Pirinen, M., Pearson, R.D., Strange, A., Su, Z., Vukcevic, D., Langford, C., Hunt, S.E., Edkins, S., Gwilliam, R., Blackburn, H., Bumpstead, S.J., Dronov, S., Gillman, M., Gray, E., Hammond, N., Jayakumar, A., McCann, O.T., Liddle, J., Potter, S.C., Ravindrarajah, R., Ricketts, M., Tashakkori-Ghanbaria, A., Waller, M.J., Weston, P., Widaa, S., Whittaker, P., McCarthy, M.I., Stefansson, K., Scolnick, E., Purcell, S., McCarroll, S. a, Sklar, P., Hultman, C.M., Sullivan, P.F., 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. 45, 1150–9. doi:10.1038/ng.2742

Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., Andreassen, O.A., Scolnick, E., Cichon, S., St. Clair, D., Corvin, A., Gurling, H., Werge, T., Rujescu, D., Blackwood, D.H.R., Pato, C.N., Malhotra, A.K., Purcell, S., Dudbridge, F., Neale, B.M., Rossin, L., Visscher, P.M., Posthuma, D., Ruderfer, D.M., Fanous, A., Stefansson, H., Steinberg, S., Mowry, B.J., Golimbet, V., De Hert, M., Jönsson, E.G., Bitter, I., Pietiläinen, O.P.H., Collier, D.A., Tosato, S., Agartz, I., Albus, M., Alexander, M., Amdur, R.L., Amin, F., Bass, N., Bergen, S.E., Black, D.W., Børglum, A.D., Brown, M.A., Bruggeman, R., Buccola, N.G., Byerley, W.F., Cahn, W., Cantor, R.M., Carr, V.J., Catts, S. V, Choudhury, K., Cloninger, C.R., Cormican, P., Craddock, N., Danoy, P.A., Datta, S., de Haan, L., Demontis, D., Dikeos, D., Djurovic, S., Donnelly, P., Donohoe, G., Duong, L., Dwyer, S., Fink-Jensen, A., Freedman, R., Freimer, N.B., Friedl, M., Georgieva, L., Giegling, I., Gill, M., Glenthøj, B., Godard, S., Hamshere, M., Hansen, M., Hansen, T., Hartmann, A.M., Henskens, F.A., Hougaard, D.M., Hultman, C.M., Ingason, A., Jablensky, A. V, Jakobsen, K.D., Jay, M., Jürgens, G., Kahn, R.S., Keller, M.C., Kenis, G., Kenny, E., Kim, Y., Kirov, G.K., Konnerth, H., Konte, B., Krabbendam, L., Krasucki, R., Lasseter, V.K., Laurent, C., Lawrence, J., Lencz, T., Lerer, F.B., Liang, K.-Y., Lichtenstein, P., Lieberman, J.A., Linszen, D.H., Lönnqvist, J., Loughland, C.M., Maclean, A.W., Maher, B.S., Maier, W., Mallet, J., Malloy, P., Mattheisen, M., Mattingsdal, M., McGhee, K.A., McGrath, J.J., McIntosh, A., McLean, D.E., McQuillin, A., Melle, I., Michie, P.T., Milanova, V., Morris, D.W., Mors, O., Mortensen, P.B., Moskvina, V., Muglia, P., Myin-Germeys, I., Nertney, D.A., Nestadt, G., Nielsen, J., Nikolov, I., Nordentoft, M., Norton, N., Nöthen, M.M., O'Dushlaine, C.T., Olincy, A., Olsen, L., O'Neill, F.A., Ørntoft, T.F., Owen, M.J., Pantelis, C., Papadimitriou, G., Pato, M.T., Peltonen, L., Petursson, H., Pickard, B., Pimm, J., Pulver, A.E., Puri, V., Quested, D., Quinn, E.M., Rasmussen, H.B., Réthelyi, J.M., Ribble, R., Rietschel, M., Riley, B.P., Ruggeri, M., Schall, U., Schulze, T.G., Schwab, S.G., Scott, R.J., Shi, J., Sigurdsson, E., Silverman, J.M., Spencer, C.C.A., Stefansson, K., Strange, A., Strengman, E., Stroup, T.S., Suvisaari, J., Terenius, L., Thirumalai, S., Thygesen, J.H., Timm, S., Toncheva, D., van den Oord, E., van Os, J., van Winkel, R., Veldink, J., Walsh, D., Wang, A.G., Wiersma, D., Wildenauer, D.B., Williams, H.J., Williams, N.M., Wormley, B., Zammit, S., Sullivan, P.F., O'Donovan, M.C., Daly, M.J., Gejman, P. V, 2011. Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. doi:10.1038/ng.940

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–7. doi:10.1038/nature13595

Simeone, J.C., Ward, A.J., Rotella, P., Collins, J., Windisch, R., 2015. An evaluation of variation in published estimates of schizophrenia prevalence from 1990–2013: a systematic literature review. BMC Psychiatry 15, 193. doi:10.1186/s12888-015-0578-7

Singh, S., Kumar, A., Agarwal, S., Phadke, S.R., Jaiswal, Y., 2014. Genetic insight of schizophrenia: Past and future perspectives. Gene. doi:10.1016/j.gene.2013.09.110

Terwisscha Van Scheltinga, A.F., Bakker, S.C., Van Haren, N.E.M., Derks, E.M., Buizer-Voskamp, J.E., Boos, H.B.M., Cahn, W., Hulshoff Pol, H.E., Ripke, S., Ophoff, R.A., Kahn, R.S., 2013. Genetic schizophrenia risk variants jointly modulate total brain and white matter volume. Biol. Psychiatry 73, 525–531. doi:10.1016/j.biopsych.2012.08.017

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 988–999. doi:10.1109/72.788640

Yang, H., Liu, J., Sui, J., Pearlson, G., Calhoun, V.D., 2010. A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. Front. Hum. Neurosci. 4, 192. doi:10.3389/fnhum.2010.00192

Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. U. S. A. 109, 1193–8. doi:10.1073/pnas.1119675109

Chapter 6

Summary and conclusion

Summary and conclusion

The main question in this thesis was: Can we apply machine learning techniques to MRI scans of the brain or genetic data to create predictive models beneficial to the clinic?

The studies conducted in this thesis used machine learning to explore brain abnormalities and genetic variation in patients with schizophrenia, bipolar patients and healthy controls.

In this final chapter a summary and discussion of the main findings, possible implications and future research suggestions are provided.

In Chapter 2, we studied the generalizability and possibility of clinical application of machine learning models to predict schizophrenia by magnetic resonance images of the brain. For this purpose, we included two large independent samples including schizophrenia patients and controls. The first sample included 239 subjects (128 patients and 111 controls) and was utilized to build a support vector machine model to discriminate between patients and controls. The other sample included 277 subjects (155 patients and 122 controls) and was utilized to validate the model.

We demonstrated that it is possible to achieve similar classification accuracy in two completely independent sets of subjects. The accuracy achieved in the sample on which the model was developed was 71.4% and by applying this model to the validation sample we achieved an accuracy of 70.4%. This showed that the model was generalizable and not overly optimized for the discovery sample.

The model consisted of a gray matter density pattern. This discriminative brain pattern is a description of the cumulative contributions of all features. The interpretation of the effects of a single brain region on the separation of the groups is not as straightforward as with classic statistical models, however it is clear that certain brain regions contributed more than others. Substantial contributions where patients had relatively larger gray matter densities were found for the basal ganglia and left occipital lobe; relatively small gray matter densities in patients were found in the frontal and superior temporal lobes and hippocampus. These findings are consistent with previously reported structural brain abnormalities in schizophrenia patients (S. Haijma et al., 2013).

In addition, we investigated the impact of sample size. Our research showed that for this particular classification problem at least 130 subjects were required to achieve a well classifying model independent of the subjects that were included. Models based on smaller samples led to large fluctuations in classification accuracies and sometimes resulted in accuracies lower than chance level.

The research in this chapter resulted in a stable model that could classify schizophrenia patients from healthy controls in a new and independent test sample. This is the first important step towards clinically useful models. The next step was to create a model related to a more complex clinical issue.

In Chapter 3, we aimed to separate bipolar patients from both healthy controls and schizophrenia patients and to confirm that we could separate schizophrenia patients from healthy controls. To facilitate this, we created three support vector machine models, each one separating two groups. The discovery sample (N = 66 per group) and validation sample (N = 46, 47, 43 schizophrenia patients, bipolar patients, and healthy controls, respectively) were acquired on scanners with different field strengths.

We confirmed that it is possible to separate schizophrenia patients and healthy subjects with an average accuracy of 75,5% in the validation sample. More importantly, we demonstrated that bipolar disorder patients could be separated from patients with schizophrenia with an accuracy of 65,5%. Bipolar patients and healthy subjects could not be classified significantly above chance. This is in line with previous reports showing less pronounced brain abnormalities in patients with bipolar disorder as compared to those with schizophrenia. Consistent with the earlier models designed to separate between schizophrenia patients and healthy individuals, the regions that contributed most in the discriminative brain pattern were in the frontal and (superior) temporal lobe. The other pattern that separated bipolar patients from schizophrenia patients was more diffuse.

Being able to separate between bipolar patients and schizophrenia patients based on MRI images of the brain is promising. Moreover, it is noteworthy that the difference between the two groups was apparent even though the samples were acquired on different scanners with different field strengths. However, there were some limitations to this study. The sample size was slightly larger than the earlier reported minimum of 130 subjects. As was demonstrated in chapter 2, the model could still benefit significantly from a larger sample. Additionally the patients in this sample were chronically ill, which is interesting for scientific purposes, but less so when investigating clinical application.

In Chapter 4, we investigated if we could classify the illness outcome of patients from an MRI scan made during or just after their first psychotic episode. For this purpose, we acquired longitudinal clinical data from recent onset psychotic patients. All patients underwent a baseline MRI scan and were assessed clinically 3 to 7 years after their baseline scan.

In order to collect a large enough sample, we relied on multi-center data. We included five longitudinal first episode patient samples from the University Medical Center Utrecht, The Netherlands (n = 67), the Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom (n = 97), the University of São Paulo, Brazil (n = 64), the University of Cantabria, Santander, Spain (n = 107), and the University of Melbourne, Australia (n = 54).

Our first important finding in this chapter was that combining multi-center neuroimaging data led to a single classifying model that performs 90% accurate when classifying a strong, unequivocal biological outcomes being gender. Moreover, when comparing single-center and multi-center accuracies the results showed that a multi-center model improved the performance in individual smaller and possibly more heterogeneous samples.

The other main finding was that when classifying illness course, defined as "continuous" illness course (no remission of symptoms of greater than 6 months); or "remitting" illness course (one or more periods of remission of at least 6 months, and no episode lasting longer than 6 months), classification accuracies were modest at best, ranging from below chance to 70%. Moreover, the results were significant only in centers with comparable definitions of outcome.

We suggested that with larger samples and standardized clinical information, multi-center-models have the potential to yield generalizable, clinically useful predictions on illness course.

To expand our search for a schizophrenia model that can increase classification accuracy, we searched for additional data sources for our analyses. Since schizophrenia is a highly heritable brain disorder, we explored the possibility to exploit common genetic variations, i.e., single nucleotide polymorphisms (SNPs), to model schizophrenia. Findings are reported in Chapter 5. We used a support vector machine to create three classification models based on a sample of schizophrenia patients and control subjects (N = 705 cases and N = 637 controls). The first model included all available SNPs, being over seven million SNPs per individual. The second model included approximately seven hundred thousand SNPs that had the best predictive value in the polygenic risk score (PGRS) (Ripke et al., 2014). The third and final model included 74 independent GWAS-significant SNPs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). The average prediction accuracies of these models were 54% 60% and 56% respectively.

Sensitivity and specificity were not adequate to be of diagnostic value. However, with much larger samples to create for instance a model including all SNPs, interesting differentiating SNP-patterns could be revealed. Machine learning is a technique that in the future could lead to an objective diagnostic tool to classify schizophrenia patients and healthy controls by genetic data.

Conclusions

Taken together, the studies presented in this thesis showed that machine learning in general is a feasible technique to detect patterns of brain abnormalities in schizophrenia patients and bipolar patients, and to a lesser extent to predict illness course in first episode patients. The use of machine learning in genetic data needs to be further investigated. Thus to answer the main question of this thesis: Yes, we expect that in the future models such as these will be able to assist in individualized diagnosis of patients with mental disorders.

Currently there is not a single method or test that continuously results in the same diagnosis for one individual, independent of the rater (Richieri et al., 2011). Continuing research to develop objective and accurate prediction models is necessary.

Solely by structural brain scan we were able to classify schizophrenia patients from healthy controls with an accuracy of 70%. Even though the number of subjects in this study was relatively large, from our investigation of the influence of sample size it is clear that larger samples could further increase the accuracy. Taken together with the results of chapter 4, where we showed that it is possible to use a multi-center design to increase sample size and obtain good prediction accuracies for clear-cut problems, it might be interesting to create a multi-center model on established schizophrenia patients. A model fashioned on well-established schizophrenia patients could possibly be used to classify more recently ill patients.

The nature of the support vector machine allows for all types of data to be combined into a single model. An important question that may arise when reading this thesis is whether the combination of imaging and genetic data can improve the accuracy of the classification. In chapter 4 we showed that combining multi-center data with only modest single-center accuracy, even with the same modality, did not benefit of increased sample size. One could speculate that combining models with modest results does not lead to better models, and that this holds true for multi-modality models. This leads us to the conclusion that before attempting to combine imaging and genetic data, both models have to perform more accurately than what has been demonstrated thus far.

A possible way to improve the genetic model is including more genetic data. In chapter 5, we only included common genetic variants in the population. More recent research shows that rare variants are also associated with schizophrenia (Olde Loohuis et al., 2015). And although, viewed from an imaging study perspective, we had a large sample in chapter 5, these numbers are still very small compared to genetic studies leading to significant findings. Possibly building support vector machine models on much larger numbers will increase the accuracy of these models enough to build combined imaging and genetics models.

References

Haijma, S. V, Van Haren, N., Cahn, W., Koolschijn, P.C.M.P., Hulshoff Pol, H.E., Kahn, R.S., 2013. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. Schizophr. Bull. 39, 1129–38. doi:10.1093/schbul/sbs118

Olde Loohuis, L., Vorstman, J. a. S., Ori, A.P., Staats, K. a., Wang, T., Richards, A.L., Leonenko, G., Walters, J.T., DeYoung, J., Kahn, R.S., Linszen, D., Os, J. Van, Wiersma, D., Bruggeman, R., Cahn, W., Haan, L. De, Krabbendam, L., Myin-Germeys, I., Cantor, R.M., Ophoff, R. a., 2015. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nat. Commun.* 6, 7501. doi:10.1038/ncomms8501

Richieri, R. (Assistance P.-H. de M.-P.P.H.S.M.-13009 M., Boyer, L., Lancon, C., 2011. [Analysis of the reliability of diagnostic criteria and classifications in psychiatry]. *Sante Publique (Paris).* 23, S31–S38.

Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P. a., Lee, P., Bulik-Sullivan, B., Collier, D. a., Huang, H., Pers, T.H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. a., Begemann, M., Belliveau Jr, R. a., Bene, J., Bergen, S.E., Bevilacqua, E., Bigdeli, T.B., Black, D.W., Bruggeman, R., Buccola, N.G., Buckner, R.L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R.M., Carr, V.J., Carrera, N., Catts, S. V., Chambert, K.D., Chan, R.C.K., Chen, R.Y.L., Chen, E.Y.H., Cheng, W., Cheung, E.F.C., Ann Chong, S., Robert Cloninger, C., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J.J., Curtis, D., Davidson, M., Davis, K.L., Degenhardt, F., Del Favero, J., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A.H., Farrell, M.S., Frank, J., Franke, L., Freedman, R., Freimer, N.B., Friedl, M., Friedman, J.I., Fromer, M., Genovese, G., Georgieva, L., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J.I., Golimbet, V., Gopal, S., Gratten, J., de Haan, L., Hammer, C., Hamshere, M.L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A.M., Henskens, F. a., Herms, S., Hirschhorn, J.N., Hoffmann, P., Hofman, A., Hollegaard, M. V., Hougaard, D.M., Ikeda, M., Joa, I., Julià, A., Kahn, R.S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M.C., Kennedy, J.L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J. a., Konte, B., Kucinskas, V., Ausrele Kucinskiene, Z., Kuzelova-Ptackova, H., Kähler, A.K., Laurent, C., Lee Chee Keong, J., Hong Lee, S., Legge, S.E., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C.M., Lubinski, J., Lönnqvist, J., Macek Jr, M., Magnusson, P.K.E., Maher, B.S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R.W., McDonald, C., McIntosh, A.M., Meier, S., Meijer, C.J., Melegh, B., Melle, I., Mesholam-Gately, R.I., Metspalu, A., Michie, P.T., Milani, L., Milanova, V., Mokrab, Y., Morris, D.W., Mors, O., Murphy, K.C., Murray, R.M., Myin-Germeys, I., Müller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D. a., Nestadt, G., Nicodemus, K.K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F.A., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Endophenotypes International Consortium, P., Pantelis, C., Papadimitriou, G.N., Papiol, S., Parkhomenko, E., Pato, M.T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D.O., Pietiläinen, O., Pimm, J., Pocklington, A.J., Powell, J., Price, A., Pulver, A.E., Purcell, S.M., Quested, D., Rasmussen, H.B., Reichenberg, A., Reimers, M. a., Richards, A.L., Roffman, J.L., Roussos, P., Ruderfer, D.M., Salomaa, V., Sanders, A.R., Schall, U., Schubert, C.R., Schulze, T.G., Schwab, S.G., Scolnick, E.M., Scott, R.J., Seidman, L.J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J.M., Sim, K., Slominsky, P., Smoller, J.W., So, H.-C., Spencer, C. a., Stahl, E. a., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R.E., Strengman, E., Strohmaier, J., Scott Stroup, T., Subramaniam, M., Suvisaari, J., Svrakic, D.M., Szatkiewicz, J.P., Söderman, E., Thirumalai, S., Toncheva, D., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B.T., Weiser, M., Wildenauer, D.B., Williams, N.M., Williams, S., Witt, S.H., Wolen, A.R., Wong, E.H.M., Wormley, B.K., Simon Xi, H., Zai, C.C., Zheng, X., Zimprich, F., Wray, N.R., Stefansson, K., Visscher, P.M., Trust Case-Control Consortium, W., Adolfsson, R., Andreassen, O. a., Blackwood, D.H.R., Bramon, E., Buxbaum, J.D., Børglum, A.D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C.M., Iwata, N., Jablensky, A. V., Jönsson, E.G., Kendler, K.S., Kirov, G., Knight, J., Lencz, T., Levinson, D.F., Li, Q.S., Liu, J., Malhotra, A.K., McCarroll, S. a., McQuillin, A., Moran, J.L., Mortensen, P.B., Mowry, B.J., Nöthen, M.M., Ophoff, R. a., Owen, M.J., Palotie, A., Pato, C.N., Petryshen, T.L., Posthuma, D., Rietschel, M., Riley, B.P., Rujescu, D., Sham, P.C., Sklar, P., St Clair, D., Weinberger, D.R., Wendland, J.R., Werge, T., Daly, M.J., Sullivan, P.F., O'Donovan, M.C., 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. doi:10.1038/nature13595

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–7. doi:10.1038/nature13595

Sham, P.C., MacLean, C.J., Kendler, K.S., 1994. A typological model of schizophrenia based on age at onset, sex and familial morbidity. *Acta Psychiatr. Scand.* 89, 135–141. doi:10.1111/j.1600-0447.1994.tb01501.x

Nederlandse samenvatting

De ziel in de machine

De hoofdvraag in dit proefschrift was: Kunnen we geautomatiseerde modellen creëren op basis van MRI-hersenscans (magnetic resonance imaging) en genetica om voorspellingen te doen waar de kliniek baat bij heeft?

In de studies in dit proefschrift hebben we gebruik gemaakt van geautomatiseerde modellen om afwijkingen in de hersenen en genetische variatie in patiënten met schizofrenie, bipolaire stoornis en gezonde controles te onderzoeken.

In dit laatste hoofdstuk wordt een samenvatting gegeven, worden de bevindingen besproken en mogelijke toekomstige implicaties van het gedane onderzoek behandeld.

In hoofdstuk 2 hebben we de generaliseerbaarheid bestudeerd en de mogelijkheid om geautomatiseerde modellen toe te passen in een klinische setting van modellen die schizofrenie-patiënten en gezonde controles van elkaar onderscheiden op basis van MRI-hersenscans.

Om dit te bewerkstelligen hebben we twee grote onafhankelijke groepen van schizofrenie-patiënten en gezonde controles opgenomen. De eerste groep bestond uit 239 participanten (128 patiënten en 111 controles) en werd gebruikt om een support vector machine-model mee te creëren waarmee onderscheid kon worden gemaakt tussen patiënten en controles. De tweede groep bestond uit 277 participanten (155 patiënten en 122 controles) en werd gebruikt om dit model te valideren.

We hebben laten zien dat het mogelijk is om een vergelijkbare classificatie-exactheid te bereiken met twee compleet onafhankelijk groepen. De exactheid die we bereikten met het model van de eerste groep was 71,4% en toepassing van dit model op de validatie-groep leidde tot een exactheid van 70,4%. Dit liet zien dat het model niet excessief geoptimaliseerd was, maar dat het resultaat generaliseerbaar was.

Het model bestaat uit een patroon van grijze stofdichtheid. Dit onderscheidende hersenpatroon geeft de cumulatieve bijdrage aan het model van alle grijze stof dichtheden in de hersenen weer. Het effect van een enkele regio in de hersenen op de groepsverdeling is echter niet zo eenvoudig te interpreteren als bij een klassieke statische model, het is wel op te maken dat bepaalde regio's meer bijdragen dan andere. Substantiële bijdrages waar patiënten relatief grotere dichtheid grijze stof hadden waren de basale ganglia en de linker occipitale kwab; een relatief kleinere grijze stof dichtheid vonden we in de frontaal- en superieur temporaalkwab en de hippocampus. Deze bevindingen zijn vergelijkbaar met eerder gerapporteerde structurele hersenafwijkingen in patiënten met schizofrenie (Haijma et al., 2013).

Hiernaast hebben we de impact van groepsgrootte onderzocht. Ons onderzoek liet zien dat er voor dit specifieke probleem minstens 130 participanten geïncludeerd toegevoegd moesten worden om een model te creëren dat goed classificeerde, onafhankelijk van de toegevoegde individuen. Modellen die gebaseerd waren op kleinere aantallen lieten grote exactheidsverschillen zien, soms zelfs onnauwkeuriger dan kans-niveau.

Het onderzoek in dit hoofdstuk heeft geresulteerd in een stabiel model dat patiënten met schizofrenie van gezonde controles uit een onafhankelijke test groep kon onderscheiden. Dit was de eerste belangrijke stap naar klinisch- relevante modellen. De volgende stap was om modellen te creëren die complexere klinische vraagstukken beantwoorden.

In hoofdstuk 3 was het doel om patiënten met bipolaire stoornis van zowel gezonde controles als van patiënten met schizofrenie te onderscheiden. Om dit te bewerkstelligen creëerden we drie support vector machine-modellen die ieder twee categorieën onderscheidden. De primaire groep (N = 66 per categorie) en de validatie-groep (N = 46, 47, 43 respectievelijk patiënten met schizofrenie, patiënten met bipolaire stoornis en gezonde controles) waren gescand met verschillende veldsterktes.

We bevestigden dat het mogelijk is om patiënten met schizofrenie en gezonde controles van elkaar te onderscheiden met een exactheid van 75,5% in de validatie-groep. Wat nog belangrijker is, is dat we demonstreerden dat patiënten met bipolaire stoornis en patiënten met schizofrenie van elkaar te onderscheidden waren met een exactheid van 65,5%. Patiënten met bipolaire stoornis en gezonde controle waren niet significant beter dan kans van elkaar te onderscheiden. Deze resultaten zijn in lijn met eerder gerapporteerde bevindingen waar ook minder evidente hersenafwijkingen werden gevonden bij patiënten met bipolaire stoornis dan bij patiënten met schizofrenie. De regio's die het meest bijdroegen in het hersenpatroon, frontaal- en (superieur) temporaalkwab, waren consistent met de eerder gevonden regio's in de modellen die waren ontworpen om patiënten met schizofrenie en gezonde controles van elkaar te onderscheiden. Het hersenpatroon dat patiënten met schizofrenie van patiënten met bipolaire stoornis van elkaar scheidde was diffuser.

De mogelijkheid om patiënten met schizofrenie van patiënten met bipolaire stoornis te onderscheiden op basis van MRI hersenscans is veelbelovend. Bovendien is het noemenswaardig dat het verschil evident was ondanks dat de participanten op verschillende scanners met verschillende veldsterktes waren vergaard. Echter waren er ook wat beperkingen aan de studieopzet. Hoewel de groepsgrootte iets groter was dan het eerder benoemde aantal van 130 participanten, tonen we in hoofdstuk 2 aan dat een model toch significant gebaat kan zijn bij nog grotere groepen. Daarbij waren de patiënten chronisch ziek, wat voor onderzoeksdoeleinden interessant is, maar minder interessant wanneer we de mogelijkheid willen onderzoeken voor een klinische toepassing.

In hoofdstuk 4 onderzochten we of we ziektebeloop konden voorspellen aan de hand van een hersenscan die gemaakt was tijdens of net na de eerste psychotische episode. Om dit te bewerkstelligen vergaarden we longitudinale data van recent zieke psychotische patiënten. Alle patiënten ondergingen bij de eerste meting een MRI hersenscan en werden bij een tweede meting, 3 tot 7 jaar later, klinisch beoordeeld.

Om een grote groep participanten te vergaren waren we afhankelijk van multicenter data. We gebruikten vijf longitudinale eerste episode patiëntengroepen van het Universitair Medisch Centrum Utrecht, Nederland (n = 67), het instituut van psychiatrie, psychologie en neurowetenschappen, Londen, Verenigd koninkrijk (n = 97), de Universiteit van São Paulo, Brazilië (n = 64), de Universiteit van Cantabria, Santander, Spanje (n = 107) en de Universiteit van Melbourne, Australië (n = 54).

Onze eerste belangrijke bevinding van dit hoofdstuk is dat het combineren van multicenter neuro-imaging data in één model tot een 90% accuraat resultaat leidde wanneer er een sterk onomstotelijk biologische bepaling wordt geclassificeerd, in dit geval geslacht. Bovendien, wanneer de exactheid van de losse modellen en het multicentermodel met elkaar worden vergeleken, kun je zien dat het multicenter model een positieve invloed heeft op de prestatie van de kleinere en mogelijk heterogenere groepen binnen dat model.

De andere centrale bevinding is dat wanneer ziektebeloop geclassificeerd wordt, gedefinieerd als ‘continue’ ziektebeloop (geen remissie van symptomen langer dan 6 maanden) of ‘remissie’ ziektebeloop (een of meer periodes van remissie van minstens 6 maanden en geen enkele episode langer dan 6 maanden), was de exactheid hoogstens bescheiden, variërend tussen onder kans en 70%. Bovendien waren de resultaten alleen significant in de centers waar de definities van ziektebeloop het meest op elkaar leken.

We vermoeden dat met grotere groepen en gestandaardiseerde klinische informatie, multicentermodellen de potentie hebben om generaliseerbare klinisch bruikbare voorspellingen te realiseren van ziektebeloop.

Om onze zoektocht naar een superieur schizofreniemodel uit te breiden hebben we gezocht naar additionele databronnen voor onze analyses. Omdat schizofrenie een erg erfelijke hersenaandoening is hebben we de mogelijkheid onderzocht om veelvoorkomende genetische variaties, d.w.z. single nucleotide polymorphisms (SNPs), te exploiteren om schizofrenie te modelleren. Deze bevindingen staan gerapporteerd in hoofdstuk 5. We gebruikten een support vector machine om drie classificatiemodellen te creëren op basis van een groep patiënten met schizofrenie en gezonde controles (N = 705 patiënten en N = 637 controles). Het eerste model bestond uit alle beschikbare

SNPs, in dit geval meer dan zeven miljoen SNPs per individu. Het tweede model bestond uit (bij benadering) zevenhonderdduizend SNPs die de beste voorspellende waarde hadden in de polygenetic risc score (PGRS) (Ripke et al., 2014). Het derde en laatste model bestond uit 74 onafhankelijke GWAS-significante SNPs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). De gemiddelde exactheid van deze modellen was respectievelijk 54%, 60% en 56%.

De sensitiviteit en specificiteit waren niet adequaat genoeg om bruikbaar te zijn voor diagnosticerende doeleinden. Echter wanneer er veel grotere groepen gebruikt worden om bijvoorbeeld een model te maken gebaseerd op alle SNPs zouden er interessante onderscheidende SNP-patronen blootgelegd kunnen worden. Geautomatiseerde modellen zouden in de toekomst kunnen leiden tot objectieve diagnostische modellen om patiënten met schizofrenie en gezonde controles aan de hand van genetische data van elkaar te kunnen onderscheiden.

Conclusie

Al met al hebben de studies die we hebben gepresenteerd in dit manuscript, laten zien dat geautomatiseerde modellen in het algemeen een goede techniek zijn om patronen te herkennen van hersenafwijkingen in patiënten met schizofrenie en bipolaire patiënten, en in mindere mate om ziektebeloop mee te voorspellen in een eerste episode groep. Het gebruik van geautomatiseerde modellen met genetische data vereist uitgebreider onderzoek dan tot noch toe. Om de hoofdvraag van dit manuscript te beantwoorden: Ja, we verwachten dat in de toekomst modellen zoals deze zullen assisteren bij het maken van individuele diagnoses van patiënten met een geestesziekte.

Op dit moment is er geen enkele methode of test die herhaaldelijk tot dezelfde diagnose voor een individu komt onafhankelijk van degene die de methode of test toepast (Richieri et al., 2011). Het is dus van belang dat onderzoek naar het ontwikkelen van een objectief, accuraat voorspellend model zich voortzet.

Aan de hand van alleen een structurele hersenscan was het mogelijk om een patiënt met schizofrenie te onderscheiden van een gezonde controle met een precisie van 70%. Ook al was het aantal participanten in de studie relatief groot, in ons onderzoek naar de invloed van groepsgrootte kwam duidelijk naar voren dat een nog grotere groep de exactheid van het model zou kunnen verbeteren. Samengenomen met de resultaten uit hoofdstuk 4, waarin we demonstreerden dat het mogelijk was om een multicenter opzet te gebruiken om groepsgrootte te vergroten, zou het interessant zijn om een multicenter model te maken op patiënten waarvan met grote zekerheid is vastgesteld dat ze schizofrenie hebben. Een model op deze patiënten zou mogelijk kunnen leiden tot een model dat patiënten die slechts recent ziek zijn geworden ook al kan classificeren.

De manier waarop een support vector machine in elkaar zit zorgt ervoor dat allerlei datatypes gecombineerd kunnen worden in een model. Een belangrijke vraag die wellicht opkomt wanneer dit manuscript gelezen wordt is of de mogelijkheid bestaat om beter classificerende modellen te maken wanneer je hersenscans en genetische data met elkaar zou combineren. In hoofdstuk 4 lieten we zien dat het combineren van multicenterdata niet per definitie leidt tot een beter classificerend model wanneer er in een center slechts beperkte exactheid is ook al gebruiken we dezelfde modaliteit. We zouden kunnen speculeren dat dit ook geldt voor het combineren van meerdere modaliteiten. Dit brengt ons tot de conclusie dat, voordat we meerdere modaliteiten samenvoegen in een model, we eerst betere resultaten moeten hebben in de afzonderlijke modellen.

Een mogelijke manier om het genetische model te verbeteren is door meer genetische data te betrekken. In hoofdstuk 5 hebben we alleen veelvoorkomende genetische variatie in de populatie onderzocht. Recenter onderzoek laat zien dat ook bijzondere variatie geassocieerd worden met schizofrenie (Olde Loohuis et al., 2015). Ook al hebben we vanuit het perspectief van een MRI-studie grote aantallen participanten gebruikt, zijn deze aantallen voor een genetische studie zeer klein om significante resultaten te behalen. Mogelijkerwijs zouden we met veel grotere aantallen een betere precisie behalen waardoor we ook modaliteiten zouden kunnen combineren.

Dankwoord

Met veel trots heb ik de afgelopen maanden toegewerkt naar dit moment, mijn proefschrift is af! Het onderzoek is natuurlijk nooit af, daar ben ik inmiddels wel achter. Er komen met een onderzoek altijd meer vragen dan antwoorden bovendrijven en de wens naar perfectere, grotere en completere datasets groeit alleen maar.

Tijdens mijn promotietraject heb ik veel meegemaakt en geleerd, naast het onderzoek heb ik mijzelf en ook anderen goed leren kennen.

Graag wil ik René Kahn bedanken, in het bijzonder omdat hij naast de rol van promotor mij ook de kans en de vrijheid heeft gegeven tijdens mijn promotietraject mijn Olympische droom na te jagen. We hebben met het Nederlands 7s rugby dames team de Olympische Spelen niet gehaald, maar het was fantastisch om mee te strijden onder de Nederlandse vlag.

Ook wil ik Hugo Schnack mijn copromotor graag bedanken. Hij heeft me geleerd van A tot Z onomstotelijk en degelijk onderzoek te verrichten. Dankzij dat we het niet altijd eens waren heb ik goed leren opkomen voor mijn eigen ideeën en belangen.

Neeltje van Haren wil ik graag bedanken voor haar enthousiasme, de steun en de begeleiding om mijn PhD tot een goed einde te brengen. Het is dankzij haar en haar contacten geweest dat ik een half jaar in London heb gewoond tijdens mijn PhD. Het is ook heel inspirerend geweest om te zien dat het moeder zijn het leven van een onderzoeker niet in de weg hoeft te staan.

Paola Dazzan, I would like to thank you for all your time and supervision. I’ve enjoyed working with you and have learned a lot from writing with you. You have such a positive vibe and even though you always have a compliment at hand your critique was always direct and accommodating.

Marinka, niet alleen mijn kamergenoot, maar ook echt mijn kameraadje! Ik hoop dat we door de jaren heen contact blijven houden. We hebben gelachen, gehuild, we hadden onze wetenschappelijke highs en ook zeker de lows. Je bent een super fijne collega, ik wens je veel plezier overseas en hoop je daarna toch weer in NL te treffen!

Natuurlijk wil ik ook graag Rachel Brouwer bedanken, ik check nog steeds altijd grondig mijn data a.d.h.v. tips die jij ooit gaf; René Mandel, die alles altijd zo makkelijk relativeert en je vol goede moed weer aan t werk laat gaan; Wiepke Cahn, die door mijn hele promotietraject altijd tijd maakte en open stond voor brainstormsessies en nieuwe ideeën om te onderzoeken. Yumas, voor het lang genoeg in leven houden van de archaïsche Fiber en Axon. Anouk van der Weiden, bedankt voor de tips als ervaringsdeskundige van het promotietraject, de laatste loodjes wogen zo minder zwaar en ik voelde me gesterkt in mijn besluiten.

Tot slot wil ik graag nog Roeland, mijn man, bedanken die altijd mijn teksten door heeft willen lezen en soms tot ’s avonds laat op mijn aanwezigheid moest wachten. Mijn lieve zoontje Lancelot wil ik ook graag bedanken, omdat hij zo goed sliep en direct al zo goed dronk, zodat ik alles af kon ronden om daarna van mijn verlof te genieten.