◆ Human Brain Mapping 38:704–714 (2017) ◆

Individual Prediction of Long-Term Outcome in Adolescents at Ultra-High Risk for Psychosis: Applying Machine Learning Techniques to Brain Imaging Data

Sanne de Wit,[1]* Tim B. Ziermans,[2,3] M. Nieuwenhuis,[1] Patricia F. Schothorst,[1] Herman van Engeland,[1] René S. Kahn,[1] Sarah Durston,[1] and Hugo G. Schnack[1]

[1]*Department of Psychiatry, University Medical Center Utrecht, Brain Center Rudolf Magnus, Utrecht, the Netherlands*
[2]*Department of Clinical Child and Adolescent Studies, Leiden University, Leiden, the Netherlands*
[3]*Leiden Institute for Brain and Cognition, Leiden, the Netherlands*

Abstract: An important focus of studies of individuals at ultra-high risk (UHR) for psychosis has been to identify biomarkers to predict which individuals will transition to psychosis. However, the majority of individuals will prove to be resilient and go on to experience remission of their symptoms and function well. The aim of this study was to investigate the possibility of using structural MRI measures collected in UHR adolescents at baseline to quantitatively predict their long-term clinical outcome and level of functioning. We included 64 UHR individuals and 62 typically developing adolescents (12–18 years old at recruitment). At six-year follow-up, we determined resilience for 43 UHR individuals. Support Vector Regression analyses were performed to predict long-term functional and clinical outcome from baseline MRI measures on a continuous scale, instead of the more typical binary classification. This led to predictive correlations of baseline MR measures with level of functioning, and negative and disorganization symptoms. The highest correlation ($r = 0.42$) was found between baseline subcortical volumes and long-term level of functioning. In conclusion, our results show that structural MRI data can be used to quantitatively predict long-term functional and clinical outcome in UHR individuals with medium effect size, suggesting that there may be scope for predicting outcome at the individual level. Moreover, we recommend classifying individual outcome on a continuous scale, enabling the assessment of different functional and clinical scales separately without the need to set a threshold. *Hum Brain Mapp 38:704–714, 2017.*

Key words: ultra-high risk; psychosis; outcome; prediction; brain imaging; machine-learning

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: ZON-MW – the Netherlands organization for health research and development; Contract grant number: 2630.0001

*Correspondence to: S. de Wit, Department of Psychiatry, NICHE laboratory, University Medical Center Utrecht, Brain Center Rudolf Magnus, HP A01.126 (B01.108), Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. E-mail: s.dewit@umcutrecht.nl

Received for publication 15 December 2015; Revised 13 September 2016; Accepted 15 September 2016.

DOI: 10.1002/hbm.23410

Published online 4 October 2016 in Wiley Online Library (wileyonlinelibrary.com)

INTRODUCTION

The introduction of criteria for individuals at ultra-high risk (UHR) for developing psychosis in the mid 1990s [Yung and McGorry, 1996] has resulted in a sizeable literature investigating the mechanisms of psychosis onset and disorder progression [for reviews see Addington and Heinssen, 2012; Fusar-Poli et al., 2013]. Besides clinical predictors, measures of brain anatomy and function have been put forward as possible neurobiological predictors for the onset of psychosis. Such potential predictors include loss of brain volume and reduced activation in insular, temporal, parietal, and superior brain areas (for review see Wood et al., 2013]. To date, such studies have focused on predicting psychosis onset. Although it is relevant to predict who will make a transition to psychosis and who will not, it is at least as important to be able to predict who will recover from their at risk state and go on to function well, i.e., who will prove to be resilient. This is particularly important as the rate of transition to psychosis reported in studies has been becoming lower in recent years, leading to a call for studies of UHR remission [Addington et al., 2011; Simon et al., 2011, 2013; de Wit et al., 2014]. Moreover, the identification of predictors of resilience could lead to a better understanding of the heterogeneity in outcome and ultimately the disorder itself.

We have previously studied remission and explored brain development in resilient and non-resilient UHR individuals, defined by long-term functional outcome [De Wit et al., 2016]. We found widespread differences in brain volume that were already present at young age, as well as differences that appeared later in development. The differences that were already present at young age (12 years) included reductions in volume of frontal, temporal and parietal cortex. As these differences were already present at first assessment, they may be promising for predicting who will recover from an at-risk state and who will not.

The majority of studies exploring neurobiological markers have used group-level statistical analysis, thereby limiting the clinical applicability of findings. Multivariate pattern recognition methods can be used to overcome these limitations. These methods provide the possibility to make inferences about a subject's health status at an individual level and thus are more suited for clinical decision making purposes [Zarogianni et al., 2013]. So far, machine-learning studies in UHR individuals have been scarce and although pattern recognition has often been used in other clinical context [see for reviews: Mourao-Miranda et al., 2011; Wolfers et al., 2015], classification using neuroimaging data in the UHR field is still in its infancy. A promising accuracy of 82% was shown in a study discriminating UHR individuals who developed psychosis from those who did not using structural MRI and four-year clinical follow-up data [Koutsouleris et al., 2009]. Only one study [Kambeitz-Ilankovic et al., 2015] used structural brain markers to predict individual outcome based on level of functioning rather than transition to psychosis and found that cortical surface patterns predicted good versus poor outcome status at two-year follow-up with an accuracy of 82%. These studies have used support vector machines (SVM) to classify groups in a binary manner (e.g., transition vs. non-transition). However, the threshold for dividing the group into transition versus non-transition subgroups (or good versus poor outcome) is arbitrary by definition, and there has been much discussion about the validity of the threshold for psychosis [Fusar-Poli and Van Os, 2013; Yung et al., 2010]. By using measures on a continuous scale one does not have to make an artificial division in the sample. The technique of Support Vector Regression (SVR) permits the quantitative prediction of a variable of interest (e.g., a clinical symptom score) without the need for a discrete categorical decision (e.g., affected vs. unaffected), allowing exploration of outcome on a gradual scale. To date, only Tognin and colleagues have attempted to predict outcome on a continuous scale [Tognin et al., 2013]. Using the Positive and Negative Syndrome Scale (PANSS), they found a correlation of 0.34 between two-year symptom progression and baseline cortical thickness. These results are encouraging, as they suggest brain measures may be useful for predicting later outcome in a clinically relevant manner. However, two years is not long in clinical terms and the long-term utility of such predictions needs to be assessed. Therefore, we followed up adolescents at UHR for psychosis over a six-year period and monitored clinical and functional outcome. We focused on SVR with baseline structural MRI data to individually predict long-term functional and clinical outcome on a continuous scale. To allow comparison with earlier studies, we performed complementary SVM analyses to separate UHR individuals from typically developing controls and resilient from non-resilient UHR individuals in a binary manner.

Abbreviations

IQ	Intelligence quotient
LGI	Local gyrification index
LOO	Leave-one-out
ROC	Receiver operating characteristic
SIPS	Structured Interview for Prodromal Syndromes
SVM	Support vector machines
TDC	Typically developing controls
TE	Echo time
TR	Repetition time
UHR	Ultra-high risk

METHODS

Participants

All data were collected at the Department of Psychiatry at the University Medical Center Utrecht, Brain Center Rudolf Magnus in the Netherlands. Participants were

between 12 and 18 years of age at the time of recruitment and were included after the nature of the experimental procedures was explained and written informed consent was obtained. The study was approved by the Dutch Central Committee on Research Involving Human Subjects.

Recruitment details have been previously described [Sprong et al., 2008; Ziermans et al., 2011]. Briefly, adolescents at UHR were referred by general practitioners or other psychiatry clinics. They had to fulfill at least one of the following criteria: (1) attenuated positive symptoms, (2) brief, limited, or intermittent psychotic symptoms, (3) genetic risk for psychosis combined with a deterioration in overall level of social, occupational/school, and psychological functioning in the past year or (4) two or more of nine basic symptoms of mild cognitive disturbances. The first three inclusion criteria were assessed using the Structured Interview for Prodromal Syndromes (SIPS) [McGlashan et al., 2001] and the Family Interview for Genetic Studies [Maxwell, 1982]. The fourth inclusion criterion was assessed using the Bonn Scale for the Assessment of Basic Symptoms-Prediction List [Schultze-Lutter and Klosterkötter, 2002]. Exclusion criteria were: a past or present psychotic episode lasting more than one week, traumatic brain injury or any known neurological disorder, and verbal intellectual IQ < 75. The typically developing control (TDC) group consisted of adolescents recruited through secondary schools in the region of Utrecht. They were excluded if they met any UHR-criterion, if they or any first degree relative had a history of any psychiatric illness, or if they had a second-degree relative with a psychotic disorder.

At baseline, 64 UHR individuals and 62 TDC completed clinical assessment and an MRI scan. Clinical follow-up assessments were conducted at nine months, 18 months, 24 months, and 72 months post-baseline. Follow-up MRI scans were collected at 24 months and six years post-baseline. For the purpose of this study only baseline MRI data were used. To investigate the predictive value of neuroimaging for long-term functional and clinical outcome, the six-year follow-up data of the SIPS interview and the mGAF scale were used. Of the 64 UHR individuals at baseline, 41 individuals consented to long-term (six-year) follow-up. Reasons for discontinuation were: (1) assessments were considered too time consuming by the individuals ($n = 18$), (2) the individual could no longer be contacted ($n = 4$), and (3) the individual had emigrated ($n = 1$).

For the complementary analyses of predicting outcome on a binary scale, six-year follow-up data was used to divide the UHR group into "resilient" and "non-resilient" subgroups. We used the outcome measure "functional outcome" to define resilience, using the modified Global Assessment of Functioning (mGAF) scale [Hall, 1995]. Poor functional outcome (non-resilient) was defined as an mGAF score of < 65 and good functional outcome (resilient) as an mGAF score of ≥ 65 [de Wit, et al., 2016]. Details of this procedure have been described previously [de Wit, et al., 2016]. Global intellectual functioning (IQ) was assessed using the Wechsler Intelligence Scales [Wechsler, 1997; Wechsler, 2002].

Image Acquisition

MRI scans were acquired on a single 1.5-T scanner (Philips, Best, The Netherlands). Whole brain T1-weighted three-dimensional fast field echo scans were made with 1.5-mm contiguous coronal slices of the whole head [256 $\times$ 256 matrix, FoV = 256 mm, echo time (TE) = 4.6 ms, repetition time (TR) = 30 ms, flip angle = 30°].

Image Processing

Scans were processed and analyzed using FreeSurfer *v 5.1.0* software. Technical details of the automated reconstruction scheme of this well-validated software program are described elsewhere [Carmona et al., 2009; Dale et al., 1999; Fischl et al., 1999]. Before quantitative analyses could be performed, output required qualitative inspection. Surface reconstruction, cortical parcellation and white matter segmentation were therefore evaluated for accuracy. Manual edits were performed as necessary by a rater blind to subject identity and group membership. Edits included removal of non-brain tissue and perfecting the white matter mask. For these manual interventions standard procedures, documented on the FreeSurfer website, were used. We calculated average volume (mm^3), cortical thickness (mm), surface area (mm^2), and gyrification for the 34 cortical structures in each hemisphere from the Desikan-Killiany atlas ($4 \times 34 \times 2 = 272$ measurements) [Desikan et al., 2006]. For volume, extra measures included cortical volume (left, right, and total), cortical white matter volume (left, right, and total) and total gray matter volume (seven measurements). For surface area, extra measures included white surface area (left and right, two measurements). Gyrification could not be estimated for seven scans, because of FreeSurfer processing errors. We also measured the volume of subcortical structures ($n = 25$, including ventricular system), as well as total subcortical gray matter volume and gray and white matter separately for the cerebellum (left and right, five measurements). This resulted in a total of 311 features that were available for the SVM and SVR models. To correct for possible influences of age and gender, effects of age and gender were regressed out, and all brain imaging data were standardized by subtracting the mean and dividing by the standard deviation. To test the robustness of results, extra analyses were performed on a matched sample of UHR individuals ($n = 53$) and TDC ($n = 53$) where individuals were matched on age and gender. Results of the latter analysis are included in the Supporting Information.

Classification Models

To solve our classification problem, we used SVM, a supervised learning algorithm that is frequently used in psychiatric neuroimaging [Orrù et al., 2012]. The SVM model is trained to classify subjects based on their features [Vapnik, 1999]. Full details of the modeling procedure have been described previously [Nieuwenhuis et al., 2012; Schnack et al., 2014]. Briefly, each subject i is represented by features congregated into a vector x_i. 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 $y(x_i)$:

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

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 (e.g., TDC -1; UHR individuals 1), and the function is optimized by requiring $y(x_i) < 0$ if $t_i = -1$, and $y(x_i) > 0$ if $t_i = 1$. 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 w provides information on feature importance, as well as whether it is either an increase or decrease of a particular feature's value that contributes to being classified as either 1 or -1 (in this example UHR individual or TDC).

There can be many surfaces that 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. The position of the OSH is determined by a subset of the subjects, the so-called support vectors. 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, which was optimized as described in Nieuwenhuis et al., 2012.

For this study we focus on SVR. Whereas SVM classifies on a binary scale, the SVM approach has also been adapted to predict numerical values through SVR [Smola and Schölkopf, 1998]. Instead of constructing a hyperplane for classification, SVR derives a function on the basis of training data to predict numerical values. It uses the same principles as the SVM for classification, but with an additional parameter, v, that controls the number of support vectors and training errors [Smola and Schölkopf, 1998].

First, as a proof of principle, we built a model to separate UHR individuals from TDCs (model A). Next, we built models to predict clinical and functional outcome in UHR individuals at six-year follow-up: one binary model to separate resilient UHR individuals from non-resilient UHR individuals at six-year follow-up (model B, binary) to be able to compare our results to previous studies. The focus of this study is on the four continuous models (models C) that we built to predict functional outcome (model C1) and clinical outcome (models C2-4) on a continuous scale:

A. UHR-TDC, to separate UHR individuals from TDCs (binary; SVM)
B. R-NONR, to separate resilient (R) UHR individuals from non-resilient (NON-R) UHR individuals on a binary scale (binary; SVM)
(C1) mGAF score at six-year follow-up (continuous, SVR)
(C2) SIPS Positive score at six-year follow-up (continuous, SVR)
(C3) SIPS Negative score at six-year follow-up (continuous, SVR)
(C4) SIPS Disorganization score at six-year follow-up (continuous, SVR)

All models were built using baseline features of a single type, respectively: baseline cortical volume, cortical thickness, surface area, gyrification [local gyrification index (LGI)] and subcortical volumes as described above. In addition, models based on combinations of different feature types were also built. As the SIPS "Disorganization" subscale score differed significantly between resilient and non-resilient UHR individuals at baseline, we also built models including MRI-based features and the baseline SIPS Disorganization score as a predictor.

Performance Measures and Statistical Significance Testing

The quality of an SVR model was assessed by the correlation coefficient (r) between true and predicted values.

The quality of a SVM model was assessed by three quantities:

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

The accuracy of the models was tested using leave-one-out (LOO) cross validation [Nieuwenhuis et al., 2012]. In this procedure each subject is taken out once and used to test the prediction model built on the other subjects. To test the statistical significance of the corresponding accuracies and weights, we randomly permuted the labels of the training sample and built models from these data. We repeated this process 1000 times to determine null-distributions of accuracies, correlation coefficients and weights. P-values of the accuracy, correlation coefficient and weights were calculated as the fraction of models from the permutation procedure that had a larger accuracy/correlation coefficient/weight than the accuracy/correlation coefficient/weight of

TABLE I. Demographic and clinical data of resilient and non-resilient UHR individuals[a]

	Resilient (R)	Non-Resilient (NON-R)	R vs. NON-R
Number of individuals (*n*)	17	24	*n.s.*
Gender, M/F (*n*)	13/4	14/10	*n.s.*
Premorbid IQ, mean (SD)	101.18 (12.89)	101.88 (10.46)	*n.s.*
Age at baseline, years			
Mean (SD)	15.42 (2.20)	15.86 (2.32)	*n.s.*
Range	12.31–19.64	12.28–19.44	
Age at 6-year follow-up, years			
Mean (SD)	21.34 (2.58)	20,96 (2.31)	*n.s.*
Range	17.88–25.79	16.84–24.80	
SIPS/SOPS baseline, mean (SD)			
Total score	21.35 (10.74)	26.25 (13.49)	*n.s.*
Positive symptoms	7.41 (4.53)	8.67 (3.86)	*n.s.*
Negative symptoms	4.24 (4.66)	4.75 (3.88)	*n.s.*
Disorganized symptoms	3.41 (3.30)	5.79 (3.78)	$U = 113{,}5$, $P = 0.016$
General symptoms	6.29 (4.33)	7.04 (4.81)	*n.s.*
mGAF baseline, mean (SD)	57.06 (13.57)	54.92 (16.59)	*n.s.*
SIPS/SOPS 6-year follow-up, mean (SD)			
Total score	11.75 (8.36)	38.71 (17.09)	$U = 22.5$, $P = <0.001$
Positive symptoms	3.88 (3.59)	10.71 (5.90)	$U = 57{,}0$, $P = <0.001$
Negative symptoms	3.88 (3.72)	12.92 (7.47)	$U = 49.0$, $P = <0.001$
Disorganized symptoms	3.19 (2.74)	7.75 (3.57)	$t = 4{,}33$, $P = <0.001$
General symptoms	1.50 (1.59)	7.33 (4.59)	$U = 42.5$, $P = <0.001$
mGAF 6-year follow-up, mean (SD)	77.94 (7.30)	44.08 (11.78)	$t = 10.49$, $P = <0.001$
Psychotropic medication baseline, any			*n.s.*
No	9	13	
Yes	8	11	
Psychotropic medication 6-year follow-up, any			*n.s.*
No	13	11	
Yes	4	13	

Notes: [a]Subgroups are based on functional outcome at 6-year follow-up; outcome was unknown for 23 UHR individuals not included here.
UHR = Individuals at ultra-high risk for psychosis; IQ = intelligence quotient; SD = standard deviation; SIPS/SOPS =.
Structured Interview for Prodromal Symptoms / Scale of Prodromal Symptoms; mGAF = Modified Global Assessment of Functioning.

our full model. The size and significance of the weights provides an indication of the relative importance of the respective features for classification and prediction. It should be noted, however, that these values should be interpreted with care [Haufe et al., 2014].

Finally, we performed a receiver operating characteristic (ROC) analysis to illustrate the performance of the best binary classifier model (SVM). The curve was created by plotting the true positive rate against the false positive rate at various threshold settings.

Implementation and Statistical Analyses

The open source machine learning library LIBSVM [Chang and Lin, 2011] was integrated with our own software to carry out the SVR and SVM classifications [Nieuwenhuis et al., 2012]. During training the model, we weighted the cases according to the inverse of the number in their respective groups, using the libsvm 'weight' option. The statistical software package IBM SPSS version 22.0 was used (1) to test for between-group differences in the demographic and clinical data (*t*-test/Fisher's exact/Mann Whitney).

RESULTS

Demographic and Clinical Characteristics

When the whole group of UHR individuals was compared to the group of TDCs, only IQ differed between groups, with lower IQ for UHR individuals than TDCs ($t = 3.53$, $P = 0.001$) [Ziermans et al., 2014]. A full list of demographic characteristics is provided in the Supporting Information (Supporting Information Table I). There were no statistical differences in the demographic variables between resilient and non-resilient UHR individuals (Table I). Clinically, there were few differences in baseline symptoms with only 'Disorganized symptoms' lower in resilient UHR individuals than non-resilient UHR individuals ($U = 113{,}5$, $P = 0.016$). At six-

TABLE II. Results of baseline MRI features separating resilient from non-resilient UHR individuals (models B) and quantitative predictions of long-term functional and clinical outcome (models C)

Binary models (B)	Sensitivity (%)	Specificity (%)	Average accuracy (%)	*P*
Cortical Volume	71	67	69	0.015*
Surface area	47	42	44	n.s.
Cortical thickness	47	50	49	n.s.
Gyrification	69	78	73	0.016*
Subcortical volume	59	75	67	0.047*
SIPS Disorganization (all)	76	71	74	0.001**
SIPS Disorganization (gyrification subset)	75	78	76	0.002**
Gyrification + Subcortical Volume	69	78	73	0.021*
SIPS Disorganization + Gyrification	69	83	76	0.005**
SIPS Disorganization + Subcortical Volume	76	71	74	0.013*
SIPS Disorganization + Cortical Volume	59	75	67	0.002**
SIPS Disorganization + Gyrification + Subcortical Volume	69	**94**	82	0.003**
Continuous models (C)	***r***	***P***		
Model C1: mGAF score at 6-yr follow-up				
Gyrification	0.382	0.048*		
Subcortical Volume	0.424	0.008**		
SIPS Disorganization + Gyrification + Subcortical Volume	0.327	n.s.		
Model C2: SIPS Positive score at 6-yr follow-up				
Cortical thickness	0.258	n.s.		
Model C3: SIPS Negative score at 6-yr follow-up				
Subcortical Volume	0.349	0.048*		
Model C4: SIPS Disorganization score at 6-yr follow-up				
Gyrification	0.411	0.044*		
Subcortical Volume	0.342	n.s.		
SIPS Disorganization + Gyrification + Subcortical Volume	0.378	n.s.		

Notes: [a]Subgroups are based on functional outcome at 6-year follow-up.
UHR = Individuals at ultra-high risk for psychosis; mGAF = Modified Global Assessment of Functioning; SIPS = Structured Interview for Prodromal Symptoms; n.s. = non-significant.
*$P < 0.05$.
**$P < 0.01$.

year follow-up, resilient UHR individuals had lower scores than non-resilient UHR individuals on all symptom scales, as well as a higher mGAF score ($P < 0.001$ for all comparisons; Table I). The groups did not differ in terms of the use of psychotropic medication.

Comparison to Previous Literature (Binary Classification)

To allow comparison to earlier studies, we first built binary classification models separating UHR individuals from TDCs as well as resilient UHR individuals from non-resilient UHR individuals:

Separation of UHR individuals and TDCs (model A)

The highest average accuracy (using LOO cross validation) was achieved using subcortical volumes (64%) with a sensitivity of 59%, a specificity of 68% and significance of $P = 0.018$. Surface area and cortical thickness features also yielded significant models with a sensitivity of 66%, a specificity of 56%, and an average accuracy of 61% ($P = 0.005$) for surface area and a sensitivity of 56%, a specificity of 63%, and an average accuracy of 60% ($P = 0.007$) for cortical thickness. Other brain measures did not discriminate between UHR individuals and TDCs. Results are provided in Supporting Information Table II.

Separation of resilient UHR individuals and non-resilient UHR individuals (model B)

Models that separated resilient from non-resilient UHR individuals on a binary scale (model B, Table II) included those using cortical volume (LOO average accuracy of 69%, $P = 0.015$), gyrification (LOO average accuracy of 73%, $P = 0.016$), and subcortical volumes (LOO average accuracy of 67%, $P = 0.047$). Figure 1 shows the weight-vector w from the best model (gyrification) mapped onto the brain. Warm colors indicate that increases in LGI contribute to being classified as non-resilient, while cool colors indicate that decreases in LGI contribute to being classified as non-

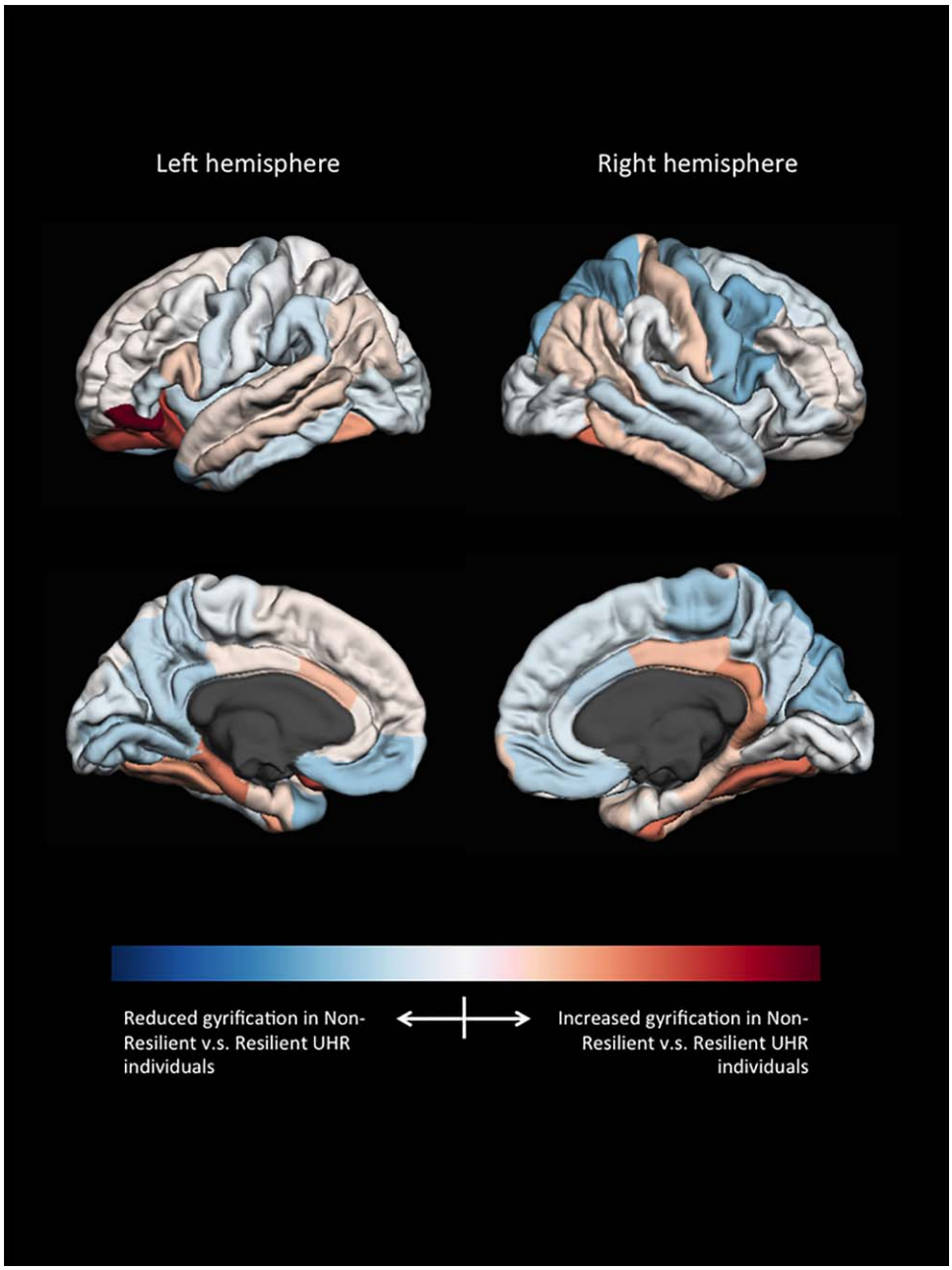

Figure 1.
Weight-vector (w-map) of the model best separating resilient from non-resilient UHR individuals (gyrification). [Color figure can be viewed at wileyonlinelibrary.com]

resilient. Significant contributions to the model's discriminative pattern were found for the inferior frontal gyrus (pars orbitalis), fusiform gyrus, lateral orbitofrontal gyrus, and precentral gyrus. For subcortical volumes, substantial contributions were found for the right cerebellum, corpus callosum, amygdala, thalamus, and basal ganglia (pallidum). The binary classification model with the highest average accuracy included a combination of gyrification, subcortical volumes, and SIPS disorganization (which differed between groups at baseline) with an average accuracy of 82% (sensitivity 69%, specificity 94%, $P = 0.003$). For the model with highest accuracy, predictive value is shown in a ROC curve in Figure 2 with an Area-Under-the-Curve of 0.753.

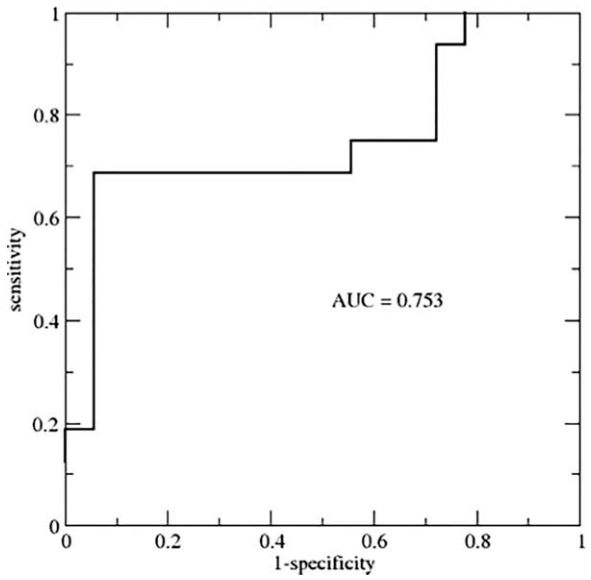

Figure 2.

Receiver-operator-curve of the class probability values obtained from the SVM model of subcortical volume, gyrification and SIPS Disorganization. *AUC = area under the curve; SVM = support vector machine; SIPS = Structured Interview for Prodromal Symptoms.*

Long-Term Outcome Prediction of UHR Individuals on a Continuous Scale (Models C)

Predictions of clinical and functional outcome at six-year follow-up on a continuous scale are shown in Table II. Models with predictive correlation coefficients close to zero are not listed. Gyrification and subcortical volumes yielded significant predictive correlations with level of functioning at six-year follow-up (Model C1: $r = 0.382$, $P = 0.048$ and $r = 0.424$, $P = 0.008$ respectively), negative symptoms (Model C3: $r = 0.349$, $P = 0.048$ for subcortical volumes), and disorganization symptoms (Model C4: $r = 0.411$, $P = 0.044$ for gyrification). Predictions and weights of the model with the highest predictive value (Model C1, baseline subcortical volumes predicting GAF score at six-year follow-up) are shown in Figure 3. Substantial contributions were primarily found for the corpus callosum, caudate nucleus, thalamus, pallidum, cerebellum, amygdala and third and lateral ventricle. Of these, only the weight of the corpus callosum (mid posterior part) reached statistical significance ($P = 0.039$). The addition of SIPS disorganization score to the brain models did not improve predictive value with correlations of 0.327 with GAF score at six-year follow-up (*n.s.*) and 0.378 with SIPS Disorganization score at six-year follow-up (*n.s.*).

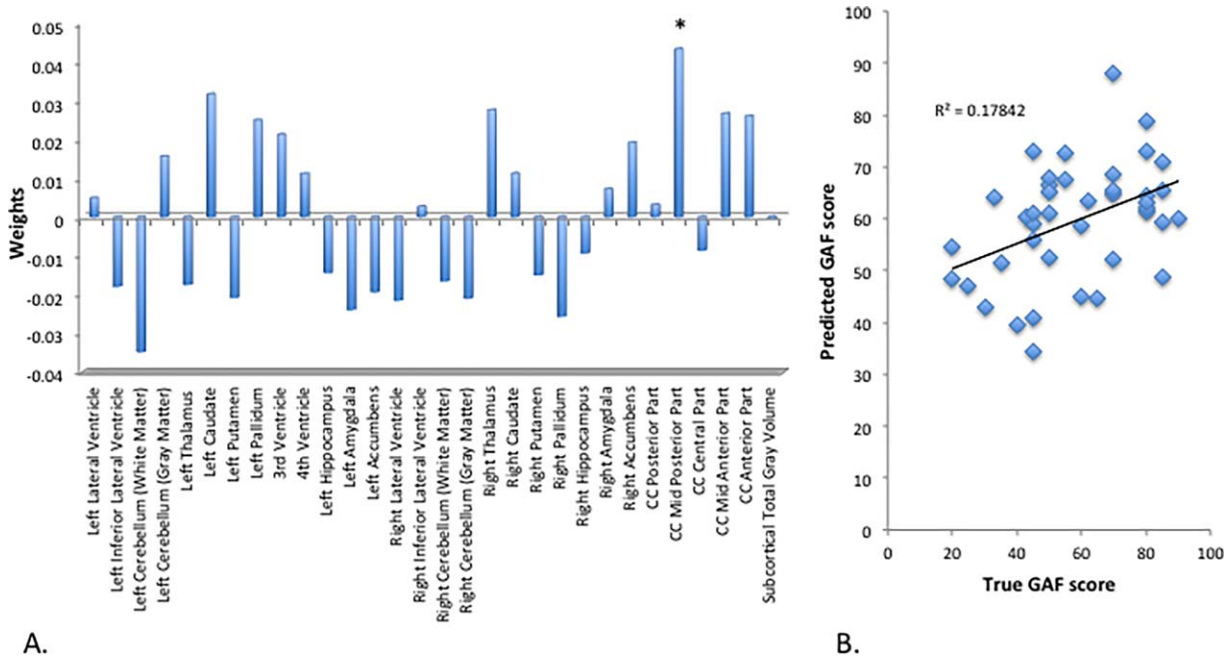

Figure 3.

SVR subcortical volumes predicting GAF-score at 6-year follow-up. (**A**) Weight-vector bar chart. (**B**) Scatter plot with linear trendline. *CC = corpus callosum; SVR = support vector regression; GAF = global assessment of functioning.* * = $P < 0.05$. [Color figure can be viewed at wileyonlinelibrary.com]

Post-Hoc Analyses on a Matched Sample of UHR Individuals and TDC

To test the robustness of results, extra analyses were performed on the same sample of UHR individuals and TDC, but, instead of regressing gender and age out, the sample was matched on age and gender. Results are provided in Supporting Information Tables III and IV and show that results are comparable with minor differences between the two strategies.

DISCUSSION

In this paper, we set out to use baseline structural MRI data to predict long-term functional and clinical outcome in adolescents at UHR for psychosis on an individual basis. Predictions of long-term functioning on a continuous scale led to predictive correlations between baseline MRI measures and level of functioning (mGAF score), and negative and disorganization symptoms at six-year follow-up. The highest correlation (0.42) was found between baseline subcortical volumes and long-term level of functioning. Subcortical volumes with substantial contributions to this model included the corpus callosum, caudate nucleus, thalamus, pallidum, and amygdala as well as cerebellum and third and lateral ventricle.

To date, only one other study has tried to make quantitative predictions of clinical symptoms in UHR individuals at follow-up [Tognin et al., 2013]. Tognin and colleagues reported a predictive correlation of 0.34 between baseline cortical thickness and progression of total positive- and negative symptoms from baseline to two-year follow-up. In this study, highest prediction accuracies were found between gyrification and disorganization symptoms ($r = 0.41$) and subcortical volumes and level of functioning ($r = 0.42$). We looked at positive, negative and disorganization symptoms separately and found that it was not positive symptoms, but rather level of functioning and disorganization symptoms that were most accurately predicted by baseline MRI data. This underscores the idea that not only positive symptoms are important for long-term outcome, but that other symptom clusters and level of functioning are equally important for the long-term outcome of UHR individuals [Carrión et al., 2013; Fusar-Poli et al., 2013; Fusar-Poli and Borgwardt, 2007]. SVR has been used more often for predicting brain age in psychiatric disorders, especially in schizophrenia [Koutsouleris et al., 2014b; Schnack et al., 2016]. By measuring the difference between chronological and neuroanatomical brain age, accelerated brain age in schizophrenia as well as the UHR state was shown. The focus in this study was on predicting clinical and functional outcome, where we attempted to remove possible influences of age (and sex). In a future study, it would be valuable to investigate if, using brain measures that are not corrected for age, these findings of accelerated aging of the brain can be replicated in our UHR cohort.

Most machine-learning studies attempting to classify UHR individuals have focused on binary classification separating UHR individuals who developed psychosis from those who did not. The accuracy of their predictions based on structural MRI data range from 80 to 84% [Koutsouleris et al., 2009; Koutsouleris et al., 2012; Koutsouleris et al., 2014a]. To allow comparison to these studies we complementary separated resilient from non-resilient UHR individuals in a binary manner. We achieved accuracies ranging from 67 to 73%, based on gyrification, cortical volume and subcortical volumes. Substantial contributions to these models came from frontal (gyrification) and temporal (cortical volume) brain regions as well as corpus callosum, amygdala, thalamus, basal ganglia and cerebellum. These areas are in agreement with earlier UHR studies [for review see: Wood et al., 2013]. The combination of SIPS disorganization symptom score and brain measures improved accuracy, with maximum accuracy of 82%. It should be noted that this model, with in total 99 features from three different modalities included, has the highest complexity as compared to the other models. More complex models are more prone to overfitting and may consequently show poorer generalization in new samples. Interestingly, specificity was particularly high for this model, with only one non-resilient individual being misclassified as resilient at six-year follow-up. This is of great importance as this will result in the inclusion of less false-positive UHR individuals and consequently, might prevent unnecessary treatment. Only one previous study has compared good and poor functional outcome instead of comparing individuals with and without transition to psychosis [Kambeitz-Ilankovic et al., 2015] and reported an accuracy of 82%. Intriguingly, their most accurate model was based on surface area data while we found that surface area did not discriminate between UHR individuals. We have previously reported that surface area did not differ between UHR groups at young age (12–18 years), but that differences appeared with development [De Wit et al., 2016]. As such, age differences of the samples could explain the discrepancy.

Also our subcortical findings overlap with those of earlier studies classifying UHR individuals with later transition to psychosis. Especially the thalamus and basal ganglia have been consistently reported to differ between groups [Koutsouleris et al., 2009; Koutsouleris et al., 2012; Mittal et al., 2010; Wood et al., 2008; Wood et al., 2013]. As our classification is based on functional outcome, this suggests that changes in subcortical structures may not be specific for the development of psychosis, but may rather be associated with general psychopathology.

Although the majority of studies have focused on binary classification, we believe it is important to predict long-term functioning on a continuous scale because of two reasons. First, it is not necessary to set a threshold, which

could be of great advantage as there has been much discussion about the validity of the threshold for psychosis [Fusar-Poli and Van Os, 2013; Yung et al., 2010]. With a continuous scale it is not only possible to separate the poor functioning from good functioning UHR individuals but also to make predictions on a gradual scale within the poor and good functioning subgroups, possibly increasing accuracy. Second, it is possible to look at different clinical scales separately (e.g., positive, negative, and disorganization symptoms). This could be relevant information when choosing appropriate intervention.

Limitations of our study include our relatively modest sample size and possible confounders. Due to considerable attrition during our long follow-up, sample sizes are modest, especially when only the UHR group is included for analysis. With relatively many parameters this may have caused overfitting. However, the possibility of overfitting (and also underfitting) affects every learning algorithm and validation on an entirely independent dataset is therefore needed to test the robustness and generazibility of our results. Second, as a large number of our UHR individuals were on psychotropic medication, this could have played a role in the classification. However, the number of individuals on medication did not differ between resilient and non-resilient UHR individuals. Another possible confounder was the skewed distribution of gender between resilient and non-resilient groups, even though the difference did not reach statistical significance. However, as we were investigating brain anatomy at one time point rather than its development over time, we chose to regress gender out rather than to match groups for it. With extra analyses on matched groups, we have shown that the results and the conclusions drawn from them are robust. Strong points of this paper include the long follow-up period (six years) and the use of SVR models to predict outcome on a continuous scale.

In conclusion, our results show that structural MRI data can be used to quantitatively predict long-term functional and clinical outcome in UHR individuals with medium effect sizes, suggesting that there may be scope for predicting outcome at the individual level. This finding is clinically important, as individual outcome predictions on a gradual scale might allow personalized medicine. In addition, of scientific interest, it permits studying different clinical and functional scales separately.

ACKNOWLEDGMENTS

The authors thank all subjects for their participation in this study. We thank Dr. Mirjam Simons-Sprong, Anneke Schouten, Nieke Kobussen, and Petra Klaassen for their help with subject recruitment and testing.

The funding agency did not have any role in study design; collection, analysis or interpretation of the data; the writing of the report; or in the decision to submit the paper for publication.

REFERENCES

Addington J, Cornblatt BA, Cadenhead KS, Cannon TD, McGlashan TH, Perkins DO, Seidman LJ, Tsuang MT, Walker EF, Woods SW, Heinssen R (2011): At clinical high risk for psychosis: Outcome for nonconverters. Am J Psychiatry 168: 800–805.

Addington J, Heinssen R (2012): Prediction and prevention of psychosis in youth at clinical high risk. Annu Rev Clin Psychol 8: 269–289.

Carmona S, Proal E, Hoekzema EA, Gispert J-D, Picado M, Moreno I, Soliva JC, Bielsa A, Rovira M, Hilferty J, Bulbena A, Casas M, Tobeña A, Vilarroya O (2009): Ventro-striatal reductions underpin symptoms of hyperactivity and impulsivity in attention-deficit/hyperactivity disorder. Biol Psychiatry 66: 972–977.

Carrión RE, McLaughlin D, Goldberg TE, Auther AM, Olsen RH, Olvet DM, Correll CU, Cornblatt BA (2013): Prediction of functional outcome in individuals at clinical high risk for psychosis. JAMA Psychiatry 70:1133–1142.

Chang CC, Lin CJ (2011): LIBSVM. A library for support vector machines. ACM Trans Intell Syst Technol 2:27.

Dale AM, Fischl B, Sereno MI (1999): Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9: 179–194.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006): An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980.

Fischl B, Sereno MI, Dale AM (1999): Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9:195–207.

Fusar-Poli P, Borgwardt S (2007): Integrating the negative psychotic symptoms in the high risk criteria for the prediction of psychosis. Med Hypotheses 69:959–960.

Fusar-Poli P, Van Os J (2013): Lost in transition: Setting the psychosis threshold in prodromal research. Acta Psychiatr Scand 127:248–252.

Fusar-Poli P, Borgwardt S, Bechdolf A, Addington J, Riecher-Rössler A, Schultze-Lutter F, Keshavan M, Wood S, Ruhrmann S, Seidman LJ, Valmaggia L, Cannon T, Velthorst E, De Haan L, Cornblatt B, Bonoldi I, Birchwood M, McGlashan T, Carpenter W, McGorry P, Klosterkötter J, McGuire P, Yung A (2013): The psychosis high-risk state: A comprehensive state-of-the-art review. JAMA Psychiatry 70:107–120.

Hall RC (1995): Global assessment of functioning. A modified scale. Psychosomatics 36:267–275.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, Bießmann F (2014): On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110.

Kambeitz-Ilankovic L, Meisenzahl EM, Cabral C, von Saldern S, Kambeitz J, Falkai P, Möller H-J, Reiser M, Koutsouleris N (2015): Prediction of outcome in the psychosis prodrome using neuroanatomical pattern classification. Schizophr Res 173:159–165.

Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, Schmitt G, Zetzsche T, Decker P, Reiser M, Möller 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.

Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Möller H-J, Riecher-Rössler A (2012): Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: Results from the FePsy study. Schizophr Bull 38: 1234–1246.

Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, Smieskova R, Studerus E, Kambeitz-Ilankovic L, von Saldern S, Cabral C, Reiser M, Falkai P, Borgwardt S (2014a): Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. Schizophr Bull 1–12.

Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, Bottlender R, Frodl T, Falkai P, Riecher-Rossler A, Moller HJ, Reiser M, Pantelis C, Meisenzahl E (2014b): Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. Schizophr Bull 40:1140–1153.

Maxwell ME (1982): Family Interview for Genetic Studies (FIGS): A Manual for FIGS. Clinical Neurogenetics Branch, Intramural Research Program, National Institute of Mental Health, Bethesda, Maryland.

McGlashan TH, Miller TJ, Woods SW (2001): Structured Interview for Prodromal Syndromes (SIPS)—version 3.0. PRIME Research Clinic, Yale School of Medicine, New Haven.

Mittal VA, Walker EF, Bearden CE, Walder D, Trottman H, Daley M, Simone A, Cannon TD (2010): Markers of basal ganglia dysfunction and conversion to psychosis: Neurocognitive deficits and dyskinesias in the prodromal period. Biol Psychiatry 68:93–99.

Mourao-Miranda J, Reinders AATS, Rocha-Rego V, Lappin J, Rondina J, Morgan C, Morgan KD, Fearon P, Jones PB, Doody GA, Murray RM, Kapur S, Dazzan P (2011): Individualized prediction of illness course at the first psychotic episode: A support vector machine MRI study. Psychol Med 42: 1037–1047.

Nieuwenhuis M, van Haren NEM, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG (2012): Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. Neuroimage 61:606–612.

Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A (2012): Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. Neurosci Biobehav Rev 36:1140–1152.

Schnack HG, Nieuwenhuis M, van Haren NEM, Abramovic L, Scheewe TW, Brouwer RM, Hulshoff Pol HE, Kahn RS (2014): Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage 84:299–306.

Schnack HG, van Haren NEM, Nieuwenhuis M, Hulshoff Pol HE, Cahn W, Kahn RS (2016): Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. Am J Psychiatry 173:607–616.

Schultze-Lutter F, Klosterkötter J (2002): Bonn Scale for the Assessment of Basic Symptoms—Prediction list (BSABS-P). Cologne: University of Cologne.

Simon AE, Borgwardt S, Riecher-Rössler A, Velthorst E, de Haan L, Fusar-Poli P (2013): Moving beyond transition outcomes: Meta-analysis of remission rates in individuals at high clinical risk for psychosis. Psychiatry Res 209:266–272. http://www.ncbi.nlm.nih.gov/pubmed/23871169.

Simon AE, Velthorst E, Nieman DH, Linszen D, Umbricht D, de Haan L (2011): Ultra high-risk state for psychosis and non-transition: A systematic review. Schizophr Res 132:8–17.

Smola, A and Schölkopf B (1998): Tutorial on Support Vector Regression.

Sprong M, Becker HE, Schothorst PF, Swaab H, Ziermans TB, Dingemans PM, Linszen D, van Engeland H (2008): Pathways to psychosis: A comparison of the pervasive developmental disorder subtype multiple complex developmental disorder and the "At Risk Mental State." Schizophr Res 99:38–47.

Tognin S, Pettersson-Yeo W, Valli I, Hutton C, Woolley J, Allen P, McGuire P, Mechelli A (2013): Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis. Front Psychiatry 4:187.

Vapnik VN (1999): An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999.

Wechsler D (1997): Wechsler Adult Intelligence Scale-III NL: Afname en scoringshandleiding [Manual]. Psychological Corporation Ltd., Harcourt Publishers, Amsterdam, the Netherlands.

Wechsler D (2002): Wechsler Intelligence Scale for Children-III NL: Handleiding en verantwoording [Manual]. Psychological Corporation Ltd., Harcourt Assessment, Amsterdam, the Netherlands.

Wit de S, Schothorst PF, Oranje B, Ziermans TB, Durston S, Kahn RS (2014): Adolescents at ultra-high risk for psychosis: Long-term outcome of individuals who recover from their at-risk state. Eur Neuropsychopharmacol 24:865–873.

Wit de S, Wierenga LM, Oranje B, Ziermans TB, Schothorst PF, van Engeland H, Kahn RS, Sarah D. (2016): Brain development in adolescents at ultra-high risk for psychosis: Longitudinal changes related to resilience. Neuroimage Clin 12:542–549.

Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57:328–349. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26254595.

Wood SJ, Pantelis C, Velakoulis D, Yücel M, Fornito A, McGorry PD (2008): Progressive changes in the development toward schizophrenia: Studies in subjects at increased symptomatic risk. Schizophr Bull 34:322–329.

Wood SJ, Reniers RLEP, Heinze K (2013): Neuroimaging findings in the at-risk mental state: A review of recent literature. Can J Psychiatry 58:13–18.

Yung AR, McGorry PD (1996): The initial prodrome in psychosis: Descriptive and qualitative aspects. Aust N Z J Psychiatry 30: 587–599.

Yung AR, Nelson B, Thompson A, Wood SJ (2010): The psychosis threshold in Ultra High Risk (prodromal) research: Is it valid? Schizophr Res 120:1–6.

Zarogianni E, Moorhead TWJ, Lawrie SM (2013): Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. NeuroImage Clin 3:279–289.

Ziermans TB, Schothorst PF, Sprong M, van Engeland H (2011): Transition and remission in adolescents at ultra-high risk for psychosis. Schizophr Res 126:58–64.

Ziermans T, de Wit S, Schothorst P, Sprong M, van Engeland H, Kahn R, Durston S (2014): Neurocognitive and clinical predictors of long-term outcome in adolescents at ultra-high risk for psychosis: A 6-year follow-up. PLoS One 9:e93994.