Early Recognition and Disease Prediction in the At-Risk Mental States for Psychosis Using Neurocognitive Pattern Classification

Nikolaos Koutsouleris*,1, Christos Davatzikos2, Ronald Bottender1, Katja Patschere-Kliche1, Johanna Scheuerecker1, Petra Decker1, Christian Gaser1, Hans-Jürgen Möller1, and Eva M. Meisenzahi1

1Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Nussbaumstrasse 7, 80336 Munich, Germany; 2Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA; 3Department of Psychiatry, Friedrich-Schiller-University, Jena, Germany
*To whom correspondence should be addressed; tel.: +49-89-5160-5717, fax: +49 89 5160-3413, e-mail: Nikolaos.Koutsouleris@med.uni-muenchen.de

Background: Neuropsychological deficits predate overt psychosis and overlap with the impairments in the established disease. However, to date, no single neurocognitive measure has shown sufficient power for a diagnostic test. Thus, it remains to be determined whether multivariate neurocognitive pattern classification could facilitate the diagnostic identification of different at-risk mental states (ARMS) for psychosis and the individualized prediction of illness transition.

Methods: First, classification of 30 healthy controls (HC) vs 48 ARMS individuals subgrouped into 20 “early,” 28 “late” ARMS subjects was performed based on a comprehensive neuropsychological test battery. Second, disease prediction was evaluated by categorizing the neurocognitive baseline data of those ARMS individuals with transition (n = 15) vs non transition (n = 20) vs HC after 4 years of follow-up. Generalizability of classification was estimated by repeated double cross-validation.

Results: The 3-group cross-validated classification accuracies in the first analysis were 94.2% (HC vs rest), 85.0% (early at-risk subjects vs rest), and, 91.4% (late at-risk subjects vs rest) and 90.8% (HC vs rest), 90.8% (converters vs rest), and 89.0% (nonconverters vs rest) in the second analysis. Patterns distinguishing the early or late ARMS from HC primarily involved the verbal learning/memory domains, while executive functioning and verbal IQ deficits were particularly characteristic of the late ARMS. Disease transition was mainly predicted by executive and verbal learning impairments.

Conclusions: Different ARMS and their clinical outcomes may be reliably identified on an individual basis by evaluating neurocognitive test batteries using multivariate pattern recognition. These patterns may have the potential to substantially improve the early recognition of psychosis.

Key words: individualized psychosis prediction/multivariate analysis/neurocognitive test battery

Introduction

Neurocognitive deficits have been described as a core feature of schizophrenic psychosis affecting cognitive performance particularly within the domains of processing speed, verbal learning, and executive functioning.1 Some of these deficits seem to emerge early in the course of the disease and may be traced even in different at-risk mental states (ARMS) for the illness. In this context, the state of ultrahigh risk (UHR) for psychosis characterized by attenuated and/or brief limited intermittent psychotic symptoms may involve abnormalities across a range of cognitive abilities including speed of processing, working memory, sustained attention, cognitive flexibility, and verbal memory.2–7 This cross-domain neurocognitive pattern overlaps with the profound deficits observed in the established disease, albeit not reaching the level of impairment seen in the full clinical picture of the disorder. This latter observation suggests that task performance may either deteriorate within specific cognitive domains during transition to psychosis as shown by cross-sectional and longitudinal findings in different ARMS populations,7–9 or alternatively may result from a higher conversion rate in the UHR compared with lower risk states. Furthermore, neuropsychological investigations of an early, ie, initial ARMS as defined by subtle cognitive-perceptual “basic” symptoms10,11 found that these individuals already exhibited executive control impairments compared with healthy volunteers.9 Taken together, these data revealed patterns of neurocognitive deficits that predate the onset of the overt disease and...
Neurocognition-based Early Recognition and Disease Prediction in ARMS

hence may substantially improve the difficult clinical detection of initial prodromal stages. This could facilitate an earlier commencement of therapeutic approaches aimed at ameliorating signs and symptoms and even preventing the onset of psychosis.12

This promising perspective has increasingly moved into the research focus because the current risk assessment strategies have to date produced only modest 1-year prediction rates ranging from 9% to 54% by relying solely on operationalized psychopathological criteria.13 In this regard, previous investigations found that reduced baseline performance in verbal IQ, verbal/working memory, and olfactory identification tasks may distinguish subsequent converters from non converters in a clinically defined ARMS.3,14,15 Furthermore, Lencz et al5 and Riecher-Rössler et al16 consistently reported that the prediction of disease transition could be substantially improved to an accuracy of up to 80% by using linear regression models that combined both psychopathological and neurocognitive measures. In contrast to these results, a recent multicenter investigation of the North American Longitudinal Prodromal Study17 found that neuropsychological data did not provide additional predictive power beyond multivariate clinical prediction models.18 Moreover, it is still unclear whether a neurocognition-based prediction of psychosis will generalize to ARMS individuals not used for deriving the respective diagnostic models. This means that diagnostic validity has to be assessed either in independent test populations or estimated using rigorous cross-validation (CV) that separates the training from validation data throughout the model generation process. Thus, it has to be elucidated whether neurocognitive abnormalities in the ARMS reliably enhance the early detection of different prodromal stages and the prediction of subsequent disease manifestation to a level allowing for single-subject inferences.

A valid answer to this question has to address major methodological obstacles. First, the considerable intradividual variability of single neuropsychological tests may blur diagnostically relevant information. This low signal-to-noise ratio of single test measures could be substantially increased by using multivariate approaches in order to extract high-dimensional discriminative patterns across comprehensive neuropsychological test batteries. However, it will be insufficient to detect discriminative patterns that perfectly describe specific diagnostic boundaries in a given training population. This overfitting problem may be resolved by using a machine-learning framework capable of detecting multivariate patterns of neuropsychological deficits that generalize well to unseen individuals.19 In this context and based on our own experiences with Magnetic Resonance Imaging (MRI)-based multivariate prediction models,20–23 we employed the support-vector machine (SVM)24 in conjunction with ensemble learning methods25,26 in order to evaluate whether potentially complex patterns of cognitive ability derived from the combination of several neuropsychological tests may facilitate the individualized recognition of different ARMS and the prediction of frank psychosis.

Methods

Study Participants

Forty-eight individuals in an ARMS for psychosis and 30 healthy controls (HC) matched group-wise for age, gender, and premorbid verbal IQ (tables 3 and 4) were recruited at the Early Detection and Intervention Center for Mental Crises, Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Germany, for neuropsychological testing using operationalized criteria as detailed in22,27–34 and table 1. These criteria were based on a 2-stage concept of the ARMS, distinguishing between (1) an early ARMS (ARMS-E), mainly defined by the presence of basic symptoms9–11 and associated with an increased, but not imminent risk of psychosis and (2) a “late” ARMS (ARMS-L), characterized by an ultrahigh risk for psychosis following the Personal Assessment of Crisis Evaluation (PACE) criteria.37,38 Candidate ARMS and HC individuals were carefully screened for the exclusion criteria listed in table 1 by evaluating the personal and familial history using a semistructured clinical interview and the structured clinical interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, (DSM-IV).39 Additionally, ARMS individuals were rated using the Global Assessment of Functioning (GAF) Scale of the DSM-IV, the Positive and Negative Symptom Scale (PANSS),40 and the Montgomery-Åsberg Depression Rating Scale (MADRS).41 It is of note that 80%/0% of the ARMS/HC subjects analyzed in the present study overlapped with the cohort previously used for a MRI-based pattern recognition analysis.22 Moreover, an overlapping population was recently assessed for neuroanatomical correlates of executive dysfunction.42

Included ARMS individuals were seen weekly in the first month, monthly in the first year, quarterly in the second year, and thereafter annually to detect possible transitions to psychosis according to the criteria of Yung et al35: PANSS scores of 4 or above on the hallucination item (P3) or scores of 5 or above on the unusual thought content (G9), suspiciousness (P6), or conceptual disorganization (P2) items. Symptoms had to occur daily and persist for more than 1 week to be deemed a transition to frank psychosis. ARMS individuals were assigned to the transition group (ARMS-T) if they met these criteria once during the follow-up period of 4 years and had a diagnosis of schizophrenia spectrum disorder 1 year after transition following the International Classification of Diseases-10 research criteria. Follow-up information could be obtained from 35 subjects (3.9 (SD: 1.2) years after study inclusion, including 15 converters (ARMS-T: n = 11,
Table 1. Inclusion Criteria for the ARMS Subjects/Exclusion Criteria

ARMS-E: ARMS subjects without Attenuated Psychotic Symptoms (APS) and/or Brief Limited Intermittent Psychotic Symptoms (BLIPS) …
(1) … having one or more of the following basic symptoms appeared first at least 12 mo prior to study inclusion and several times per week during the last 3 mo.
   Thought interferences
   Thought perseveration
   Thought pressure
   Thought blockages
   Disturbances of receptive language, either heard or read
   Decreased ability to discriminate between ideas and perception, fantasy and true memories
   Unstable ideas of reference (subject-centrism)
   Derealization
   Visual perception disturbances
   Acoustic perception disturbances
   and/or
(2) … showing a reduction in the Global Assessment of Functioning Score (DSM-IV) of at least 30 points (within the past year) combined with at least one of the following trait markers:
   First-degree relative with a lifetime-diagnosis of schizophrenia or a schizophrenia spectrum disorder
   Pre- or perinatal complications

ARMS-L: ARMS subjects with or without basic symptoms, with or without global functioning & trait markers …
(1) … having at least one of the following APS within the last 3 mo, appearing several times per week for a period of at least 1 wk:
   Ideas of reference
   Odd beliefs or magical thinking
   Unusual perceptual experiences
   Odd thinking and speech
   Suspiciousness or paranoid ideation and/or
(2) … having at least one of the following BLIPS, defined as the appearance of one of the following psychotic symptoms for less than 1 wk (interval between episodes at least 1 wk), resolving spontaneously:
   Hallucinations
   Delusions
   Formal thought disorder
   Gross disorganized or catatonic behavior

Exclusion criteria
Disease transition as defined by Yung et al\textsuperscript{15}
A past or present diagnosis of schizophrenia spectrum and bipolar disorders, as well as delirium, dementia, amnestic, or other cognitive disorders, mental retardation, and psychiatric disorders due to a somatic factor, following the DSM-IV criteria
Developmental and personality disorders, following the DSM-IV criteria
Alcohol or drug abuse within 3 mo prior to examination, following the DSM-IV criteria
A past or present inflammatory, traumatic or epileptic diseases of the central nervous system
Any previous treatment with antipsychotics prior to neurocognitive assessment
Healthy controls: positive familial history of schizophrenic or affective psychoses in the first-degree relatives

Note: ARMS, At-Risk Mental State for psychosis; ARMS-E, early ARMS subgroup; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Adopted From Häfner et al.\textsuperscript{36}
Cognitive domains were defined according to Schultze-Lutter et al.50

Table 2. Neuropsychological Test Battery

<table>
<thead>
<tr>
<th>Cognitive Domain</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premorbid verbal IQ</td>
<td>1. Raw score correct</td>
</tr>
<tr>
<td>Mehrfach-Wortschatztest B (MWT-B)43</td>
<td></td>
</tr>
<tr>
<td>Processing speed</td>
<td>2. Time to completion (s)</td>
</tr>
<tr>
<td>Trail-Making Test, part A (TMT-A)44</td>
<td>3. Raw score correct</td>
</tr>
<tr>
<td>Digit Symbol Test (DST, WAIS-III)45</td>
<td></td>
</tr>
<tr>
<td>Working Memory</td>
<td>4. Raw score correct</td>
</tr>
<tr>
<td>Digit Span Test (DS, WAIS-III)45</td>
<td>5. Raw score correct</td>
</tr>
<tr>
<td>Letter Number Span Test (LNS)46</td>
<td>6. Error score</td>
</tr>
<tr>
<td>Subject-Ordered Pointing Task (SOPT)57</td>
<td></td>
</tr>
<tr>
<td>Verbal Learning and Memory</td>
<td>7. Sum of raw score correct after trials 1–5 (RAVLT-IR)</td>
</tr>
<tr>
<td>Rey Auditory Verbal Learning Test (RAVLT)48</td>
<td>8. Raw score correct after delayed recall (RAVLT-DR)</td>
</tr>
<tr>
<td>Executive Functions</td>
<td>9. Retention: difference between raw score correct in trial 5 and delayed recall (RAVLT-Ret)</td>
</tr>
<tr>
<td>Trail-Making Test, part B (TMT-B)44</td>
<td></td>
</tr>
<tr>
<td>Verbal Fluency (letters) (VF)49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10. Time to completion (s)</td>
</tr>
<tr>
<td></td>
<td>11. Difference between TMT-B and TMT-A (TMT-[B–A])</td>
</tr>
<tr>
<td></td>
<td>12. Sum of correct responses</td>
</tr>
</tbody>
</table>

Note: Cognitive domains were defined according to Schultze-Lutter et al.50

because of their primary strength, which is providing optimal methods for classifying single individuals, rather than simply describing statistical group differences (see also the methodological comparison of different classification algorithms in table 5 of the online supplementary material). In our case, the 12 different neuropsychological test scores were used by the SVM to determine the best nonlinear classification model that reliably predicted the study participants’ group membership in 3 different analyses: (1) HC vs ARMS, (2) HC vs ARMS-E vs ARMS-L, and (3) HC vs ARMS-NT vs ARMS-T. As customary in predictive analytics, the SVM model is constructed from one set of subjects (the training sample) and applied to a different set of subjects (the test sample), using CV (see Methods in online supplementary material). This process produces an unbiased estimate of the expected diagnostic and prognostic accuracy of the SVM model on new individuals, rather than merely fitting the current patient group. The principles of generating and validating predictive SVM models on separate training and testing samples of a study population have been detailed in our previous work.22

In summary, the neurocognitive data of the training sample were first adjusted for age and gender effects using partial correlations. The residuals were projected to a high-dimensional feature space using the radial basis functions in order to account for nonlinear relations between the neurocognitive data and the group membership of the training subjects. In this feature space, the SVM algorithm found the optimal between-group boundary by maximizing the geometric distance between the most similar subjects of opposite groups (the support vectors [SV]).24,52,53

It has been shown that this “maximum-margin” principle in conjunction with the nonlinear projection generates classification rules that are adaptive to subtle between-group differences and therefore generalize well to new individuals.24

The group membership of unseen test subjects was predicted by first applying the adjustment and nonlinear transformation parameters of the training sample to the test sample. Then, the trained binary SVM models (eg.: HC vs ARMS-E) determined the geometric positions of the test subjects relative to the “learned” decision boundary, resulting in a decision value and a group membership prediction for each test subject. In the 3-group analyses 1 and 2, we used these decision values to construct multigroup classifiers, where the binary SVM model with the maximum decision value decided about the test subject’s group membership.

This SVM training and validation scheme was wrapped in a repeated nested CV analysis (see Methods in online supplementary material).23,54 On the outer loop of this analysis, we performed 50 repetitions of the following CV cycle. First, the order of the subjects was permuted within each group and then the entire population was split into ten nonoverlapping samples. Each of these samples was iteratively held back as validation data, while the 9 remaining samples entered the inner CV loop. This outer CV loop produced a robust and unbiased estimate of classification generalizability based on a large number of validation samples that were strictly separated from the entire training process performed on the inner loop. At this inner loop, we used 10-fold CV with 10 repetitions to generate ensembles of SVM models. More specifically, for each validation sample at the outer CV level, 100 different training data
Table 3. Analysis of Sociodemographic, Clinical and Global Anatomical Variables

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>ARMS</th>
<th>$T/\chi^2$</th>
<th>ARMS-E</th>
<th>ARMS-L</th>
<th>$F/\chi^2$</th>
<th>ARMS-NT</th>
<th>ARMS-T</th>
<th>$F/\chi^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sociodemographic Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>30</td>
<td>48</td>
<td>20</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age: mean (SD) (years)</td>
<td>26.0 (2.7)</td>
<td>24.7 (5.8)</td>
<td>25.0 (5.6)</td>
<td>24.5 (5.9)</td>
<td>0.80 .456</td>
<td>25.8 (6.8)</td>
<td>22.8 (3.8)</td>
<td>2.88 .063</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender: male/female (%)</td>
<td>18/12 (60/40)</td>
<td>32/16 (66.7/33.3)</td>
<td>0.36 .630</td>
<td>13/7 (65/35)</td>
<td>0.40 .834</td>
<td>14/6 (70/30)</td>
<td>11/4 (73.3/26.7)</td>
<td>1.00 .662</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School education: mean (SD) (years)</td>
<td>12.4 (1.2)</td>
<td>12.0 (1.2)</td>
<td>1.72 .090</td>
<td>12.3 (1.0)</td>
<td>11.8 (1.2)</td>
<td>2.55 .085</td>
<td>11.9 (1.2)</td>
<td>11.9 (1.1)</td>
<td>1.69 .192</td>
<td></td>
</tr>
<tr>
<td>Number (%) of first-degree relatives with schizophrenic psychoses</td>
<td>–</td>
<td>6 (12.5)</td>
<td>–</td>
<td>4 (20)</td>
<td>2 (7.1)</td>
<td>1.64 .379</td>
<td>4 (20)</td>
<td>1 (7.1)</td>
<td>1.19 .379</td>
<td></td>
</tr>
<tr>
<td>Number (%) of first-degree relatives with affective psychoses</td>
<td>–</td>
<td>9 (18.8)</td>
<td>–</td>
<td>3 (15)</td>
<td>6 (22.2)</td>
<td>0.39 .713</td>
<td>4 (20)</td>
<td>4 (28.6)</td>
<td>0.34 .689</td>
<td></td>
</tr>
<tr>
<td><strong>Clinical Variables: mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAF score</td>
<td>–</td>
<td>58.6 (10.8)</td>
<td>–</td>
<td>60.1 (8.4)</td>
<td>57.3 (12.5)</td>
<td>0.76 .451</td>
<td>58.6 (10.9)</td>
<td>59.9 (14.2)</td>
<td>-0.24 .814</td>
<td></td>
</tr>
<tr>
<td>PANSS total score</td>
<td>–</td>
<td>62.6 (18.7)</td>
<td>–</td>
<td>59.9 (14.5)</td>
<td>64.9 (21.6)</td>
<td>-0.81 .422</td>
<td>52.9 (11.8)</td>
<td>65.5 (20.5)</td>
<td>-1.96 .062</td>
<td></td>
</tr>
<tr>
<td>PANSS positive score</td>
<td>–</td>
<td>12.8 (4.5)</td>
<td>–</td>
<td>10.8 (3.5)</td>
<td>14.5 (4.7)</td>
<td>-2.80 .008*</td>
<td>11.4 (4.2)</td>
<td>13.5 (4.2)</td>
<td>-1.20 .242</td>
<td></td>
</tr>
<tr>
<td>PANSS negative score</td>
<td>–</td>
<td>16.5 (7.7)</td>
<td>–</td>
<td>15.9 (6.7)</td>
<td>17.0 (8.6)</td>
<td>-0.42 .676</td>
<td>12.1 (4.9)</td>
<td>19.4 (8.4)</td>
<td>-2.78 .011*</td>
<td></td>
</tr>
<tr>
<td>PANSS general score</td>
<td>–</td>
<td>33.3 (9.5)</td>
<td>–</td>
<td>33.2 (7.8)</td>
<td>33.4 (11.0)</td>
<td>-0.61 .952</td>
<td>29.5 (6.1)</td>
<td>32.6 (11.3)</td>
<td>-0.74 .477</td>
<td></td>
</tr>
<tr>
<td>MADRS score</td>
<td>–</td>
<td>17.5 (9.4)</td>
<td>–</td>
<td>19.7 (7.8)</td>
<td>15.8 (10.5)</td>
<td>1.18 .247</td>
<td>17.7 (5.4)</td>
<td>11.1 (9.8)</td>
<td>2.02 .057</td>
<td></td>
</tr>
</tbody>
</table>

Note: ARMS, At-Risk Mental State for psychosis; ARMS-E, early ARMS subgroup; ARMS-L, late ARMS subgroup; ARMS-NT, nontransition subgroup; ARMS-T, transition subgroup; GAF, Global Assessment of Functioning; HC, Healthy Control subjects; PANSS, Positive and Negative Symptom Scale; F, main effect’s F value; T, Student’s t-test value; \(\chi^2\), Pearson \(\chi^2\) value; MADRS, Montgomery-Åsberg Depression Rating Scale. Schooling years, clinical, and global anatomical variables were assessed using ANCOVA designs, with group entered as main effect and age and gender defined as covariates of no interest. All \(P\) values are two-sided and exact in case of nonparametric tests.

*Significant at \(P<0.05\).
Table 3. Sociodemographic Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>HC</th>
<th>ARMS</th>
<th>ARMS-E</th>
<th>ARMS-L</th>
<th>ARMS-NT</th>
<th>ARMS-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td>30</td>
<td>48</td>
<td>20</td>
<td>28</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Gender</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Education</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Note: Values are mean ± SD. *Significant at P < 0.05.

Table 4. Statistical Analysis of Between-Group Differences in the 12 Neurocognitive Test Measures

<table>
<thead>
<tr>
<th>Test Variable</th>
<th>HC vs ARMS (t-test)</th>
<th>HC vs ARMS-E vs ARMS-L: ANOVA &amp; Post-Hoc Analyses</th>
<th>HC vs ARMS-NT vs ARMS-T: ANOVA &amp; Post-Hoc Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>Raw</td>
<td>Z</td>
</tr>
<tr>
<td>MWT-B (IQ)</td>
<td>109.7 (14.3)</td>
<td>106.7 (14.3)</td>
<td>–0.2</td>
</tr>
<tr>
<td>DST</td>
<td>67.8 (12.5)</td>
<td>60.0 (12.5)</td>
<td>–0.8</td>
</tr>
<tr>
<td>DS</td>
<td>17.4 (2.1)</td>
<td>17.0 (2.1)</td>
<td>–0.1</td>
</tr>
<tr>
<td>LNS</td>
<td>17.8 (2.8)</td>
<td>16.8 (2.8)</td>
<td>–0.7</td>
</tr>
<tr>
<td>TMT-A</td>
<td>24.8 (1.5)</td>
<td>28.3 (1.5)</td>
<td>–0.5</td>
</tr>
<tr>
<td>TMT-B</td>
<td>48.2 (13.3)</td>
<td>69.2 (13.3)</td>
<td>–1.3</td>
</tr>
<tr>
<td>TMT-(B–A)</td>
<td>23.2 (1.3)</td>
<td>40.9 (1.3)</td>
<td>–1.2</td>
</tr>
<tr>
<td>SOFT</td>
<td>1.5 (1.0)</td>
<td>4.7 (1.0)</td>
<td>–2.2</td>
</tr>
<tr>
<td>RAVLT-IR</td>
<td>64.0 (10.8)</td>
<td>55.3 (10.8)</td>
<td>–1.3</td>
</tr>
<tr>
<td>RAVLT-DR</td>
<td>14.1 (2.5)</td>
<td>11.8 (2.5)</td>
<td>–1.6</td>
</tr>
<tr>
<td>RAVLT-Rel</td>
<td>0.3 (2.0)</td>
<td>1.2 (2.0)</td>
<td>–0.9</td>
</tr>
<tr>
<td>VF</td>
<td>35.1 (10.2)</td>
<td>33.3 (10.2)</td>
<td>–0.2</td>
</tr>
</tbody>
</table>

Note: Neurocognitive test scores were adjusted for the effects of age and gender and standardized according to respective means and SDs of the HC data. For each neurocognitive test variable, statistical comparisons were conducted to evaluate group-level differences between HC vs ARMS (t-test) as well as HC vs ARMS-E vs ARMS-L and HC vs ARMS-NT vs ARMS-T (ANOVA). The Holm-Bonferroni correction was employed to correct the P values for multiple comparisons and significant between-group differences were flagged with an asterisk. In these cases, a Bonferroni post-hoc analysis was carried out to determine the significance of pairwise group differences. Abbreviations of neuropsychological test variables are detailed in table 2.

*Significant at P<0.05.
partitions were created at the inner CV level. In each of these 100 training partitions, the most discriminative sets of neurocognitive features were determined. Each of these sets was used to train a separate SVM model. Then, each of these models predicted the group membership of the unseen validation subjects on the outer loop. These predictions were averaged across all 100 training partitions to yield an ensemble decision. Finally, for each validation subject, all SVM ensemble decisions were aggregated across those outer training partitions, in which this subject had not been involved in the training process. Majority voting was used to determine the final out-of-training group membership for each validation subject (tables 5 and 6).

This ensemble learning approach has proven to achieve robust classification because it greatly reduces the risk of unfortunate selections of poorly performing single classifiers by averaging the diagnostic decisions of numerous predictive models. Furthermore, this approach improves classification generalizability particularly in small samples because it increases the classifiers’ ability to detect complex decision boundaries by means of training sample variation. 25

The performance of these binary and multigroup ensemble classifiers on the unseen validation data was measured in terms of sensitivity, specificity, balanced accuracy (BAC), positive/negative predictive value (PPV/NPV), and false positive rate. The discriminative neurocognitive patterns that formed the basis for the classifiers’ diagnostic decisions were first evaluated by computing each neurocognitive measure’s probability of being selected as discriminative feature across each classification analysis (figure 2). Second, the nonlinear discriminative neurocognitive patterns were approximated by computing for each neurocognitive measure, the difference between the respective model’s SV. 22 Then, the mean and SE of these SV differences was computed across all SVM models to form discriminative neurocognitive profiles (figure 3). Neurocognitive features with zero-crossing SEs were considered unreliable at the 95% confidence level. Third, we quantified the importance of single neurocognitive variables in the ARMS-NT vs ARMS-T analysis by removing one neurocognitive measure at a time and repeating the entire classification experiment (see figure 1 and table 3 in online supplementary material).

Results

Age, gender, education, and premorbid verbal IQ (tables 3 and 4) did not differ significantly in ARMS vs HC, HC vs ARMS-E vs ARMS-L, and HC vs ARMS-NT vs ARMS-T. A trend difference was observed in the mean age of the HC, ARMS-E, and ARMS-L groups (table 3). Thus, in order to minimize age- and also possible gender-related effects, we adjusted the data for these covariates in all classification analyses using partial correlations. Furthermore, we studied the impact of using unadjusted data in the ARMS-NT vs ARMS-T analysis by removing one neurocognitive measure at a time and repeating the entire classification experiment (see table 4 in online supplementary material).

The ARMS samples did not differ with respect to the prevalence of schizophrenic or affective psychosis in the first-degree relatives. Global functioning was similarly reduced in the ARMS-E and ARMS-L, ARMS-T and

Table 5. Two-group classification performance: The performance of the binary SVM ensemble classifiers (group ‘‘+1’’ vs group ‘‘−1’’) was evaluated (1) by constructing a binary SVM ensemble from all SVM base learners of a CV1 partition, in which the respective CV2 test subjects had not been included, (2) by computing the average decision value in each of these binary CV1 ensembles in order to determine the group membership (average decision value > 0 or < 0) of the respective CV2 test subjects and (3) through majority voting across those binary CV1 SVM ensembles, in which the CV2 test subjects had not participated in the training process (see also the Methods section for a detailed explanation of the employed ensemble learning framework).

<table>
<thead>
<tr>
<th>Binary classifiers</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>BAC (%)</th>
<th>FPR (%)</th>
<th>PPV (%)</th>
<th>NPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurocognitive SVM analysis: 30 HC vs 48 ARMS</td>
<td>24</td>
<td>46</td>
<td>2</td>
<td>6</td>
<td>95.8</td>
<td>80.0</td>
<td>87.9</td>
<td>4.2</td>
<td>92.3</td>
<td>88.5</td>
</tr>
<tr>
<td>HC vs ARMS-E vs ARMS-L</td>
<td>30</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>90.0</td>
<td>100</td>
<td>95.0</td>
<td>10.0</td>
<td>93.8</td>
<td>100</td>
</tr>
<tr>
<td>ARMS-E vs ARMS-L</td>
<td>26</td>
<td>25</td>
<td>3</td>
<td>4</td>
<td>89.3</td>
<td>86.7</td>
<td>88.0</td>
<td>10.7</td>
<td>89.7</td>
<td>86.2</td>
</tr>
<tr>
<td>Neurocognitive SVM analysis: 30 HC vs 20 ARMS-NT vs 15 ARMS-NT</td>
<td>28</td>
<td>16</td>
<td>4</td>
<td>2</td>
<td>80.0</td>
<td>93.3</td>
<td>86.7</td>
<td>20.0</td>
<td>87.5</td>
<td>89.9</td>
</tr>
<tr>
<td>HC vs ARMS-NT vs ARMS-T</td>
<td>15</td>
<td>12</td>
<td>3</td>
<td>5</td>
<td>80.0</td>
<td>75.0</td>
<td>77.5</td>
<td>20.0</td>
<td>83.3</td>
<td>70.6</td>
</tr>
</tbody>
</table>

Note: ARMS, At-Risk Mental State for psychosis; ARMS-E, early ARMS subgroup; ARMS-L, late ARMS subgroup; ARMS-NT, nontransition subgroup; ARMS-T, transition subgroup; HC, Healthy Control subjects, SVM, support-vector machine, CV, cross-validation.

Balanced accuracy (BAC), false positive rate (FPR), positive predictive value (PPV) and negative predictive value (NPV) were calculated from the confusion matrix containing the number of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP).
Neurocognition-based Early Recognition and Disease Prediction in ARMS

ARMS-NT groups. A significantly higher PANSS positive score was observed in ARMS-L vs ARMS-E individuals. The ARMS-T group was characterized by a significantly higher PANSS total score and a lower MADRS score compared with ARMS-NT. Correlations between PANSS total and negative scores and rey auditory verbal learning test (RAVLT)-DR and RAVLT-Ret measures were detected in our supplementary analysis using an exploratory threshold of \( P < .05 \) (See table 1 in online supplementary material). However, these correlations did not survive correction for multiple comparisons.

**Neurocognitive Analysis of HC vs ARMS Individuals**

**Univariate Results.** After correction for multiple comparisons, cognitive set shifting (trail-making test [TMT]-B, TMT-[B–A]), visual working memory (subject-ordered pointing task [SOPT]), and verbal learning/memory (RAVLT-IR, RAVLT-DR) was significantly impaired in ARMS vs HC subjects (table 4).

**SVM Classification Analysis.** The binary HC vs ARMS classification performance in the CV2 test data was high (\( \text{BAC} = 87.9\% \), sensitivity = 95.8\%, specificity = 80\%, PPV = 92.3\%, and NPV = 88.5\%; table 5, figure 1A). Figures 2A and 3A show that the highest feature selection probabilities and reliable group differences were observed for premorbid verbal IQ (Mehrfach-Wortschatztest [MWT]-B), cognitive set shifting (TMT-B, TMT-[B–A]), visual working memory (SOPT), and verbal learning (RAVLT-IR, RAVLT-DR).

**Neurocognitive Analysis of HC vs ARMS-E vs ARMS-L**

**Univariate Results.** The performance reductions detected in the entire ARMS group vs HC were also found in the HC vs ARMS-E vs ARMS-L analysis, with the digit symbol test (DST) being additionally involved (table 4). The post-hoc analysis revealed that the ARMS-L group showed significant performance deficits across these neurocognitive measures, ranging 1–2 SD below the HC
significant performance reductions in ARMS-E vs HC were confined to the TMT-B/C0A and the SOPT. No significant group-level differences were observed in ARMS-E vs ARMS-L.

**SVM Classification Analysis.** The 3-group classification accuracy was 88.5% (table 6; figure 1C, left). Only one HC was mislabeled as ARMS-L (HC vs rest: sensitivity = 96.7%, specificity = 91.7%, BAC = 94.2%). Two ARMS-E were mislabeled as HC and 4 as ARMS-L (ARMS-E vs rest: 70.0%, 100%, and 85.0%). Two ARMS-L were misclassified as HC (ARMS-L vs rest: 92.6%, 90%, and 91.4%). Among the binary classifiers, the highest BAC (sensitivity, specificity) of 95% (90% and 100%) was observed in HC vs ARMS-E,
followed by 88% (89.3% and 86.7%) in HC vs ARMS-L and 75.0% (100% and 50%) in ARMS-E vs ARMS-L (table 5; figure 1B, left).

The discriminative pattern of HC vs ARMS-E (figure 2B) showed high selection probabilities in the working memory (Letter number span [LNS], SOPT: >90% of...
N. Koutsouleris et al.

constituted the optimal separating decision boundary of a SVM features of all nearest-neighbor support-vector pairs that between the scaled (0,1) and adjusted (age, gender) neurocognitive patterns were approximated (1) by computing the difference vector method. Therefore, these nonlinear discriminative neurocognitive patterns were difficult to visualize due to the nonlinearity of the classification experiments. The support-vector machine (SVM) classification results of our study were obtained based on complex and subtle classification experiments. The support-vector machine (SVM) model, trained on a given inner cross-validation (CV1) training sample and (2) by calculating the arithmetic mean and SE of the mean for these difference vectors obtained across all CV1 training samples in each of three classification experiments: A: healthy vs ARMS-L, C: HC vs ARMS-NT vs ARMS-T. 

**Neurocognitive Analysis of HC vs ARMS-NT vs ARMS-T Individuals**

Univariate Results. After correcting for multiple comparisons, the significant performance differences found in the HC vs ARMS-NT vs ARMS-T analysis overlapped with the HC vs ARMS comparison, except for the RAVLT-DR (table 4). Compared with HC, both ARMS-NT and ARMS-T individuals had significant performance deficits across the TMT-B, TMT-(B–A), SOPT, and RAVLT-IR but they did not differ from each other in these measures (table 4).

**SVM Classification Analysis.** The 3-group classification accuracy was 87.7% (table 6; figure 1C, right). One HC was misclassified as ARMS-NT (HC vs rest: sensitivity = 96.7%, specificity = 85.7%, and BAC = 90.8%). Three ARMS-NT were wrongly assigned to the HC and 1 to the ARMS-T group (ARMS-NT vs rest: 80.0%, 95.6%, and 90.8%). Two ARMS-T were miscalibrated as HC and 1 as ARMS-NT (ARMS-T vs rest: 80%, 98%, 89%). The binary HC vs ARMS-T classifier attained the highest performance (BAC, sensitivity, and specificity: 93.3%, 86.7%, and 100%), followed by HC vs ARMS-NT (86.7%, 80.0%, and 93.3%) and ARMS-NT vs ARMS-T (77.5%, 80%, and 75%) (table 5; figure 1B, right).

The discriminative pattern of HC vs ARMS-NT (figure 2C) was similar to the HC vs ARMS-E profile and showed high selection probabilities (>80%) in the working memory and delayed verbal learning domain (LNS, SOPT, RAVLT-DR), which were reliable for the SOPT and RAVLT-DR measures (figure 3C). The HC vs ARMS-T pattern was characterized by frequently selected executive and working memory measures (digit span, TMT-B,
TMT-(B−A), verbal fluency: >75% models). Reliable between-group SV differences were found in cognitive flexibility, visual working memory, and verbal learning abilities (TMT-B, TMT-[B−A], SOPT, and RAVLT).

Finally, the discriminative ARMS-NT vs ARMS-T pattern particularly involved verbal and executive functions, as measured by the MWT-B, DST, TMT-B, RAVLT-DR, and RAVLT-Ret (>60% of the classification models included these measures). The cognitive flexibility and delayed verbal learning measures were reliable at the 95% confidence interval. Balanced classification accuracy dropped 2%–7% when one of these neurocognitive variables was removed from the test battery (see figure 1 and table 3 in online supplementary material).

Discussion

To our knowledge, this is the first study to assess the feasibility of an individualized early recognition of different ARMS and a prediction of disease transition by analyzing neurocognitive data using state-of-the-art machine-learning techniques. These data were acquired from a neuroleptic-naïve at-risk population recruited using operationalized high-risk criteria that were previously employed to study neurobiological and neurocognitive correlates of the ARMS for psychosis.9,22,23,27–34.55 Furthermore, the transition rate of 42.9% in our follow-up cases indicates that the ARMS individuals’ risk for transition to psychosis lies within the range of previous UHR studies, eg, the initial findings from the PACE clinic in Australia,37,38 the earlyTreatment of Pre-Psychosis clinic in Norway,56 the Swiss FEPSY study,16 or the North American Prodrome Longitudinal Study.17,18,57

We could demonstrate that a neurocognition-based classification system may achieve high diagnostic accuracies in (1) distinguishing unseen ARMS from HC, (2) recognizing their level of vulnerability, as defined by the operationalized ARMS-E and ARMS-L criteria, and (3) predicting their subsequent outcome regarding transition or nontransition to psychosis over 4 years. Furthermore, the repeated nested CV framework provided an unbiased estimate of classification generalizability to new, unseen cases. Hence, the observation of overall high cross-validated classification accuracies in the 3 SVM experiments may support that the underlying neurocognitive patterns indeed represent distinct and homogeneous neurocognitive signatures of the ARMS and the emerging illness. Therefore, these signatures may be interpreted as neurocognitive markers that allow for a reliable early recognition of the ARMS and the prodromal phase of psychosis at the individual level.

Neurocognitive Markers of ARMS-E and ARMS-L

The (visual) working memory and verbal learning domains were particularly involved in the discriminative pattern underlying the high classification accuracy (96%) of ARMS-E vs HC individuals (table 5). This observation partly overlaps with reports of verbal memory alterations in at-risk populations characterized by predictable basic symptoms.3,5,11 However, these investigations did not find deficits in the (visual) working memory domain (SOPT, LNS) to be associated with an “early” ARMS for psychosis. Furthermore, Simon et al55 did not detect any significant neurocognitive differences in at-risk individuals recruited for basic symptoms compared with help-seeking patient controls. These inconsistencies may be attributed to differences in statistical power and employed methodology.58 Nevertheless, our binary classification results suggest that a pattern of altered verbal and mnemonic functions may reliably distinguish at-risk individuals experiencing predictive basic symptoms11,28 from healthy volunteers on a single-subject basis.

Compared with this pattern, the neurocognitive signature used by the HC vs ARMS-L classifier relied strongly on premorbid verbal IQ, processing speed, cognitive set shifting, and to a lesser degree verbal learning abilities. This pattern facilitated a cross-validated BAC of 88% in the classification of HC vs ARMS-L individuals. The pronounced involvement of executive functioning in this discriminative profile agrees with previous neuropsychological studies of UHR subjects.2,8,9,14,59,60 Furthermore, in keeping with our findings, premorbid verbal IQ,5,55 processing speed,9,60 and verbal learning impairments,5,9,55 have been shown to characterize the UHR for psychosis, albeit not as a consistent deficiency pattern throughout the literature. Again, this heterogeneity may be due to differences in sample characteristics, employed neuropsychological measures, and statistical strategies across these studies.58 In this regard, it is noteworthy that the neurocognitive profile of the ARMS-L obtained using univariate statistical procedures (table 4) differs from the discriminative patterns detected by our multivariate methodology. This discrepancy reflects the SVM’s ability to perform multivariate prediction, meaning that the algorithm analyzes the relationships of different cognitive features with respect to the conjoint discriminative power they provide. Thus, single predictive features may be redundant when combined into a discriminative profile. In contrast, features that appear to be irrelevant may aid in reliably separating the groups within a high-dimensional nonlinear discriminative space.

This phenomenon was particularly observed in the ARMS-E vs ARMS-L analysis, where univariate methods did not reveal any significant differences. However, the binary SVM classifier was able to separate unseen ARMS-E and ARMS-L with a BAC of 75% by relying on a discriminative pattern that involved premorbid verbal IQ, processing speed, visual working memory, and delayed verbal learning. This discriminative pattern is in keeping with existing studies that showed that these
neurocognitive domains are more impaired in ARMS-L vs ARMS-E individuals.\(^7,9,55\)

In summary, our binary and multigroup classification results (table 6) suggest that an early recognition of ARMS individuals may be feasible on a single-subject basis. If replicated in larger, independent populations, these findings could support the diagnostic application of neurocognitive pattern classification in the early stages of psychosis development. These at-risk states are likely to be more amenable to therapeutic intervention than the established disorder, but at the same time, they elude clinical detection due to the absence of frank psychotic symptoms.

In this regard, prospective studies are needed to trace neurocognitive deficits from the initial prodrome to the first-episode of the disease in order to evaluate whether the discriminative profiles of the ARMS-E and ARMS-L evolve on a single disease trajectory, or alternatively, whether they represent 2 distinct vulnerability states of psychosis, as suggested by the divergent transition rates in both subgroups (see methods section).

Neurocognitive Markers of Transition to Psychosis

Our second classification analysis demonstrated that neurocognition-based SVM classifiers may be capable of predicting whether a subclinical ARMS will evolve into frank psychosis at the single-subject level. The neurocognitive patterns extracted by the binary classifiers partly overlapped with the profiles identified in the HC vs ARMS-E vs ARMS-L analysis. More specifically, the classification of converters vs nonconverters achieved a BAC of 77.5% based on a discriminative pattern, which mainly involved premorbid verbal IQ, executive functions, and (verbal) learning abilities. Furthermore, the multigroup classifier (table 6) separated the converters/nonconverters from the other study groups with a BAC of 89%/90.8%, suggesting that larger training samples and additional neurocognitive information provided by the HC group facilitated the modeling of the complex separating boundary between the conversion and nonconversion samples.

Linear classification models combining clinical with neurocognitive data have previously shown to improve disease prediction beyond the levels of purely clinical recognition strategies. In the study of Lencz et al,\(^5\) the authors found that regression models using verbal memory performance and positive symptom severity achieved an accuracy of 80% in distinguishing subsequent converters from the nonconverters. Similarly, Riecher-Rössler et al\(^16\) found that logistic regression integrating suspiciousness and anhedonia scores with executive functioning correctly predicted transition in 80% of the cases. However, none of these studies estimated the generalizability of prediction using a rigorous CV procedure that separated the training from the validation data at all steps of the model creation process. Moreover, a recent multicenter study,\(^17\) which involved a total of 269 clinical high-risk subjects with follow-up information and 193 HC, did not find neurocognitive data to provide predictive power beyond clinical prediction models.\(^18\)

In the context of these conflicting findings, our results suggest that multivariate machine learning methods may be capable of extracting those discriminative patterns from cross-domain neuropsychological test batteries that indeed facilitate a reliable prediction of psychosis in unseen ARMS individuals. It is of note that our multigroup classification system outperformed (89%) the PACE-based UHR criteria in predicting psychosis as 14 of 21 followed ARMS-L subjects (66.7%), which were recruited using these criteria, subsequently developed the disease. However, we have to point out that our classification results were obtained in a clinically defined high-risk population and hence cannot be generalized to asymptomatic individuals at-risk for psychosis due to eg. genetic reasons. Furthermore, it is unknown how the SVM method would perform in ARMS samples with lower conversion rates, as reported recently.\(^61\)

In keeping with Lencz et al\(^5\) and Riecher-Rössler et al,\(^16\) our findings support that neurocognitive pattern recognition may substantially enhance the diagnostic reliability of existing clinical early recognition strategies.\(^18,28,38\) Furthermore, the higher PANSS negative scores observed in the conversion vs the nonconversion group (table 3) as well as the correlation between PANSS negative scores and RAVLT-DR (\(r=.56\)) as well as RAVLT-Ret (\(r=.56\)) suggests an association between prodromal negative symptoms and cognitive deficits. In this regard, multimodal diagnostic applications combining psychopathological and neurocognitive data may outperform classifiers relying only on neurocognitive information.\(^5,16\) However, these multimodal diagnostic systems would essentially depend on skilled clinicians trained to reliably detect subtle prodromal symptoms across individuals and over time, which may confine such systems to highly specialized clinical centers.

Our current and previous results\(^22\) suggest that neurocognitive and neuroanatomical pattern classification may achieve equivalent prediction results. This is important because the applicability and availability of MRI-based procedures in clinical real-world scenarios may be limited due to concomitant psychiatric, medical, and economic conditions. Furthermore, the similar classification accuracies obtained by neurocognition- and MRI-based pattern classification may point to a link between structural brain alterations\(^22,23,30,31,36,62–64\) and neurocognitive abnormalities in the ARMS. To date, these potential associations have been assessed only by 2 studies that found correlations between hippocampal volume and verbal learning deficits\(^34\) as well as between prefronto-callosal volumes and executive impairment.\(^42\) Therefore, further research is needed to evaluate the performance of
multimodal classifiers that integrate both neuroanatomical and neurocognitive data.

Finally, 2 further limitations have to be considered. Although the repeated double CV framework provided a reliable estimate of classification generalizability, we cannot rule out that the classification accuracies observed in our analyses may be due an accidental recruitment of an easy to categorize ARMS population. This possibility could be precluded in future multicenter studies that recruit and follow significantly larger samples of ARMS individuals. Furthermore, the encouraging results of our classification analyses do not imply that similar levels of specificity and sensitivity in predicting psychosis could be achieved in individuals presenting with ARMS-like symptoms in a normal clinical setting. This is because the classifiers were not trained to detect other psychiatric conditions like depression, bipolar disorder, or borderline personality disorder that may present with overlapping clinical and neurocognitive abnormalities. In this regard, the ARMS-NT group’s clinical outcome involved affective spectrum diagnoses, such as major depression, bipolar, and adjustment disorder, possibly pointing to a specificity of the neurocognitive patterns in separating schizophrenia spectrum psychosis from affective disorders. This important issue should be further examined in transnosological ARMS cohorts that cover ARMS for different psychiatric disorders.

**Funding**

Funding was provided by Ludwig-Maximilian-University to Nikolaos Koutsouleris for the recruitment and examination of the ARMS subjects. The funding source had no further role in the study design, the collection, analysis and interpretation of data, the writing of the report and in the decision to submit the article for publication.

**Supplementary Material**

Supplementary material is available at http://schizophreniabulletin.oxfordjournals.org.

**Acknowledgments**

We would like to thank Dr Reinhold Bader, Linux Cluster Systems for the Munich and Bavarian Universities, for his support in integrating the LIBSVM software into the batch system of the Linux cluster. Conflicts of Interest: None to declare

**References**


47. Gold JM, Carpenter C, Randolph C, Goldberg TE, Weinberger DR. Auditory working memory and Wisconsin Card Sorting Test performance in schizophrenia. *Arch Gen Psychiatry.* 1997;54:159–165.


