# Supplementary Online Content

Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, et al; the PRONIA Consortium. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. Published online September 26, 2018. *JAMA Psychiatry*. doi:10.1001/jamapsychiatry.2018.2165

**eTable 14.** Assessment of expert raters' prediction performance as a function of decreasing categorization thresholds applied to the CHR individuals and ROD patients follow-up GF:S scores.

**eFigure 1.** Observational study design of PRONIA.

**eFigure 2.** CONSORT Chart (A) and design of the main machine learning analyses performed in the multi-site discovery database of PRONIA (B).

**eFigure 3.** GF:S (left) and GF:R (right) score distributions of the CHR and ROD study group at the baseline and follow-up examinations.

**eFigure 4.** Histogram analysis of GF:S (blue) and GF:R (orange) changes over time in the CHR (left) and ROD (right) groups.

**eFigure 5.** Image quality assessment performed in the T1-weighted images of 412 study participants using the quality assessment functionality of the CAT12 toolbox.

**eFigure 6.** Correlation analyses were conducted to assess whether the prognostic decision scores generated by the clinical and sMRI classifiers were influenced by the follow-up intervals in the CHR and ROD groups.

**eFigure 7.** The predictive signature underlying the original GF:S outcome prediction model (A) was qualitatively compared to the signature produced by the site effects analysis (B).

**eFigure 8.** Results of voxel-based analysis of variance between 67 CHR persons with predicted impaired ($CHR_{Impaired}$) GF:S outcome vs. 49 CHR persons with predicted good ($CHR_{Good}$) GF:S outcomes (A) and 67 CHR persons with predicted impaired GF:S outcome vs. 116 healthy volunteers matched for site, age, and sex to the CHR group (B).

**eFigure 9.** Results of voxel-based analysis of variance between 120 healthy volunteers and 60 ROD persons with predicted good ($ROD_{Good}$) social functioning outcome (A), 120 healthy volunteers and 60 ROD patients with predicted impaired ($ROD_{Impaired}$) social functioning outcomes, (B) and the ROD patients with predicted good vs. impaired outcomes at follow-up (C).

**eFigure 10.** Prognostic generalization of the sMRI-based classifiers across 4 psychometric factors detected by Non-Negative Matrix Factorization (NNMF).

**eFigure 11.** Scatter plots in the upper panel show associations between observed GF:S scores ($GF:S_{obs}$) at follow-up and classification probabilities produced by combined clinical-sMRI based GF:S outcome predictor in the CHR (left) and ROD (right) group.

**eFigure 12.** Analysis of a sequential clinical-combined outcome predictor algorithm in the CHR group.

**eFigure 13.** Analysis of a clinical-combined outcome predictor algorithm in the ROD group.

This supplementary material has been provided by the authors to give readers additional information about their work.

**eMethods**

### 1.1. PRONIA recruitment infrastructure

The 236 study participants (116 CHR persons and 120 ROD patients) analyzed in the present study were recruited following a standardized recruitment and ascertainment protocol (see **eFigure 1** and **eTable 3**). The observational study protocol involved follow-up examinations every three months after the index ascertainment and was implemented by the 7 PRONIA sites described in **eTable 2**.

Upon study enrolment, the participants were pseudonymized twice, locally at each site and centrally at the level of the PRONIA portal. The PRONIA portal consists of a multi-user database hosting the clinical and neurocognitive information, and defaced MR images obtained from the study participants. The data are organized into digital questionnaires, visits, and cases. The portal provides the case managers with a controlled web-based interface to enter and upload the different data into the respective questionnaires. Furthermore, the PRONIA consortium has implemented a PRONIA@home mobile device interface that allows the study participants to securely log into the portal and fill out the self-rating questionnaires of given visit. Upon completion of the data entry across all questionnaires of given visit, the data are checked by an automatic quality control procedure which executes approximately 1600 data integrity and dependency rules. These rules include (1) basic checking of missing data and data ranges, (2) checking of dependency within one questionnaire, (3) dependencies between two questionnaires within one visit, and (4) dependencies between two consecutive visits (such as consistency of dates). Detected errors are fed back to the respective case managers allowing for a manual correction of the respective issues. This process is re-iterated until the quality of the clinical questionnaires in the given visit is sufficient for the entire visit to be locked.

The clinical data analyzed in the present study consist of the quality-checked and locked information of the study participants recruited until the 1st of May 2016, who had (1) received a structural MRI scan at baseline (see **eTable 4**), and (2) were assessed with the Global Functioning: Social and Role scales (GF:S, GF:R)[1] at least on one of the 3-, 6-, 9- and 12-month follow-up visits (see **eFigure 1**).

### 1.2. PRONIA study design and examination instruments

A comprehensive battery of assessment tools was used within a longitudinal observational study design to generate a multi-modal phenotypic profile of each study participant (**eFigure 1** and **eTable 3**). The clinical part of the battery compiled instruments that capture sociodemographic, somatic, environmental, diagnostic, psychopathological, functional and quality-of-life related variables in the PRONIA study population. This clinical battery was complemented by neuroimaging examinations as well as blood sampling for later genetic characterization, which were carried out at the baseline and 9-month follow-up time points.

### 1.3. Study inclusion and exclusion criteria

General inclusion criteria of the PRONIA study were age between 15 and 40 years, sufficient language skills for participation as well as capacity to provide informed consent/assent. General exclusion criteria were an IQ below 70, current or past head trauma with loss of consciousness (> 5 minutes), current or past known neurological or somatic disorders potentially affecting the structure or functioning of the brain, current or past alcohol dependence, or polysubstance dependence within the past six months, and any medical indication against MRI. In addition, HC exclusion criteria were: (1) any current or past DSM-IV axis disorder; (2) a positive familial history (1st degree relatives) for affective or non-affective psychoses; and (3) an intake of psychotropic medications or drugs more than 5 times/year

and in the month before study inclusion. ROD patients had to meet criteria for major depression fulfilled within the past 3 months, as established by the Structured Clinical Interview for DSM-IV-TR (SCID).[7] Specific ROD exclusion criteria were: (1) a previous episode of DSM-IV-TR major depression prior to the current or recent episode, and (2) a duration of the current episode exceeding 24 months.

The CHR state was defined by: (1) cognitive disturbances (COGDIS), as assessed by the Schizophrenia Proneness Instrument (SPI-A)[9]; and/or (2) ultra-high-risk (UHR) criteria for psychosis, as operationalized by the Structured Interview for Psychosis-Risk Syndromes (SIPS)[27]. The COGDIS criterion requires at least 2 of 9 cognitive basic symptoms of at least moderate severity (≥3) during the last 3 months, provided that the symptom has not always been present at the same severity level. The basic symptoms comprise: inability to divide attention; thought interference, pressure, and blockage; and disturbances of receptive and expressive speech, disturbance of abstract thinking, unstable ideas of reference, and captivation of attention by details of the visual field. Inclusion based on the adapted PRONIA UHR definition required at least one of the following three criteria: (I) 1 of 5 *attenuated psychotic symptoms* (unusual thought content/ delusional ideas, suspiciousness/persecutory ideas, grandiosity, perceptual abnormalities/hallucinations, and disorganized communication) with a moderate to severe, but not psychotic, severity (SIPS score 3-5) that (1) began within the past year or was rated one or more scale points higher compared to 12 month ago, and (2) occurred at an average frequency of at least once per week for at least several minutes per event in the past month; (II) *brief limited intermittent psychotic symptoms* (BLIPS) as defined by one of the symptoms listed above (1) reaching a psychotic level of intensity in each of the past 3 months for at least several minutes per day, OR (2) reaching a psychotic level of intensity in the past month, occurring at an average frequency of at least once per week for at least several minutes per event in the past month, or occurring at least for a cumulative period of more than one hour within the past month, AND (1+2) remitting spontaneously within one week (i.e. without antipsychotic medication); (III) a *genetic risk and functional deterioration* (GRFD) state was defined by a current 30% or greater reduction in the functional disability score of the split version of the Global Assessment of Functioning Scale (GAF-F)[28] compared with the highest lifetime level of functioning, and having a first-degree relative with a history of any psychotic disorder, or having a DSM-IV-TR schizotypal personality disorder.

Specific exclusion criteria for CHR and ROD patients were (1) an intake of antipsychotic medication for more than 30 cumulative days at or above the minimum dosage threshold defined by the DGPPN S3 Guidelines for the treatment of first-episode psychosis (**eTable 5** and [29]), and (2) any intake of antipsychotic drugs within the past 3 months before psychopathological baseline assessments at or above the minimum dosage threshold.

### 1.4. CONSORT Chart (Figure 2A)

Between 2/15/2014 and 5/1/2016, the PRONIA consortium consisting of seven academic early recognition services (**eTable 2**) and covering a European catchment population of 5,384,000 persons, screened 3416 individuals for study eligibility. Among those, 152 individuals met inclusion criteria for the CHR state and 156 patients met inclusion criteria for recent-onset episode of major depression, as described above. Ten candidate CHR individuals and 8 candidate ROD patients had to be excluded from this cohort because (1) CHR inclusion criteria could not be validated in the monthly PRONIA case conference, (2) subjects met exclusion criteria, or (3) did not finish the baseline examination, as described in **eTable 3**. Finally, 287 CHR and ROD individuals could be examined at baseline using structural MRI (**eTable 4**). Of those participants, 82.9% (238) could be naturalistically followed as described in **eFigure 2A** to ascertain changes of clinical symptoms, diagnostic criteria, and levels of functioning. From this

cohort, we had to exclude 2 further CHR individuals because of artifacts / poor quality of their baseline MRI scans (see below).

### 1.5. Inter-rater reliability analysis of global functioning social and roles scales

The PRONIA investigators received repeated training by one of the GF: Social and Role scale's authors, Andrea Auther (AA). Reliability testing was conducted independently of the PRONIA consortium by AA on 4 written transcripts of interviews from Zucker Hillside Hospital. We performed an Intra Class Correlation (ICC) analysis to measure the between-rater agreement on the target measures. For each test case raters had to generate 6 functioning scores: current, lowest and highest in the past year for the social and role functioning domain (GF:S/GF:R). Thirty-six PRONIA raters participated in the reliability testing. ICC analysis results are presented in **eTable 6**. Cicchetti (1994) gives the following guidelines for interpretation for kappa or ICC inter-rater agreement measures: Less than 0.40 = Poor; 0.40 and 0.59 = Fair; 0.60 and 0.74 = Good; 0.75 and 1.00 = Excellent.[30]

### 1.6. MRI harmonization and data acquisition

When setting up the PRONIA study, we decided to generate a MRI database that would represent the MR scanner sequence heterogeneity encountered in clinical real-world. The aim of this strategy was to strengthen the generalizability and clinical applicability of the predictive models developed by our machine learning analyses. Thus, we agreed on a minimal harmonization protocol that required the PRONIA sites to only (1) acquire isotropic or nearly isotropic voxel sizes of preferably 1 mm resolution, (2) set the Field Of View (FOV) parameters accordingly to guarantee the full 3D coverage of the brain including all parts of the cerebellum, and (3) define the relaxation time (TR) and echo time (TE) as well as other imaging parameters in a way that would maximize the contrast between cortical ribbon and the white matter and enhance the signal-to-noise ratio in the images. **eTable 4** lists the parameters defining the structural MR sequences used to examine in the PRONIA discovery sample participants.

### 1.7. MRI processing

At each PRONIA site, all images were visually inspected, automatically defaced, and anonymized using an in-house Freesurfer-based script prior to data centralization. Then, the open-source CAT12 toolbox (version r1155; http://dbm.neuro.uni-jena.de/cat12/), an extension of SPM12[31] designed for the processing and analysis of structural brain images, was used to segment images into gray matter (GM), white matter, and cerebrospinal fluid maps, and then to high-dimensionally register them to the stereotactic space of the Montreal Neurological Institute (MNI-152 space). The manual of the CAT12 toolbox (http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf) details the processing steps applied to the structural images, consisting of (1) the 1st denoising step based on Spatially Adaptive Non-Local Means (SANLM) filtering;[32] (2) an Adaptive Maximum A Posteriori (AMAP) segmentation technique, which models local variations of intensity distributions as slowly varying spatial functions and thus achieves a homogeneous segmentation across cortical and subcortical structures;[33] (3) the 2nd denoising step using Markov Random Field approach which incorporates spatial prior information of adjacent voxels into the segmentation estimation generated by AMAP;[33] (4) a Local Adaptive Segmentation (LAS) step, which adjusts the images for white matter (WM) inhomogeneities and varying gray matter (GM) intensities caused by differing iron content in e.g. cortical and subcortical structures. The LAS step is carried out before the final AMAP segmentation; (5) a partial volume segmentation algorithm that is capable of modeling tissues with intensities between GM and WM, as well as GM and cerebrospinal fluid (CSF) and is applied to the AMAP-generated tissue segments; (6) a high-dimensional

DARTEL registration of the image to a MNI-template generated from the MRI data of 555 healthy controls in the IXI database (http://www.braindevelopment.org). The registered GM images were multiplied with the Jacobian determinants obtained during registration to produce GM volume (GMV) maps. Images were smoothed with 10 mm before entering the subsequent analysis steps.

The Quality Assurance framework of CAT12 was used to empirically check the quality of the GMV maps. By computing the correlation of each image to all other 414 images (116 CHR patients, 120 ROD patients, 176 unique matched HC), we found 11 (2.66%) images whose correlation exceeded 2 standard deviations from the sample mean. These images were inspected and 2 were removed because of MRI artefacts. Notably, 99.5% of the images achieved a good overall weighted quality (B), and 83.0% the data quality was even rated with a B+ (**eFigure 5A**). Furthermore, we employed a two-sided $\chi^2$ test to assess whether the images' overall weighted quality scores were significantly associated with the classification error of the MRI-based GF:S outcome prediction models described below. This was not the case as shown in **eFigure 5B**.

The 412 GMV maps were smoothed with a 10 mm Gaussian kernel and entered the subsequent analyses. Finally, we employed generalization theory[34,35] to compute a between-site voxel reliability map (*G* coefficient map) by analyzing the GMV maps of the 6 travelling HC participants sent to each PRONIA scanner during the calibration study. This *G* coefficient map was used for reliability-based voxel masking in all our MRI-based machine learning analyses.

### 1.8. Machine learning analyses (Figure 2B)

We used our open-source software NeuroMiner (www.pronia.eu/neurominer/) to train three overall machine learning models to predict the GF:S or GF:R outcome targets in the CHR and ROD groups (**eFigure 2B**). The first model was trained on the 8 GF:S and GF:R variables acquired at baseline (highest lifetime GF:S/GF:R scores, highest/lowest GF:S/GF:R scores in the past year, lowest GF:S/GF:R scores in the week before study inclusion). The second model learned to predict the targets using the GMV data. The third was trained to optimally combine the predictions of the former two models (combined model).[36] We tested the geographic generalizability of these three models to new patient cohorts using nested leave-site-out cross-validation (LSO-CV, **eFigure 2B**): at the outer CV ($CV_2$) level, we iteratively held back every study site as validation sample, while the rest of the data entered the inner CV ($CV_1$) cycle, where cases were again iteratively assigned to training data and test site samples used to identify optimally predictive hyperparameter combinations. The MRI-based model optimization consisted of the following preprocessing steps which aimed at maximizing the models' cross-site generalizability: First, reliable voxels were extracted from the $CV_1$ training cases' GMV maps at the 15%, 25% or 50% percentiles of the inter-site *G* coefficient map ($T_G$). Second, the dimensionality of the training cases' thresholded GMV images was reduced by means of using Principal Component Analysis (PCA). To achieve a low generalization error[37], we trained PCA models with a restricted range of 11 to 19 Principal Components (*PC*) in the $CV_1$ training data (see also [38,39]). Then, the resulting PC scores were scaled to [-1, 1] and forwarded to a greedy sequential backward elimination (SBE) algorithm[40] that used the linear class-weighted Support Vector Machine (SVM)[41] to detect a set of PCs that optimally predicted the training and test cases' labels in a given $CV_1$ partition. The SBE starts with the full candidate feature pool and iteratively tests whether deletion of each variable in the pool improves, or at least does not change, a chosen model optimization criterion (in our case the Prognostic Summary Index, see below). Following the completed iteration, all features are ranked based on their recorded optimization criteria and the top-ranked feature (or block of top-ranked features) is discarded from the feature pool. Then, the algorithm reiterates through the remaining feature pool until the performance starts decreasing

or a stopping criterion is reached. In our case, we stopped the SBE when 20% of the PCs had been discarded from the feature pool based on our empirical experience that this early stopping criterion improves both performance and generalizability.

We used the mean Prognostic Summary Index (PSI) penalized by SVM model complexity[42] as criterion for the hyperparameter optimization: $\overline{\text{PSI}_{\text{reg}}} = \sum_{i=1}^{k=6}(n_{TP_i}/n_{P_i} + n_{TN_i}/n_{N_i} - 1 - n_{SV_i}/n_i) * 100/k$ was computed at given parameter combination across all *k* $CV_1$ partitions, where $n_{TP_i}/n_{P_i}$ was the Positive Predictive Value (PPV), $n_{TN_i}/n_{N_i}$ the Negative Predictive Value (NPV), $n_{SV_i}/n_i$ the fraction of the training population serving as support vectors in the $i^{\text{th}}$ $CV_1$ partition. The PSI was introduced by Linn and Grunau[43] to measure the total net gain in prognostic certainty provided by a dichotomous test given the known pre-test prevalence of the target condition. The PSI is computed as $\text{PSI} = \text{PPV} - \text{Prevalence}) + (\text{NPV} - (1 - \text{Prevalence})) = \text{PPV} + \text{NPV} - 1$. Based on this measure, optimization aimed at finding a combination of $T_G$, *PCs*, and the SVM's regularization parameter *C* [range: $2^{[-3\underset{\in\mathbb{Z}}{\rightarrow}+4]}$] that maximized $\overline{\text{PSI}_{\text{reg}}}$ within a 3 ($T_G$) × 5 (*PC*) × 8 (*C*) hyperparameter cube (**eFigure 2B**). Twenty percent of the most optimal ensembles within the cube were selected and then applied without any further modification to the $CV_2$ validation data of given held-back PRONIA site. This produced a mean decision score ($\overline{D_{ens}}$) and majority voting-based class membership probabilities ($P_{ens}$) for each $CV_2$ validation subject (see also[44]). For the clinical machine learning analysis, we only scaled the 8 predictors feature-wise to [-1, 1] and forwarded them to the SVM analysis, which was conducted as described above. Here, optimization involved only the SVM's *C* parameter.

After training the clinical and MRI-based models, we implemented a stacking-based data fusion framework[36,45] to assess whether the combination of these two unimodal classifiers would generate superior predictive systems for given GF:S/GF:R outcome targets. To rule out any information leakage between the training and test samples, we employed the identical nested leave-site-out cross-validation scheme for the unimodal and combined classification experiments. The stacking procedure started by combining the decision scores of the MRI-based and clinical SVM classifier committees within given $CV_1$ partition, standardizing the resulting matrices and subsequently using them as new sets of predictive features, which replaced the sMRI and clinical data in a given $CV_1$ partition. These data were then submitted to a greedy sequential forward search algorithm[40], which used L2-regularized logistic regression (L2LR)[46] to find a parsimonious combination of clinical and sMRI-based decision scores maximizing $\overline{\text{PSI}_{\text{reg}}}$ across the *C* parameter range. As for the unimodal predictors described above, we determined an ensemble of optimized L2LR models across the *C* range that conjointly maximized $\overline{\text{PSI}_{\text{reg}}}$ in given $CV_1$ training and test data. Then, the $CV_2$ validation predictions of the previously trained clinical and MRI-based SVM ensembles were combined and standardized. Each L2LR ensemble was then applied to this standardized $CV_2$ decision score matrix to generate probability estimates $P_{ens}$. Majority voting on $P_{ens}$was used to predict the $CV_2$ outcome targets and this procedure was repeated until all $CV_2$ cases had received a combined classifier prediction.

Furthermore, we explore whether the addition of MRI to sequential prognostic algorithms consisting of clinical and combined predictive tools could stabilize or even improve prediction performance in CHR and ROD samples selected for increasing clinical model ambiguity. To this end, we examined the performance of the CHR and ROD-specific combined GF:S outcome predictors as a function of increasingly ambiguous decision score ranges of the respective clinical models (**eFigures 12 and 13**).

Finally, we compared the sMRI-based, clinical and combined models to our raters' performance in correctly predicting social or role functioning. At the end of the clinical baseline characterization, raters were asked to reply to the question 'Do you think the patient will likely have a poor functional

outcome?'. This resulted in a binary prognostic variable consisting of poor vs. good outcome functional predictions. These prognostic categorizations were treated as estimates for the CHR/ROD individuals' social and role functioning outcomes and compared to the machine learning predictions using pairwise McNemar tests with Yates' correction and the obtained P values were adjusted for multiple comparisons using the False Discovery Rate (FDR, **eTable 9**).[47,48] A statistically significant difference between raters and machine learning-based pattern recognition was determined at α=0.05. We also assessed whether the raters' misclassification rate was related to their clinical experience as defined by the question 'How many months of experience in early recognition of psychosis do you have?'. We did not find any significant difference of the months of experience between correct or wrong prognostic assignments (CHR: t=-0.42, *P*=0.673; ROD: t=-1.17, *P*=0.248). Finally, as 'poor functional outcome' was not operationalized in the question to the raters, we performed a sensitivity analysis of the raters' estimates at lower GF:S cutoff definitions of 'poor functioning' to assess whether estimates followed a more severe internal heuristic of functional impairment (**eTable 14**).

Model performance was measured using sensitivity, specificity, balanced accuracy (BAC), PPV, NPV, PSI and Area-Under-the Curve (AUC) based on the class membership probability scores generated through ensemble-based majority voting in the outer leave-site-out (LSO) validation cycle. The performances of the sMRI-based, clinical and combined models were compared using two-sided McNemar tests with Yates' correction and resulting pairwise P values were corrected for multiple comparisons using FDR (**eTable 9**). Furthermore, we determined whether the models' performance measured as $PSI_{observed}$ at the $CV_2$ level differed significantly from a null distribution of respective ensemble predictors trained on *n*=1000 random label permutations.[49] Significance was defined at α=0.05 as $P = \sum_{i=1}^{n=1000}\left(PSI_{observed} < PSI_{permuted_i}\right)/n$. P values were adjusted for multiple comparisons per study group using FDR. Finally, we quantified the association between the $CV_2$ validation individuals' ensemble-based outcome probabilities and their GF:S/GF:R scores at follow-up. Two-sided significant Spearman correlations were identified at α=0.05.

We performed a series of analyses to validate the predictive value of the sMRI-based CHR and ROD classifiers: First, we explored whether the site effects found in the CHR individuals' outcome distributions could have biased our analysis results (**Table 1**, main manuscript). Therefore, we replaced the CHR subjects' GMV images with the data of age- and sex-matched healthy volunteers from the same catchment area and retrained the GF:S outcome classifier described above. An analogous technique was used in [50] to assess whether age and sex effects biased an MRI classifier that distinguished between patients with schizophrenia and major depression. Second, as the GF:S/R scores of the CHR/ROD samples with subsequently impaired outcomes differed from those with good outcomes (**Table 1**, main manuscript), we explored the possibility that the sMRI-based machine learning pipelines were informed by the GF variation at baseline rather than at follow-up. Therefore, we replaced the follow-up GF:S targets with their baseline counterparts and repeated the analysis described above. Third, to facilitate the interpretation of the multivariate signatures consisting of relative GMV reductions and increments observed in impaired vs. good social functioning outcome groups, we used VBM with threshold-free cluster enhancement[51] to compare the predicted outcome samples to site-, age- and sex-matched healthy control groups (**eFigures 8** and **9**).

### 1.9. Assessment of prognostic generalization

To test whether clinical, MRI-based, and combined models generalized transdiagnostically across the CHR and ROD groups, the $CV_2$ cycle was changed from a leave-site-out to a leave-group-out validation

design: First, we held back the CHR persons for validation, while using the ROD sample for model optimization. Then, the ROD patients served as validation sample, while the CHR cohort entered the leave-site-out $CV_1$ cycle. Transdiagnostic prediction performances were reported in **Table 3** of the main manuscript and tested for significance using the permutation-based approach described in 1.8.

Second, we explored whether clinical, MRI-based, and combined models' predictions of impaired vs. good social functioning were differentially associated with diagnostic outcomes in the CHR and ROD groups. More specifically, we tested whether the decision scores of these three models also predicted (1) the transition to psychosis in the CHR sample, (2) the presence of a DSM-IV-TR diagnosis of symptomatic major depression at the 9-month follow-up examination, and (3) presence of at least one symptomatic SCID-IV diagnosis across the mood, anxiety, and substance abuse domains of the DSM-IV at the 9-month follow-up examination. The GF:S classifiers' prognostic generalization to these labels was assessed using the permutation-based approach described above and were reported in **eTable 10**.

Finally, we used orthogonal Non-Negative Matrix Factorization (NNMF)[52] as implemented in the NNMF toolbox for MATLAB[53] to detect parsimonious multivariate clinical baseline patterns associated with predicted membership of the impaired vs. good functioning groups (**eFigure 10**). We selected core clinical and functional domains for the NNMF analysis, parametrized by the total scores of the Positive and Negative Syndrome Scale (PANSS),[4] and the Beck Depression Inventory II (BDI-II),[10] single items of the Structured Interview for Psychosis-Risk Syndromes (SIPS),[8] the nine items of COGDIS of the Schizophrenia Proneness Instrument, adult version (SPI-A)[9,54], GAF-S/-F, GF:S/:R, Domain scores of the Functional Remission in General Schizophrenia (FROGS), and the WHO Quality of Life Questionnaire – Brief Version (WHOQUOL-BREF). Two baseline NNMF models were derived from a training population of 99 CHR and 99 ROD participants and were applied to the respective data recorded at the 9-month follow-up examination. Using repeated-measures ANOVAs, the resulting NNMF score *trajectories* were compared between CHR/ROD persons predicted with impaired vs. good social functioning outcomes by the respective MRI-based predictors.

**eResults**

### 1.10. Analysis of social and role function distributions at baseline and follow-up

A qualitative pie chart analysis (**eFigure 3**) was conducted to visualize the changes of the social and role functioning scores observed between the baseline and follow-up examination in each study group. The paired Wilcoxon signed rank test was used to assess these changes for statistical significance at an α level of 0.05. Results indicate significant improvement of social and role functioning occurring between the baseline and follow-up in both study groups.

### 1.11. Comparing the variability of social and role functioning changes over time

Due to the differential predictability of the social and role functioning outcomes (**Table 2**, main manuscript) we compared the variability of the GF:S and GF:R changes between baseline and follow-up timepoints ($\Delta GF = GF_{follow\text{-}up} - GF_{baseline}$). **eFigure 4** shows the histograms of the ΔGF:S (blue) and ΔGF:R (orange) scores in the CHR (left) and ROD (right) groups, which indicates an increased variability of the GF:R changes over time, despite a similar overall improvement of the social and role functioning levels in both study groups. Furthermore, we performed a log-likelihood ratio test of equality of variances in SPSS (v23, IBM Inc.) to compare the temporal variance of GF:S to the temporal variance of GF:R. We observed a significant variance difference of the GF trajectories in both study groups (CHR: $\chi^2$=8.63, *P*=0.002; ROD: $\chi^2$=5.46, *P*=0.010) supporting a more pronounced temporal variability of role compared to social functioning.

### 1.12. Pairwise classifier comparisons

We used McNemar's tests (see e.g. [56]) to perform pairwise comparisons of the misclassification rates between the three types of machine learning models and the expert raters (**eTable 9**). Comparisons were conducted separately for each study group and outcome label. Statistical significance was defined at α=0.05 and P values were for corrected for multiple comparisons each in study group using the False Discovery Rate.

### 1.13. Assessment of potential MRI scanner and site effects on classification performance

We observed significant site-by-outcome class interactions in the CHR group and in both GF:S and the GF:R outcome categories (**Table 1**, main manuscript). As these interaction effects could have biased the training of the MRI-based outcome predictors, we conducted a validation analysis to further explore this possibility: we replaced the GMV maps of the CHR subjects with the maps of site-, age- and sex-matched HC (**eTable 7**) and repeated the machine learning experiment as in the original analysis. This approach directly tested the hypothesis that site- and scanner-related effects could have aided the SVM algorithm in detecting discriminative neuroanatomical information between the outcome classes despite our G-theory-based voxel selection strategy. As such, we would expect that the algorithm trained to predict the CHR subjects' GF:S outcomes using the HCs' baseline MR images would (1) provide a significant model as determined by permutation analysis, and (2) potentially achieve a prediction performance close to the model trained on the CHR persons' images. Based on the findings reported in **eFigure 7** and **eTable 7**, we rejected the hypothesis of site or population effects in our CHR outcome classifiers at α=0.05 (*P*=0.401). We also observed that the HC-based classifier did perform

close to randomness (**eTable 7**). Finally, we did not observe any significant overlap between the neuroanatomical signature of the CHR-based GF:S outcome prediction model and the model trained on the HC data (**eFigure 7**).

### 1.14. Predicting the baseline GF:S class of the CHR group using MRI data

In our group-level analysis we found significant differences in the baseline global, social and role functioning levels of CHR and ROD persons with subsequent impaired vs. good GF:S/GF:R outcomes. This observation raised the possibility that the MRI-based outcome classifiers learned to predict baseline variations of social and role functioning rather than the functional outcome categories measured at the follow-up examination. To investigate this possibility further, we replaced the follow-up GF:S outcome labels of the CHR and ROD groups with the respective classes derived from the baseline GF:S scores and repeated the respective machine learning analyses with the same parameter setup as described in the main manuscript. The MRI classifiers predicted baseline functioning classes with a BAC (sensitivity, specificity) of 58.1% (67.8%, 48.3%) in the CHR group and 46.5% (63.4%, 29.6%) in the ROD patients. Both models did not survive permutation testing and did not significantly explain the variance of the respective subjects' baseline GF:S scores (**eTable 8**). To facilitate the comparison between outcome and baseline prediction results, we added the performance measures of the sMRI-based predictors to **eTable 8**.

### 1.15. Comparing prognostic groups to matched healthy controls using voxel-based morphometry

To improve the interpretability of the relative volumetric differences shown in **Figure 2** of the main manuscript, we used voxel-based morphometry (VBM) to compare the 66/50 CHR individuals assigned to impaired ($CHR_{Impaired}$) /good ($CHR_{Good}$) GF:S outcomes by the sMRI classifier with the age- and sex-matched 116 healthy controls used for the site effects analysis. As in our machine learning analyses, the GMV maps of these 232 study participants were smoothed with a 10 mm full-width-at-half-maximum Gaussian kernel and proportionally scaled to global gray matter volume before entering an univariate analysis of variance in SPM12 (**http://www.fil.ion.ucl.ac.uk/spm/software/spm12/**). The statistical analysis was restricted to the brain regions that showed absolute Cross-Validation Ratios (CVR) ≥ 2 in the respective machine learning analysis (see **Figure 2A**; main manuscript). We used the Threshold Free Cluster Enhancement (TFCE) technique[51] as implemented in the TFCE toolbox for SPM12 (**http://dbm.neuro.uni-jena.de/tfce/**) to sensitize the statistical inference for spatial contiguous clusters of GMV differences between the study groups. Using the TFCE toolbox, we used non-parametric permutation testing with 1000 permutations to test for volume differences between (1) healthy volunteers and CHR persons with good GF:S prognosis, (2) healthy volunteers and CHR persons with poor GF:S prognosis, and (3) the two CHR prognostic groups. Statistical significance of the resulting TFCE maps was determined at $P<0.05$, FDR-corrected.

We did not find any volumetric differences between HC and CHR individuals with good outcome prognoses. In contrast, significant group-level GMV differences were detected between the HC group and the $CHR_{Impaired}$ persons; **eFigure 8B**): Compared to HC, $CHR_{Impaired}$ showed increased GMV in dorsolateral prefrontal, lateral parietal, and cerebellar brain regions, while decreased GMV was found with a spatial pattern involving temporo-occipital, intrasylvian, and inferior frontal cortical areas. The spatial distribution of these GMV increments and reductions highly overlapped with the pattern observed in the analysis of volumetric differences between CHR persons with predicted impaired vs. good GF:S outcomes (**eFigure 8A**). It was also similar to the pattern of reliable voxel elements in the multivariate decision boundary of the sMRI-based outcome classifier (**Figure 2**; main manuscript).

Furthermore, we conducted a similar VBM analysis in the prognostic outcome groups defined by the ROD classifier (**eFigure 9**). First, 120 HC drawn from the PRONIA discovery sample were matched for site, age, and sex to the ROD group. As described above, we used a voxel-level analysis of variance design to explore the quantitative relationship between the pattern of relative GMV increments and reductions used by the MRI-based classifier to predict one-year social functioning in the ROD group and the GMV variation of the healthy control group. To this end, the CVR map shown in **Figure 2B** was thresholded at an absolute value of 2 and thus defined the search volume for the VBM analysis. Then, T contrasts testing for GMV differences between the HC, $ROD_{Good}$, and $ROD_{Impaired}$ were defined in SPM12 and then tested using the permutation-based inference engine of the TFCE toolbox for SPM12 (see above). The results are shown in **eFigure 9A-C**. In summary, we observed patterns of volumetric differences, which were to some degree similar to the findings obtained for the CHR group: $ROD_{Impaired}$ patients showed extended volumetric increments in medial, lateral, and orbitofrontal cortices compared to healthy volunteers and $ROD_{Good}$ patients (**eFigure 9B** and **9C**). In contrast, $ROD_{Good}$ patients showed volumetric increments in parahippocampal, hippocampal and occipital brain regions compared to HC (**eFigure 9A**). Volumetric reductions in $ROD_{Good}$ compared to HC were detected in the prefrontal brain areas (**eFigure 9A**), where $ROD_{Impaired}$ patients had volume increments (**eFigure 9B**). A similar reversal of the directionality of the HC differences was not evident in the $ROD_{Impaired}$ patients concerning the pattern of limbic and occipital GMV increments observed in $ROD_{Good}$ (compare **eFigures 9A** and **9B**). In summary, these GMV reduction and increments highly overlapped with the reliable volumetric effects characterizing the multivariate neuroanatomical decision function of the ROD-specific social functioning predictor.

### 1.16. DSM-IV diagnoses in affective, anxiety and substance disorder categories

We re-assessed the CHR and ROD patients at the T1 examination using the SCID-IV interview and performed Fisher's exact tests to compare diagnostic frequencies in the impaired vs. the good functional outcome samples of the CHR and ROD groups (**eTable 11**). This group-level analysis showed that CHR patients with impaired social or role functioning at follow-up were significantly more likely to have at least one symptomatic DSM-IV-TR diagnosis at the T1 examination compared with good outcomes subjects. A similar trend associated with poor vs. good social functioning at follow-up was observed in the ROD group. These effects were driven by the association between major depression and poor functioning in both study groups.

### 1.17. Diagnostic breakdown of the transitions to psychoses observed during the follow-up period.

The SCID-IV interview was used to define the diagnoses of psychotic disorders in the 10 cases who met criteria for disease transition within the observational period. **eTable 12** lists the diagnoses of these cases.

### 1.18. Analysis of possible effects of age, sex, and ethnicity on classification performance.

We explored whether the sociodemographic variables age, sex, and ethnicity (Caucasian vs. non-caucasian) interacted with the performance of our predictive models. To this end, we compared misclassified to correctly classified subjects for each of these covariates using t tests and $\chi^2$ tests. P values were corrected in each study group for multiple comparisons using the False-Discovery-Rate. **eTable 13** summarizes the results of this analysis.

### 1.19. Prognostic generalization of the MRI-based classifiers across different psychometric domains

We analyzed how the sMRI-based outcome classifiers stratified individuals across 4 different psychometric factors derived from the clinical baseline data using Orthogonal Non-Negative Matrix Factorization (**eFigure 10**). In the CHR group, we interpreted these factors as (1) negative/affective symptom burden, (2) depression and previous functioning, (3) current functioning and quality of life, and (4) perceptual and cognitive disturbances. The factors were similar in the ROD group, with the difference that current functioning paired with depressive symptoms, while quality of life formed a separate factor. We then applied the same factor solutions to the follow-up data and assessed factor score changes between time-points in each group. This analysis corroborated the different degrees of prognostic generalizability of the CHR- vs. the ROD-specific classifiers (**eFigure 10**): CHR persons with predicted social functioning impairments showed significant follow-up differences in all four factor scores, and additional time x factor effects in the negative-affective and depression-previous functioning factors. In contrast, the trajectories of the 2 ROD samples differed only in the depression-functioning factor.

### 1.20. Approximating GF:S outcome scores using post hoc ordinal regression

As the outcome distributions of the ordinal GF:S and GF:R outcome scores were highly unbalanced and involved only few samples in the lower and upper tails (**eFigure 11**), regression models such as Support Vector Regression[55] could not be effectively used in the current study population. However, we observed that the combined classifiers' probability estimates were highly predictive of the study participants' GF:S scores (see $R^2$ values in **Table 2**, main manuscript). Therefore, we tested in this supplementary analysis whether the post hoc calibration of classification probabilities could yield a regression model that produces accurate estimates of the follow-up GF:S scores. To this end, an ordinal regression model was fitted in SPSS (version 23, IBM Inc.) with the dependent variable being the observed GF:S follow-up scores of the CHR or ROD patients and the independent variable being the respective outcome probabilities generated by the combined L2-regularized logistic regression algorithm. The post hoc regression models performed with a mean absolute error of 0.64 GF:S points in the CHR and 0.83 points in the ROD group in (**eFigure 11**). Calibrated estimates were most predictive in a range of 6 to 9 GF:S points in the CHR group and 7 to 9 points in the ROD group. Overestimation effects were present in the lower tails of the GF outcome distributions.

### 1.21. Analysis of GF:S outcome predictor algorithms involving clinical and combined models.

We tested the predictive performance of prognostic chains that used the functional GF:S outcome predictors as a first-line risk assessment tool and passed CHR/ROD persons to the downstream combined sMRI-clinical predictors at clinical decision score cutoffs increasingly close to the SVM decision boundary. To explore the optimal entry point of MRI in the prognostic chain, we analyzed the performance (PSI, PPV, and NPV) of the downstream combined predictors conditional on the ambiguity of the respective clinical model (see **eFigures 12** and **13**).

**References**


1. Cornblatt BA, Auther AM, Niendam T, et al. Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. *Schizophrenia Bulletin*. 2007;33(3):688-702.
2. Yung A, Phillips L, McGorry P, Ward J, Donovan K, Thompson K. Comprehensive assessment of at-risk mental states (CAARMS). *Melbourne, Australia, University of Melbourne, Department of Psychiatry, Personal Assessment and Crisis Evaluation Clinic*. 2002.
3. Llorca P-M, Lançon C, Lancrenon S, et al. The "Functional Remission of General Schizophrenia"(FROGS) scale: development and validation of a new questionnaire. *Schizophrenia Research*. 2009;113(2):218-225.
4. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*. 1987;13(2):261-276.
5. Alvarez E, Garcia-Ribera C, Torrens M, Udina C, Guillamat R, Casas M. Premorbid adjustment scale as a prognostic predictor for schizophrenia. *The British Journal of Psychiatry*. 1987.
6. Andreasen NC. The Scale for the Assessment of Negative Symptoms (SANS): conceptual and theoretical foundations. *Br J Psychiatry Suppl*. 1989;(7):49-58.
7. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P)*. New York: Biometrics Research, New York State Psychiatric Institute; 2002.
8. McGlashan TH, Miller TJ, Woods SW, Hoffman RE, Davidson L. Instrument for the assessment of prodromal symptoms and states. In: Miller T, Madnick SA, McGlashan TH, Libiger J, Johannessen JO, eds. *Early Intervention Psychiatric Disorders*. Dordrecht: Kluwer Academic; 2001:135-149.
9. Schultze-Lutter F, Addington J, Ruhrmann S, Klosterkötter J. *Schizophrenia Proneness Instrument, Adult Version (SPI-A)*. Giovanni Fioriti Editore. 2007.
10. Beck AT, Steer RA, Ball R, Ranieri W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*. 1996;67(3):588-597.
11. Endler NS, Parker JD, Ridder DT de, Heck GL van. *CISS Coping Inventory for stressful situations*. Harcourt; 2004.
12. Bernstein D, Fink L. CTQ: Childhood Trauma Questionnaire: a retrospective self-report. *San Antonio, TX: Psychological Corp*. 1998.
13. Veale JF. Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis. *Laterality: Asymmetries of Body, Brain and Cognition*. 2014;19(2):164-177.
14. Williams DR, Yu Y, Jackson JS, Anderson NB. Racial Differences in Physical and Mental Health. 1997;2:335-351.
15. Cole JD, Kazarian SS. The level of expressed emotion scale: A new measure of expressed emotion. *Journal of Clinical Psychology*. 1988;44:392-397.
16. Zimet GD, Powell SS, Farley GK, Werkman S, Berkoff KA. Psychometric characteristics of the Multidimensional Scale of Perceived Social Support. *Journal of Personality Assessment*. 1990;55(3-4):610-617.
17. Costa P, McCrae R. *Revised NEO Personality Inventory (NEOPIR). NEO five-factor inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Incorporated; 1992.
18. Friborg O, Hjemdal O, Rosenvinge J, Martinussen M. A new rating scale for adult resilience: what are the central protective resources behind healthy adjustment? *International Journal of Methods in Psychiatric Research*. 2003;12(2):65-76.
19. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. *The British Journal of Psychiatry : the journal of mental science*. 2000;176:379-386.
20. WHO. *WHOQOLBREF Introduction, Administration, Scoring and Generic version of the Assessment. Field Trial Version*. Geneva, Switzerland: WHO Division of Mental Health; 1996.

21. Cornblatt BA, Risch NJ, Faris G, Friedman D, Erlenmeyer-Kimling L. The Continuous Performance Test, identical pairs version (CPT-IP): I. New findings about sustained attention in normal families. *Psychiatry Research*. 1988;26(2):223-238.
22. Nowicki S, Duke MP. Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*. 1994;18(1):9-35.
23. Rey A. *L'examen psychologique Dans les Encéphalopathies traumatiques*.; 1943.
24. Roiser JP, Stephan KE, Ouden HE den, Friston KJ, Joyce EM. Adaptive and aberrant reward prediction signals in the human brain. *Neuroimage*. 2010;50(2):657-664.
25. Petrides M, Milner B. Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*. 1982;20:249-262.
26. Reitan RM. TMT, Trail Making Test A & B. 1992.
27. McGlashan T, Walsh B, Woods S. *The Psychosis Risk Syndrome: Handbook for Diagnosis and Follow-Up*. Oxford University Press; 2010.
28. Pedersen G, Hagtvet KA, Karterud S. Generalizability studies of the Global Assessment of Functioning–Split version. *Comprehensive Psychiatry*. 2007;48(1):88-94.
29. Psychiatrie P und P und N (DGPPN) Deutsche Gesellschaft für. *DGPPN S3 Treatment Guideline Schizophrenia and Psychotic Disorders*. AWMF; 2006.
30. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994;6(4):284.
31. Neuroimaging WTC for. Statistical Parametric Mapping 12. 2014.
32. Manjón JV, Tohka J, García-Martí G, et al. Robust MRI brain tissue parameter estimation by multistage outlier rejection. *Magnetic Resonance in Medicine*. 2008;59(4):866-873.
33. Rajapakse JC, Giedd JN, Rapoport JL. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Trans Med Imaging*. 1997;16(2):176-186.
34. Mushquash C, o Connor BP. SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*. 2006;38(3):542-547.
35. Brennan RL. Generalizability theory and classical test theory. *Applied Measurement in Education*. 2010;24(1):1-21.
36. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5:241-259.
37. Hansen LK, Larsen J, Nielsen FA, et al. Generalizable patterns in neuroimaging: how many principal components? *Neuroimage*. 1999;9(5):534-544.
38. Koutsouleris N, Meisenzahl E, Davatzikos C, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry*. 2009;66(7):700-712.
39. Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, et al. Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin*. 2015;41(2):471-482.
40. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517.
41. Vapnik VN. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*. 1999;10(5):988-999.
42. Boardman M, Trappenberg T. A Heuristic for Free Parameter Optimization with Support Vector Machines. In: *Proc. Int. Joint Conf. Neural Networks IJCNN'06*; 2006:610-617.
43. Linn S, Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiologic Perspectives & Innovations*. 2006;3(1):11.
44. Koutsouleris N, Kahn RS, Chekroud AM, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry*. 2016;3(10):935-946.
45. Polikar R. Ensemble based systems in decision making. *IEEE Circuits and Systems*. 2006;6(3):21-45.
46. Fan R, Chang K, Hsieh C, Wang X, Lin C. LIBLINEAR: A Library for Large Linear Classification. 2008.

47. Edwards AL. Note on the “correction for continuity” in testing the significance of the difference between correlated proportions. *Psychometrika*. 1948;13(3):185-187.
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995:289-300.
49. Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. *Inf Process Med Imaging*. 2003;18:330-341.
50. Koutsouleris N, Meisenzahl EM, Borgwardt S, et al. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain*. 2015;138(Pt 7):2059-2073.
51. Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*. 2009;44(1):83-98.
52. Ding C, Li T, Peng W, Park H. Orthogonal Nonnegative Matrix T-factorizations for Clustering. In: *Proceedings 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. Philadelphia, PA, USA: ACM; 2006:126-135.
53. Li Y, Ngom A. The non-negative matrix factorization toolbox for biological data mining. *Source Code for Biology and Medicine*. 2013;8(1):10.
54. Schultze-Lutter F, Koch E. Schizophrenia Proneness Instrument, Children and Youth Version (SPI-CY). Giovanni Fioriti Editore. 2010.
55. Smola A, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing*. 2004;14:199-222.
56. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*. 1998;10(7):1895-1923.

**eTable 1: Descriptive analysis of the sociodemographic and functioning** data of the HC samples matched for site, age and sex to the CHR and ROD groups. **Abbreviations:** *GAF Global Assessment of Functioning, GF:S* Global Functioning: Social scale*, GF:R* Global Functioning: Role scale.

| Variables | HC matched to CHR | HC matched to ROD |
|---|---|---|
| N | 116 | 120 |
| Age [ mean (SD) years ] | 25.0 (5.5) | 26.1 (6.1) |
| Sex [ M/F, % ] | 58/58 (50/50) | 54 / 66 (45/55) |
| Edinburgh Handedness Score [mean (SD)] | 75.5 (47.6) | 79.7 (41.7) |
| BMI [kg/m2] [mean (SD)] | 23.2 (3.0) | 23.3 (3.1) |
| Education [mean (SD) years] | 15.7 (2.9) | 16.2 (3.2) |
| Educational problems [mean (SD) years repeated] | 0.15 (0.40) | 0.2 (0.4) |
| Having a partnership most of the time in the year before study inclusion [N/Y] (%) | 79 (68.1) | 83 (69.2) |
| GAF Disability Highest Lifetime [ Mean (SD) ] | 87.1 (5.1) | 86.6 (4.8) |
| GAF Symptoms Highest Lifetime [ Mean (SD) ] | 88.2 (4.9) | 87.9 (5.5) |
| GAF Disability Past Year [ Mean (SD) ] | 86.2 (5.6) | 85.5 (5.4) |
| GAF Symptoms Past Year [ Mean (SD) ] | 87.0 (5.6) | 86.6 (5.9) |
| GAF Disability Past Month [ Mean (SD) ] | 85.7 (5.8) | 85.1 (5.5) |
| GAF Symptoms Past Month [ Mean (SD) ] | 86.7 (5.8) | 86.2 (6.2) |
| GF:S Highest Lifetime [ Mean (SD) ] | 8.8 (0.6) | 8.7 (0.7) |
| GF:S Lowest Past Year [ Mean (SD) ] | 8.2 (0.9) | 8.2 (0.9) |
| GF:S Highest Past Year [ Mean (SD) ] | 8.7 (0.7) | 8.6 (0.8) |
| GF:S Baseline [ Mean (SD) ] | 8.5 (0.8) | 8.5 (0.8) |
| GF:R Highest Lifetime [ Mean (SD) ] | 8.7 (0.7) | 8.7 (0.7) |
| GF:R Lowest Past Year [ Mean (SD) ] | 8.3 (0.8) | 8.2 (0.8) |
| GF:R Highest Past Year [ Mean (SD) ] | 8.7 (0.7) | 8.6 (0.7) |
| GF:R Baseline [ Mean (SD) ] | 8.6 (0.7) | 8.5 (0.7) |

**eTable 2: Characteristics of the recruiting institutions in the PRONIA consortium.**

| PRONIA Site | Institution Name | Country | Type of Service | Catchment Population | Screening population / year |
|---|---|---|---|---|---|
| Munich | Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich | DE | Academic outpatient services including specialized service for early recognition of psychosis; tertiary care academic hospital | 1,200,000 | 700 |
| Basel | Department of Psychiatry and Psychotherapy, University of Basel | CH | Academic inpatient and outpatient services including specialized service for early recognition and intervention of psychosis; tertiary care academic hospital | 500,000 | 200 |
| Milan Niguarda | Department of Pathophysiology and Transplantation, University of Milan. Four recruitment hospitals: Niguarda, Policlinico, San Paolo, Villa San Benedetto Menni in Albese con Cassano | IT | Psychiatric outpatient services including specialized services for early recognition of psychosis and persons at high risk; Academic hospital, providing psychiatric inpatient services, psychiatric outpatient services and local services; | 600,000 | 1,000 |
| Cologne | Department of Psychiatry and Psychotherapy, University of Cologne | DE | Academic outpatient services including specialized service for early recognition of psychosis; tertiary care academic hospital | 1,000,000 | 600 |
| Birmingham | The University of Birmingham | UK | Academic specialised Early Intervention Service for Psychosis covering Birmingham and Solihull. Community and Inpatient | 1,200,000 | 800 |
| Turku | Department of Psychiatry, University of Turku | FI | Psychiatric outpatient and hospital services responsible for treatment of psychiatric patients in their catchment areas in the South-Western Finland. | 284,000 | 2,300 |
| Udine | Department of Psychiatry, University of Udine | IT | Psychiatic outpatient services, academic hospital and local services. Tertiary care neuropsychiatric service | 600,000 | 500 |

**eTable 3: Clinical and neurocognitive examinations performed in the CHR, ROD, recent-onset psychosis (ROP), and HC groups during the 18-month follow-up period of the study**. **Clinical assessment types:** OR Clinician-based rating instrument, SR Self-rating-based instrument, NPT Neuropsychological test. **Examination timepoints:** T0 Baseline examination, IV3/IV6/IV12/IV15 3, 6, 12, 15-month examinations conducted only in the clinical study participants, T1 9-month examination, T2 18-month follow-up examination. **Clinician-based instruments:** CAARMS Comprehensive Assessment of the At-Risk Mental States[2], CHR Criteria Clinical High-Risk criteria summary checklist, FROGS Functional Remission in General Schizophrenia[3], GAF Global Assessment of Functioning, GF:S/R Global Functioning: Social / Role[1], PANSS Positive and Negative Symptom Scale[4], PAS Premorbid Adjustment Scale[5], SANS Scale for the Assessment of Negative Symptoms[6], SCID-IV Screening/Summary Structured Clinical Interview for DSM-IV[7], SIPS Structured Interview for Psychosis-Risk Syndromes (modified version 5.0)[8], SPI-A [COGDIS/COPER] Schizophrenia Proneness Instrument [Cognitive disturbances (COGDIS) / Cognitive-Perceptual (COPER) disturbances][9], Transition Criteria Interval questionnaire for the assessment of transition criteria, UHR - Schizotypy, Genetic Risk SIPS interview for the assessment of schizotypal personality traits, and familial risk for psychosis. **Self-rating instruments:** BDI-II Beck Depression Inventory II[10], CISS-24 Coping Inventory for Stressful Situations – 24 items[11], CTQ Childhood Trauma Questionnaire[12], EHI-SR Edinburgh Handedness Inventory – Short Version[13], EDS Everyday Discrimination Scale – Modified Version[14], LEE Level of Expressed Emotions[15], MSPSS the Multidimensional Scale for Perceived Social Support[16], NEO-FFI NEO Five Factor Inventory of Personality Traits[17], RSA Resilience Scale for Adults[18], SPIN Social Phobia Inventory[19], WHO-QOL-BREF WHO Quality of Life Questionnaire – Brief Version[20]. Neurocognitive tests: CPT-IP Continuous-Performance Test – Identical Pairs (adapted tablet version)[21], DANVA Diagnostic Analysis of Non-Verbal Accuracy 2 (adapted tablet version)[22], DS Auditory Digit Span (Forward/Backward) adapted from the PEBL battery, DSST Digit-Symbol-Substitution Test from the BACS battery, ROCF Rey-Osterrieth complex figure[23], SAT Salience Attribution Task (adapted version)[24], SOPT self-ordered pointing task (adapted version)[25], TMT-A/-B Trail-Making Test A and B[26], VF phonemic/semantic verbal fluency test.

| Instrument | Form | Screening | | T0 | | IV3 | IV6 | T1 | | IV12 | IV15 | T2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PAT | HC | PAT | HC | PAT | PAT | PAT | HC | PAT | PAT | PAT | HC |
| General Data | OR | X | X | | | | | X | X | | | X | X |
| Reasons for Referral | OR | X | | | | | | | | | | | |
| Treatment Documentation | OR | X | X | | | X | X | X | X | X | X | X | X |
| Somatic state and Health History | OR | X | X | | | | | X | X | | | X | X |
| SPI-A COGDIS/COPER | OR | X | X | | | X | X | X | X | X | X | X | X |
| SIPS positive symptoms | OR | X | X | | | X | | X | X | | | X | X |
| CAARMS | OR | X | X | | | X | | X | X | | | X | X |
| GAF | OR | X | X | | | X | | X | X | | | X | X |
| UHR – Schizotypy, Genetic Risk | OR | X | X | | | X | | X | X | | | X | X |
| CHR Criteria | OR | X | X | | | | | X | X | | | X | X |
| Transition Criteria | OR | | | | | X | X | | | X | X | | |
| SCID-IV Screening | OR | X | X | | | | | X | X | | | X | X |
| SCID-IV Summary | OR | X | X | | | | | X | X | | | X | X |
| Demographic and Biographic Data | OR | | | X | X | | | X | X | | | X | X |
| PAS | OR | | | X | X | | | X | | | | X | |
| SPI-A | OR | | | X | X | | | X | | | | X | |
| SIPS negative, disorganized and general symptoms | OR | | | X | X | | | X | | | | X | |
| PANSS | OR | | | X | | X | X | X | | X | X | X | |
| SANS | OR | | | X | | | | X | | | | X | |
| Chart of Life Events | OR | | | X | X | X | X | X | X | X | X | X | X |
| FROGS | OR | | | X | | | | X | | | | X | |
| GF: Social & Role | OR | | | X | X | X | X | X | X | X | X | X | X |
| Prognostic evaluation | OR | | | X | | | | X | | | | X | |
| MSPSS | SR | | | X | X | | | X | X | | | X | X |
| RSA | SR | | | X | X | | | X | X | | | X | X |
| CISS 24 | SR | | | X | X | | | X | X | | | X | X |
| SPIN | SR | | | X | X | | | X | X | | | X | X |
| BDI-II | SR | | | X | X | X | X | X | X | X | X | X | X |
| WHO-QOL-BREF | SR | | | X | X | | | X | X | | | X | X |
| EHI-SR | SR | | | X | X | | | | | | | | |
| LEE | SR | | | X | X | | | X | X | | | X | X |
| Wisconsin Scales | SR | | | X | X | | | | | | | | |
| EDS | SR | | | X | X | | | | | | | | |
| Bullying Scale T0 | SR | | | X | X | | | | | | | | |
| CTQ | SR | | | X | X | | | | | | | | |
| NEO-FFI | SR | | | X | X | | | | | | | | |
| DS backward (BACS) | NPT | | | X | X | | | X | X | | | | |
| DS forward (BACS) | NPT | | | X | X | | | X | X | | | | |
| CPT-IP (BACS) | NPT | | | X | X | | | X | X | | | | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DANVA | NPT | | | X | X | | | X | X | | | | |
| DSST | NPT | | | X | X | | | X | X | | | | |
| RAVLT | NPT | | | X | X | | | X | X | | | | |
| ROCF | NPT | | | X | X | | | X | X | | | | |
| SAT | NPT | | | X | X | | | X | X | | | | |
| SOPT | NPT | | | X | X | | | X | X | | | | |
| TMT-A | NPT | | | X | X | | | X | X | | | | |
| TMT-B | NPT | | | X | X | | | X | X | | | | |
| VF phonetic | NPT | | | X | X | | | X | X | | | | |
| VF semantic | NPT | | | X | X | | | X | X | | | | |
| WAIS-III | NPT | | | X | X | | | X | X | | | | |

**eTable 4: MR scanner systems and structural MRI sequence parameters** used at the respective PRONIA sites.

| PRONIA Site | Model | Field Strength | Coil Channels | Flip Angle | TR [ms] | TE [ms] | Voxel Size [mm] | FOV | Slice Number |
|---|---|---|---|---|---|---|---|---|---|
| Munich | Philips Ingenia | 3T | 32 | 8 | 9.5 | 5.5 | 0.97 x 0.97 x 1.0 | 250 x 250 | 190 |
| Milan Niguarda | Philips Achieva Intera | 1.5T | 8 | 12 | Short-est (8.1) | Short-est (3.7) | 0.93 x 0.93 x 1.0 | 240 x 240 | 170 |
| Basel | SIEMENS Verio | 3T | 12 | 8 | 2000 | 3.4 | 1.0 x 1.0 x 1.0 | 256 x 256 | 176 |
| Cologne | Philips Achieva | 3T | 8 | 8 | 9.5 | 5.5 | 0.97 x 0.97 x 1.0 | 250 x 250 | 190 |
| Birmingham | Philips Achieva | 3T | 32 | 8 | 8.4 | 3.8 | 1.0 x 1.0 x 1.0 | 288 x 288 | 175 |
| Turku | Philips Ingenuity | 3T | 32 | 7 | 8.1 | 3.7 | 1.0 x 1.0 x 1.0 | 256 x 256 | 176 |
| Udine | Philips Achieva | 3T | 8 | 12 | Short-est (8.1) | Short-est (3.7) | 0.93 x 0.93 x 1.0 | 240 x 240 | 170 |

**eTable 5: DGPPN S3 Guidelines for the treatment of first-episode psychosis and schizophrenia** (translated English version of Table 4.1 stated in the short version of the guideline manual available in https://www.dgppn.de/_Resources/Persistent/a6e04aa47e146de9e159fd2ca1e6987853a055d7/S3_Schizo_Kurzversion.pdf)**.** Candidate CHR and ROD patients were excluded if they had received antipsychotic medication (1) for more than 30 cumulative days at or above the minimum target dosage threshold for the treatment of first-episode psychosis, or (2) within the past 3 months before psychopathological baseline assessments at or above the minimum target dosage threshold for the treatment of first-episode psychosis. Abbreviations: DI dosage interval, [2]maximum recommended dosage according to prescribing information.

| Substance | Recommended starting dosage (mg/d) | DI' | Target dosage first-episode psychosis (mg/d) | Target dosage relapsing schizophrenia (mg/d) | Maximum dosage recommended (mg/d)[2] |
|---|---|---|---|---|---|
| **Atypical Antipsychotics** | | | | | |
| Amisulpride | 200 | (1)-2 | 100-300 | 400-800 | 1200 |
| Aripiprazole | (10)-15 | 1 | 15-(30) | 15-30 | 30 |
| Olanzapine | 5-10 | 1 | 5-15 | 5-20 | 20 |
| Quetiapine | 50 | 2 | 300-600 | 400-750 | 750 |
| Risperidone | 2 | 1-2 | 1-4 | 3-6-(10) | 16 |
| Ziprasidone | 40 | 2 | 40-80 | 80-160 | 160 |
| **Typical Antipsychotics** | | | | | |
| Fluphenazine | 0.4-10 | 2-3 | 2.4-10 | 10-20 | 20-(40) |
| Flupentixole | 2-10 | 1-3 | 2-10 | 10-60 | 60 |
| Haloperidole | 1-10 | (1)-2 | 1-4 | 3-15 | 100 |
| Perazine | 50-150 | 1-2 | 100-300 | 200-600 | 1000 |
| Perphenazine | 4-24 | 1-3 | 6-36 | 12-42 | 56 |
| Pimozide | 1-4 | 2 | 1-4 | 2-12 | 16 |
| Zotepine | 25-50 | 2-(4) | 50-150 | 75-150 | 450 |
| Zuclopenthixole | 2-50 | 1-3 | 2-10 | 25-50 | 75 |

**eTable 6: Intra-class correlation analysis** of the GF:S / GF:R scores generated by the PRONIA raters on the test cases.

| | **PRE-TRAINING ICCs**<br>For all Sites<br>(95% CI) | **POST-TRAINING ICCs**<br>For all sites<br>(95% CI) |
|---|---|---|
| Social & Role | **.836**<br>(.751 - .910) | **.871**<br>(.801 - .931) |

**eTable 7: Comparison of the performance measures** obtained from the leave-site-out validation of the original MRI-based GF:S outcome predictor and the predictor trained on the GMV data of age- and sex-matched healthy controls.

| | TP | TN | FP | FN | Sens | Spec | BAC | PPV | NPV | PSI | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **GF:S outcome predictor trained on the original CHR data [*P* = 0.001]** | | | | | | | | | | | |
| Full sample performance | 53 | 36 | 14 | 13 | 80.3 | 72.0 | 76.2 | 79.1 | 73.5 | 52.6 | 0.78 |
| Mean (SD) cross-site performance | 7.6 (6.0) | 5.1 (4.0) | 2.0 (1.5) | 1.9 (1.2) | 73.0 (23.0) | 67.4 (36.0) | 72.0 (10.3) | 80.3 (17.8) | 68.0 (15.0) | 52.0 (21.8) | 0.77 (0.13) |
| **GF:S outcome predictor trained on the data of age- and sex-matched HC** [*P* = 0.401] | | | | | | | | | | | |
| Full sample performance | 37 | 25 | 25 | 29 | 56.1 | 50.0 | 53.0 | 59.7 | 46.3 | 5.97 | 0.52 |
| Mean (SD) cross-site performance | 5.3 (3.7) | 3.6 (1.8) | 3.6 (2.8) | 4.1 (3.7) | 62.5 (22.3) | 59.3 (24.9) | 60.9 (15.0) | 64.0 (21.7) | 50.7 (28.8) | 14.7 (21.1) | 0.64 (0.21) |

**eTable 8: Comparison of the performance measures of the MRI-based GF:S outcome predictors and sMRI-based classifiers predicting the baseline GF:S classes of the same CHR or ROD individuals. Abbreviations:** TP number of true positives, TN number of true negatives, FP number of false positives, FN number of false negatives, Sens Sensitivity, Spec Specificity, BAC Balanced Accuracy, PPV Positive Predictive Value, NPV Negative Predictive Value, PSI Prognostic Summary Index, AUC Area-under-the Curve.

| Prediction performance | TP | TN | FP | FN | Sens | Spec | BAC | PPV | NPV | PSI | AUC | $P_{PERM}$ | $R^2_{GF}$ | $P_{GF}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CHR group** | | | | | | | | | | | | | | |
| GF:S outcomes | 53 | 36 | 14 | 13 | 80.3 | 72.0 | 76.2 | 79.1 | 73.5 | 52.6 | 0.78 | .001 | .224 | <.001 |
| GF:S baseline | 59 | 14 | 15 | 28 | 67.8 | 48.3 | 58.1 | 79.7 | 33.3 | 13.1 | 0.55 | .205 | .003 | 0.578 |
| **ROD group** | | | | | | | | | | | | | | |
| GF:S outcomes | 42 | 36 | 19 | 23 | 64.6 | 65.5 | 65.0 | 68.9 | 61.0 | 29.9 | 0.70 | .013 | .079 | .002 |
| GF:S baseline | 59 | 8 | 19 | 34 | 63.4 | 29.6 | 46.5 | 75.6 | 19.1 | -5.3 | 0.49 | .881 | .002 | .643 |

**eTable 9: Pairwise comparisons of classifier performance using McNemar's tests.**

| | Social Functioning | | | | Role Functioning | | | |
|---|---|---|---|---|---|---|---|---|
| | CHR | | ROD | | CHR | | ROD | |
| **Classifier comparisons** | $\chi^2$ | $P_{FDR}$ | $\chi^2$ | $P_{FDR}$ | $\chi^2$ | $P_{FDR}$ | $\chi^2$ | $P_{FDR}$ |
| MRI vs. clinical classifiers | 3.3 | .080 | 0.0 | .899 | **5.0** | **.038** | 0.6 | .600 |
| MRI vs. combined classifiers | 0.2 | .739 | 2.5 | .168 | **7.0** | **.019** | 0.2 | .752 |
| MRI vs. classifiers | **16.0** | **<.001** | **27.2** | **<.001** | **19.3** | **<.001** | **21.8** | **<.001** |
| Clinical vs. combined classifiers | 4.1 | .059 | 3.6 | .101 | 0.0 | 1.000 | 0.2 | .752 |
| Clinical classifiers vs. raters | **6.3** | **.021** | **33.2** | **<.001** | **6.3** | **.021** | **35.6** | **<.001** |
| Combined classifiers vs. raters | **17.4** | **<.001** | **45.2** | **<.001** | **7.3** | **.019** | **31.6** | **<.001** |

**eTable 10: Prognostic generalization performance of clinical, imaging-based, and combined models trained to predict social functioning outcomes in the CHR and ROD groups.** The significance of prognostic generalization was assessed by computing the PSI in 1000 permutations ($P_{PERM}$)of the respective outcome label (transition to psychosis, symptomatic depression at follow-up, ≥1 DSM-IV-TR diagnosis at follow) and comparing them to the observed PSI of the respective model. P values were adjusted for multiple comparisons in each group separately using the False-Discovery Rate. Performance measure abbreviations are defined in Table 2. **Abbreviations:** *TP* true positives, *TN* true negatives, *FP* False Positives, *FN* False Negatives, *Sens* Sensitivity, *Spec* Specificity, *BAC* Balanced Accuracy, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *AUC* Area-under-the Curve.

| Modality | Study group | TP | TN | FP | FN | Sens | Spec | BAC | PPV | NPV | PSI | AUC | $P_{PERM}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Models trained to predict impaired (≤7) vs. good (>7) GF:S outcomes at follow-up** | | | | | | | | | | | | | |
| **Generalization to the prediction of transition to psychosis** | | | | | | | | | | | | | |
| Clinical classifiers | CHR | 2 | 58 | 50 | 6 | 25.0 | 53.7 | 39.4 | 3.9 | 90.6 | -5.5 | 0.39 | 0.910 |
| sMRI classifiers | CHR | **8** | **49** | **59** | **0** | **100** | **45.4** | **72.7** | **11.9** | **100** | **11.9** | **0.73** | **0.014** |
| Combined classifiers | CHR | 5 | 49 | 59 | 3 | 62.5 | 45.4 | 53.9 | 7.8 | 94.2 | 2.0 | 0.54 | 0.238 |
| **Generalization to the prediction of symptomatic depression** | | | | | | | | | | | | | |
| Clinical classifiers | CHR | **13** | **52** | **36** | **0** | **100** | **59.1** | **79.5** | **26.5** | **100** | **26.5** | **0.79** | **<.001** |
| Clinical classifiers | ROD | 15 | 45 | 30 | 10 | 60.0 | 60.0 | 60.0 | 33.3 | 81.8 | 15.2 | 0.58 | 0.243 |
| sMRI classifiers | CHR | 9 | 38 | 50 | 4 | 69.2 | 43.2 | 56.2 | 15.3 | 90.5 | 5.73 | 0.63 | 0.195 |
| sMRI classifiers | ROD | 15 | 40 | 35 | 10 | 60.0 | 53.3 | 56.7 | 30.0 | 80.0 | 10.0 | 0.56 | 0.243 |
| Combined classifiers | CHR | **12** | **43** | **45** | **1** | **92.3** | **48.9** | **70.6** | **21.1** | **97.7** | **18.8** | **0.75** | **0.032** |
| Combined classifiers | ROD | 17 | 34 | 41 | 8 | 68.0 | 45.3 | 56.7 | 29.3 | 81.0 | 10.3 | 0.57 | 0.243 |
| **Generalization to the prediction of having at least 1 DSM-IV-TR diagnosis at the T1 examination (mood, anxiety, substance abuse disorders)** | | | | | | | | | | | | | |
| Clinical classifiers | CHR | **25** | **47** | **24** | **6** | **80.7** | **65.7** | **73.2** | **51.0** | **88.5** | **39.5** | **0.77** | **0.027** |
| Clinical classifiers | ROD | 20 | 38 | 25 | 17 | 54.1 | 60.3 | 57.2 | 44.4 | 69.1 | 13.5 | 0.53 | 0.281 |
| sMRI classifiers | CHR | 22 | 33 | 37 | 9 | 71.0 | 47.1 | 59.1 | 37.3 | 78.6 | 15.9 | 0.65 | 0.138 |
| sMRI classifiers | ROD | 19 | 32 | 31 | 18 | 51.4 | 50.8 | 51.1 | 38.0 | 64.0 | 2.0 | 0.50 | 0.425 |
| Combined classifiers | CHR | 25 | 38 | 32 | 6 | 80.7 | 54.3 | 67.5 | 43.9 | 86.4 | 30.2 | 0.72 | 0.054 |
| Combined classifiers | ROD | 24 | 29 | 34 | 13 | 64.9 | 46.0 | 55.4 | 41.4 | 69.1 | 10.4 | 0.53 | 0.243 |

**eTable 11: Prevalence comparisons of the DSM-IV-TR diagnoses in the CHR and ROD samples characterized by impaired vs. good social and role functioning at baseline (T0) and at the T1 follow-up examination.** The analyses were carried out for diagnoses in the mood, anxiety, and substance abuse domain. Presence of threshold diagnostic criteria in the past month before respective timepoint was examined using Fisher exact tests. For dysthymic disorder lifetime presence of threshold and subthreshold criteria were combined and compared against absence of lifetime criteria. P values were group- and timepoint-wise corrected for multiple comparisons using the False-Discovery Rate. Significance was defined at α=0.05. **Abbreviations:** *Tp* Time point, *Y/N* meeting diagnostic criteria/not meeting diagnostic criteria.

| DSM-IV-TR diagnostic categories | Tp | Global Functioning: Social at follow-up | | | | | | | | Global Functioning: Role at follow-up | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clinical High-Risk Group | | | | Recent-Onset Depression Group | | | | Clinical High-Risk Group | | | | Recent-Onset Depression Group | | | |
| | | GF:S ≤ 7 | GF:S > 7 | $\chi^2$ | $P_{FDR}$ | GF:S ≤ 7 | GF:S > 7 | $\chi^2$ | $P_{FDR}$ | GF:R ≤ 7 | GF:R > 7 | $\chi^2$ | $P_{FDR}$ | GF:R ≤ 7 | GF:R > 7 | $\chi^2$ | $P_{FDR}$ |
| ≥1 DSM-IV diagnosis: Y/N (%Y/%N) | T0 | 46/20 (69.7/30.3) | 29/21 (58.0/42.0) | 1.70 | .528 | 57/8 (87.7/12.4) | 48/7 (87.3/12.7) | 0.01 | 1.000 | 50/19 (72.5/27.5) | 25/22 (53.2/46.8) | 4.54 | .297 | 58/6 (90.6/9.4) | 47/9 (83.9/16.1) | 1.22 | 1.000 |
| | T1 | **27/35 (43.5/56.5)** | **4/35 (10.3/89.7)** | **12.5** | **.007** | *25/27 (48.1/51.9)* | *12/36 (25.0/75.0)* | *5.70* | *.090* | **27/34 (44.3/55.7)** | **4/36 (10.0/90.0)** | **13.3** | **<.001** | 24/27 (47.1/52.9) | 13/36 (26.5/73.5) | 4.52 | .120 |
| Bipolar I disorder | T0 | 1/65 (1.5/98.5) | 0/50 (0.0/100.0) | 0.76 | 1.00 | 0/65 (0.0/100.0) | 0/55 (0.0/100.0) | - | - | 1/68 (1.4/98.6) | 0/47 (0.0/100.0) | 0.69 | 1.00 | 0/64 (0.0/100.0) | 0/56 (0.0/100.0) | - | - |
| | T1 | 1/61 (1.6/98.4) | 0/39 (0.0/100) | 0.64 | 1.00 | 0/52 (0.0/100) | 0/48 (0.0/100) | - | - | 1/60 (1.6/98.4) | 0/40 (0/100) | 0.66 | 1.00 | 0/51 (0/100) | 0/49 (0/100) | - | - |
| Bipolar II disorder | T0 | 0/66 (0.0/100.0) | 2/48 (4.0/96.0) | 2.69 | .450 | 0/65 (0/100) | 0/55 (0/100) | - | - | 0/69 (0.0/100.0) | 2/45 (4.3/95.7) | 3.00 | .450 | 0/64 (0.0/100.0) | 0/56 (0.0/100.0) | - | - |
| | T1 | 0/62 (0.0/100) | 0/39 (0.0/100) | - | - | 0/52 (0.0/100) | 0/48 0.0/100) | - | - | 0/61 (0/100) | 0/40 (0/100) | - | - | 0/51 (0/100) | 0/49 (0/100) | - | - |
| Major depressive disorder | T0 | 27/39 (40.9/59.1) | 21/29 (42.0/58.0) | 0.01 | 1.00 | 55/10 (84.6/15.4) | 46/9 (83.6/16.4) | 0.02 | 1.00 | 34/35 (49.3/50.7) | 14/33 (29.8/70.2) | 4.40 | .297 | 57/7 (89.1/10.9) | 44/12 (78.6/21.4) | 2.47 | 1.00 |
| | T1 | *12/50 (19.4/80.6)* | *1/38 (2.6/97.4)* | *6.02* | *.083* | **20/32 (38.5/61.5)** | **5/43 (10.4/89.6)** | **10.5** | **.036** | **13/48 (21.3/78.7)** | **0/40 (0/100)** | **9.78** | **.007** | *18/33 (35.3/64.7)* | *7/42 (14.3/85.7)* | *5.88* | *.090* |
| Dysthymic disorder | T0 | 5/61 (7.6/92.4) | 3/47 (6.0/94.0) | 0.11 | - | 1/64 (1.5/98.5) | 0/55 (0.0/100.0) | 0.85 | 1.00 | 3/66 (4.3/95.7) | 0/47 (0.0/100.0) | 1.72 | .532 | 1/63 (1.6/98.4) | 0/56 (0.0/100.0) | 0.88 | 1.00 |
| | T1 | 4/58 (6.5/93.5) | 0/39 (0/100) | 2.62 | .345 | 3/49 (5.8/94.2) | 1/47 (2.1/97.9) | .883 | .929 | 3/58 (4.9/95.1) | 1/39 (2.5/97.5) | 0.37 | .761 | 4/47 (7.8/92.2) | 0/49 (0/100) | 4.00 | .480 |
| Panic disorder | T0 | 8/58 (12.1/87.9) | 3/47 (6.0/94.0) | 1.24 | .586 | 3/62 (4.6/95.4) | 2/53 (3.6/96.4) | 0.07 | 1.00 | 10/59 (14.5/85.5) | 1/46 (2.1/97.9) | 4.98 | .297 | 4/60 (6.3/93.8) | 1/55 (1.8/98.2) | 1.49 | 1.00 |
| | T1 | 3/59 (4.8/95.2) | 1/38 (2.6/97.4) | 0.33 | .761 | 2/50 (3.8/96.2) | 0/48 (0.0/100) | 1.88 | .812 | 4/57 (6.6/93.4) | 0/40 (0.0/100) | 2.73 | .345 | 1/50 (2.0/98.0) | 1/48 (2.0/98.0) | 0.00 | 1.00 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Agoraphobia (AWOPD) | T0 | 3/63 (4.5/95.5) | 2/48 (4.0/96.0) | 0.02 | - | 0/65 (0.0/100.0) | 1/54 (1.8/98.2) | 1.19 | 1.00 | 2/67 (2.9/97.1) | 3/44 (6.4/93.6) | 0.82 | .890 | 1/63 (1.6/98.4) | 0/56 (0.0/100.0) | 0.88 | 1.00 |
| | T1 | 4/58 (6.5/93.5) | 0/39 (0.0/100) | 2.62 | .345 | 0/52 (0.0/100) | 2/46 (4.2/95.8) | 2.21 | .480 | 3/58 (4.9/95.1) | 3/44 (2.5/97.5) | 0.37 | .761 | 0/51 (0.0/100) | 2/47 (4.1/100) | 2.12 | .480 |
| Social Phobia | T0 | 15/51 (22.7/77.3) | 4/46 (8.0/92.0) | 4.51 | .297 | 6/59 (9.2/90.8) | 4/51 (7.3/92.7) | 0.15 | 1.00 | 15/54 (21.7/78.3) | 4/43 (8.5/91.5) | 3.57 | .330 | 5/59 (7.8/92.2) | 5/51 (8.9/91.1) | 0.05 | 1.00 |
| | T1 | 6/56 (9.7/90.3) | 0/39 (0.0/100) | 4.01 | .345 | 3/49 (5.8/94.2) | 4/45 (6.3/93.8) | 0.01 | 1.00 | 4/57 (6.6/93.4) | 2/38 (5.0/95.0) | 0.11 | 1.00 | 2/49 (3.9/96.1) | 4/45 (8.2/91.8) | 0.80 | .778 |
| Specific Phobia | T0 | 2/64 (3.0/97.0) | 0/50 (0.0/100.0) | 1.54 | .754 | 0/65 (0.0/100.0) | 4/51 (7.3/92.7) | 4.89 | 1.00 | 2/67 (2.9/97.1) | 0/47 (0.0/100.0) | 1.39 | .754 | 1/63 (1.6/98.4) | 3/53 (5.4/94.6) | 1.34 | 1.00 |
| | T1 | 0/62 (0.0/100) | 1/38 (2.6/97.4) | 1.61 | .622 | 0/52 (0.0/100) | 5/43 (10.4/89.6) | 5.70 | .090 | 0/61 (0.0/100) | 1/39 (2.5/97.5) | 1.54 | .622 | 0/51 (0.0/100) | 5/44 (10.2/89.8) | 5.49 | .090 |
| Obsessive compulsive disorder | T0 | 7/59 (10.6/89.4) | 2/48 (4.0/96.0) | 1.74 | .543 | 2/63 (3.1/96.9) | 1/54 (1.8/98.2) | 0.19 | 1.00 | 6/63 (8.7/91.3) | 3/44 (6.4/93.6) | 2.09 | .954 | 2/62 (3.1/96.9) | 1/55 (1.8/98.2) | 0.22 | 1.00 |
| | T1 | 5/57 (8.1/91.9) | 0/39 (0.0/100) | 3.31 | .345 | 2/50 (3.8/96.2) | 2/46 (4.2/95.8) | 0.01 | 1.00 | 5/56 (8.2/91.8) | 0/40 (0.0/100) | 3.45 | .345 | 2/49 (3.9/96.1) | 1/55 (4.1/95.9) | 0.00 | 1.00 |
| Posttraumatic Stress disorder | T0 | 1/65 (1.5/98.5) | 0/50 (0.0/100.0) | 0.76 | 1.00 | 2/63 (3.1/96.9) | 1/54 (1.8/98.2) | 0.19 | 1.00 | 1/68 (1.4/98.6) | 0/47 (0/100) | 0.69 | 1.00 | 2/62 (3.1/96.9) | 1/55 (1.8/98.2) | 0.22 | 1.00 |
| | T1 | 0/62 (0.0/100) | 0/39 (0.0/100) | - | - | 0/52 (0.0/100) | 0/48 (0.0/100) | - | - | 0/61 (0.0/100) | 0/40 (0.0/100) | - | - | 0/51 (0.0/100) | 0/49 (0.0/100) | - | - |
| Alcohol Dependence | T0 | 0/66 (0.0/100.0) | 0/50 (0.0/100.0) | - | - | 1/64 (1.5/98.5) | 2/53 (3.6/96.4) | 0.54 | 1.00 | 0/69 (0.0/100.0) | 0/47 (0.0/100.0) | - | - | 0/64 (0.0/100.0) | 3/53 (5.4/94.6) | 3.52 | 1.00 |
| | T1 | 3/59 (4.8/95.2) | 0/39 (0.0/100) | 1.95 | .517 | 0/52 (0.0/100) | 0/48 (0.0/100) | - | - | 3/58 (4.9/95.1) | 0/40 (0.0/100) | 2.03 | .517 | 0/51 (0.0/100) | 0/49 (0.0/100) | - | - |
| Sedative-Hypnotic-Anxiolytic Dep. | T0 | 0/66 (0.0/100.0) | 0/50 (0.0/100.0) | - | - | 1/64 (1.5/98.5) | 0/55 (0.0/100.0) | 0.85 | 1.000 | 0/69 (0.0/100.0) | 0/47 (0.0/100.0) | - | - | 1/63 (1.6/98.4) | 0/56 (0.0/100.0) | 0.88 | 1.00 |
| | T1 | 0/62 (0.0/100) | 0/39 (0.0/100) | - | - | 0/52 (0.0/100) | 0/48 (0.0/100) | - | - | 0/61 (0.0/100) | 0/40 (0.0/100) | - | - | 0/51 (0.0/100) | 0/49 (0.0/100) | - | - |
| Cannabis Dependence | T0 | 1/65 (1.5/98.5) | 4/46 (8.0/92.0) | 2.90 | .450 | 2/63 (3.1/96.9) | 0/55 (0.0/100.0) | 1.72 | 1.00 | 1/68 (1.4/98.6) | 4/43 (8.5/91.5) | 3.38 | .450 | 1/63 (1.6/98.4) | 1/55 (1.8/98.2) | 0.01 | 1.00 |
| | T1 | 2/60 (3.2/96.8) | 0/39 (0.0/100) | 1.28 | .716 | 1/51 (1.9/98.1) | 0/48 (0.0/100) | 0.93 | 1.00 | 2/59 (3.3/96.7) | 0/40 (0.0/100) | 1.34 | .716 | 1/50 (2.0/98.0) | 0/49 (0.0/100) | 0.97 | 1.00 |

**eTable 12: List of the SCID-IV diagnoses in CHR and ROD patients** who developed a psychotic disorder during the follow-up period. Furthermore, the groups' social /role functioning scores at baseline and follow-up were compared using t tests. Significant P values and associated group differences were highlighted with bold face numbers. *excluding brief limited intermittent psychotic symptoms.

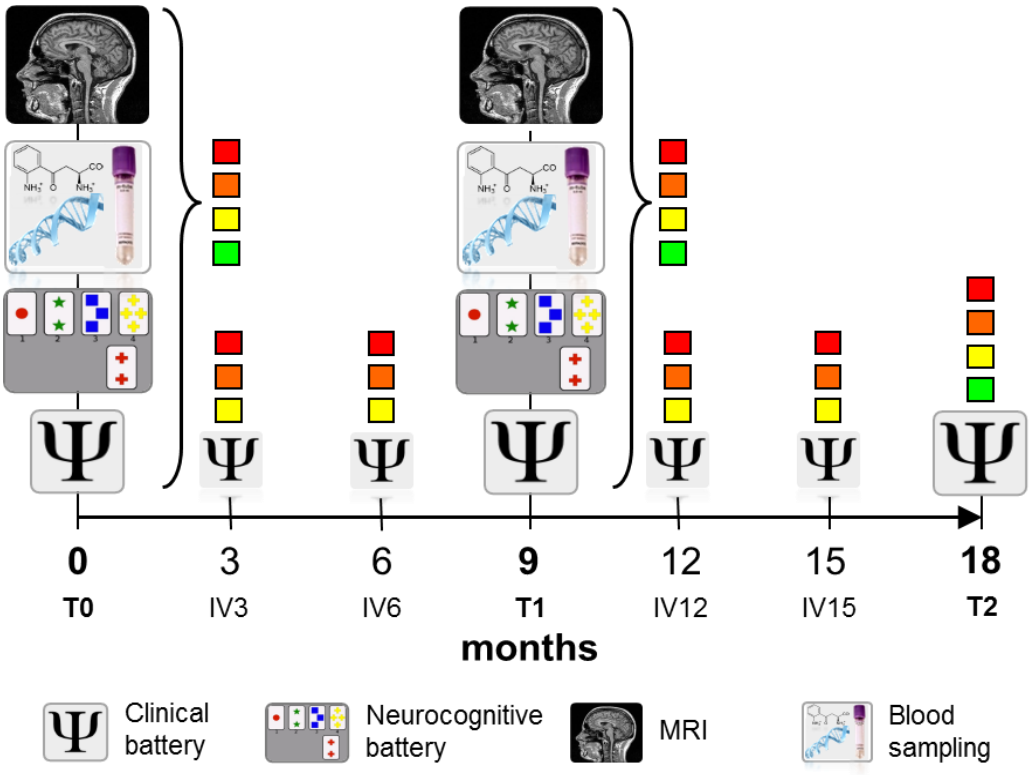
| | **Transition group** | **Non-transition group** | ***T/χ²*** | ***P*** |
|---|---|---|---|---|
| N | 10 | 226 | | |
| CHR / ROD | 8/2 (80.0/20.0) | 108/118 (47.8/52.2) | 3.98 | .056 |
| Schizophrenia | 6 | - | | |
| Schizophreniform Psychosis | - | - | | |
| Schizoaffective Psychosis | 1 | - | | |
| Delusional disorder | - | - | | |
| Brief psychotic disorder | - | - | | |
| Psychosis NOS* | 3 | - | | |
| Major depression with mood incongruent psychotic features | 1 | - | | |
| GF:S score at baseline | 6.70 (1.64) | 6.41 (1.38) | -0.64 | .523 |
| GF:S score at follow-up | **6.00 (1.25)** | **7.09 (1.27)** | **2.67** | **.008** |
| GF:R score at baseline | 6.20 (1.48) | 6.22 (1.54) | 0.04 | .966 |
| GF:R score at follow-up | **5.40 (1.78)** | **7.00 (1.59)** | **3.09** | **.002** |

**eTable 13: Interactions between classification performance and age, sex, and ethnicity.**

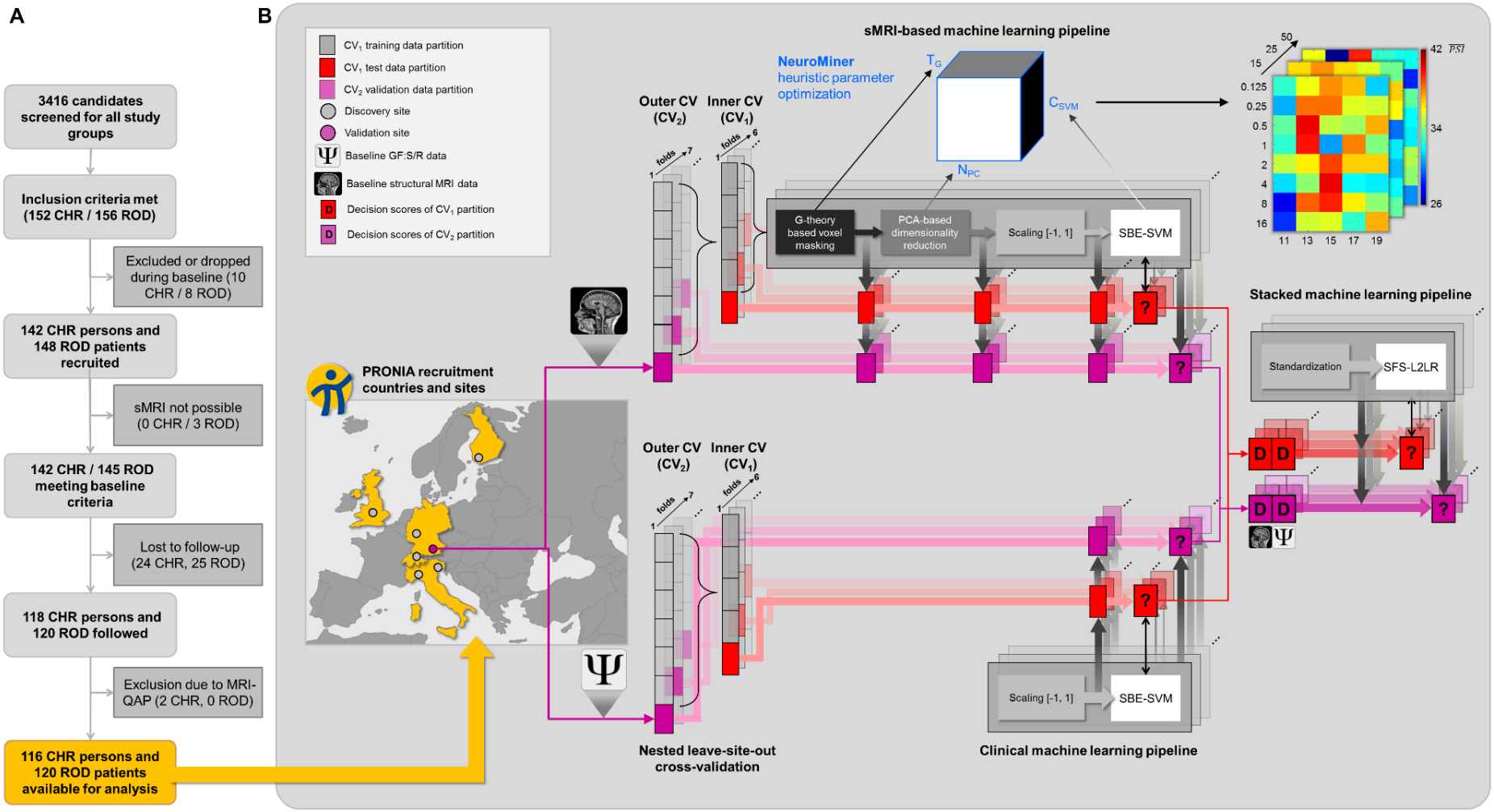
| Covariate effects on mis-classification (correctly vs. wrongly classified patients) | Social functioning | | | | Role functioning | | | |
|---|---|---|---|---|---|---|---|---|
| | CHR | | ROD | | CHR | | ROD | |
| | $T / \chi^2$ | $P_{FDR}$ | $T / \chi^2$ | $P_{FDR}$ | $T / \chi^2$ | $P_{FDR}$ | $T / \chi^2$ | $P_{FDR}$ |
| **Clinical classifiers** | | | | | | | | |
| Age | -1.05 | .969 | -0.58 | .890 | -0.47 | 1.00 | -0.78 | .890 |
| Sex | 0.00 | 1.00 | 0.09 | .898 | 1.41 | .969 | 0.26 | .890 |
| Ethnicity | 0.24 | 1.00 | 0.17 | .890 | 2.62 | .969 | 0.25 | .890 |
| **sMRI classifiers** | | | | | | | | |
| Age | 1.67 | .882 | -1.35 | .810 | 0.82 | 1.00 | -0.40 | .969 |
| Sex | 0.05 | 1.00 | 0.45 | .890 | 0.04 | 1.00 | 6.24 | .306 |
| Ethnicity | 0.18 | 1.00 | 3.00 | .810 | 0.17 | 1.00 | 0.88 | .890 |
| **Combined classifiers** | | | | | | | | |
| Age | -0.03 | 1.00 | -0.36 | .890 | -1.01 | .969 | 1.08 | .890 |
| Sex | 0.97 | 1.00 | 0.15 | .898 | 0.00 | 1.00 | 3.43 | .810 |
| Ethnicity | 0.57 | 1.00 | 1.94 | .843 | 7.02 | .306 | 0.63 | .890 |

**eTable 14: Assessment of expert raters' prediction performance** as a function of decreasing categorization thresholds applied to the CHR individuals and ROD patients follow-up GF:S scores. **Abbreviations:** TP number of true positives, TN number of true negatives, FP number of false positives, FN number of false negatives, Sens Sensitivity, Spec Specificity, BAC Balanced Accuracy, PPV Positive Predictive Value, NPV Negative Predictive Value, PSI Prognostic Summary Index, AUC Area-under-the Curve.

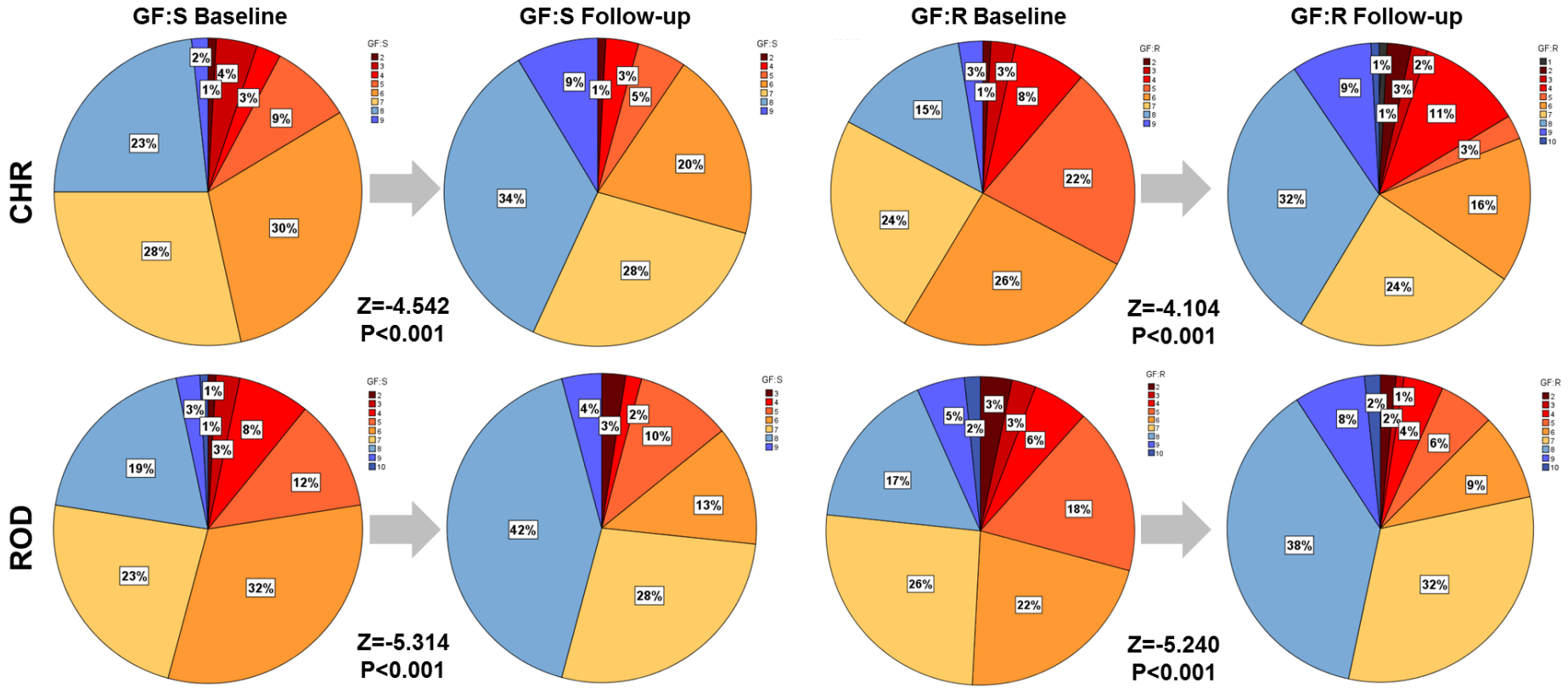
| Prediction performance | TP | TN | FP | FN | Sens | Spec | BAC | PPV | NPV | PSI | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **CHR: Expert prognosis of GF:S outcomes** | | | | | | | | | | | |
| GF:S≤7 vs GF:S>7 | 34 | 46 | 4 | 32 | 51.5 | 92.0 | 71.8 | 89.5 | 59.0 | 48.4 | 0.72 |
| GF:S≤6 vs GF:S>6 | 18 | 62 | 20 | 16 | 75.6 | 52.9 | 64.3 | 47.4 | 79.5 | 26.9 | 0.64 |
| GF:S≤5 vs GF:S>5 | 6 | 73 | 32 | 5 | 54.5 | 69.5 | 62.0 | 15.8 | 93.6 | 9.4 | 0.62 |
| **ROD: Expert prognosis of GF:S outcomes** | | | | | | | | | | | |
| GF:S≤7 vs GF:S>7 | 17 | 50 | 4 | 47 | 26.6 | 92.6 | 59.6 | 81.0 | 51.5 | 32.5 | 0.60 |
| GF:S≤6 vs GF:S>6 | 14 | 79 | 7 | 18 | 43.8 | 91.9 | 67.8 | 66.7 | 81.4 | 48.1 | 0.68 |
| GF:S≤5 vs GF:S>5 | 9 | 89 | 12 | 8 | 52.9 | 88.1 | 70.5 | 42.9 | 91.8 | 34.6 | 0.71 |



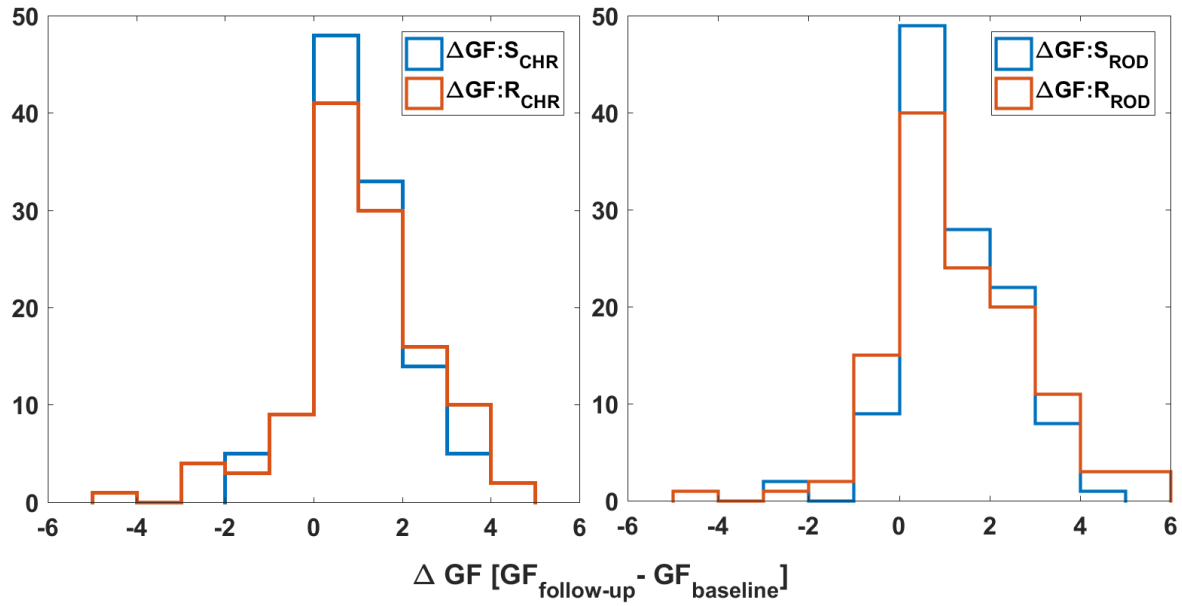

**eFigure 1: Observational study design of PRONIA.** Colored boxed indicate type of assessment / visits conducted in each of the study groups: Healthy controls (green), patients with recent-onset depression (yellow), persons with a clinical high-risk for psychosis (orange), patients with recent-onset psychosis (red).



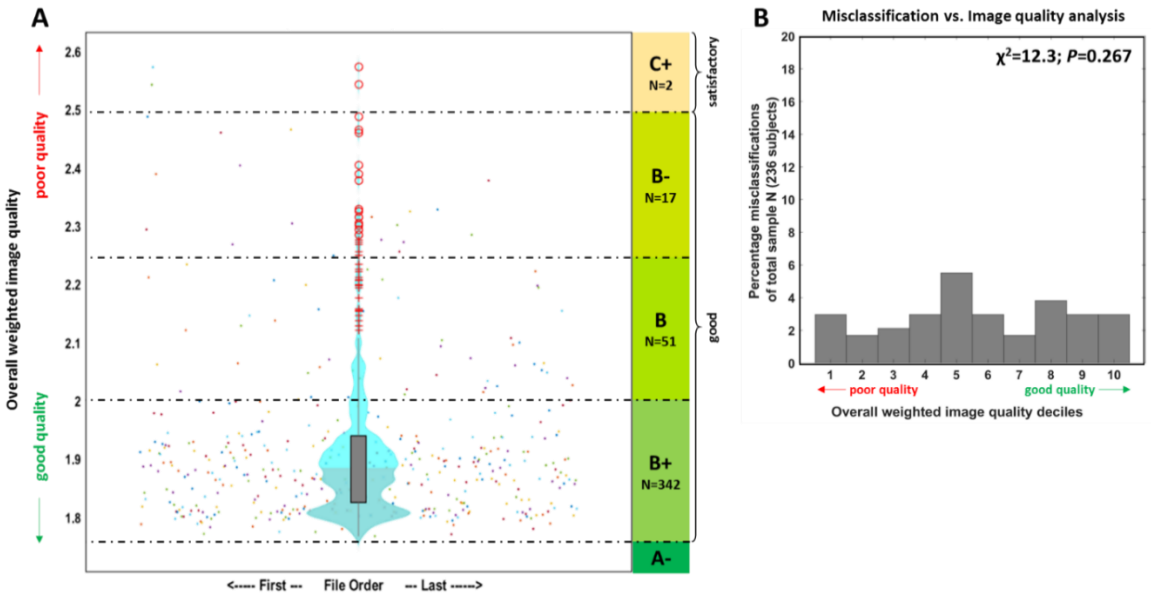

**eFigure 2: CONSORT Chart (A) and design of the main machine learning analyses performed in the multi-site discovery database of PRONIA (B**). The data analyzed in the current work was collected from 7 different sites in 5 European countries covering diverse mental healthcare systems and pathways to early recognition (see **eTable 1**). A nested leave-site-out (LSO) cross-validation scheme was used to optimize the sMRI-based, clinical and combined machine learning systems for the generalizable prediction of GF:S/R-derived outcome classes: The test data entering the inner LSO cross-validation cycle ($CV_1$) was used to search for parameter combinations that maximized the out-of-site generalizability of the sequential backward elimination support vector machines (SBE-SVM). The search is illustrated by the $\overline{PSI}$ hyperparameter cube computed for the training of the sMRI-based GF:S outcome predictor in the CHR group (top right). Optimized processing steps (grey arrows) were then applied without any modification to the held-back validation data at the outer LSO cycle ($CV_2$) to estimate the population-level generalizability of given predictive model. Stacking was performed by concatenating the decision scores (D) of the MRI-based and clinical prediction model at the $CV_1$ level and using them as feature pool for sequential forward selection employing L2-regularized logistic regression (SFS-L2LR).



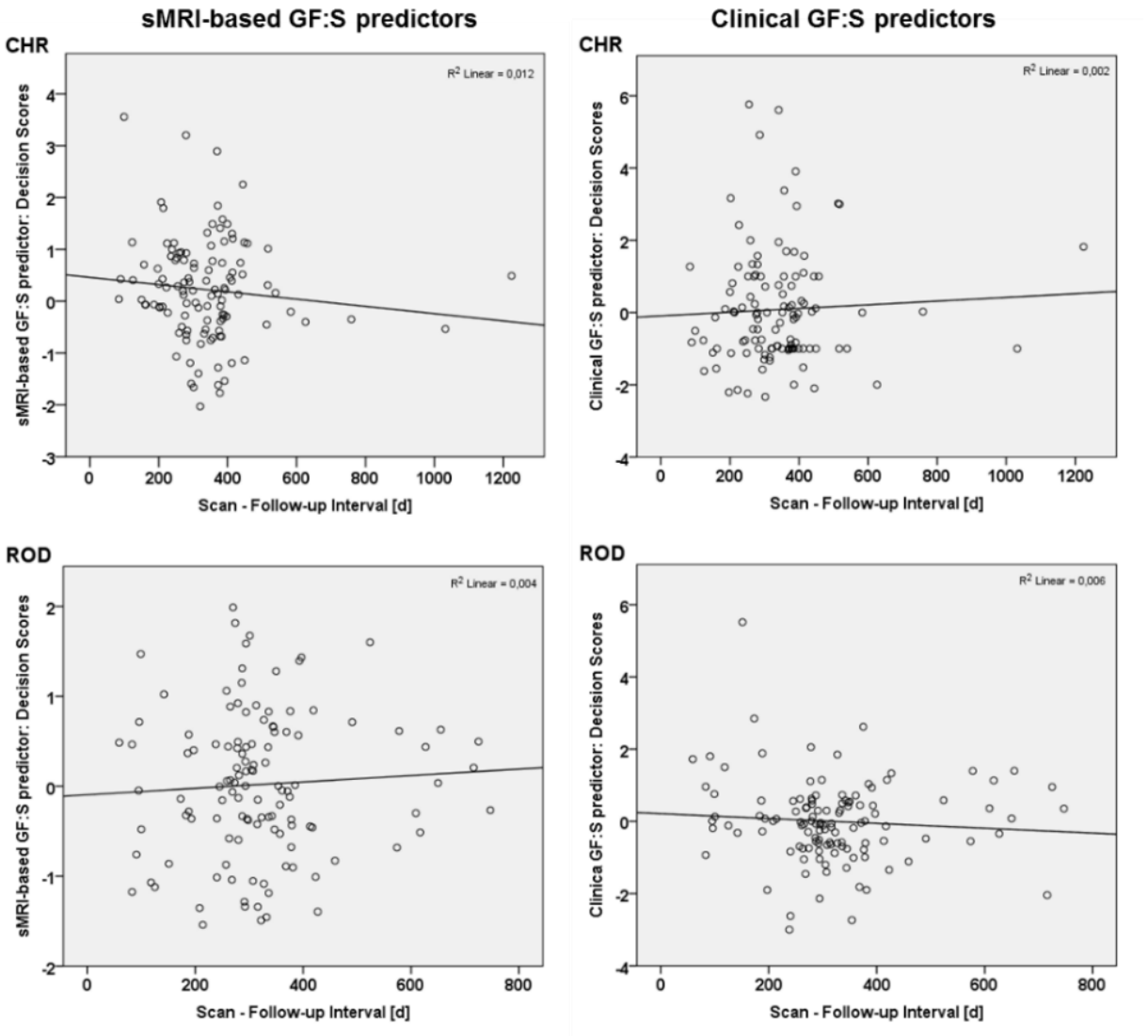

**eFigure 3: GF:S *(left)* and GF:R *(right)* score distributions of the CHR and ROD study group at the baseline and follow-up examinations.** Paired Wilcoxon signed rank test were used to assess the distribution changes for statistical significance at α=0.05.



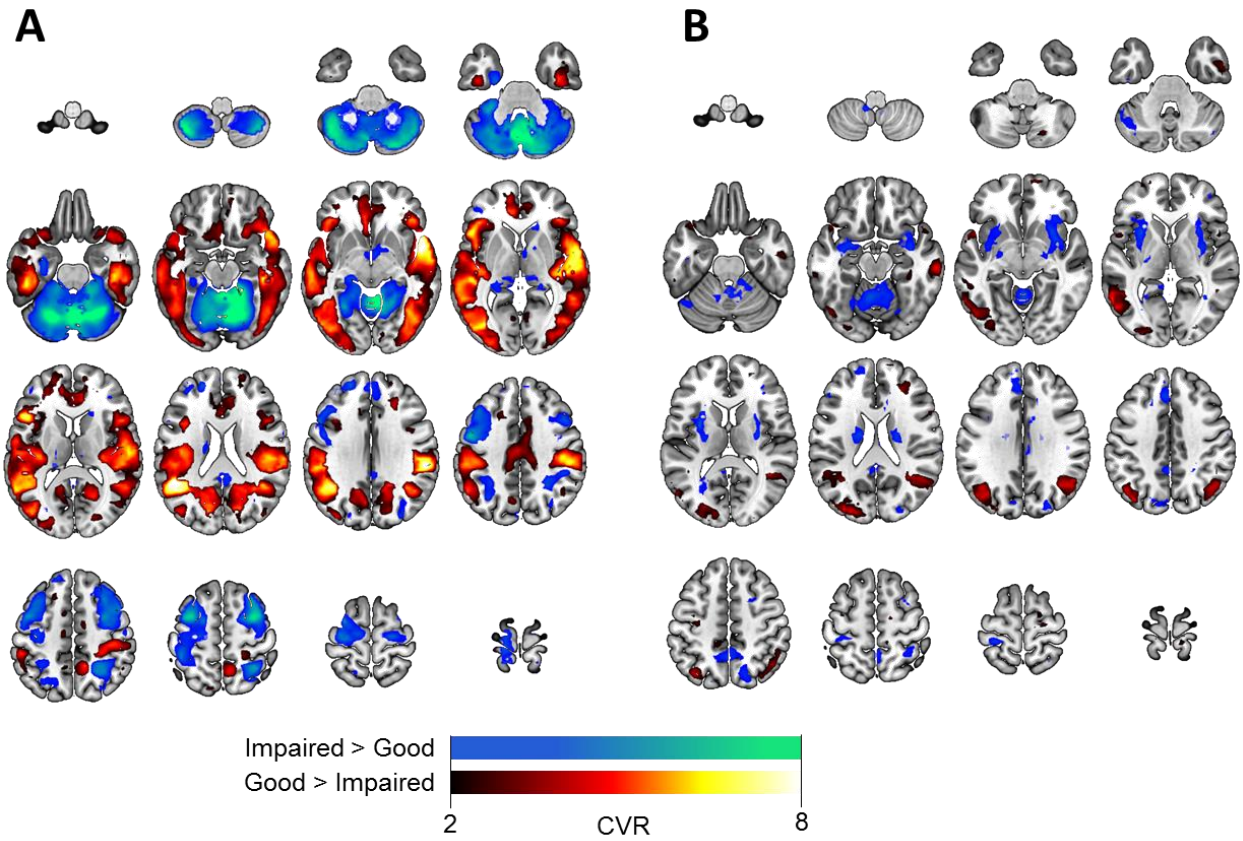

**eFigure 4: Histogram analysis of GF:S (blue) and GF:R (orange) changes over time in the CHR (left) and ROD (right) groups.** A likelihood ratio analysis of the differences of changes over time between the Global Functioning Social and Role scores of the CHR and ROD patients is described in the eResults 2.12.



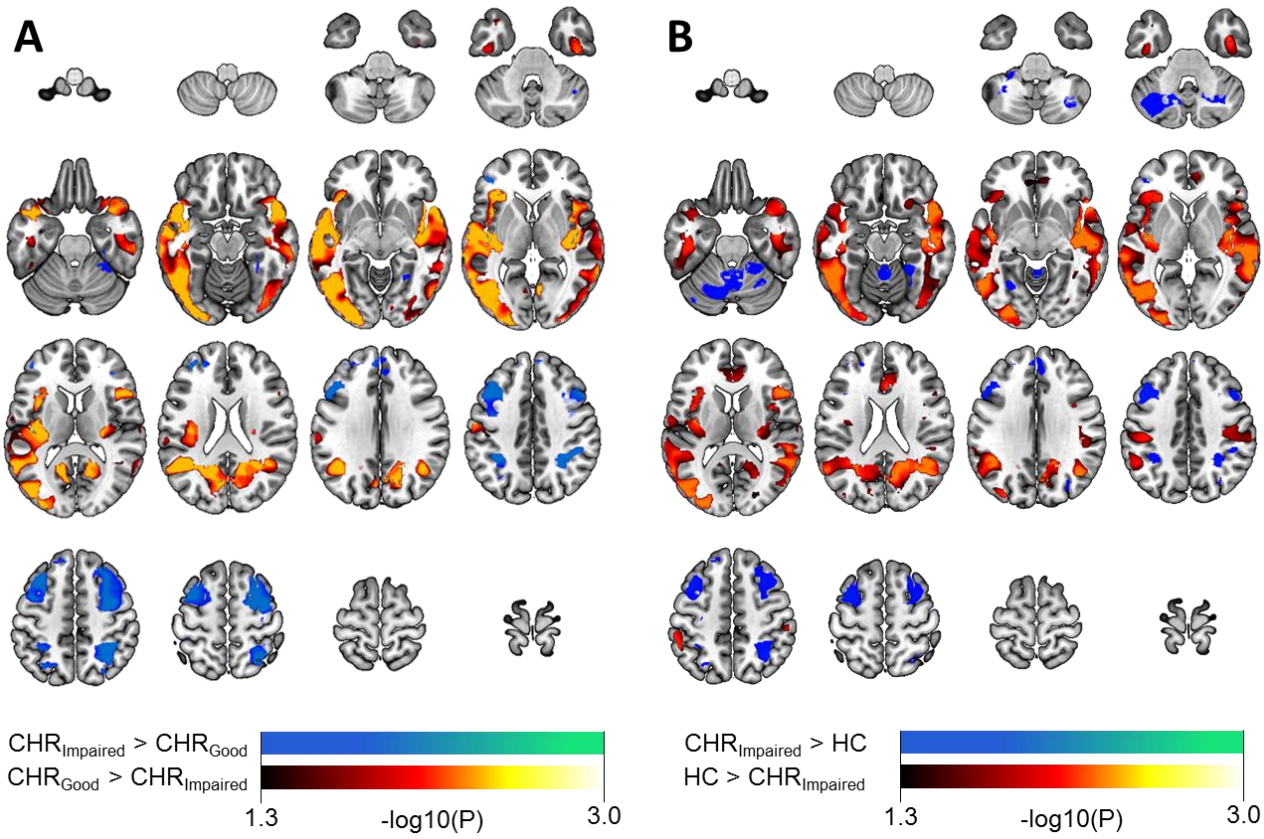

**eFigure 5: Image quality assessment performed in the T1-weighted images of 412 study participants** using the quality assessment functionality of the CAT12 toolbox. **A:** Violin plot showing distribution of the overall weighted image quality computed by the quality ascertainment procedure (QAP) of CAT12 using the noise and bias information of each scan. The QAP maps the rating scores to image quality grades (A-F) shown on the right side of the Figure. **B:** A $\chi^2$ analysis of the effects of overall image quality (measured in terms of image quality deciles) on the misclassifications rate in the combined CHR and ROD cohorts. The $\chi^2$ test did not find any significant effect between classification error and image quality decile at $\alpha$=0.05.



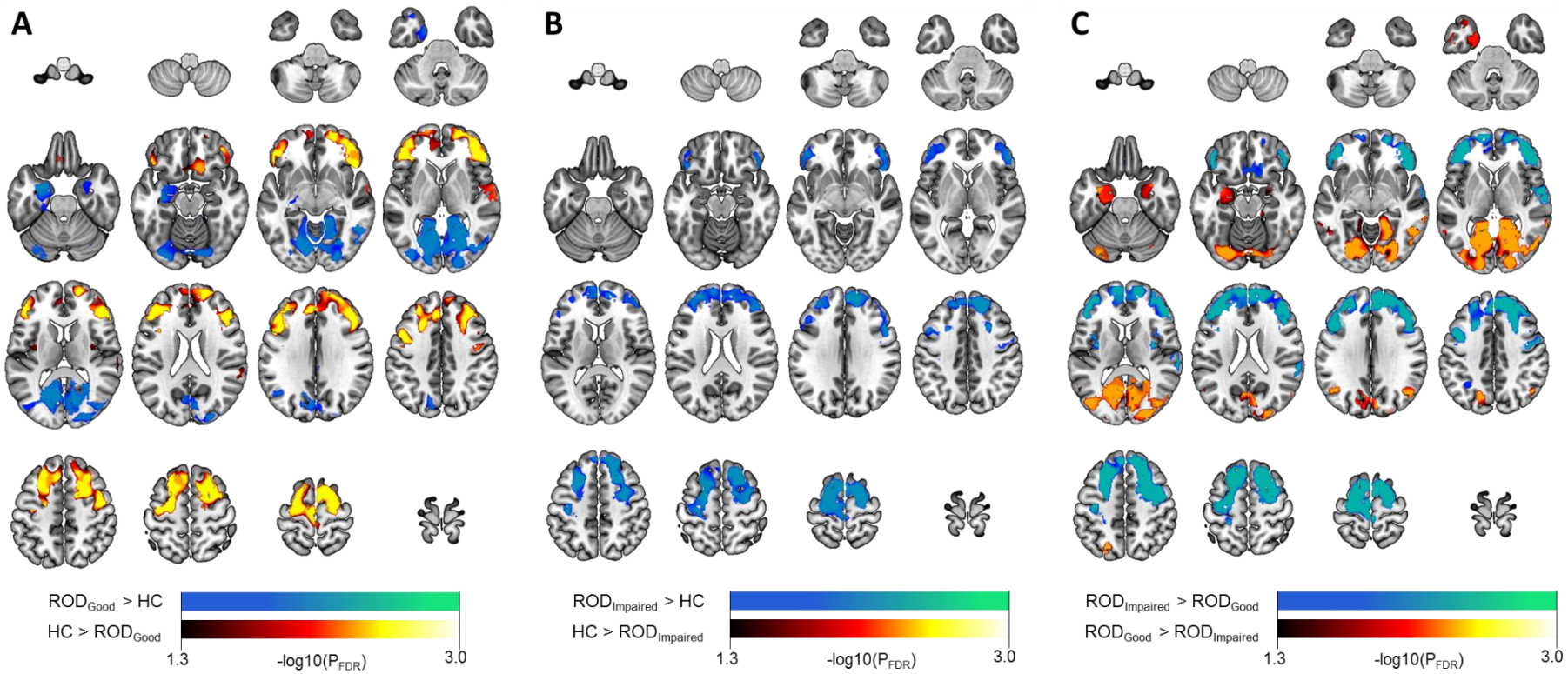

**eFigure 6:** Correlation analyses were conducted to assess whether the prognostic decision scores generated by the clinical and sMRI classifiers were influenced by the follow-up intervals in the CHR and ROD groups. The results were all non-significant as shown in scatter plots, suggesting the follow-up interval variation did neither impact on the clinical models' nor on the sMRI models' classification performance.



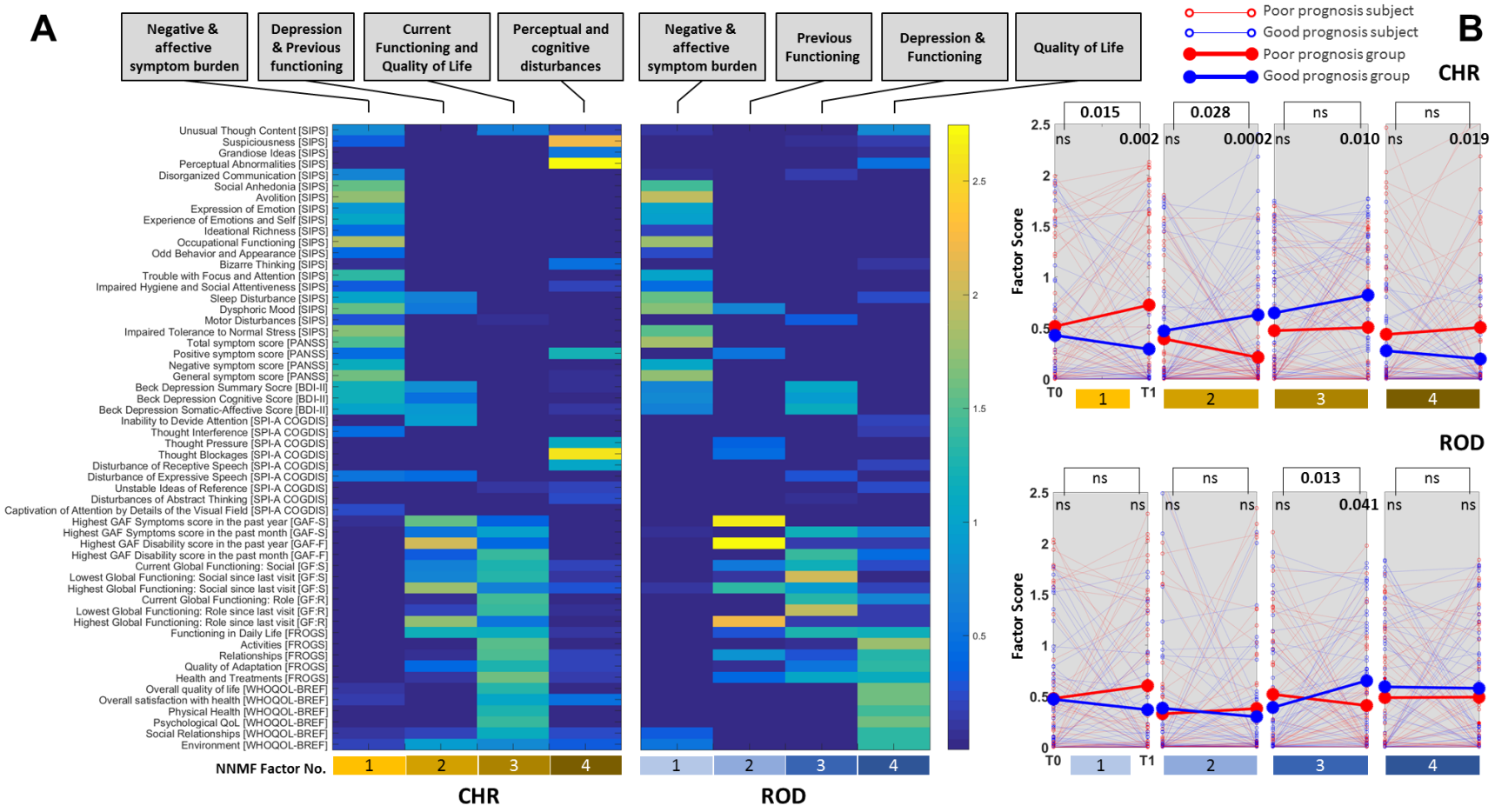

**eFigure 7:** The predictive signature underlying the original GF:S outcome prediction model (**A**) was qualitatively compared to the signature produced by the site effects analysis (**B**). **Abbreviations:** CVR Cross-Validation Ratio (see **Figure 1**, main manuscript, for methodological descriptions).



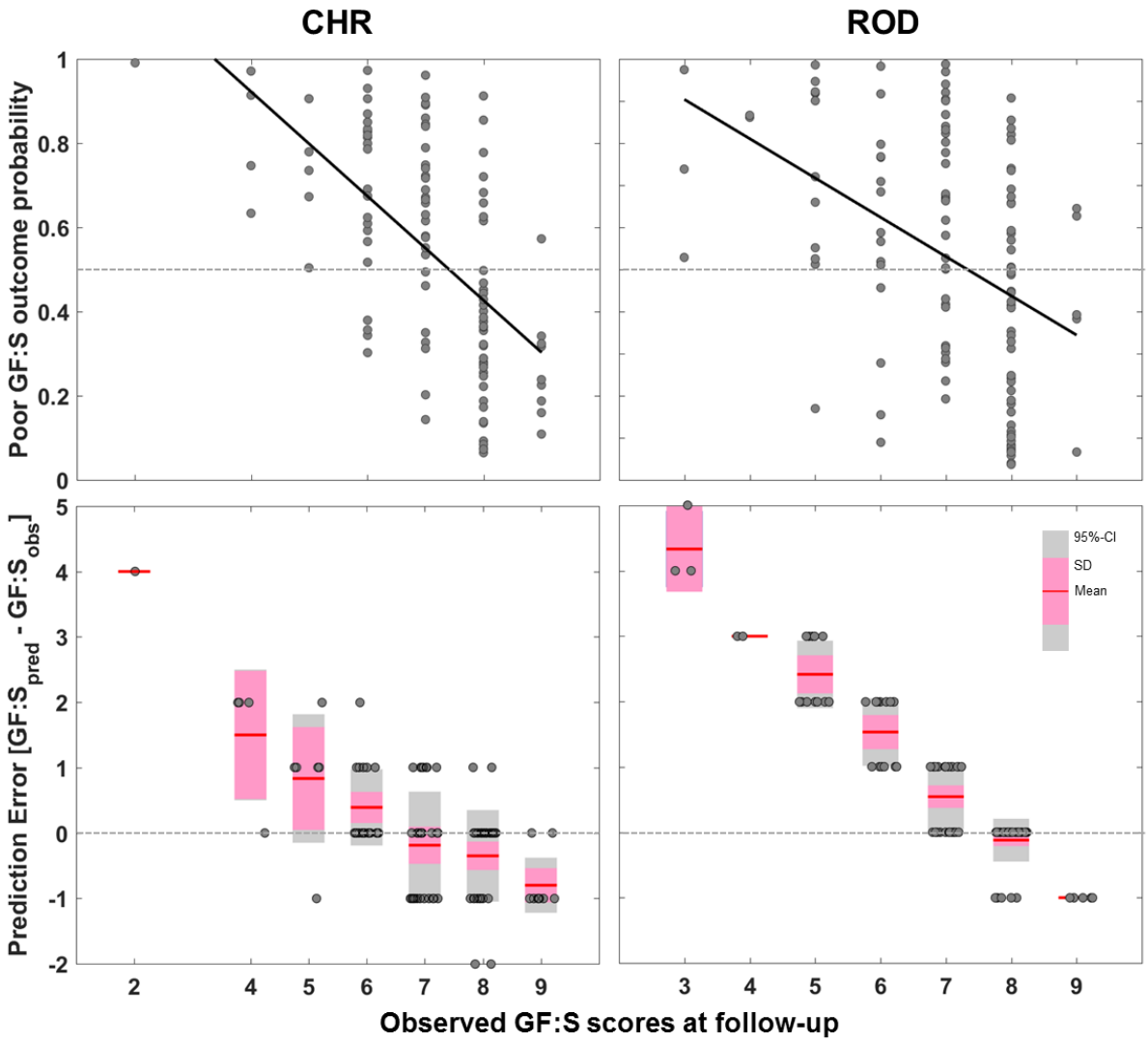

**eFigure 8: Results of voxel-based analysis of variance** between 67 CHR persons with predicted impaired ($CHR_{Impaired}$) GF:S outcome vs. 49 CHR persons with predicted good ($CHR_{Good}$) GF:S outcomes (**A**) and 67 CHR persons with predicted impaired GF:S outcome vs. 116 healthy volunteers matched for site, age, and sex to the CHR group (**B**).



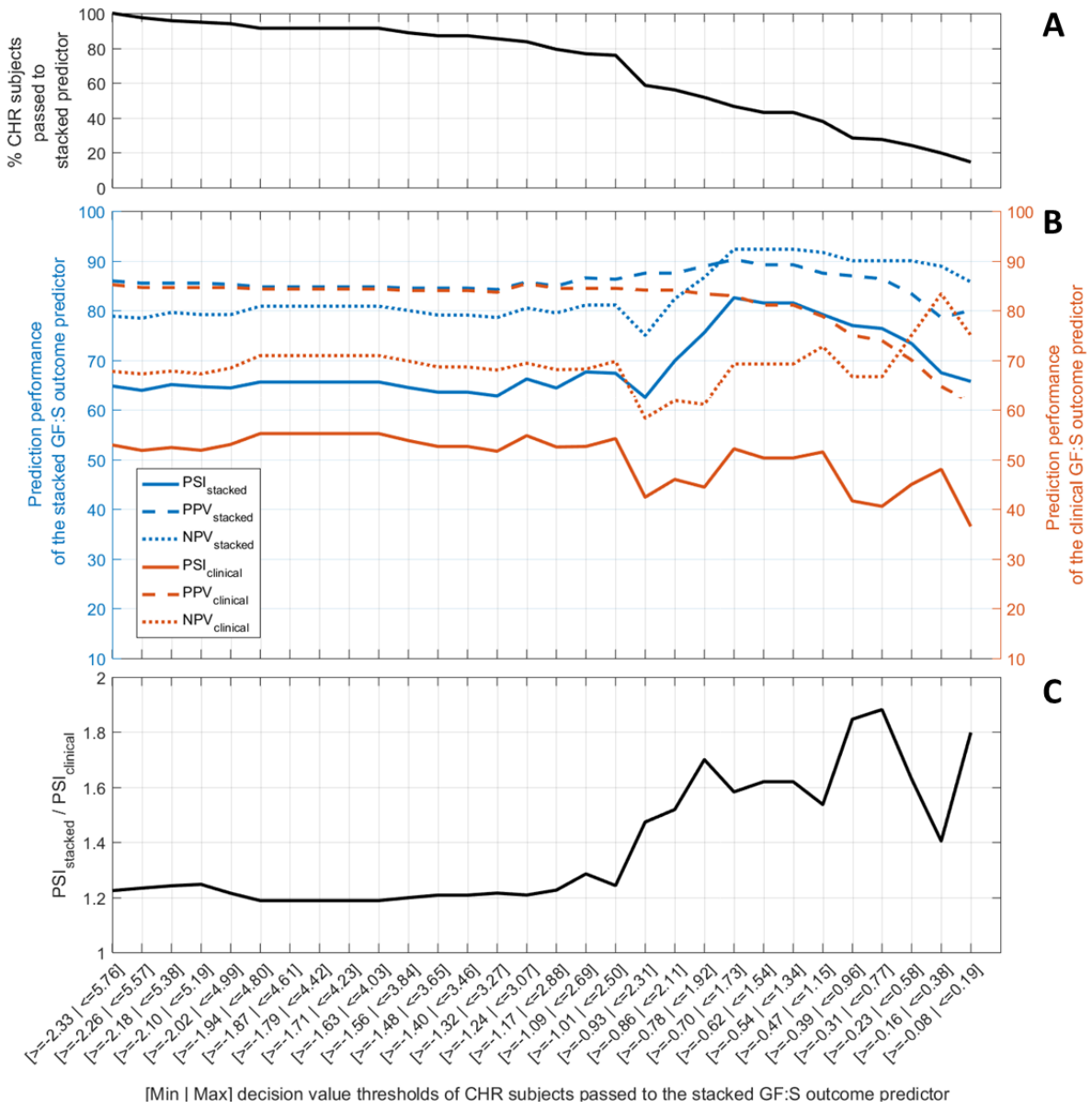

**eFigure 9: Results of voxel-based analysis of variance** between 120 healthy volunteers and 60 ROD persons with predicted good ($ROD_{Good}$) social functioning outcome (**A**), 120 healthy volunteers and 60 ROD patients with predicted impaired ($ROD_{Impaired}$) social functioning outcomes, (**B**) and the ROD patients with predicted good vs. impaired outcomes at follow-up (**C**). The 120 healthy volunteers were matched one-to-one for site, age, and sex to the patients in the ROD group.



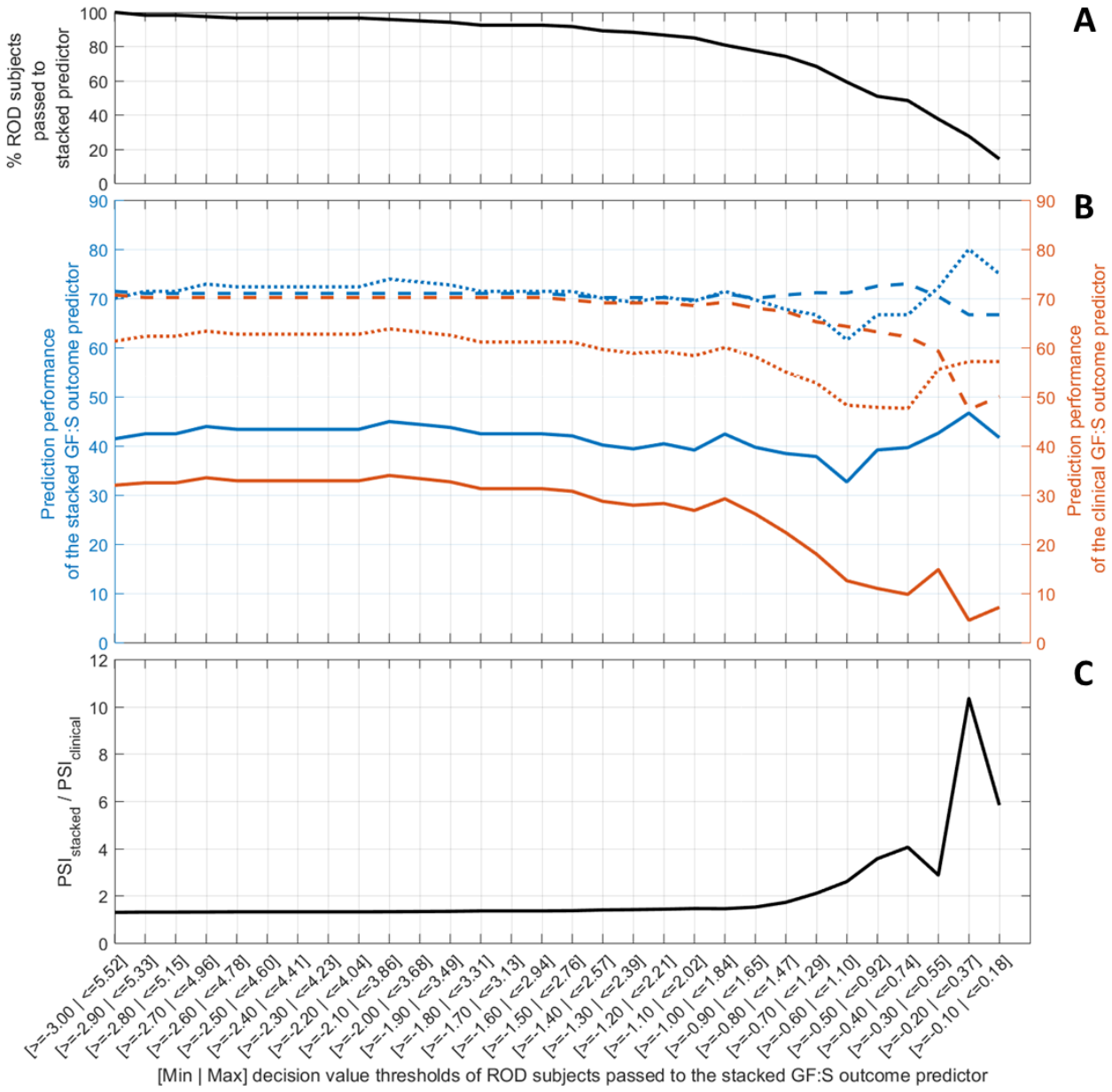

**eFigure 10: Prognostic generalization of the sMRI-based classifiers across 4 psychometric factors detected by Non-Negative Matrix Factorization (NNMF). A:** For each study group, the phenomenological baseline variables capturing attenuated positive, negative, and general symptoms, cognitive basic symptoms, and depressive symptoms, as well as measures of general, social and occupational functioning, and quality of life were projected to four factors using Orthogonal Non-Negative Matrix T-Factorization[52]. The resulting sparse factor matrices were inspected, and the factors were interpreted according to the variables showing non-negative loadings on given factor (see descriptor boxes above the factor loading matrices). **B:** The follow-up data available for the 9-month T1 examination ($N_{CHR}$=99; $N_{ROD}$=99; see also **eFigure 1**) were projected into the respective NNMF model. The obtained follow-up factor scores were used to compute factor trajectories for each CHR and ROD person in the analysis. Trajectories were plotted as colored lines depending on the predicted GF:S outcome class of given subject (red/blue: impaired/good functioning prognosis). For each factor, a repeated-measures analysis of variance with the factors TIME and PROGNOSIS CLASS was conducted to explore cross-sectional differences between PROGNOSIS CLASS as well as TIME x PROGNOSIS CLASS interactions of sMRI-based social outcome predictors. Significance was defined at α=0.05.




**eFigure 11:** Scatter plots in the upper panel show associations between observed GF:S scores ($GF:S_{obs}$) at follow-up and classification probabilities produced by combined clinical-sMRI based GF:S outcome predictor in the CHR (left) and ROD (right) group. Using post hoc ordinal regression these probabilities were calibrated to the observed GF outcome score range. The prediction errors of these regression models were separately analyzed in the CHR and ROD group in the box plots of the lower panel. Grey box plot areas represent 95% confidence intervals, purple areas the respective standard deviations, while red lines depict the mean of the predicted scores at given observed GF:S level.




**eFigure 12: Analysis of a sequential clinical-combined outcome predictor algorithm in the CHR group**. **A:** Percentage of CHR persons passed to the combined sMRI-clinical GF:S outcome prediction model at increasingly ambiguous clinical decision score cutoffs (see x axis in **C**). **B:** The left y axis shows the combined classifier's prediction performance measured in terms of positive predictive value (PPV), negative predictive value (NPV) and prognostic summary index (PSI). The right y axis shows the respective performance metrics of the clinical prediction model. **C:** PSI ratio of the two types of models as a function of clinical model ambiguity. Combined sMRI-clinical predictions seem to be particularly superior to the purely clinical algorithm in 52% of the CHR group having decision scores ≥-0.78 and ≤1.92.




**eFigure 13*:* Analysis of a clinical-combined outcome predictor algorithm in the ROD group**. See legend of **eFigure 11** for details.