Supplementary Online Content

Dwyer DB, Kalman JL, Budde M, et al. An investigation of psychosis subgroups with prognostic validation and exploration of genetic underpinnings: the PsyCourse study. *JAMA Psychiatry*. Published online February 12, 2020. doi:10.1001/jamapsychiatry.2019.4910

© 2020 American Medical Association. All rights reserved.

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

© 2020 American Medical Association. All rights reserved.

eMethods 1. Recruitment and sample characteristics

Details regarding participant recruitment can be found in Budde et al. (2018). Adult patients (≥18 years) with an ICD-10 lifetime diagnosis of schizophrenia (SZ; F20.x), brief psychotic disorder (F23.x), schizo-affective disorder (SZA; F25.x), bipolar disorder (F31.x), manic episode (F30.x), or recurrent major depression (reMDD; F33.x) were identified based on recommendations of the clinical staff or by querying patient registries of the participating clinical sites. Eligible individuals were invited to participate in the first study visit. Written informed consent was obtained from study participants according to European and German law, specifically permitting the use of pseudonymised data for broad research usage and the release of prior medical data from health records that was used to obtain information on medical history. A specialised data protection system was developed in order to protect patient confidentiality in the context of such broad consent (Budde et al., 2018). Identified participants were re-assessed using an adapted version of the Structured Clinical Interview for DSM-IV; Axis I Disorders (SCID-I). Participants were assessed for the following lifetime diagnoses: SZ (295.10/295.20/295.30/295.60/295.90), schizophreniform disorder (295.40), brief psychotic disorder (298.8), SZA (295.70), BD (296.0x/296.4x/296.5x/296.6x/296.8x), reMDD (296.3x). Participants were excluded if they did not meet the DSM-IV criteria for each of these categories or the they had insufficient language or intellectual capacities.

As reported in Budde et al. (2018), the recruitment procedure resulted in a discovery sample of 891 individuals. A total of 2.7% (n=24) were not patients at the time of entry into the study, 47.5% (n=420) were enrolled in psychiatric day clinics, 5.4% (n=48) were inpatients, and 44.3% (n=392) were outpatients. The high number of day clinic patients reflects a common treatment pathway for patients in Germany and Austria as part of a transitional model of care between inpatient and outpatient treatment. Day patients commonly attend a clinic between Monday to

© 2020 American Medical Association. All rights reserved.

Friday during working hours for psychiatric, psychological, and psychosocial treatment, but do not stay overnight. A total of 89% of participants were born in Germany or Austria and there was almost complete genetic overlap with individuals from the European ancestry reference population (see eMethods 6 and eFigure 8).

Raters were provided with written instructions for all instruments and each new rater was extensively trained by an existing experienced interviewer. The training was adapted to the ability of each rater. Depending on the rater experience, training included discussing instructions in detail, watching an experienced investigator conduct an assessment, and performing assessments under supervision. Formal training for all investigators was also held on a regular basis. Formal rater reliability testing was not conducted and this is a limitation of the current study.

© 2020 American Medical Association. All rights reserved.

eMethods 2. Baseline measures

Broad inclusion criteria were used to select measures based on clinical information that is usually considered during repeated real-world clinical assessments (eTable 1). These included variables assessing demographic details, medical history, the positive and negative symptom scale (PANSS[1]), the inventory of depressive symptomatology (IDS-C_{30}[2]), the young mania rating scale (YMRS[3]), verbal intelligence (Mehrfachwahl-Wortschatz-Test[4]), cognitive switching (Trail Making Test), speed of processing (Digit Symbol Substitution Test), attention and working memory (Digit Span forwards and backwards), general functioning (e.g., the global assessment of functioning; GAF[5]), WHO Quality of Life Questionnaire brief version (WHOQoL-BREF[6]), and a short version of the Big-Five personality test[7] (eTable 1). For questionnaires, all items were included rather than only summary scores in order to fully capture variance associated with illness.

© 2020 American Medical Association. All rights reserved.

eMethods 3. Analysis overview

The analysis overview is presented in eFigure 1. The analyses were divided into two main procedures associated with the discovery sample and the replication sample. In the discovery sample, subjects and baseline features that exhibited high missing values (>25%) were excluded, and the data were scaled and imputed using a k-nearest neighbours' approach (see eMethods 4). The sNMF procedure was then applied to the baseline data consisting of the application of sNMF 1000 times to randomly drawn samples in order to derive a consensus matrix that determined the optimal number of subgroups and the subject assignment (eMethods 5). The subgroup assignments were then forwarded to three main analyses: a) polygenic score analyses to determine genetic differences between the subgroups (eMethods 6); b) mixed modelling of longitudinal course of symptoms and functioning (eMethods 7); and c) multigroup supervised machine learning in order to address the limitations of the sNMF procedure and to obtain the subgroup labels in the replication sample (eMethods 8). Following the application of the supervised learning models to the replication sample, the resulting replication sample subgroups were characterised by their loading on each of the discovery sample sNMF factors, by their clinical and polygenic risk score differences, and the mixed-models from the discovery sample were applied without modification. Each of these steps is described below.

© 2020 American Medical Association. All rights reserved.

eMethods 4. Participant and feature filtering

As stated in the Main Document, a total of 126 cases were excluded because they had 25% or more missing baseline data. Following exclusion, there was no difference in the treatment status of the remaining 765 cases used in this analysis compared to the original sample with 2.9% inpatients (n=22), 51% day-patients (n=388), 5.8% inpatients (n=44), and 40.3% outpatients (n=307). Further analyses were then conducted to determine differences with the excluded cases across the main exploratory variables used throughout this study (eTable 3). Results demonstrated that the individuals were receiving less intensive treatment (i.e., less day and inpatients), had higher psychosis symptoms, and lower functioning.

Based on the inclusive feature list, features were excluded if 95% or more of values were equal (i.e., if the variable contained limited variance; n=10) or if the there were ≥25% missing values across participants (n=32). A total of 188 variables remained representing the key domains of demographics, clinical history, medical history, symptoms, general functioning, and cognition (eTable 1 and 2). Resulting variables were scaled [0,1] and imputed using a nearest neighbour approach[8] with the median of the 7 nearest neighbour cases.

© 2020 American Medical Association. All rights reserved.

eMethods 5. Clustering analyses

Sparse Non-Negative Matrix Factorisation

A consensus clustering approach was used incorporating non-negative matrix factorisation (NMF) that has been used with high dimensional genetics datasets to address the problems of a high variable-to-case ratio, the delineation of complex cluster boundaries, and interpretability of the factor solution within high dimensional spaces[9-14]. NMF is ideally suited to high dimensional data and is often used in genetic analyses to automatically extract sparse and interpretable factors. It reduces the data by splitting it into smaller subsets such that the distance between the original matrix and the subset matrix is minimal. For this study, we used a sparse variant of the NMF algorithm with default settings (beta=0.1; 200 iterations; tolerance = 0.0001) in order to induce sparsity within the coefficient matrices (i.e., in order to further separate individuals)[15, 16]. In particular, this technique uses L_1-norm minimization with alternating non-negativity constrained least squares in order to derive sparse coefficient matrices. A *k* range of 2-9 clusters was investigated and *k* was chosen using stability criteria (see below).

Consensus Clustering

Sparse NMF was embedded within a consensus clustering procedure[13, 14], which involves iterating the decomposition 1000 times and each time randomly resampling (without replacement) 80% of the original sample. This produces 1000 clustering solutions for each predefined *k* value that specifies the number of subgroups. For each iteration, an $n \times n$ co-consistency matrix was generated (n = subjects) wherein each element contained 1 if two individuals were classified in the same cluster and 0 otherwise. We summed the matrices across the 1000 iterations for each *k* solution to generate a consensus matrix reflecting the

© 2020 American Medical Association. All rights reserved.

frequency in which individuals were included together in each cluster. Consensus matrices from all *k* clustering solutions were summed and agglomerative hierarchical clustering (Euclidean distance; average linkage) was then applied to this ensemble to generate final partition of the data and clustering solution. This is an established method that has been outlined in previous methodological papers[13, 14] and employed in oncology[9-12].

Selection of Number of Factors/Subgroups

The main criterion for the determination of factors/clusters was cluster stability. Quantification of stability involves the assessment of the number of times that two individuals are either included, or not included, in a cluster together. This can be visualised in as a heatmap (eFigure 2) and the ideal solution is one in which two individuals are either always placed into the same cluster or are never placed into the same cluster—i.e., there are clean and distinctive boundaries between clusters. As per defined guidelines[9-14], we used the Cophenetic correlation coefficient to quantify the degree of stability of cluster assignment across iterations of the clustering procedure and determined the solution by selecting the highest Cophenetic correlation coefficient prior to a declining trend[13]. Additionally, because limitations of the Cophenetic correlation coefficient have been recently raised[17], we used a new additional measure that also quantifies cluster stability according to the cumulative density function of the consensus matrices for each *k* called the proportion of ambiguous clustering (PAC)[17]. For the PAC, we selected the cluster solutions that minimised the PAC value.

In addition to applying the standard methods used in oncology to define the number of clusters[9-14], we extended the previous genetics methods by employing a permutation based approach to obtain a non-parametric *p*-value for the winning *k* solution. This was in order to confirm that the cluster solution did not occur by chance. Specifically, for each permutation, we shuffled the

© 2020 American Medical Association. All rights reserved.

coefficient scores on each factor within individuals and recalculated the Cophenetic and PAC scores. We then iterated this procedure 1000 times in order to construct null distributions each score and calculated the significance by determining the number of times that the permuted values exceeded the observed values. Significance was determined at $p<0.05$. Notably, the original matrices were not shuffled and NMF repeated in this analysis due to the computational time cost of the NMF clustering procedure—i.e., (1000 iterations × (k-1) × 1000 permutations = 80 000 NMF iterations.

Factor Solutions and Variable Interpretation

In order to interpret the feature weights of the factor solutions across iterations of the consensus clustering approach, we also extended previous consensus clustering research[9-14] by establishing a stability criterion for the feature weights using a previously published technique[18]. We first aligned the factors of each of the NMF iterations within the winning k solution by correlating factors with the first NMF iteration and matching the factors with the highest correlation coefficients (matching factors $r > 0.8$). The mean and standard error of each variable weight across all iterations of the NMF procedure was then computed. For each iteration, variables with an absolute weight value greater than their respective standard error multiplied by 1.96 were set to one. This identified weights that reliably contributed to the average factor weight score using a 95% confidence interval[19, 20]. These binary matrices were then summed, and the variables that passed the threshold in 95% of iterations were interpreted as being stable. We observed that the NMF solution was strongly stable across iterations despite the resampling scheme used.

Factor Loadings and Radar Plots

© 2020 American Medical Association. All rights reserved.

In order to graphically represent the coefficient loadings of each subgroup on each factor, the average weighting for each subgroup taken from the mean coefficient matrix (i.e., arising from the 1000 repetitions of the sNMF algorithm) was calculated and plotted using radar plots.

All analyses were conducted using MATLAB R2015b and code is available on request and at https://github.com/domdwyer.

© 2020 American Medical Association. All rights reserved.

eMethods 6. Genotyping and calculation of schizophrenia polygenic risk scores

PsyCourse samples were genotyped using the Infinium CoreExome-24+Human PsyChip Consortium, versions 1.0 and 1.1 (Illumina, USA). Quality control steps, using PLINK 1.9[21], have been previously described[22]. Principal components analysis (PCA) on the genetic relationship matrix modelled ancestry differences between study participants using the EIGENSOFT package (smartPCA[23]). It uses a principal component analysis based on a pruned subset of approximately 50 000 autosomal SNPs, after excluding regions with high linkage disequilibrium. Genotype imputation in these samples was carried out with IMPUTE2 / SHAPEIT[24, 25], using the 1000 Genomes project Phase 3 reference panel. Genetic variants with a poor imputation quality (INFO < 0.8) and a MAF<1% were excluded in downstream analyses. Polygenic risk scores for schizophrenia, bipolar disorder, major depressive disorder and educational attainment were calculated using the summary statistics of the latest respective GWAS[26-29]. PRS were calculated with PLINK1.9 by multiplying the imputation dosage of each risk allele by the log(OR) for each genetic variant in the respective GWAS (training data). The resulting values were summed up to an individual estimate of the schizophrenia/bipolar disorder/major depressive disorder/educational attainment genetic risk burden in each individual at 10 thresholds between $p<5\times10^{-8}$ and $p=1$ based on previous research[30].

Data from this study was not included in the 1000 Genomes Project or any of the latest GWAS used to calculate PRS for schizophrenia, bipolar disorder, major depressive disorder, or educational attainment[26-29].

In order to investigate genetic variation, the first two ancestry principal components from the discovery and validation samples were overlayed onto the European reference population of the 1000 Genomes project[31] with results demonstrating substantial similarity (eFigure 8) that

© 2020 American Medical Association. All rights reserved.

reinforced the descriptive results showing that 89% of the sample was born in Germany or Austria.

© 2020 American Medical Association. All rights reserved.

eMethods 7. Longitudinal analyses using mixed models

Longitudinal outcome measures were specifically chosen based on their importance for prognoses in terms of symptoms and functional outcomes. A further requirement was that they were easily interpretable as prognostic markers and that they could be analysed with conventional statistics. Towards these goals, measures included PANSS total, IDS-C_{30} total, YMRS total, GAF, and the WHOQoL-BREF. Total scores were used in order to increase the interpretability of results and to restrict the amount of multiple comparisons to avoid false positives. Symptom and functioning variables were extracted for the baseline and each of the three follow-up periods conducted at 6-mth intervals.

Illness course was investigated over the three longitudinal time-points and four primary outcome measures of symptoms and functioning using mixed models in the discovery sample (eMethods). (*R* package *lme4*[32]; REML; Satterthwaite approximation of degrees of freedom; cases and intercepts modelled as random effects). Main fixed effects of subgroup and time, linear trends, quadratic trends, and trend interactions with subgroup were tested. Mixed model parametric bootstrap methods for the Likelihood Ratio Test (LRT) assessed model complexity (*R* package *PBmodcomp*[33]; 200 iterations). Post-hoc tests (*EMMEANS*) were used to compare trends of each measure between subgroups (corrected for multiple testing using false-discovery rates[34]). Effect size for each complete model (marginal R^2) was calculated following Nakagawa & Schielzeth[35] and Johnson[36] as implemented in the *R* package *r2glmm*[37]. Data were not imputed and all cases had completed the study. Additional analyses accounting for missing data were also conducted (see eResults 8 below).

© 2020 American Medical Association. All rights reserved.

eMethods 8. Identification of critical variables and replication analyses

We used robust resampling procedures within a consensus clustering framework that have been extensively used in other medical fields with high dimensional data[9-12], but the approach has three limitations related to its potential for replication and external application: 1) the feature set is not reduced, which implies that exact replication would require all 188 variables to be included in further studies and this is not practically feasible; 2) external validity in another sample is not assessed; 3) metrics regarding the accuracy of individual classification into each group are not provided. We addressed each of these limitations by performing additional supervised machine learning analyses in a new sample of individuals who were collected as part of ongoing PsyCourse data collection (n=458; see Main Document).

The aim of the machine learning analyses was to first classify individuals in the discovery sample on the basis of the 10 variables with the highest sNMF weight within each factor (Table 2, eTable 4) before applying these models to the held-out subjects in order to categorise individuals into the subgroups. The raw data from a total of 45 unique features (Table 2, eTable 4) were forwarded to a machine learning pipeline using the open source tool NeuroMiner (http://www.pronia.eu/neurominer). A nested cross-validation framework was used with the aim to predict subgroup membership in 10 pair-wise comparisons (i.e., 1 vs. 2, 1 vs. 3, 1 vs. 4, etc.) and pairwise decoding was used to determine multigroup classifications[38]. In the inner loop of the cross-validation, the training data were separated into 10 folds. Training data were preprocessed by pruning variables with no variance (i.e., all one value), followed by variable scaling [0,1] and imputation using the nearest-neighbour approach; i.e., for each missing value, 7 nearest-neighbour cases were determined (Euclidean distance) and median of neighbours imputed. Training samples were analysed with an L2-regularised logistic regression algorithm (LIBLINEAR; https://www.csie.ntu.edu.tw/~cjlin/liblinear/) and we selected models that

© 2020 American Medical Association. All rights reserved.

maximized the balanced accuracy across a range of hyperparameters (geometric progression from $C=2^{-6}$ to 2^{6}). To account for uneven sample sizes of each subgroup pair we employed class weighting of the *C* hyperparameters using the inverse ratio of the training group sizes. For each training set, this procedure was repeated 10 times. The analysis chains were applied to the held-out test sample, resulting in 10 x 10 = 100 decision scores reflecting the degree to which an individual was classified into one of the subgroups. Each individual's class membership was determined through majority voting.

The models, including all preprocessing and classification steps of the analysis chains for all cross-validation folds, were then applied without modification to the baseline data of the PsyCourse validation sample (n=458; see Main Document). Each individual's class membership was determined using majority voting.

Following the determination of the subgroup membership in the validation sample using the supervised learning procedure, the subgroups were characterised in terms of the clinical differences on the *a priori* selected variables and also according to their factor loadings (Figure 1, Main Text). In order to determine the factor loadings, the sNMF models from the discovery sample analyses were applied to the replication cohort baseline data in order to obtain individual subject loadings on the five factors (i.e., the coefficient matrices). The mean for each subgroup defined by the supervised learning procedure described above was then calculated for each factor and the results were overlayed onto the discovery sample factor loading plot (Figure 1). PRS differences were also quantified following the procedure listed above for the discovery sample in eMethods 6. Finally, the longitudinal mixed models (eMethods 7) were also applied without modification to the replication sample in order to compare the illness trajectories over the 18-month time period.

© 2020 American Medical Association. All rights reserved.

eResults 1. Subgroup determination

Consensus clustering sNMF indicated potential clustering solutions of *k*=2,5,6 based on the Cophenetic correlation coefficient and the PAC (eFigure 2). A five cluster solution was chosen as the most stable, parsimonious, explanatory, and predictive on the basis of: 1) the two cluster solution did not demonstrate strong consensus cluster boundaries and did not result in optimal separation across key variables; 2) the six subgroup solution contained a subgroup with only 7 individuals in it and this was not considered sufficient for further analyses; 3) the retention of a predominantly five subgroup solution even when selecting six subgroups was apparent in eFigure 2.

© 2020 American Medical Association. All rights reserved.

eResults 2. Factor solutions

Five sNMF factors were dictated by the subgroup stability solution (eFigure 2). The top features for each factor consisted based on the 95% confidence criterion are presented in eTable 4. Factors are discussed in the main text and are labelled as: factor 1, good quality of life; factor 2, suicide history; factor 3, depression symptoms and deficits; factor 4, environmental risk and male gender; factor 5, psychosis related symptoms and deficits. As expected, the subgroups preferentially loaded on each of the five factors such that each group could be explained with reference to the corresponding factor—e.g., 51% of the total factor loadings of subgroup 1 were explained by the first quality of life factor (eFigure 2). However, it was notable that each of the other four factors jointly contributed approximately half (mean(SD) = 54% (6%)) of the remaining variance for each group, implying that all five factors influenced the group separation and characterised the groups to varying degrees. Indeed, the suicide factor explained 36% of the total factor loadings for the second group, while the other factors explained the remaining 64%. This shared variance highlights the importance of each of the other four factors to the derivation of the suicide subgroup and also the contributions of all factors across the subgroups.

© 2020 American Medical Association. All rights reserved.

eResults 3. Site and experimental rater effects

Because PsyCourse was a naturalistically designed study containing multiple sites with differing populations, catchment areas, intake criteria, and services, it was possible that subgroups were systematically influenced by the site that the participants were derived from. In addition, there was the possibility of systematic experimental rater effects, despite the use of commonly employed questionnaires that typically have sufficiently high inter-rater reliability[39].

In order to address these potential sources of bias, we conducted further analyses. For many studies, the effect of site and rater are interlinked because raters are commonly associated with specific sites. However, in PsyCourse 9 raters assessed 294 cases across 14 sites, which provided an opportunity to test site, rater, and site x rater interactions. Specifically, we analysed the mean sNMF coefficient scores for the five-factor solution because these scores are ultimately used as a basis to determine the cluster membership of individuals and provide more variance than cluster memberships alone. For each of the five factors, we performed an ANOVA investigating site, rater, and site x rater effects. Results demonstrated site effects across multiple factors, but rater and site x rater effects were not found at $p<0.05$, uncorrected (eTable 8).

In order to further explore the effects of site across subgroups, we performed χ^2 analyses testing the relationship between subgroup membership and site (eTable 7). These tests were also significant ($\chi^2=441.7$, $p<0.001$), and post-hoc tests demonstrated that this was predominantly driven by 59.3% (n=51; $\chi^2=315.8$, $p<0.001$) of subgroup 5 (severe psychosis) being collected from one site in Germany (Günzburg). This raised the possibility of a systematic rater effect, which directed the identification of the subgroup 5 (e.g., the rater at the site systematically rating higher than other raters in the consortium). In order to exclude the

© 2020 American Medical Association. All rights reserved.

possibility that a site bias was mediating the formation of subgroup 5, we conducted two main further analyses that repeated the main sNMF procedure while: 1) excluding the cases from Günzburg; and 2) excluding the Günzburg cases and replacing them with matched cases from the replication cohort. Replacement with matched cases was a critical step because it removes the influence of the Günzburg site without biasing the analyses when removing approximately 59% of a specific subgroup. We then compared these results with the findings from the original discovery sample.

Individuals from the Günzburg site (n=67) were matched based on the following variables: educational status, current paid employment, duration of illness, DSM-IV schizophrenia diagnosis, the global assessment of functioning (GAF), the WHO Quality of Life Scale global score, and the positive and negative syndrome scale (PANSS) sum scores for the positive, negative, general, and total domains. We then identified the 188 variables used in the discovery sNMF analysis for these individuals in order to merge the cases with the original discovery sample. During this process, we discovered that scoring changes had occurred between the discovery (i.e., PsyCourse data release v1.1.2) and replication samples (i.e., PsyCourse data release v3.0). Specifically, the level of severity reported for 15 binary somatic health variables (e.g., “History infectious diseases”, yes/no) that had been increased such that more severe instances of the diseases or conditions were required to fulfil the criterion (e.g., more severe infectious diseases) (eTable 14). This necessitated the removal of these variables so that the discovery and replication samples could be merged. As such, we then conducted conducted three analyses: a) the original analysis but without the somatic variables; b) a new analysis excluding the Günzburg site and the somatic variables; c) an analysis replacing the Günzburg cases with matched cases from the replication sample. We then compared the factor solutions and specifically the individuals in the fifth severe psychosis subgroup.

© 2020 American Medical Association. All rights reserved.

Visual inspection of the factor solution was virtually identical to the original discovery analysis (eFigure 9), which was supported by an average (SD) correlation between the factors of r=0.98 (0.03), p<0.001. The fifth subgroup representing cases with severe psychosis was then specifically compared across the main clinical variables of interest analysed in the main document (eTable 15). Removing individuals from the Günzburg site resulted in decreased numbers of cases in the fifth subgroup (i.e., 42 versus 78), but this increased with the addition of matched replacement cases (i.e., 65 versus 78), which demonstrates the specificity of the consensus clustering procedure in identifying individuals with a severe psychosis phenotype. Clinically, the solutions that did not differ on the most characteristic variables that separated this group, such as the educational attainment and PANSS symptoms. However, some differences to the discovery analysis were notable including: a decrease in treatment status (i.e., less people in inpatient settings), an increase in the percentage of individuals with illicit drug use, an increase in depression and mania symptoms, and a decrease in the global assessment of functioning score (eTable 15). These differences occurred despite the same factor solution, suggesting that they are not critically involved in the formation of this subgroup. Importantly, there were no remaining site biases for the fifth subgroup across these additional analyses (eTables 16 & 17).

When combined, the findings suggest that the convenience and referral sampling method used in the study to naturalistically assess a broad range of psychiatric presentations resulted in an enrichment of subtypes at specific sites, but that this enrichment did not mediate the subgroup determination. Despite this finding, the results emphasise the importance of conducting multi-site studies in order to adequately sample the full range of clinical presentations, especially for the detection of low prevalence subgroups in specialised programmes at specific hospitals.

© 2020 American Medical Association. All rights reserved.

Finally, the identification of a difference in the measured severity of the somatic variables raised the possibility that changes in the measurement of the severity of the infectious diseases variable included in the supervised learning replication analysis may have altered the results. We investigated this possibility by repeating the supervised learning analysis without the infectious diseases variable. Specifically, we trained a new multigroup model to classify individuals in each of the five subgroups using the same feature set as in the original analysis (eMethods 8, above), but without the infectious diseases variable. We then applied this modified model to the validation sample and conducted tests related to the correspondence with the factor loadings, the longitudinal illness course, and the polygenic risk analysis. Results demonstrated the same patterns as found in the original analysis and manuscript (eFigure 10). These results suggest that replication of the subgroups using the machine learning models is not dependent on changes in the level of severity for items such as the infectious diseases variable.

© 2020 American Medical Association. All rights reserved.

eResults 4. Validation of subgroups: further explanation

Highly overlapping mean factor loadings were demonstrated (Figure 1, Main Text), indicating a high degree of similarity between the groups. Further machine learning analyses were conducted to quantify the degree of similarity between the discovery and replication groups on the basis of the top 10 sNMF variables identified in from the main analysis (Table 2, Main Text). Specifically, we repeated the supervised machine learning analysis described in eMethods 8 but instead of separating individuals into subgroups we attempted to classify individuals from the discovery and replication samples within each subgroup (e.g., we attempted to distinguish subgroup 1 discovery individuals in the discovery sample from subgroup 1 individuals from the replication sample). Results demonstrated an average classification accuracy of 62% (eTable 13). When compared to an average classification accuracy in the discovery sample of 90.6% in separating the groups from each other (eTable 11), the results highlight a degree of replication that agreed with the visual representation of the factor score overlap depicted in Figure 1 (Main Text).

Further analyses also indicated substantial overlap between the discovery and validation subgroups. As represented in eTable 12 and eFigure 4, validation subgroup 1 (n=149) consisted of a greater proportion of females (65.8%) with a later-onset of affective psychosis diagnoses, lower symptom burden, and a stable illness course. Subgroup 3 (n=134) contained individuals with predominantly depressive symptoms and quadratic illness course, subgroup 4 exhibited similarly high-functioning individuals with psychosis with largely stable illness course, and subgroup 5 consisted of individuals with low educational achievement, severe symptoms, quadratic illness courses, and reduced education polygenic score (eFigure 7).

© 2020 American Medical Association. All rights reserved.

eResults 5. Supplementary analysis: clustering stability for different imputation settings

As described in eMethods 4, the data were scaled and imputed using a k nearest neighbour (knn) approach prior to the sNMF procedure. For this study, 7 nearest neighbours were selected based on prior experience in supervised and unsupervised learning. However, the possibility remained that different numbers of nearest neighbours could have resulted in a different solution. We tested whether this was the case by repeating the main analysis for knn=[3,5,9] in order to compare this with the original solution and determine the stability of the clustering solution. As shown in eFigure 11, the knn value did not change the factor solutions. The average correlation (SD) between each factor across knn numbers was r=0.98 (0.0001).

© 2020 American Medical Association. All rights reserved.

eResults 6. Supplementary analysis: analysis of separability based on cognitive variables

Although we found a robust clinical solution that did not require the use of cognitive variables, the relationship between the fifth severe psychosis subgroup and the level of education raised the possibility that separating boundaries between the subgroups could be achieved with cognitive variables. To investigate this possibility, in the discovery sample, we forwarded the cognitive variables used in the main analysis (Trail Making Test, A and B; Digit Span Forwards; Digit Span Backwards; Digit Symbol Test; and the Mehrfachwahl-Wortschatz-Test) to a multigroup supervised learning pipeline that mirrored what is described in eMethods 8 above. We used a wrapper technique to determine feature importance. These results showed that the only subgroup that could be reliably separated was the fifth subgroup, labelled as the severe psychosis group, with a balanced accuracy of 70.6% (eTable 18) and backwards digit span (a test of attention and working memory) was selected most frequently (eFigure 12). All other subgroups were separated at chance levels. These results highlight the cognitive impairment found in this group while also demonstrating that the other subgroups are poorly separated based on this short cognitive battery despite having distinctive clinical and outcome signatures. A limitation of these analyses was that a limited number of cognitive variables available in this cohort and further studies are required to determine if additional cognitive measures (e.g., declarative or visual memory) could better separate the subgroups.

© 2020 American Medical Association. All rights reserved.

eResults 7. Supplementary analysis: analysis of differences in lifetime illness course

This study was specifically interested in short- to medium-term illness courses (i.e., symptoms and functioning) over an 18-month period because this is relevant to more immediate treatment planning and could serve as valuable information to a treating team. However, a question that remains is how these medium-term illness dynamics are embedded into the long-term course. In order to address this question, we used the Operational Criteria Checklist for Psychotic Illness and Affective Illness (OPCRIT)[40] Item 90 which is a specific operational criteria checklist designed to separate individuals into four subgroups based on illness course from onset to the time of the assessment: 1, multiple episodes with good remission; 2, multiple episodes with partial remission; 3, ongoing chronic disease; 4, ongoing chronic disease with deterioration. All available data was used to answer the item, including: medical records, previous PsyCourse data, and during interviews. This measure was collected at the third follow-up time period at 18-months.

Lifetime course differences on the OPCRIT Item 90 were analysed using Kruskal-Wallis tests for both the subgroup solution and for diagnostic labels as a comparison. Significant differences were found for subgroups ($\chi^2_{(4,371)}$=47, p<0.001) and diagnoses ($\chi^2_{(3,372)}$=32.8, p<0.001) (eFigure 13). An increase of episodic courses was found across analyses for subgroup 1 (i.e., affective psychosis subgroup; p<0.001) and also the diagnostic label of affective psychosis (i.e., bipolar disorder I and II; p<0.001) when compared to groups containing more individuals with schizophrenia. Importantly, however, separability of the subgroups based on a higher prevalence of a chronic lifetime course of illness was only found for the unbiased subgroup solution in the severe psychosis subgroup (p<0.001). This finding highlights that the subtyping approach can also identify subtypes of long-term illness, but is most important when compared to the 18-month illness course. Specifically, the severe psychosis group exhibited an inverted U-

© 2020 American Medical Association. All rights reserved.

shaped illness course as measured by the global assessment of functioning (see Main Text and Figure 2 in the Main Document). These results highlight the medium-term illness dynamics that are embedded within a long-term course that is clinically described as chronic and ongoing.

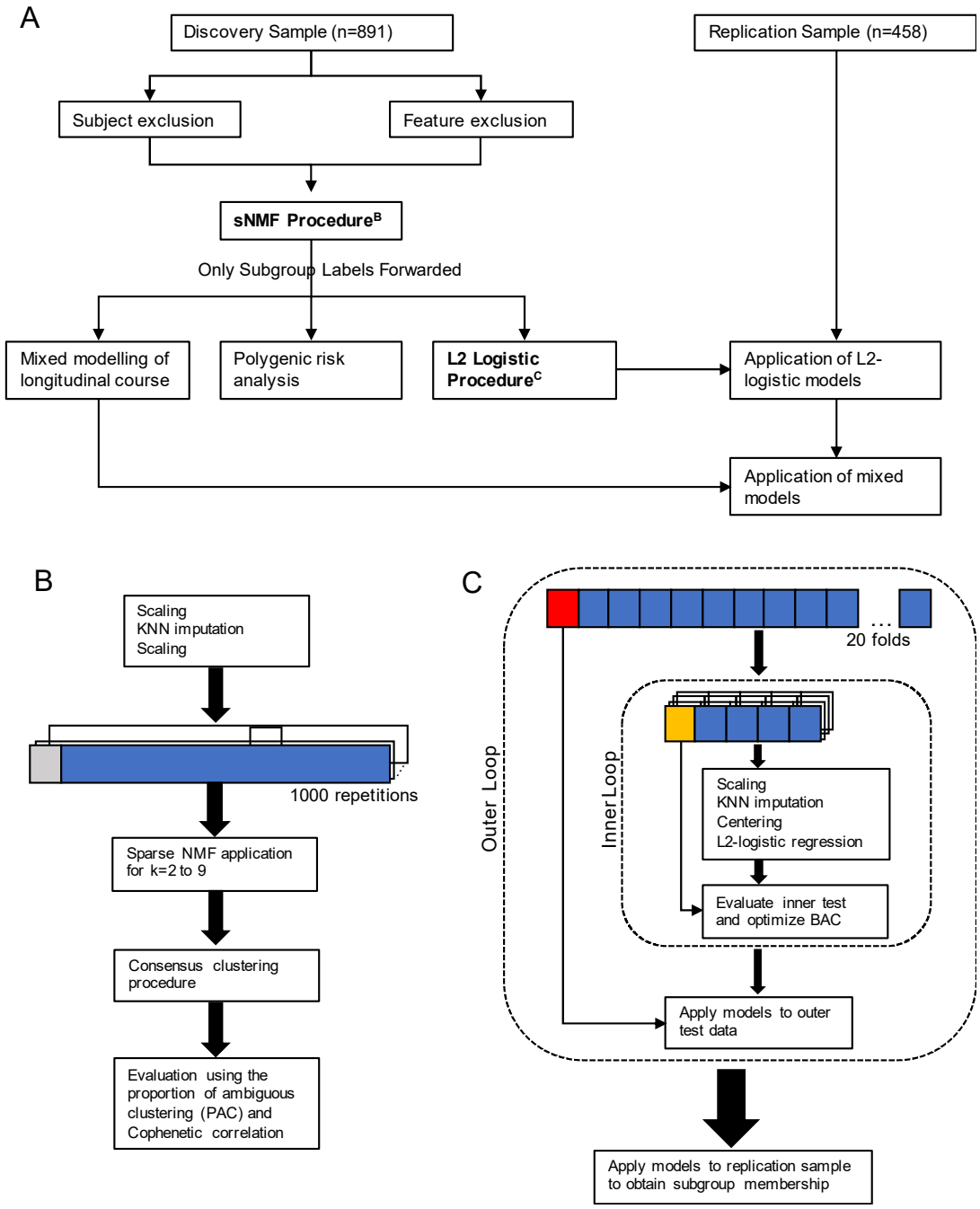
© 2020 American Medical Association. All rights reserved.

eResults 8. Supplementary analysis: analysis of missing participants in longitudinal data and mixed models

As stated in the eMethods, discovery subjects had an average of 2.5 follow-up timepoints and replication subjects had 1.8. Mixed models were used that account for missing data over timepoints, but the possibility remained that an interaction between attrition and subgroup could have mediated the results. To address this problem in the variables assessed using mixed models, we first quantified the data retention across each subgroup for all timepoints associated with the discovery and replication samples (eTable 19). We then repeated the mixed model analysis while controlling for the number of longitudinal missing data-points for each individual and also the interaction of missing data with subgroup. At the fourth time-point, there was an average of 54% missing data in the discovery set and 83%. Missingness did not significantly predict any of the longitudinal variables of interest (eTable 20), the interaction of missingness with subgroup was non-significant (eTable 20), and the significant trend differences outlined in the main text were maintained (eTable 21). These results support the observation that attrition increased over the four time-points, but that this did not affect the significance of the mixed model illness course difference. This result also supports the observation that the inverted U-shaped patterns were found for variables, such as the GAF and IDSC, despite more linear rates of attrition that did not decrease at the 18-month time-point (Figure 2, Main Text; eFigure 4).

© 2020 American Medical Association. All rights reserved.

eFigure 1. Analysis flowchart and overview



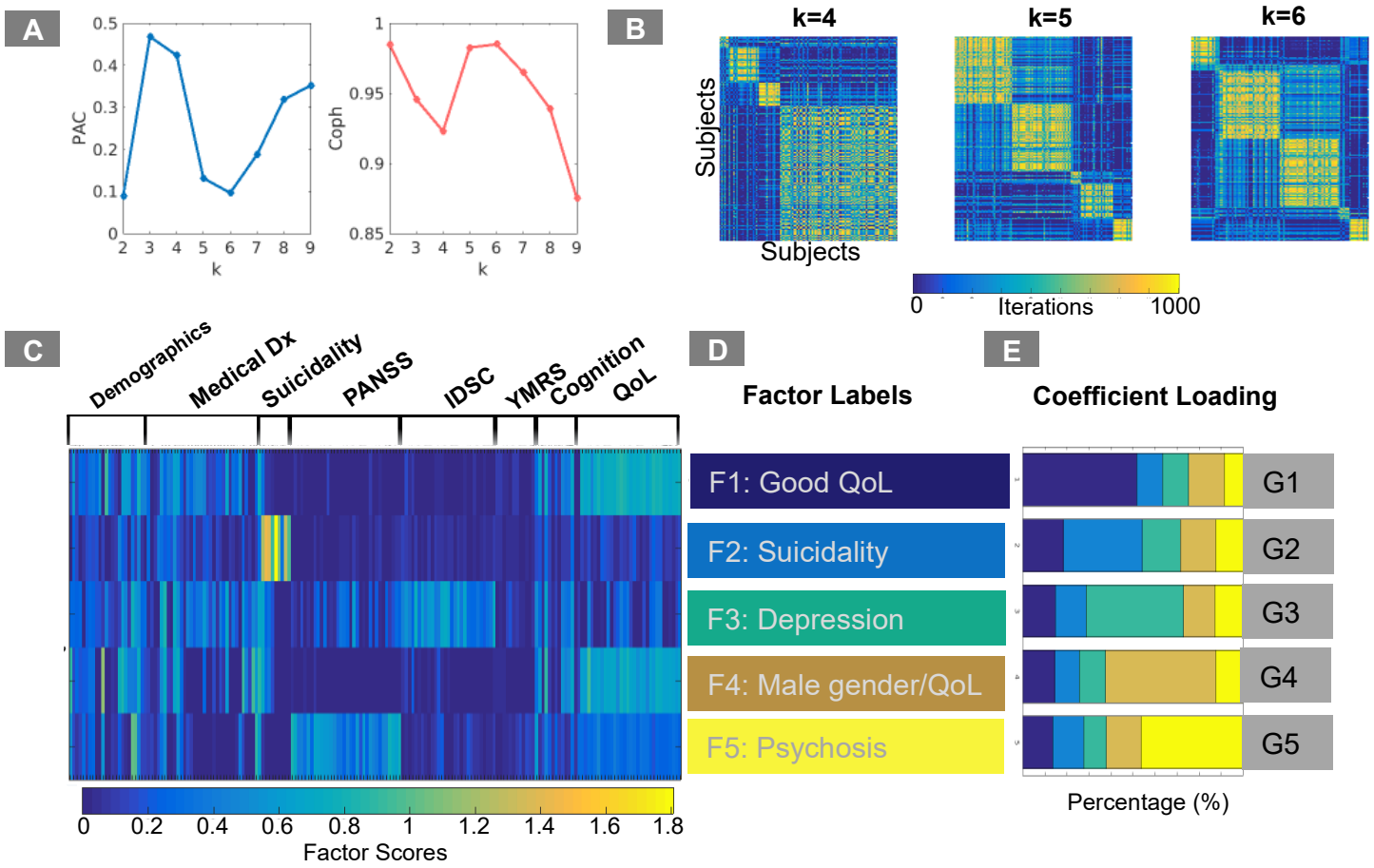
The discovery sample was first processed to exclude subjects and features with more than 25% missing values before being passed to the sparse non-negative matrix factorisation (sNMF) procedure (A) to define the subgroups. Following the subgroup definition, mixed models were used to characterise longitudinal trajectories for pre-selected symptom and functioning variables, differences polygenic risk scores were calculated, and the labels were passed to a supervised learning procedure (B) in order to create subgroup models. The supervised learning models were then applied without modification to the replication data in order to assign individuals to subgroups and the discovery cohort mixed models for each of the subgroups were then applied to the longitudinal data from the replication sample. **B.** the sNMF was adapted from established research and involved data scaling, k-nearest neighbours imputation, and scaling before being separated into 1000 randomly selected samples (80% of total). The sNMF algorithm is applied to each, consensus clustering

© 2020 American Medical Association. All rights reserved.

is used, and then the solutions are evaluated for stability of cluster assignment (see eMethods). **C**. The supervised learning procedure involves repeated, nested cross validation where parameters are optimised in an inner loop before being applied to the held out data (see eMethods). The models were then applied to the replication data.

© 2020 American Medical Association. All rights reserved.

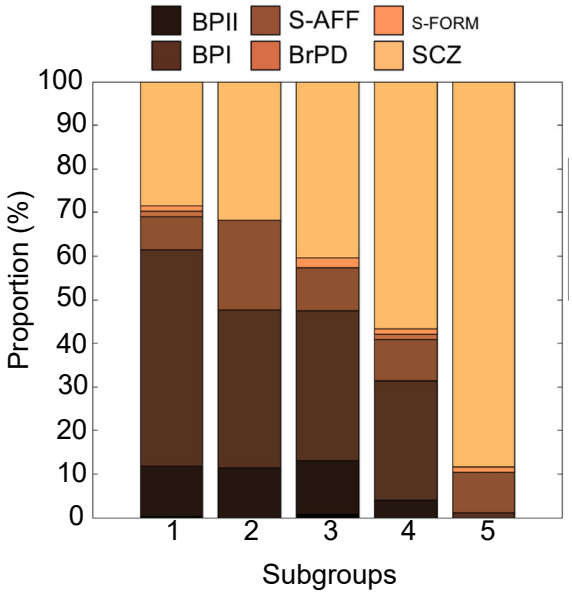
eFigure 2. Sparse nonnegative matrix factorization (sNMF) consensus clustering results



A. The proportion of ambiguous clustering (PAC[17]; left) and the Cophenetic correlation coefficient[13] (right) indicate factor/group solutions of 2, 5, and 6 groups. **B.** Consensus heatmaps display the amount of times that two individuals are clustered together (high consensus = yellow colour). The purpose of the plots is to find the cleanest cluster partition where cluster boundaries are well defined and the clusters demonstrate a high consensus density. A 5 factor/group solution was chosen because: 1) the 2 factor solution had less dense consensus areas and did not demonstrate meaningful separation across the outcome measures; 2) the 6 factor/group solution contained a subgroup of only 7 individuals. Permutation-based analyses testing the possibility of the observed Cophenetic correlation coefficient occurring by chance (eMethods) demonstrated that the five subgroup solution was significant (median permuted c=0.15, observed c=0.98, *P*=0.001). **C.** The mean basis factor matrix (presented horizontally) for a k=5 solution across all iterations highlighting variable loadings across the 188 demographic, clinical, cognitive, and quality of life measures—see eTable 4. **D.** Approximate factor labels were given to each of the groups taking into account the variables associated with each factor in eTable 4. **E.** The proportion of factor loadings for each of the 5 subgroups corresponding to the factors is plotted based on the sNMF coefficient matrix. Results highlight that while groups preferentially load on their respective factors (e.g., life quality 1 factor loading on group 1), there is a substantial contribution (>50%) from the remaining factors within each group. Notes: PANSS, positive and negative symptom questionnaire; IDSC, inventory of depressive symptomatology (30 items); YMRS, young mania rating scale; QoL, World Health Organisation quality of life questionnaire brief version.

© 2020 American Medical Association. All rights reserved.

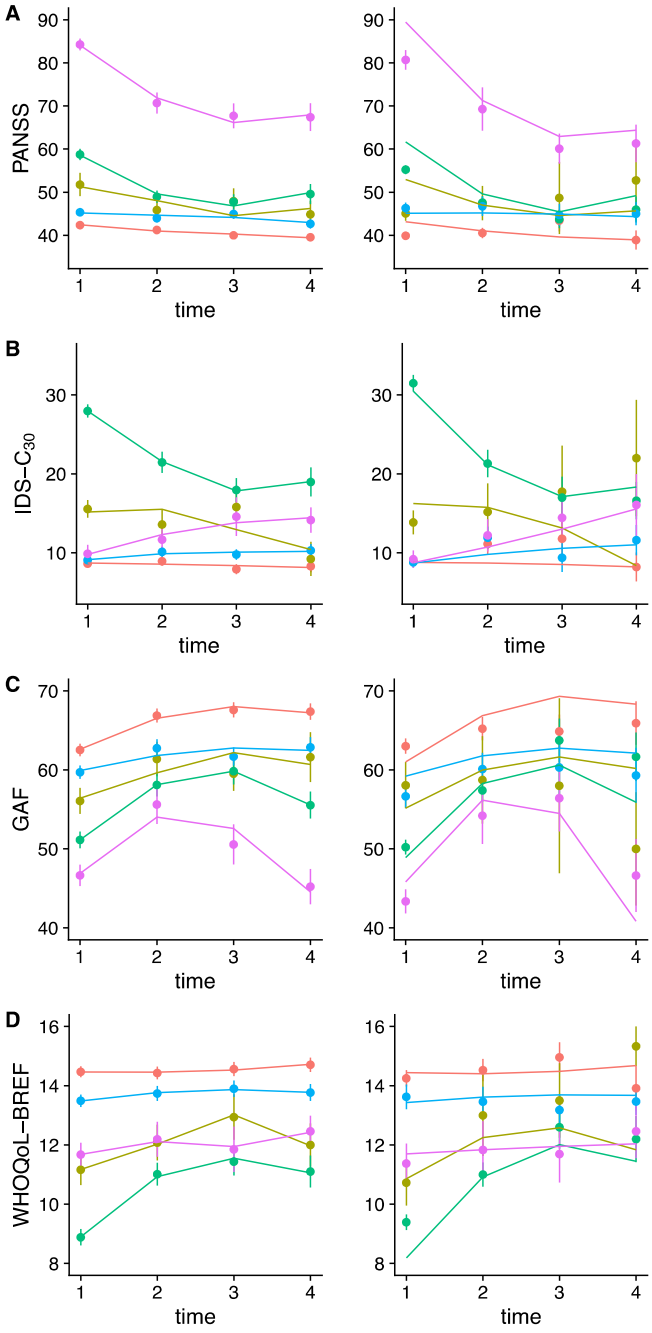
eFigure 3. Proportion of diagnoses across the subgroups



Distribution of DSM-IV diagnoses across the subgroups ordered according to the prevalence of schizophrenia diagnosis from the first (29%) to the fifth groups (88%). Figure demonstrates increasing proportion of schizophrenia diagnoses and evidence of transdiagnostic subgroups. For exact proportions and significance tests see Table 1 and eTable 5. Significant differences in diagnostic proportions were found for schizophrenia, bipolar I, and bipolar II (p<0.001). BPII, bipolar II; BPI, bipolar I; S-AFF, schizoaffective disorder; BrPD, brief psychotic disorder; S-FORM, schizophreniform; SCZ, schizophrenia.

© 2020 American Medical Association. All rights reserved.

eFigure 4. Illness course comparisons of discovery and validation cohorts



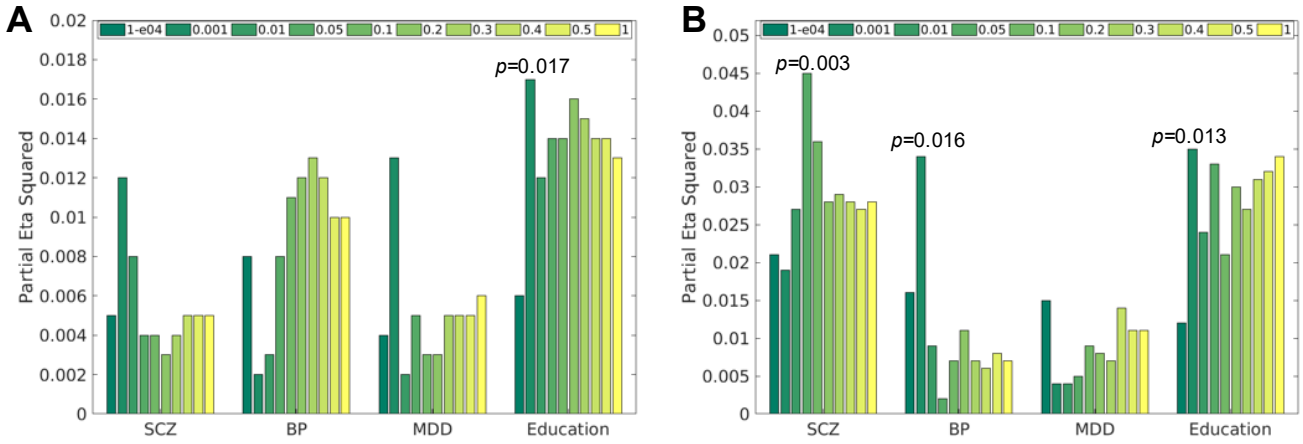
Illness course comparisons of the discovery (left) and external validation (right) samples in 6-month intervals. Linear mixed models from the discovery sample (left) were applied without modification to the external validation sample (right). In both cases, lines represent the fitted predicted values and dots are observed mean (SE) values. **A**) Psychosis symptoms as measured by the positive and negative symptom scale (PANSS) broadly demonstrated the same trajectories compared to the discovery sample with the severe psychosis group demonstrating a quadratic curve that was higher than the other subgroups; **B)** Depression symptoms measured with the Inventory for Depressive Symptoms Clinician version (IDS-C_{30}), also demonstrated similar trajectories for the depressive psychosis subgroup and the severe psychosis subgroup; **C)** Global Assessment of Functioning (GAF) demonstrated a similar quadratic trajectory for the severe psychosis group; D) the World Health Organisation

© 2020 American Medical Association. All rights reserved.

Quality of Life scale (WHOQoL-BREF) demonstrated similar trajectories overall, except for the suicidal psychosis subgroup.

© 2020 American Medical Association. All rights reserved.

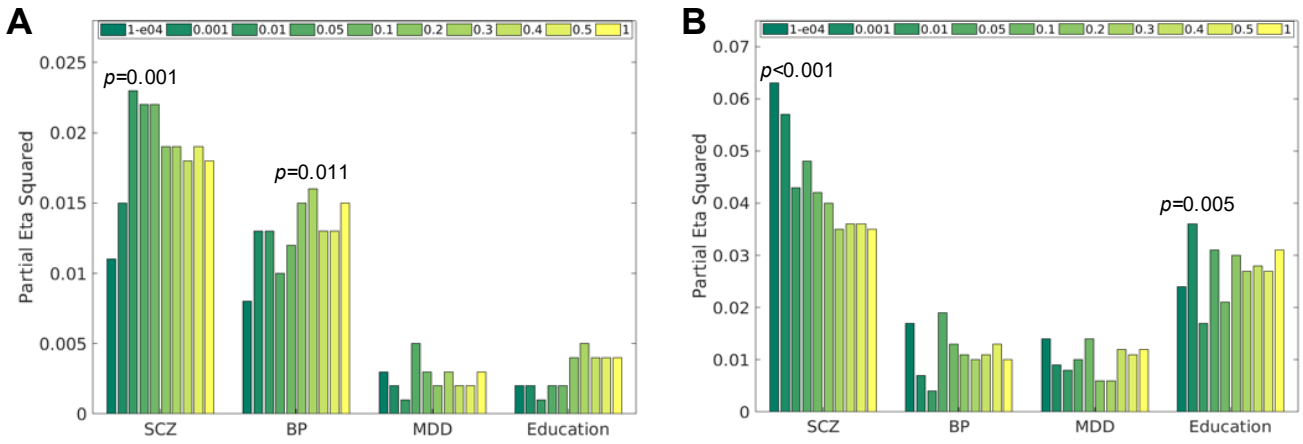
eFigure 5. Effect sizes of polygenic scores differentiating subgroups



Effect size (partial η^2) of polygenic scores differentiating subgroups in the discovery (left) and validation (right) samples. Values for differentiating between the data-driven subgroups (ANCOVA) samples based on polygenic scores based on genome-wide association studies for schizophrenia (SCZ), bipolar disorder (BP), major depressive disorder (MDD), and education. For each polygenic score, ten values are presented that reflect commonly used *p*-value cut-offs for single-nucleotide polymorphism (SNPs) included in the making the score from $p<0.0001$ to $p<1$ (see legend). For each polygenic score, the most significant effect is indicated. **A)** In the discovery sample, the highest effect sizes were found for the education polygenic score, which was significant at uncorrected levels across all SNP *p*-value cut-offs ($p<0.05$), except $p<0.0001$ and 0.01. **B)** In the validation sample, increased education effect sizes were also found that were significant at uncorrected levels across all SNP *p*-value cut-offs except 0.0001, 0.01, and 0.1.

© 2020 American Medical Association. All rights reserved.

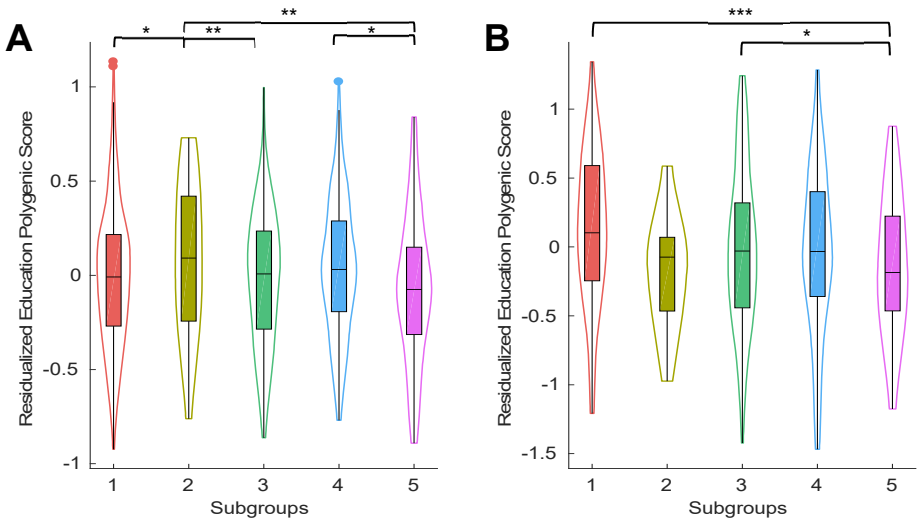
eFigure 6. Effect sizes of polygenic scores differentiating diagnostic groups



Effect size (partial η^2) of polygenic scores differentiating diagnostic groups in the discovery (left) and validation (right) samples. Diagnostic categories of schizophrenia, affective psychosis (bipolar I/II), schizoaffective disorder, and schizophreniform/brief psychotic disorder were assessed. ANCOVA conducted on polygenic scores based on genome-wide association studies for schizophrenia (SCZ), bipolar disorder (BP), major depressive disorder (MDD), and education. For each polygenic score, ten values are presented that reflect commonly used *p*-value cut-offs for single-nucleotide polymorphism (SNPs) included in the making the score from $p<0.0001$ to $p<1$ (see legend). For each polygenic score, the most significant effect is indicated. **A)** In the discovery sample, the highest effect sizes were found for the SCZ (significant at uncorrected levels across all SNP *p*-value thresholds except $p<0.0001$) and BP (significant across all *p*-value thresholds except $p=0.05$). **B)** In the validation sample, significantly increased SCZ polygenic score was found across all SNP cut-offs in addition to education, which was significant at uncorrected levels except for the SNP cut-off value of $p=0.1$.

© 2020 American Medical Association. All rights reserved.

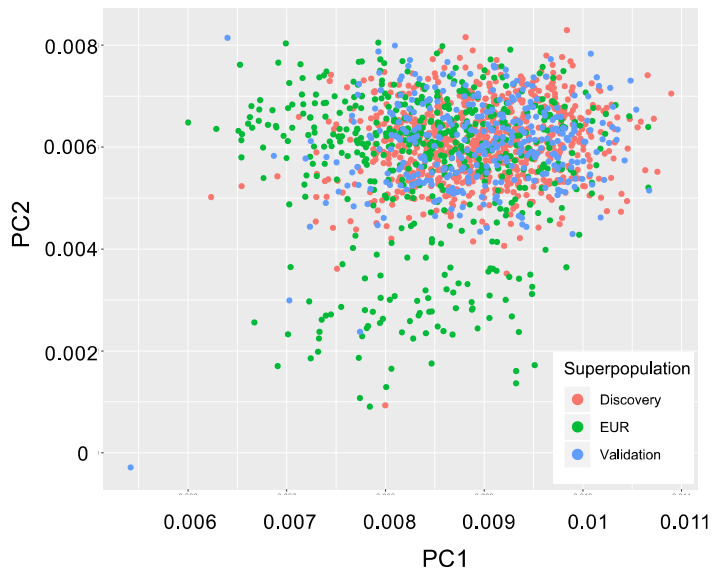
eFigure 7. Violin plots of education polygenic scores



Comparison of the education polygenic score residuals after controlling for the first four principal components at the single-nucleotide polymorphism (SNP) cut-off of *p*<0.001. A) in the discovery sample the significant main effect was influenced by a difference between subgroup 5 (severe psychosis) when compared respectively to subgroup 2 (suicidal psychosis) and subgroup 4 (higher functioning psychosis) using least significant difference (LSD) criterion. The difference between subgroup 5 and subgroup 2 remained after Bonferroni correction for multiple comparisons (p=0.027, corrected). **B)** in the validation sample differences were found between subgroup 5 (severe psychosis) when compared respectively to subgroup 3 (depressive psychosis) and subgroup 1 (higher functioning affective psychosis) using LSD and the difference between subgroup 1 and subgroup 5 remained after Bonferroni correction (p=0.005, corrected). Significance values: *, p<0.05; **, p<0.01, ***, p<0.001.

© 2020 American Medical Association. All rights reserved.

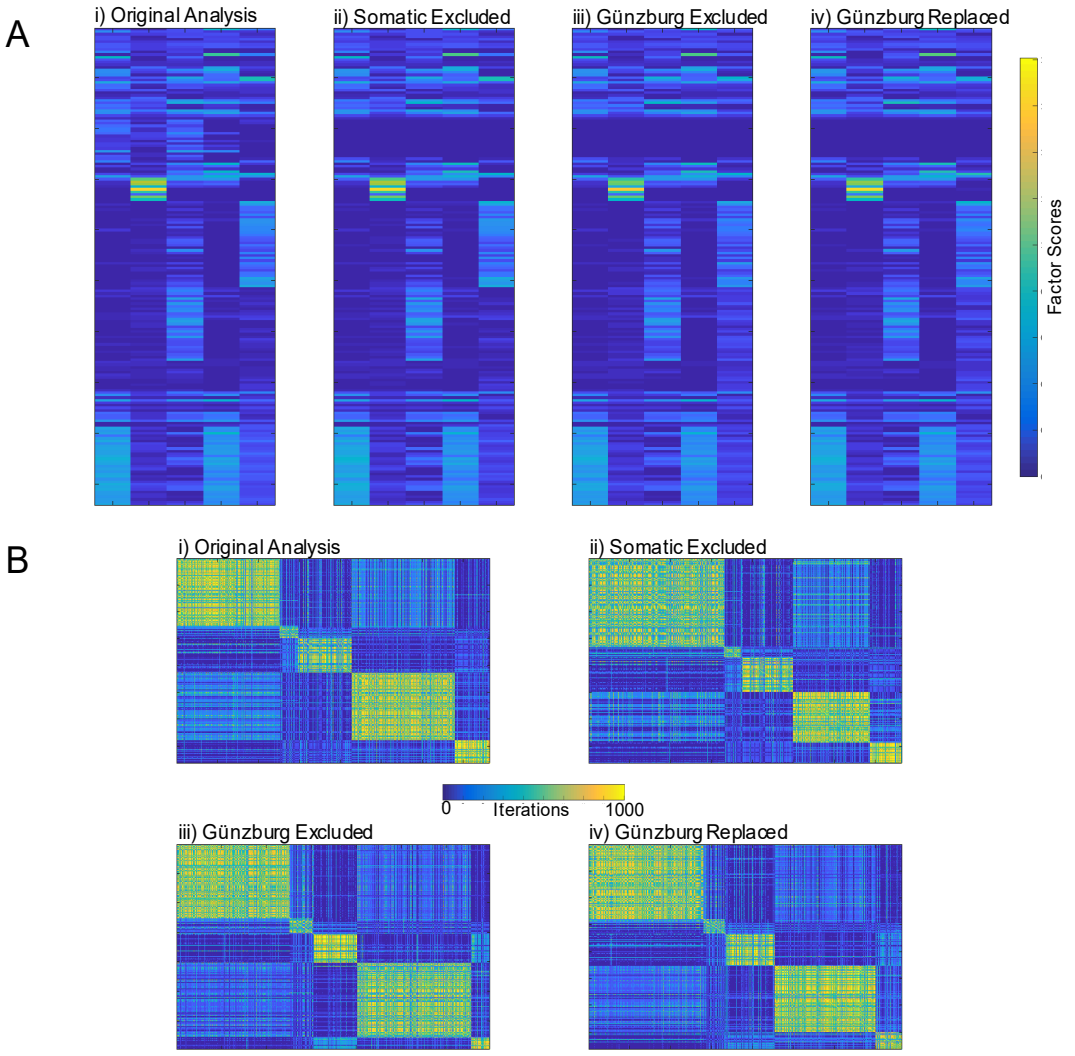
eFigure 8. Genetic ancestry overlap with the European reference population



Plot of the first two ancestry principal components for the discovery (n=765) and validation (n=458) samples overlayed onto the European ancestry (EUR) reference superpopulation from the 1000 Genomes Project. Results demonstrate substantial overlap with the reference sample indicating homogeneous European ancestry of the discovery and validation samples. Notes: EUR, European ancestry reference population.

© 2020 American Medical Association. All rights reserved.

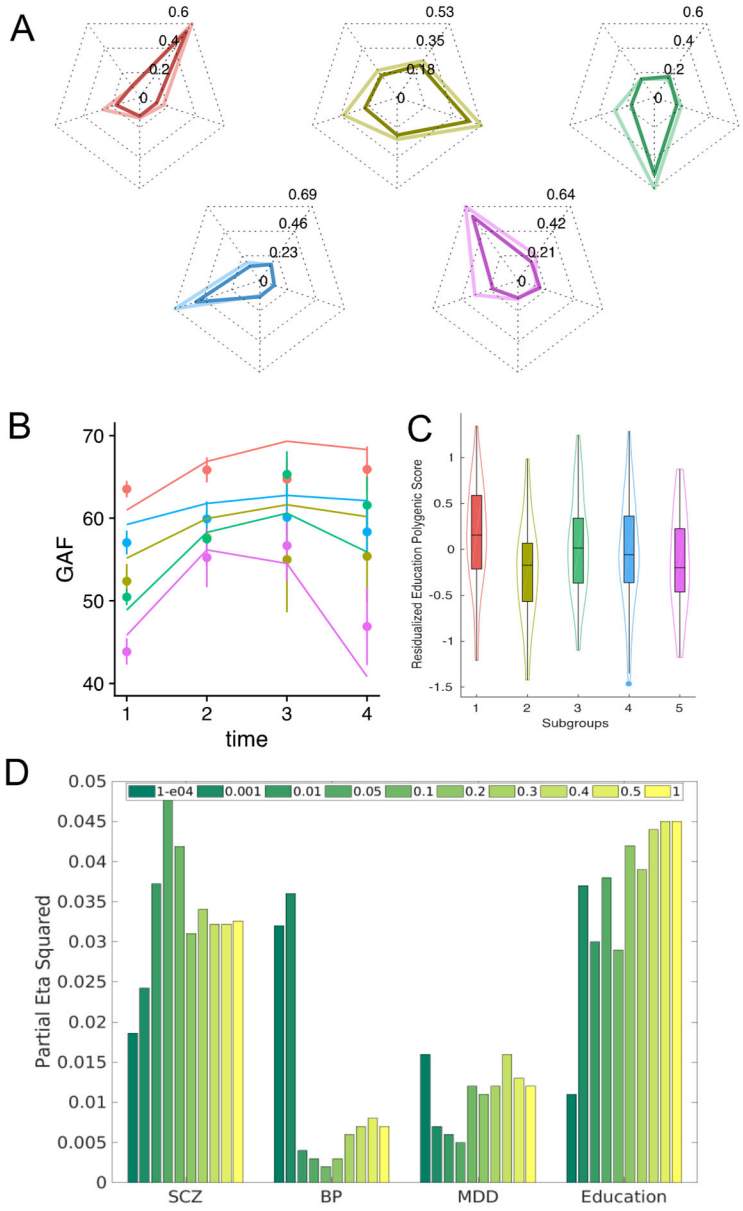
eFigure 9. Günzburg site exclusion factor and consistency matrix results



Comparison of factor matrices and consensus matrices for: i) the original discovery analysis; ii) the discovery analysis when somatic variables are excluded; iii) the discovery analysis when the Günzburg site is excluded; iv) the discovery analysis when the Günzburg site is excluded and replaced with matched cases from the replication cohort. **A)** The factor matrices demonstrate virtually indistinguishable factor solutions across all analyses that were highly correlated ($r=0.98$, $p<0.001$); **B)** The consensus matrices demonstrate highly similar solutions across all analyses. Note: somatic variables indicated in a blue region (zero valued) across the somatic excluded, Günzburg excluded, and Günzburg replaced analyses.

© 2020 American Medical Association. All rights reserved.

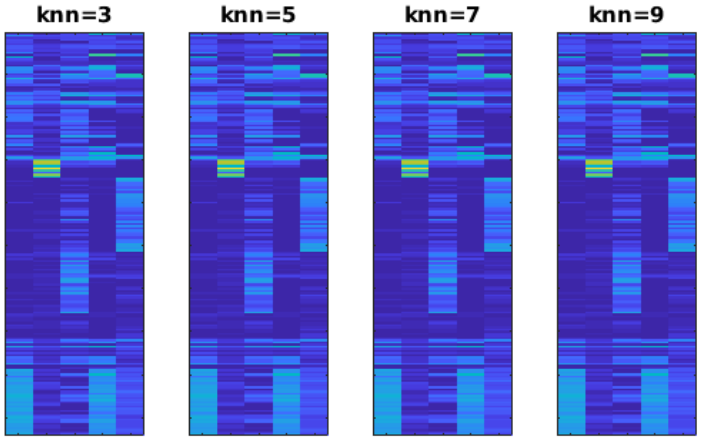
eFigure 10. Repetition of validation analyses after excluding infectious diseases variable



Validation sample results repeated after excluding the infectious diseases variable due to a change in the measurement of disease severity. Results showed minimal changes to the original replication analyses (above eFigures 4-7 and Figure 1 in the main text) across: a) factor loadings when compared to the discovery sample; b) illness course for variables such as the global assessment of functioning; c) polygenic scores for educational attainment; and D) comparisons between polygenic risk scores for schizophrenia (SCZ), bipolar (BP), major depressive disorder (MDD), and educational attainment (Education).

© 2020 American Medical Association. All rights reserved.

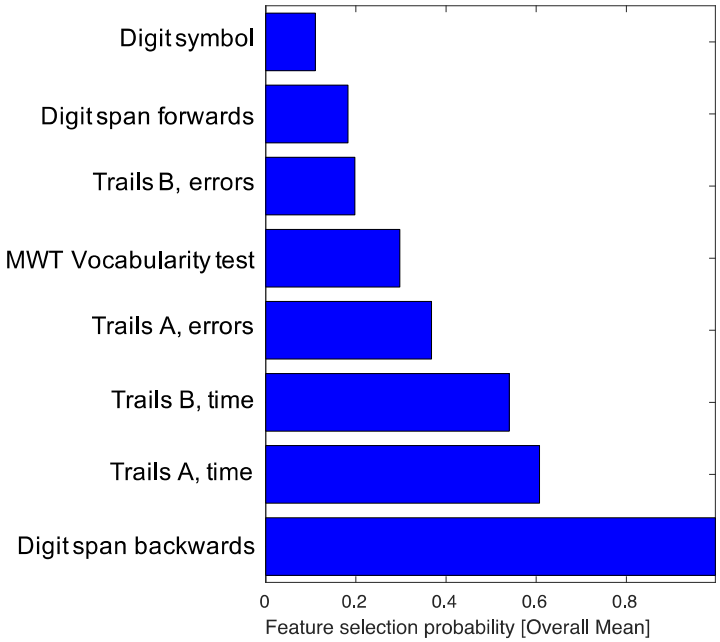
eFigure 11. Factor matrices of clustering solution across different K-nearest neighbors



Factor matrices showing the feature loading across different k nearest neighbours imputation parameters. Results show stability of clustering solution across different imputation parameters.

© 2020 American Medical Association. All rights reserved.

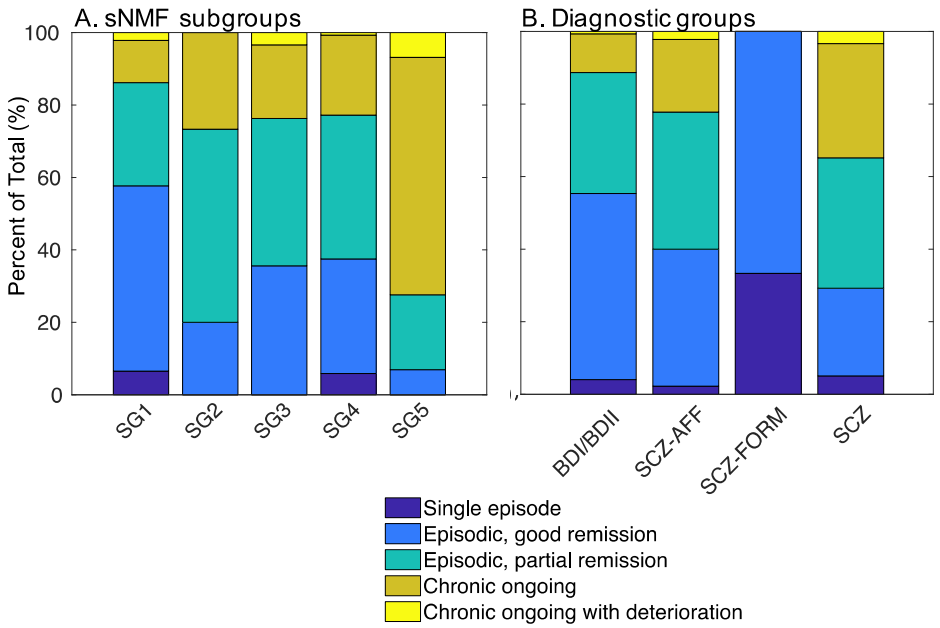
eFigure 12. Feature importance for the supplementary analysis of cognition



Plot depicts the proportion of times that each cognitive feature was selected across cross-validation folds. Results indicate that backwards digit span was most frequently selected.

© 2020 American Medical Association. All rights reserved.

eFigure 13. Assessment of lifetime illness course using the Operational Criteria Checklist for Psychotic Illness and Affective Illness (OPCRIT)



A) the distribution of illness courses across the subgroups shows that the first subgroup largely consists of individuals with an episodic course and good remission, the second group consists of individuals with largely an episodic course and partial remission, and the severe psychosis group consists of individuals with chronic ongoing illness. **B**) a diagnostic group separation is similar for affective psychosis, but does not capture the individuals in the second episodic partial remission group or separate individuals with a chronic ongoing course. Notes: SG, subgroup; BDI, bipolar disorder I; BDII, bipolar disorder II; SCZ-AFF, schizoaffective disorder; SCZ-FORM, schizophreniform disorder and brief psychotic disorder; SCZ, schizophrenia.

© 2020 American Medical Association. All rights reserved.

eBox. Abbreviations used throughout supplementary materials

Abbreviation	Description
BigFive	Big Five Personality Test
CGI	Clinical global impression scale
DSM	Diagnostic and statistical manual of mental disorders, version IV
DST	Digit symbol test
GAF	Global assessment of functioning
IDS-C_{30}	Inventory of depressive symptomatology
PANSS	Positive and negative symptom scale
SCID	Structured clinical interview for DSM-IV
TMT	Trail making task
YMRS	Young mania rating scale
WHOQoL-BREF	World Health Organisation Quality of Life brief version rating scale
WMT	Mehrfachwahl-Wortschatz-Test

© 2020 American Medical Association. All rights reserved.

eTable 1. Unfiltered features originally selected for analyses

Red font indicates where variables were later excluded due to insufficient variance (>95% same value; 10 excluded; see eTable 2) or lack of data (>25% missing; 31 excluded; see eTable 2).

Number	Variable Description
1	Sex (dichotomous; Female=0, Male=1)
2	Age at first interview (yrs; continuous)
3	Season of birth, Autumn (dichotomous)
4	Season of birth, Spring (dichotomous)
5	Season of birth, Summer (dichotomous)
6	Season of birth, Winter (dichotomous)
7	Mother's age at birth (yrs; continuous)
8	Father's age at birth (yrs; continuous)
9	Marital status, Divorced (dichotomous)
10	Marital status, Married (dichotomous)
11	Marital status, Married but living separately (dichotomous)
12	Marital status, Single (dichotomous)
13	Marital status, Widowed (dichotomous)
14	Currently has a partner (dichotomous)
15	Number of biological children (continuous)
16	Number of adoptive children (continuous)
17	Number of step children (continuous)
18	Number of brothers (continuous)
19	Number of sisters (continuous)
20	Living alone (dichotomous)
21	High-school level education (ordinal [0,1,2,3])
22	Educational status scale (ordinal [0,1,2,3,4,5,6])
23	Currently in paid employment (dichotomous)
24	Disability pension due to psychiatric illness (dichotomous)
25	Supported employment (dichotomous)
26	Work absence due to psychiatric reasons for past five years (continuous)
27	Work impaired by psychiatric symptoms (dichotomous)
28	Current psychiatric treatment (ordinal [1,2,3,4])
29	Ever treated as outpatient (ordinal [1,2,3,4])
30	Age at first outpatient treatment (continuous)
31	Ever treated as a daypatient or inpatient (dichotomous)
32	Age at first inpatient treatment (continuous)
33	Duration of illness (continuous)
34	Current first episode of psychiatric illness (dichotomous)
35	Times treated as day- or inpatient (continuous)
36	Times treated as day- or inpatient (ordinal [1,2,3,4])
37	Adverse events under current medication (dichotomous)
38	Psychiatric medication change in the past six months (dichotomous)
39	Did they ever take lithium (dichotomous)
40	If they took lithium, then for how long (ordinal [1,2,3])
41	Family member ever affected by a psychiatric disorder (dichotomous)
42	Self-reported height (continuous [centimeters])
43	Self-reported Weight (continuous [kilograms])
44	Body mass index (continuous)
45	Ever diagnosed with a cardiovascular disease (dichotomous)
46	Ever diagnosed with a metabolic disorder (dichotomous)
47	Ever diagnosed with a thyroid disorder (dichotomous)

© 2020 American Medical Association. All rights reserved.

48	Ever diagnosed with a rheumatological or musculoskeletal disorder (dichotomous)
49	Ever diagnosed with a lung disorder (dichotomous)
50	Ever diagnosed with allergies (dichotomous)
51	Ever diagnosed with a gastrointestinal disorder (dichotomous)
52	Ever diagnosed with a gallbladder or liver disorders (dichotomous)
53	Ever diagnosed with a skin disorder (dichotomous)
54	Ever diagnosed with a kidney or urinary bladder disorder (dichotomous)
55	Ever diagnosed with a neurological or psychiatric disorder except the study diagnosis (dichotomous)
56	Ever diagnosed with a disorder of the eyes or ears (dichotomous)
57	Ever diagnosed with infectious diseases (dichotomous)
58	Ever diagnosed with cancer (dichotomous)
59	Ever diagnosed with any other disorders (dichotomous)
60	Participant is a former smoker (dichotomous)
61	Participant never smoked (dichotomous)
62	Participant currently smokes (dichotomous)
63	Number of cigarettes per year (continuous)
64	Did they ever drink alcohol? (dichotomous)
65	How often did they consume alcoholic beverages during the past 12 months (ordinal [1-7])
66	How many times in past 6mths did they drink five (men)/four (women) alcoholic drinks (ordinal)
67	Lifetime alcohol dependence (dichotomous)
68	Did they ever take illicit drugs? (dichotomous)
69	SCID: Ever fulfilled DSM-IV major depressive disorder criteria? (dichotomous)
70	SCID: Age at first major depressive disorder episode (continuous)
71	SCID: Ever fulfilled DSM-IV criteria for manic episode? (dichotomous)
72	SCID: Age at first manic episode (continuous)
73	SCID: Number of manic episodes (continuous)
74	SCID: Ever fulfilled DSM-IV criteria for a mixed manic/depressive episode? (dichotomous)
75	SCID: Ever fulfilled DSM-IV criteria for hypomanic episode? (dichotomous)
76	SCID: Age at first hypomanic episode (continuous)
77	SCID: Number of hypomanic episodes (continuous)
78	SCID: Ever hallucinate (dichotomous)
79	SCID: Ever have delusions (dichotomous)
80	SCID: Ever had a psychotic episode (dichotomous)
81	SCID: Ever have suicidal ideation (dichotomous)
82	SCID: Severity of past or present suicidal ideation (ordinal [1,2,3,4)
83	SCID: Thoughts about methods of suicide (SCID assessment; ordinal [1,2,3])
84	SCID: Thoughts about leaving a suicide note (SCID assessment; ordinal [1,2,3])
85	SCID: Past suicide attempt (ordinal [1,2,3])
86	SCID: Number of suicide attempts (ordinal [1-6])
87	SCID: Preparation for a suicide attempt (dichotomous)
88	SCID: Preparation for a suicide attempt (ordinal)
89	SCID: Preparation for suicide involved a suicide note (ordinal)
90	PANSS items P1 - P16
123	PANSS: Total score (continuous [30-210])
124	IDS-C_{30} items 1 - 30
152	IDS-C_{30}: Total score (continuous [0-84])
153	YMRS items 1 - 11
164	YMRS total score (continuous [0-60])
165	Clinical Global Impression (CGI) - illness severity scale (ordinal [1,2,3,4,5,6,7])
166	Global Assessment of Functioning (GAF) (continuous [1-100])
167	Good language proficiency for cognitive testing (dichotomous)
168	Native German speaker (dichotomous)
169	Language skills not sufficient for neuropsychological testing (dichotomous)

© 2020 American Medical Association. All rights reserved.

170	Language skills sufficient for neuropsychological testing (dichotomous)
171	TMT Part A, time (continuous [seconds])
172	TMT Part A, errors (continuous [number of errors])
173	TMT Part B, time (continuous [seconds])
174	TMT Part B, errors (continuous [number of errors])
175	Digit span: forward (continuous [number of items])
176	Digit span: backward (continuous [number of items])
177	Digit symbol test (DST) (continuous)
178	Multiple choice vocabulary test (Mehrfachwahl-Wortschatz-Test; MWT; continuous)
179	Religious identification with Christianity
180	Religious identification with Islam
181	Does not identify with any religion
182	How actively do they practice their religious belief? (ordinal)
183	Self-reported medication adherence in past seven days (ordinal [1,2,3,4,5,6])
184	Self-reported medication adherence in past six months (ordinal [1,2,3,4,5,6])
185	WHOQoL-BREF: 1 - 26
211	WHOQoL-BREF global: global WHOQoL-BREF score (continuous)
212	WHOQoL-BREF physical: physical health WHOQoL-BREF score (continuous)
213	WHOQoL-BREF psychological: psychological health WHOQoL-BREF score (continuous)
214	WHOQoL-BREF social: social health WHOQoL-BREF score (continuous)
215	WHOQoL-BREF environmental: environmental health WHOQoL-BREF score (continuous)
216	BigFive items 1 - 10
226	BigFive-Extra
227	BigFive-Neuro
228	BigFive-Openness
229	BigFive-Consc
230	BigFive-Agree

Notes: for abbreviations see Table S1.

© 2020 American Medical Association. All rights reserved.

eTable 2. Variables excluded from clustering analyses

Variables excluded from clustering analyses due to limited variance (left) or missing values (right).

Limited variance (≥95% same value)	**Proportion missing (≥25% missing)**
Marital status, married but living separately	Disability pension due to psychiatric illness (dichotomous)
Marital status, widowed	Supported employment (dichotomous)
Number of adoptive children	Work absence due to psychiatric reasons for past five years (continuous)
Step children, dichotomous	Number of cigarettes per year (continuous)
Daypatient or inpatient treatment	How many times in past 6mths did they drink five (men)/four (women) alcoholic drinks (ordinal)
Ever had alcoholic beverage	SCID: Age at first major depressive disorder episode (continuous)
SCID: Ever fulfilled DSM-IV major depressive disorder criteria? (dichotomous)	SCID: Ever fulfilled DSM-IV criteria for manic episode? (dichotomous)
SCID: ever had a psychotic illness	SCID: Age at first manic episode (continuous) SCID: Number of manic episodes (continuous)
Language skills insufficient for neuropsychological assessment	SCID: Ever fulfilled DSM-IV criteria for a mixed manic/depressive episode? (dichotomous)
Language skills sufficient for neuropsychological assessment	SCID: Ever fulfilled DSM-IV criteria for hypomanic episode? (dichotomous)
	SCID: Age at first hypomanic episode (continuous)
	SCID: Number of hypomanic episodes (continuous)
	Religious identification with Christianity
	Religious identification with Islam
	Does not identify with any religion
	How actively do they practice their religious belief? (ordinal) BigFive questions 1 – 10 (all questions) BigFive-Extraversion sum score
	BigFive-Neuroticism sum score
	BigFive-Openness sum score
	BigFive-Conscientiousness sum score
	BigFive-Agreeableness sum score

© 2020 American Medical Association. All rights reserved.

eTable 3. Comparisons of discovery sample with excluded participants and the validation sample

Comparisons of the filtered discovery sample with: a) the excluded cases; b) the validation sample. False discovery rate corrected and only p-values for significant differences shown.

	Variable	Discovery		Excluded	T / χ^2 (df)	*p*-value	*d* / Phi		Validation	T / χ^2 (df)	*p*-value	*d* / Phi
N		765		126					458			
Demographics												
	Baseline age	42.7(12.9)		42.8(12.8)	-0.07(889)	--	--		42.6(13.2)	0.12(1221)	--	--
	Sex [n(%) male]	424(55.4)		79(62.7)	2.33(4)	--	--		233(50.9)	2.39(4)	--	--
	Education Status	3.5(1.6)		3.1(1.8)	2.60(864)	9.36E-03	0.25		3.7(1.7)	-1.12(1178)	--	--
	Single marital status [n(%)]	424(55.6)		70(56.5)	0.03(4)	--	--		240(54.2)	0.24(4)	--	--
	Paid Employment [n(%)]	287(38.0)		47(38.5)	0.01(4)	--	--		191(42.8)	2.77(4)	--	--
Psychiatric history												
	Current psychiatric treatment [n(%)]	2.8(1.0)		3.4(0.9)	-5.84(882)	7.50E-09	-0.57		3.0(1.0)	-1.98(1203)	--	--
	Psychiatric family history [n(%)]	532(73.7)		70(63.1)	5.42(4)	--	--		330(76.4)	1.05(4)	--	--
	Onset (1st inpatient year)	30.2(11.5)		29.7(11.7)	0.49(847)	--	--		30.3(11.8)	-0.11(1154)	--	--
	Duration of illness	12.5(10.5)		13.1(12.5)	-0.53(846)	--	--		12.4(11.0)	0.16(1148)	--	--
	Ever hallucinate (SCID) [n(%)]	433(58.4)		76(62.8)	0.82(4)	--	--		175(39.6)	39.35(4)	3.55E-10	0.18
	Ever delusional (SCID) [n(%)]	597(93.4)		97(99.0)	4.77(4)	--	--		263(59.6)	183.66(4)	7.69E-42	0.39
	Ever take illicit drugs [n(%)]	318(43.1)		43(37.1)	1.49(4)	--	--		190(45.0)	0.41(4)	--	--
	Ever smoked cigarettes [n(%)]	416(54.4)		80(66.7)	6.36(4)	1.17E-02	0.08		241(54.5)	<0.00(4)	--	--
Diagnoses												
	Major depressive disorder [n(%)]	2(0.3)		3(2.4)	8.71(4)	3.17E-03	0.1		89(19.4)	152.88(4)	4.07E-35	0.35
	Bipolar II disorder [n(%)]	60(7.8)		8(6.3)	0.34(4)	--	--		50(10.9)	3.31(4)	--	--
	Bipolar I disorder [n(%)]	256(33.5)		38(30.2)	0.53(4)	--	--		171(37.3)	1.89(4)	--	--
	Schizoaffective disorder [n(%)]	73(9.5)		10(7.9)	0.33(4)	--	--		28(6.1)	4.45(4)	--	--
	Brief psychotic disorder [n(%)]	6(0.8)		0(0.0)	--	--	--		0(0.0)	--	--	--
	Schizophreniform disorder [n(%)]	10(1.3)		1(0.8)	0.23(4)	--	--		2(0.4)	2.23(4)	--	--
	Schizophrenia	358(46.8)		66(52.4)	1.35(4)	--	--		118(25.8)	53.31(4)	2.84E-13	0.35
Symptoms												
	PANSS Positive	11.8(5.3)		15.4(6.8)	-6.59(871)	7.60E-11	-0.66		11.1(5.4)	2.36(1193)	--	--
	PANSS Negative	13.2(6.2)		16.8(7.2)	-5.60(860)	2.84E-08	-0.56		13.2(6.1)	-0.00(1179)	--	--
	PANSS Total	51.7(18.4)		65.0(22.2)	-6.91(831)	9.69E-12	-0.7		52.0(17.9)	-0.31(1135)	--	--
	IDSC Sum	12.6(10.3)		12.7(12.0)	-0.08(763)	--	--		16.3(13.6)	-4.92(1062)	9.98E-07	-0.31
	YMRS Sum	3.0(5.2)		3.0(5.5)	-0.06(849)	--	--		3.3(5.2)	-0.85(1167)	--	--
Functioning and Intelligence												
	CGI	4.0(1.0)		4.6(1.0)	-5.47(870)	5.73E-08	-0.54		4.3(1.1)	-3.84(1196)	1.29E-04	-0.23
	GAF	57.5(13.9)		51.0(13.3)	4.74(873)	2.44E-06	0.47		54.9(13.6)	3.09(1195)	2.03E-03	0.19
	WHO-QoL Global	12.7(3.7)		11.8(3.3)	0.83(732)	--	--		12.2(3.7)	1.97(1049)	--	--
	Verbal IQ (MWTB)	27.9(5.1)		27.7(5.7)	0.24(670)	--	--		28.4(4.9)	-1.59(981)	--	--
Medications												

© 2020 American Medical Association. All rights reserved.

	Antidepressants [n(%)]	282(36.9)		40(31.7)	1.23(4)	--	--		210(45.9)	9.63(4)	1.92E-03	0.09
	Antipsychotics [n(%)]	665(86.9)		109(86.5)	0.02(4)	--	--		320(69.9)	53.19(4)	3.02E-13	0.21
	Mood stabilizers [n(%)]	276(36.1)		42(33.3)	0.36(4)	--	--		182(39.7)	1.64(4)	--	--
	Tranquilizers [n(%)]	152(19.9)		29(23.0)	0.66(4)	--	--		81(17.7)	0.89(4)	--	--

Notes: for all variable acronyms see the eBox. T, t-test T score; χ^2, chi-squared value; df, degrees of freedom; p-value, p-value associated with T or χ^2 (only false-discovery rate values shown); *d*, Cohen's *d* effect size measure associated with T tests; Phi, phi coefficient measure of effect size for χ^2 tests. T used for all variables displayed with the mean and standard deviation [mean(SD)] and chi-squared tests used for all nominal variables displayed with the total number and percentage of total [n(%)].

© 2020 American Medical Association. All rights reserved.

eTable 4. Top 10 features and mean factor weights

Top ten features and mean factor weights (SD) derived from 1000 iterations of the NMF algorithm. Features displayed are those that reliably contributed to the average factor weight score at a 95% confidence interval derived from permutations analyses (eMethods). Variables sorted from highest to lowest mean weight. For all acronyms see the eBox.

	Features	Mean	SD
Factor 1	Currently has a partner (dichotomous)	0.94	0.07
	WHOQoL-BREF 13: How available is the information that you need in your day-to-day life? (ordinal [1,2,3,4,5])	0.85	0.02
	WHOQoL-BREF 23: How satisfied are you with the conditions of your living place? (ordinal [1,2,3,4,5])	0.83	0.02
	WHOQoL-BREF 14: To what extent do you have the opportunity for leisure activities (ordinal [1,2,3,4,5])	0.81	0.02
	Ever treated as outpatient (ordinal [1,2,3,4])'	0.80	0.02
	WHOQoL-BREF 25: How satisfied are you with your mode of transportation? (ordinal [1,2,3,4,5])	0.80	0.02
	WHOQoL-BREF 6: How well are you able to concentrate? (ordinal [1,2,3,4,5])	0.79	0.03
	WHOQoL-BREF 12: Have you enough money to meet your needs? (ordinal [1,2,3,4,5])	0.79	0.02
	WHOQoL-BREF environmental: environmental health WHOQoL-BREF score (continuous)	0.79	0.02
	WHOQoL-BREF 20: How satisfied are you with your personal relationships? (ordinal [1,2,3,4,5])	0.79	0.02
Factor 2	Past suicide attempt (ordinal [1,2,3])	1.81	0.13
	Severity of past or present suicidal ideation (SCID assessment; ordinal [1,2,3,4)	1.45	0.07
	Thoughts about methods of suicide (SCID assessment; ordinal [1,2,3])	1.40	0.06
	Preparation for a suicide attempt (SCID assessment; ordinal)	1.29	0.08
	Ever have suicidal ideation (SCID assessment; dichotomous)	1.22	0.05
	Number of suicide attempts (ordinal [1-6])	1.09	0.08
	Preparation for suicide involved a suicide note (ordinal)	0.98	0.06
	Thoughts about leaving a suicide note (SCID assessment; ordinal [1,2,3])	0.91	0.05
	Family member ever affected by a psychiatric disorder (dichotomous)	0.49	0.07
	Ever treated as outpatient (ordinal [1,2,3,4])'	0.44	0.04
Factor 3	Psychiatric medication change in the past six months (dichotomous)	0.83	0.05
	Work impaired by psychiatric symptoms (dichotomous)	0.77	0.04
	IDS-C_{30}: Item 5 Mood (sad) (ordinal [0,1,2,3])	0.75	0.04
	Family member ever affected by a psychiatric disorder (dichotomous)	0.73	0.04
	IDS-C_{30}: Item 16 Outlook (self) (ordinal [0,1,2,3])	0.73	0.04
	IDS-C_{30}: Total score (continuous [0-84])	0.72	0.03

© 2020 American Medical Association. All rights reserved.

	Ever diagnosed with infectious diseases (dichotomous)	0.70	0.07
	Ever have delusions (SCID assessment; dichotomous)	0.69	0.03
	IDS-C_{30}: Item 15 Concentration/decision making (ordinal [0,1,2,3])	0.68	0.03
	Adverse events under current medication (dichotomous)	0.68	0.05
Factor 4	Marital status, Single (dichotomous)	1.15	0.06
	Did they ever take illicit drugs? (dichotomous)	0.95	0.06
	WHOQoL-BREF 3: Physical pain preventing activities (ordinal [1,2,3,4,5])	0.94	0.02
	Participant currently smokes (dichotomous)	0.93	0.05
	Sex (dichotomous; Female=0, Male=1)	0.90	0.05
	Native German speaker (dichotomous)	0.90	0.02
	Ever have delusions (SCID assessment; dichotomous)	0.88	0.01
	High-school level education (ordinal [0,1,2,3])	0.84	0.02
	WHOQoL-BREF 15: How well are you able to get around? (ordinal [1,2,3,4,5])	0.81	0.02
	WHOQoL-BREF 13: How available is the information that you need in your day-to-day life? (ordinal [1,2,3,4,5])	0.76	0.02
Factor 5	Work impaired by psychiatric symptoms (dichotomous)	1.02	0.03
	Current psychiatric treatment (ordinal [1,2,3,4])	1.01	0.02
	Ever hallucinate (SCID assessment; dichotomous)	0.89	0.04
	PANSS: P1 Delusions (ordinal)	0.82	0.03
	PANSS: Total score (continuous [30-210])	0.79	0.02
	PANSS Positive sum score (continuous [7-49])	0.77	0.02
	PANSS: P2 Conceptual disorganization (ordinal)	0.71	0.02
	PANSS Negative sum score (continuous [7-49])	0.69	0.02
	PANSS: N1 Blunted affect (ordinal [1,2,3,4,5,6,7])	0.65	0.02
	PANSS General Psychopathology sum score (continuous [16-112])	0.64	0.02

Notes: PANSS, positive and negative symptom questionnaire; IDS-C_{30} inventory of depressive symptomatology;

© 2020 American Medical Association. All rights reserved.

eTable 5. Differences between the sparse nonnegative matrix factorization–derived subgroups across six clinical domains

	Variable	Full sample	Group 1	Group 2	Group 3	Group 4	Group 5	F / χ^2 (df)	*p*-value	Eta^2/Phi
N		765	252	44	131	252	86			
Demographics										
	Baseline age [mean(SD)]	42.7(12.9)	50.3(11.6)	45.5(11.0)	41.8(11.8)	36.6(11.2)	38.5(11.9)	49.09(4,760)	9.41E-37	0.21
	Sex [n(%) male]	424(55.4)	105(41.7)	13(29.5)	62(47.3)	192(76.2)	52(60.5)	79.58(10)	2.14E-16	0.32
	Education Status [mean(SD)]	3.5(1.6)	3.9(1.6)	3.7(1.3)	3.4(1.5)	3.5(1.5)	2.6(1.5)	12.88(4,739)	3.87E-10	0.07
	Single marital status [n(%)]	424(55.6)	53(21.1)	20(45.5)	56(42.7)	231(92.4)	64(74.4)	281.05(10)	1.33E-59	0.61
	Paid Employment [n(%)]	287(38.0)	122(49.0)	14(31.8)	34(26.2)	96(38.7)	21(24.7)	27.68(10)	1.45E-05	0.19
Psychiatric treatment and history										
	Current psychiatric treatment [n(%)]	2.8(1.0)	2.5(0.9)	2.9(1.0)	3.1(1.0)	2.7(1.0)	3.8(0.6)	35.19(4,756)	5.62E-27	0.16
	Psychiatric family history [n(%)]	532(73.7)	184(77.0)	39(92.9)	104(81.9)	171(71.8)	34(44.7)	46.97(10)	1.54E-09	0.25
	Onset (1st inpatient year) [mean(SD)]	30.2(11.5)	35.6(13.0)	28.3(8.6)	31.4(10.9)	25.7(8.8)	27.2(9.0)	28.50(4,728)	5.07E-22	0.14
	Duration of illness [mean(SD)]	12.5(10.5)	14.9(12.1)	16.6(11.1)	10.2(9.7)	11.2(9.0)	11.1(9.0)	7.72(4,727)	4.28E-06	0.04
	Ever hallucinate (SCID) [n(%)]	433(58.4)	106(44.2)	25(56.8)	70(55.6)	158(64.2)	74(87.1)	52.67(10)	1.00E-10	0.26
	Ever delusional (SCID) [n(%)]	597(93.4)	166(87.4)	37(97.4)	90(92.8)	220(96.5)	84(97.7)	18.40(10)	1.03E-03	0.16
	Ever take illicit drugs [n(%)]	318(43.1)	58(23.7)	10(23.8)	63(49.2)	169(70.4)	18(21.7)	134.58(10)	4.08E-28	0.42
Current DSM-IV diagnoses										
	Major depressive disorder [n(%)]	2(0.3)	1(0.4)	0(0.0)	1(0.8)	0(0.0)	0(0.0)	2.44(10)	--	--
	Bipolar II disorder [n(%)]	60(7.8)	29(11.5)	5(11.4)	16(12.2)	10(4.0)	0(0.0)	21.45(10)	2.57E-04	0.17
	Bipolar I disorder [n(%)]	256(33.5)	125(49.6)	16(36.4)	45(34.4)	69(27.4)	1(1.2)	74.18(10)	2.97E-15	0.31
	Schizoaffective disorder [n(%)]	73(9.5)	19(7.5)	9(20.5)	13(9.9)	24(9.5)	8(9.3)	7.27(10)	--	--
	Brief psychotic disorder [n(%)]	6(0.8)	3(1.2)	0(0.0)	0(0.0)	3(1.2)	0(0.0)	3.13(10)	--	--
	Schizophreniform disorder [n(%)]	10(1.3)	3(1.2)	0(0.0)	3(2.3)	3(1.2)	1(1.2)	1.63(10)	--	--
	Schizophrenia	358(46.8)	72(28.6)	14(31.8)	53(40.5)	143(56.7)	76(88.4)	109.42(10)	9.66E-23	0.38
Current psychiatric symptoms										
	PANSS Positive [mean(SD)]	11.8(5.3)	9.7(3.6)	11.3(4.2)	11.9(4.7)	11.0(4.3)	20.4(4.6)	109.28(4,753)	1.98E-73	0.37
	PANSS Negative [mean(SD)]	13.2(6.2)	10.5(4.3)	13.0(6.2)	14.9(5.6)	11.7(5.0)	22.7(5.5)	105.41(4,743)	4.09E-71	0.36
	PANSS Total [mean(SD)]	51.7(18.4)	42.4(11.0)	51.8(17.3)	58.7(14.6)	45.3(12.5)	84.3(12.8)	195.84(4,717)	2.04E-113	0.52

© 2020 American Medical Association. All rights reserved.

	IDSC Sum [mean(SD)]	12.6(10.3)	8.6(6.4)	15.6(7.2)	28.0(8.9)	9.1(6.6)	9.9(10.1)	148.70(4,660)	1.31E-90	0.47
	YMRS Sum [mean(SD)]	3.0(5.2)	3.3(5.9)	3.6(5.3)	3.2(4.4)	2.4(4.2)	3.4(6.2)	1.49(4,737)	--	--
Current functioning										
	CGI [mean(SD)]	4.0(1.0)	3.6(1.0)	4.1(1.1)	4.5(0.9)	3.9(1.1)	4.8(0.7)	30.15(4,750)	2.73E-23	0.14
	GAF [mean(SD)]	57.5(13.9)	62.5(12.6)	56.1(11.0)	51.1(12.1)	59.7(13.6)	46.7(12.5)	35.17(4,753)	5.89E-27	0.16
	WHO-QoL Global [mean(SD)]	12.7(3.7)	14.5(3.0)	11.2(3.4)	8.9(3.1)	13.5(3.2)	11.7(3.6)	70.49(4,716)	2.43E-50	0.28
	Verbal IQ (MWTB) [mean(SD)]	27.9(5.1)	29.1(4.7)	27.9(6.5)	27.5(4.5)	27.8(4.9)	24.1(5.8)	10.28(4,631)	4.58E-08	0.06
Psychiatric medications (binary)										
	Antidepressants [n(%)]	282(36.9)	105(41.7)	13(29.5)	77(58.8)	73(29.0)	14(16.3)	52.95(10)	8.73E-11	0.26
	Antipsychotics [n(%)]	665(86.9)	198(78.6)	43(97.7)	117(89.3)	223(88.5)	84(97.7)	29.94(10)	5.03E-06	0.2
	Mood stabilizers [n(%)]	276(36.1)	122(48.4)	23(52.3)	45(34.4)	75(29.8)	11(12.8)	46.38(10)	2.05E-09	0.25
	Tranquilizers [n(%)]	152(19.9)	45(17.9)	10(22.7)	38(29.0)	31(12.3)	28(32.6)	25.50(10)	3.99E-05	0.18
	Clozapine [n(%)]	58(7.6)	11(4.4)	3(6.8)	9(6.9)	22(8.7)	13(15.1)	11.29(10)	2.34E-02	0.12

Notes: for all variable acronyms see the eBox. F, ANOVA F-score; χ^2, chi-squared value; df, degrees of freedom; p-value, p-value associated with ANOVA or χ^2 (only false-discovery rate values shown); Eta^2, partial eta-squared effect size measure associated with ANOVA tests; Phi, phi coefficient measure of effect size for χ^2 tests. ANOVA used for all variables displayed with the mean and standard deviation [mean(SD)] and chi-squared tests used for all nominal variables displayed with the total number and percentage of total [n(%)].

© 2020 American Medical Association. All rights reserved.

eTable 6. Differences between sparse nonnegative matrix factorization–derived subgroups (schizophrenia only)

Differences between sNMF-derived subgroups across six clinical domains only in individuals with schizophrenia.

	Variable	Full sample	Group 1	Group 2	Group 3	Group 4	Group 5	F / χ^2 (df)	*p*-value	Eta^2/Phi
N		765	72	14	53	143	76			
Demographics										
	Baseline age	42.7 (12.9)	48.3 (12.0)	46.4 (12.5)	38.7 (11.4)	37.2 (11.3)	36.9 (11.2)	14.09 (4,353)	1.12E-10	0.14
	Sex [n (%) male]	424 (55.4)	26 (36.1)	4 (28.6)	32 (60.4)	118 (82.5)	48 (63.2)	53.32 (10)	7.29E-11	0.39
	Education Status	3.5(1.6)	3.9(1.7)	4.1(1.2)	3.2(1.5)	3.2(1.4)	2.5(1.3)	10.10(4,346)	9.71E-08	0.1
	Single marital status [n(%)]	424(55.6)	18(25.0)	8(57.1)	29(54.7)	135(95.1)	59(77.6)	120.45(10)	4.29E-25	0.58
	Paid Employment [n(%)]	287(38.0)	35(50.0)	6(42.9)	7(13.5)	52(36.9)	17(22.7)	23.23(10)	1.14E-04	0.25
Psychiatric treatment and history										
	Current psychiatric treatment [n(%)]	2.8(1.0)	2.5(0.9)	3.1(1.0)	3.3(1.0)	2.8(1.0)	3.8(0.5)	25.90(4,353)	7.77E-19	0.23
	Psychiatric family history [n(%)]	532(73.7)	47(70.1)	13(92.9)	42(82.4)	97(70.8)	29(43.9)	27.03(10)	1.96E-05	0.27
	Onset (1st inpatient year)	30.2(11.5)	31.9(11.0)	30.9(9.4)	27.0(9.7)	25.2(8.4)	26.5(8.8)	6.67(4,348)	3.51E-05	0.07
	Duration of illness	12.5(10.5)	16.3(11.4)	15.5(13.8)	11.9(9.4)	11.9(9.2)	10.3(8.8)	4.08(4,348)	3.03E-03	0.04
	Ever hallucinate (SCID) [n(%)]	433(58.4)	56(77.8)	12(85.7)	45(84.9)	111(77.6)	65(86.7)	3.83(10)	--	--
	Ever delusional (SCID) [n(%)]	597(93.4)	68(95.8)	13(92.9)	53(100.0)	140(97.9)	74(97.4)	3.53(10)	--	--
	Ever take illicit drugs [n(%)]	318(43.1)	15(21.7)	0(0.0)	34(65.4)	98(70.5)	16(21.6)	85.31(10)	1.31E-17	0.49
Current DSM-IV diagnoses										
	Major depressive disorder [n(%)]	2(0.3)	--	--	--	--	--	--	--	--
	Bipolar II disorder [n(%)]	60(7.8)	--	--	--	--	--	--	--	--
	Bipolar I disorder [n(%)]	256(33.5)	--	--	--	--	--	--	--	--
	Schizoaffective disorder [n(%)]	73(9.5)	--	--	--	--	--	--	--	--
	Brief psychotic disorder [n(%)]	6(0.8)	--	--	--	--	--	--	--	--
	Schizophreniform disorder [n(%)]	10(1.3)	--	--	--	--	--	--	--	--
	Schizophrenia	358(46.8)	72(100.0)	14(100.0)	53(100.0)	143(100.0)	76(100.0)	--	--	--
Current psychiatric symptoms										
	PANSS Positive	11.8(5.3)	10.5(3.9)	13.0(5.3)	14.3(5.2)	12.2(4.5)	20.5(4.3)	55.73(4,350)	2.42E-36	0.39

© 2020 American Medical Association. All rights reserved.

	PANSS Negative	13.2(6.2)	12.3(5.2)	15.3(6.9)	17.9(5.7)	12.9(5.4)	22.8(5.3)	50.65(4,344)	1.67E-33	0.37
	PANSS Total	51.7(18.4)	46.4(13.3)	59.0(21.6)	67.6(16.2)	48.9(12.9)	84.3(12.7)	99.73(4,332)	1.17E-55	0.55
	IDSC Sum	12.6(10.3)	8.5(6.8)	15.2(8.0)	29.5(8.8)	9.6(6.5)	9.6(10.0)	63.87(4,311)	2.27E-39	0.45
	YMRS Sum	3.0(5.2)	2.0(3.5)	1.6(2.9)	2.5(2.8)	1.8(3.2)	2.8(5.4)	1.03(4,344)	--	--
Current functioning										
	CGI	4.0(1.0)	3.7(0.9)	3.9(1.2)	4.7(0.9)	4.1(1.0)	4.7(0.7)	15.38(4,348)	1.35E-11	0.15
	GAF	57.5(13.9)	60.1(13.0)	55.9(10.6)	45.7(12.2)	56.1(12.6)	47.1(12.7)	16.40(4,349)	2.52E-12	0.16
	WHO-QoL Global	12.7(3.7)	14.6(2.8)	10.6(3.6)	9.3(3.5)	13.5(3.2)	11.7(3.6)	24.36(4,338)	9.94E-18	0.22
	Verbal IQ (MWTB)	27.9(5.1)	29.2(5.2)	29.1(3.2)	26.4(5.3)	27.2(4.9)	24.0(5.9)	7.08(4,284)	1.92E-05	0.09
Psychiatric medications (binary)										
	Antidepressants [n(%)]	282(36.9)	16(22.2)	3(21.4)	25(47.2)	34(23.8)	12(15.8)	17.77(10)	1.37E-03	0.22
	Antipsychotics [n(%)]	665(86.9)	70(97.2)	14(100.0)	49(92.5)	138(96.5)	74(97.4)	3.22(10)	--	--
	Mood stabilizers [n(%)]	276(36.1)	6(8.3)	1(7.1)	2(3.8)	9(6.3)	7(9.2)	1.73(10)	--	--
	Tranquilizers [n(%)]	152(19.9)	17(23.6)	4(28.6)	21(39.6)	22(15.4)	25(32.9)	15.85(10)	3.23E-03	0.21
	Clozapine [n(%)]	58(7.6)	6(8.3)	2(14.3)	9(17.0)	19(13.3)	12(15.8)	2.56(10)	--	--

Notes: for all variable acronyms see the eBox. F, ANOVA F-score; χ^2, chi-squared value; df, degrees of freedom; p-value, p-value associated with ANOVA or χ^2 (only false-discovery rate values shown); Eta^2, partial eta-squared effect size measure associated with ANOVA tests; Phi, phi coefficient measure of effect size for χ^2 tests. ANOVA used for all variables displayed with the mean and standard deviation [mean(SD)] and chi-squared tests used for all nominal variables displayed with the total number and percentage of total [n(%)].

© 2020 American Medical Association. All rights reserved.

eTable 7. Proportions (No. [%]) of individuals in each group across PsyCourse sites

Site	Full Group	Group 1	Group 2	Group 3	Group 4	Group 5	X^2(df)	*p*-value	Phi
Augsburg	36(4.7)	14(5.6)	2(4.5)	1(0.8)	18(7.1)	1(1.2)	10.69(10)	3.02E-02	0.12
Bad Zwischenahn	55(7.2)	18(7.1)	4(9.1)	13(9.9)	15(6.0)	5(5.8)	2.53(10)	--	0.06
Bochum	78(10.2)	14(5.6)	5(11.4)	27(20.6)	23(9.1)	9(10.5)	21.83(10)	2.17E-04	0.17
Bremen Ost	26(3.4)	5(2.0)	3(6.8)	2(1.5)	15(6.0)	1(1.2)	10.82(10)	2.87E-02	0.12
Eschwege	7(0.9)	7(2.8)	0(0.0)	0(0.0)	0(0.0)	0(0.0)	14.38(10)	6.17E-03	0.14
Göttingen	11(1.4)	9(3.6)	0(0.0)	1(0.8)	1(0.4)	0(0.0)	12.34(10)	1.50E-02	0.13
Graz	101(13.2)	44(17.5)	7(15.9)	24(18.3)	26(10.3)	0(0.0)	22.17(10)	1.85E-04	0.17
Günzburg	67(8.8)	3(1.2)	5(11.4)	0(0.0)	8(3.2)	51(59.3)	315.78(10)	4.28E-67	0.64
Hildesheim	17(2.2)	7(2.8)	0(0.0)	3(2.3)	5(2.0)	2(2.3)	1.43(10)	--	0.04
LMU München	82(10.7)	15(6.0)	5(11.4)	13(9.9)	40(15.9)	9(10.5)	13.09(10)	1.08E-02	0.13
Liebenburg	9(1.2)	5(2.0)	0(0.0)	2(1.5)	2(0.8)	0(0.0)	3.42(10)	--	0.07
Lüneburg	33(4.3)	13(5.2)	1(2.3)	2(1.5)	17(6.7)	0(0.0)	10.83(10)	2.85E-02	0.12
Münster	3(0.4)	2(0.8)	0(0.0)	1(0.8)	0(0.0)	0(0.0)	3.01(10)	--	0.06
Osnabruck	38(5.0)	20(7.9)	0(0.0)	6(4.6)	11(4.4)	1(1.2)	9.88(10)	4.25E-02	0.11
Rotenburg/Wamme	28(3.7)	14(5.6)	1(2.3)	2(1.5)	11(4.4)	0(0.0)	8.12(10)	--	0.1
Tiefenbrunn	4(0.5)	0(0.0)	0(0.0)	2(1.5)	2(0.8)	0(0.0)	4.90(10)	--	0.08
UMG Göttingen	157(20.5)	59(23.4)	9(20.5)	28(21.4)	55(21.8)	6(7.0)	11.29(10)	2.35E-02	0.12
Wilhelmshaven	13(1.7)	3(1.2)	2(4.5)	4(3.1)	3(1.2)	1(1.2)	4.50(10)	--	0.08

Notes: χ^2, chi-squared tests; p-value, p-value associated with the chi-squared test; Phi, phi-coefficient effect size measurement for the chi-squared analysis.

© 2020 American Medical Association. All rights reserved.

eTable 8. Site, rater, and site × rater analysis of variance analyses

Site, rater, and site x rater ANOVA analyses of the NMF coefficient matrices

	Term	Mean Sq	df	F	p-value	sig
Factor 1						
	Site	0.112	13, 264	2.59	0.002	**
	Rater	0.023	8, 264	0.53	--	
	Site x rater	0.008	8, 264	0.18	--	
Factor 2						
	Site	0.014	13, 264	0.66	--	
	Rater	0.003	8, 264	0.15	--	
	Site x rater	0.011	8, 264	0.51	--	
Factor 3						
	Site	0.069	13, 264	2.18	0.011	**
	Rater	0.053	8, 264	1.67	--	
	Site x rater	0.058	8, 264	1.83	--	
Factor 4						
	Site	0.093	13, 264	2.23	0.009	**
	Rater	0.018	8, 264	0.02	--	
	Site x rater	0.036	8, 264	0.87	--	
Factor 5						
	Site	0.015	13, 264	1.05	--	
	Rater	0.017	8, 264	0.02	--	
	Site x rater	0.009	8, 264	0.59	--	

Significance codes: <0.001, '***'; <0.01, '**'; <0.05, '*' ; <0.1, '.'

© 2020 American Medical Association. All rights reserved.

eTable 9. Mixed-model analysis of illness course

Degrees of freedom approximated using Satterthwaite method. Additional parametric permutation-based Likelihood Ratio Tests (LRT) conducted (200 permutations) to determine significance of adding the $Time^2$ x Group term.

	Term	Mean Sq	df	F	P-value	Sig	LRT χ^2(df)	LRT p-value
PANSS								
	Group	10453.80	4,790.3	123.40	<2.2e-16	***	--	--
	Time (linear)	5727.44	1,1413.3	67.61	4.46e-16	***	--	--
	Time (quadratic)	2123.74	1,1307.8	25.07	6.28e-07	***	--	--
	Time x Group	1349.17	4,1390.9	15.93	9.37e-13	***	--	--
	$Time^2$ x Group	654.91	4,1301.21	7.73	3.72e-06	***	44.00(5)	0.005
IDS-C_{30}								
	Group	3154.585	4,819.9	73.554	<2.2e-16	***	--	--
	Time (linear)	286.896	1,1367.6	6.689	0.009	**	--	--
	Time (quadratic)	36.706	1,1282.1	0.856	--		--	--
	Time x Group	757.337	4,1354.8	17.658	3.87e-14	***	--	--
	$Time^2$ x Group	197.177	4,1283.0	4.597	0.001	**	19.80(5)	0.009
YMRS								
	Group	19.714	4,716.5	1.550	--		--	--
	Time (linear)	25.758	1,1461.0	2.026	--		--	--
	Time (quadratic)	49.909	1,1325.9	3.925	0.048	*	--	--
	Time x Group	11.428	4,1434.0	0.899	--		--	--
	$Time^2$ x Group	33.605	4,1322.1	2.643	0.032	*	14.59(5)	0.034
GAF								
	Group	3228.516	4,830.9	39.122	<2.2e-16	***	--	--
	Time (linear)	738.710	1,1480.6	8.951	0.003	**	--	--
	Time (quadratic)	3516.076	1,1382.0	42.606	9.38e-11	***	--	--
	Time x Group	359.576	4,1461.4	4.357	0.002	**	--	--

© 2020 American Medical Association. All rights reserved.

	Term	Mean Sq	df	F	P-value	Sig	LRT χ^2(df)	LRT p-value
	$Time^2$ x Group	467.227	4,1377.1	5.662	0.0001	***	63.40(5)	0.005
	WHOQoL-BREF							
	Group	246.489	4,820.6	46.956	< 2.2e-16	***	--	--
	Time (linear)	58.702	1,1432.1	11.183	0.0008	***	--	--
	Time (quadratic)	32.495	1,1326.3	6.190	0.013	*	--	--
	Time x Group	30.450	4,1416.1	5.801	0.0001	***	--	--
	$Time^2$ x Group	19.830	4,1319.9	3.778	0.005	**	18.61(5)	0.019

Notes: PANSS, positive and negative symptom total score; IDSC, inventory of depressive symptomatology total score; YMRS, young mania rating scale total score; GAF, global assessment of functioning; WHOQoL, WHO Quality of Life Questionnaire global score; Mean Sq, mean square error; df, degrees of freedom; F, F statistic; LRT, likelihood ratio test. Significance codes: <0.001, '***'; <0.01, '**'; <0.05, '*' ; <0.1, '.'.

© 2020 American Medical Association. All rights reserved.

eTable 10. Post hoc analysis of mixed-model quadratic trends

Post-hoc analysis of mixed model quadratic trends corrected for multiple comparisons using the false-discovery rate (FDR).

	Contrast	Estimate	SE	df	T	p (FDR)	Sig.
PANSS							
	1 - 2	-119.089	134.010	1323.849	-0.889	--	
	1 - 3	-305.657	84.091	1313.539	-3.635	< 0.001	***
	1 - 4	46.854	65.545	1307.083	0.715	--	
	1 - 5	-387.905	105.978	1360.614	-3.660	< 0.001	***
	2 - 3	-186.568	144.364	1326.752	-1.292	--	
	2 - 4	165.943	134.410	1327.243	1.235	--	
	2 - 5	-268.816	158.117	1345.689	-1.700	--	
	3 - 4	352.511	84.727	1322.215	4.161	< 0.001	***
	3 - 5	-82.248	118.802	1357.399	-0.692	--	
	4 - 5	-434.759	106.483	1365.725	-4.083	< 0.001	***
IDS-C_{30}							
	1 - 2	86.287	98.818	1234.250	0.873	--	
	1 - 3	-231.076	62.741	1267.586	-3.683	0.001	**
	1 - 4	8.490	48.579	1238.267	0.175	--	
	1 - 5	-17.482	78.301	1295.529	-0.223	--	
	2 - 3	-317.363	106.359	1244.653	-2.984	0.010	**
	2 - 4	-77.797	98.670	1233.878	-0.788	--	
	2 - 5	-103.769	116.219	1260.973	-0.893	--	
	3 - 4	239.566	62.508	1266.901	3.833	0.001	**
	3 - 5	213.594	87.626	1298.630	2.438	0.037	*
	4 - 5	-25.972	78.114	1295.222	-0.332	--	
GAF							
	1 - 2	-12.413	128.339	1368.785	-0.097	--	
	1 - 3	154.374	81.327	1374.208	1.898	--	
	1 - 4	-77.868	63.047	1358.801	-1.235	--	
	1 - 5	367.096	103.420	1425.815	3.550	0.002	**
	2 - 3	166.787	138.471	1373.917	1.204	--	
	2 - 4	-65.456	128.589	1370.160	-0.509	--	
	2 - 5	379.508	152.498	1397.615	2.489	0.032	*
	3 - 4	-232.243	81.720	1377.568	-2.842	0.015	*

© 2020 American Medical Association. All rights reserved.

	Contrast	Estimate	SE	df	T	p (FDR)	Sig.
	3 - 5	212.721	115.754	1421.680	1.838	--	
	4 - 5	444.964	103.729	1427.619	4.290	< 0.001	***
WHOQoL-BREF							
	1 - 2	50.518	32.876	1289.915	1.537	--	
	1 - 3	75.767	20.956	1309.731	3.616	0.003	**
	1 - 4	9.349	16.216	1293.117	0.577	--	
	1 - 5	6.470	28.431	1384.855	0.228	--	
	2 - 3	25.248	35.567	1298.325	0.710	--	
	2 - 4	-41.169	32.997	1292.481	-1.248	--	
	2 - 5	-44.048	40.425	1337.796	-1.090	--	
	3 - 4	-66.417	21.145	1315.672	-3.141	0.009	**
	3 - 5	-69.296	31.504	1378.026	-2.200	--	
	4 - 5	-2.879	28.571	1387.426	-0.101	--	

Notes: PANSS, positive and negative symptom total score; IDSC, inventory of depressive symptomatology total score; YMRS, young mania rating scale total score; GAF, global assessment of functioning; WHOQoL, WHO Quality of Life Questionnaire global score. Significance codes: <0.001, '***'; <0.01, '**'; <0.05, '*' ; <0.1, '.'.

© 2020 American Medical Association. All rights reserved.

eTable 11. Multigroup classification performance in the discovery set

Multigroup classification performance in the discovery set using the top-ten features from the sNMF outlined in eTable 4. Accuracies provided to demonstrate the ability for the top-ten variables to separate the groups.

SG	TP	FP	TN	FN	Sp(%)	Se(%)	BA(%)	AUC	PPV	NPV	DOR
1	192	487	26	60	94.93	76.19	85.56	0.86	88.07	89.03	225.99
2	40	687	34	4	95.28	90.91	93.10	0.93	54.05	99.42	371.64
3	109	592	42	22	93.38	83.21	88.29	0.88	72.19	96.42	157.76
4	187	497	16	65	96.88	74.21	85.54	0.86	92.12	88.43	566.08
5	81	641	38	5	94.40	94.19	94.29	0.94	68.07	99.23	283.23

Notes: SG, subgroup; TP, true positive; FP, false positive; TN, true negative; FN, false negative; Sp(%), specificity percentage correct; Se(%), sensitivity percentage correct; BA, balanced accuracy; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; DOR, diagnostic odds ratio.

© 2020 American Medical Association. All rights reserved.

eTable 12. Differences between subgroups in the validation sample across 6 clinical domains

	Variable	Full sample	Group 1	Group 2	Group 3	Group 4	Group 5	F / χ^2 (df)	*p*-value	Eta^2/Phi
N		458	149	22	134	87	66			
Demographics										
	Baseline age	42.6(13.2)	49.1(13.4)	45.3(11.3)	39.1(12.9)	36.5(9.4)	42.4(11.7)	18.88(4,453)	2.28E-14	0.14
	Sex [n (%) male]	233(50.9)	51(34.2)	14(63.6)	64(47.8)	67(77.0)	37(56.1)	42.96(10)	1.05E-08	0.31
	Education Status	3.7(1.7)	4.2(1.6)	3.5(2.0)	3.6(1.6)	3.5(1.6)	2.7(1.8)	9.26(4,431)	3.46E-07	0.08
	Single marital status [n(%)]	240(54.2)	35(24.1)	10(52.6)	78(58.6)	77(93.9)	40(62.5)	107.70(10)	2.25E-22	0.48
	Paid Employment [n(%)]	191(42.8)	78(53.4)	5(26.3)	51(38.1)	37(44.0)	20(31.7)	13.27(10)	1.00E-02	0.17
Psychiatric treatment and history										
	Current psychiatric treatment [n(%)]	3.0(1.0)	2.5(1.0)	3.0(1.0)	3.2(1.0)	2.6(1.0)	3.8(0.6)	26.32(4,439)	1.41E-19	0.19
	Psychiatric family history [n(%)]	330(76.4)	119(83.8)	14(66.7)	107(82.9)	60(75.0)	30(50.0)	31.75(10)	2.15E-06	0.26
	Onset (1st inpatient year)	30.3(11.8)	35.2(12.9)	27.3(10.6)	29.4(11.6)	25.2(7.7)	28.8(11.0)	11.39(4,418)	8.78E-09	0.1
	Duration of illness	12.4(11.0)	14.0(10.7)	17.5(13.6)	10.1(9.9)	11.2(9.1)	13.4(13.6)	3.67(4,413)	6.01E-03	0.03
	Ever hallucinate (SCID) [n(%)]	175(39.6)	28(19.3)	14(66.7)	40(31.0)	39(47.6)	54(83.1)	88.92(10)	2.24E-18	0.44
	Ever delusional (SCID) [n(%)]	263(59.6)	54(38.3)	17(85.0)	63(48.8)	64(75.3)	65(98.5)	88.30(10)	3.02E-18	0.44
	Ever take illicit drugs [n(%)]	190(45.0)	30(22.1)	11(57.9)	68(54.0)	66(83.5)	15(24.2)	92.55(10)	3.79E-19	0.45
Current DSM-IV diagnoses										
	Major depressive disorder [n(%)]	89(19.4)	32(21.5)	1(4.5)	50(37.3)	6(6.9)	0(0.0)	55.53(10)	2.52E-11	0.35
	Bipolar II disorder [n(%)]	50(10.9)	29(19.5)	0(0.0)	15(11.2)	6(6.9)	0(0.0)	23.43(10)	1.04E-04	0.23
	Bipolar I disorder [n(%)]	171(37.3)	71(47.7)	12(54.5)	43(32.1)	39(44.8)	6(9.1)	35.73(10)	3.29E-07	0.28
	Schizoaffective disorder [n(%)]	28(6.1)	5(3.4)	2(9.1)	10(7.5)	6(6.9)	5(7.6)	3.08(10)	5.45E-01	0.08
	Brief psychotic disorder [n(%)]	--	--	--	--	--	--	--	--	--
	Schizophreniform disorder [n(%)]	2(0.4)	0(0.0)	0(0.0)	0(0.0)	2(2.3)	0(0.0)	--	--	--
	Schizophrenia	118(25.8)	12(8.1)	7(31.8)	16(11.9)	28(32.2)	55(83.3)	154.49(10)	2.23E-32	0.58
Current psychiatric symptoms										
	PANSS Positive	11.1(5.4)	8.5(2.4)	10.1(2.2)	9.8(3.7)	10.7(4.3)	20.0(6.1)	104.10(4,432)	5.18E-62	0.49
	PANSS Negative	13.2(6.1)	9.9(4.0)	11.3(4.0)	14.2(5.2)	11.8(4.9)	20.7(6.5)	55.75(4,428)	7.87E-38	0.34
	PANSS Total	52.0(17.9)	39.9(8.5)	45.1(7.0)	55.2(11.5)	46.3(11.3)	80.7(18.2)	137.62(4,410)	1.92E-74	0.57

© 2020 American Medical Association. All rights reserved.

	IDSC Sum	16.3(13.6)	9.0(7.5)	13.9(5.7)	31.5(11.7)	8.8(5.8)	9.2(8.2)	136.77(4,394)	3.72E-73	0.58
	YMRS Sum	3.3(5.2)	3.4(5.0)	2.2(2.9)	2.9(4.3)	3.9(6.2)	3.3(6.4)	0.60(4,422)	6.60E-01	0.01
Current functioning										
	CGI	4.3(1.1)	3.6(1.1)	4.1(0.8)	4.7(0.8)	4.1(1.1)	5.1(0.8)	38.57(4,438)	1.16E-27	0.26
	GAF	54.9(13.6)	63.0(11.6)	58.1(12.4)	50.2(11.0)	56.7(12.7)	43.4(12.2)	38.56(4,434)	1.28E-27	0.26
	WHO-QoL Global	12.2(3.7)	14.3(3.1)	10.7(2.6)	9.4(2.8)	13.6(3.4)	11.4(2.7)	42.76(4,325)	8.22E-29	0.34
	Verbal IQ (MWTB)	28.4(4.9)	29.8(4.5)	28.2(4.2)	27.8(5.1)	27.4(5.2)	27.1(4.8)	4.03(4,342)	3.29E-03	0.05

Notes: for all variable acronyms see the eBox. F, ANOVA F-score; χ^2, chi-squared value; df, degrees of freedom; p-value, p-value associated with ANOVA or χ^2 (only false-discovery rate values shown); Eta^2, partial eta-squared effect size measure associated with ANOVA tests; Phi, phi coefficient measure of effect size for χ^2 tests. ANOVA used for all variables displayed with the mean and standard deviation [mean(SD)] and chi-squared tests used for all nominal variables displayed with the total number and percentage of total [n(%)].

© 2020 American Medical Association. All rights reserved.

eTable 13. Classification of the discovery and replication samples for each subgroup

Table displaying machine learning analyses classifying the discovery and replication samples for each subgroup.

SG	TP	FP	TN	FN	Se(%)	Sp(%)	BAC	PPV	NPV	AUC	DOR
1	141	44	94	111	55.95	68.12	62.03	76.22	45.85	0.66	3.08
2	29	15	22	15	65.91	59.46	62.68	65.91	59.46	0.64	2.64
3	86	46	83	45	65.65	64.34	65.00	65.15	64.84	0.71	3.39
4	147	36	54	105	58.33	60.00	59.17	80.33	33.96	0.64	2.13
5	52	19	45	34	60.47	70.31	65.39	73.24	56.96	0.62	4.15

Notes: SG, subgroup; TP, true positive; FP, false positive; TN, true negative; FN, false negative; Sp(%), specificity percentage correct; Se(%), sensitivity percentage correct; BA, balanced accuracy; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value.

© 2020 American Medical Association. All rights reserved.

eTable 14. Somatic variables requiring exclusion for the Günzburg replacement analyses

Variables excluded in order to conduct the Günzburg site exclusion analysis

Number	Variable Description
45	Ever diagnosed with a cardiovascular disease (dichotomous)
46	Ever diagnosed with a metabolic disorder (dichotomous)
47	Ever diagnosed with a thyroid disorder (dichotomous)
48	Ever diagnosed with a rheumatological or musculoskeletal disorder (dichotomous)
49	Ever diagnosed with a lung disorder (dichotomous)
50	Ever diagnosed with allergies (dichotomous)
51	Ever diagnosed with a gastrointestinal disorder (dichotomous)
52	Ever diagnosed with a gallbladder or liver disorders (dichotomous)
53	Ever diagnosed with a skin disorder (dichotomous)
54	Ever diagnosed with a kidney or urinary bladder disorder (dichotomous)
55	Ever diagnosed with a neurological or psychiatric disorder except the study diagnosis (dichotomous)
56	Ever diagnosed with a disorder of the eyes or ears (dichotomous)
57	Ever diagnosed with infectious diseases (dichotomous)
58	Ever diagnosed with cancer (dichotomous)
59	Ever diagnosed with any other disorders (dichotomous)

© 2020 American Medical Association. All rights reserved.

eTable 15. Günzburg exclusion and replacement clinical comparison table

Comparison of the fifth subgroup across for the discovery sample, when excluding cases from the Günzburg site (G-excluded), and when replacing the Günzburg cases using individuals from the replication sample (G-replaced).

	Variable	Discovery	G-excluded	G-replaced	F / χ^2 (df)	*p*-value	Eta^2/Phi
N		78	42	65			
	Baseline age	37.8(12.0)	40.4(12.8)	41.5(12.6)	1.63(2,182)	--	--
	Sex [n (%) male]	46(59.0)	28(66.7)	34(52.3)	2.18(6)	--	--
	Education Status	2.6(1.5)	2.8(1.8)	2.7(1.7)	0.23(2,178)	--	--
	Single marital status [n(%)]	59(75.6)	33(78.6)	44(67.7)	1.86(6)	--	--
	Paid Employment [n(%)]	16(20.8)	7(16.7)	13(20.0)	0.30(6)	--	--
	Current psychiatric treatment [n(%)]	3.8(0.5)	3.6(0.8)	3.5(0.9)	4.94(2,182)	8.13E-03	0.05
	Psychiatric family history [n(%)]	31(44.9)	24(61.5)	34(55.7)	3.12(6)	--	--
	Onset (1st inpatient year)	27.0(9.0)	27.1(8.8)	28.0(9.2)	0.27(2,176)	--	--
	Duration of illness	10.6(8.6)	12.9(9.8)	13.5(12.0)	1.52(2,176)	--	--
	Ever hallucinate (SCID) [n(%)]	67(87.0)	39(92.9)	61(93.8)	2.25(6)	--	--
	Ever delusional (SCID) [n(%)]	76(97.4)	42(100.0)	65(100.0)	2.77(6)	--	--
	Ever take illicit drugs [n(%)]	16(21.3)	20(48.8)	32(51.6)	15.71(6)	3.88E-04	0.29
	Major depressive disorder [n(%)]	--	--	--	--	--	--
	Bipolar II disorder [n(%)]	--	--	--	--	--	--
	Bipolar I disorder [n(%)]	1(1.3)	1(2.4)	4(6.2)	2.81(6)	--	--
	Schizoaffective disorder [n(%)]	8(10.3)	5(11.9)	6(9.2)	0.20(6)	--	--
	Brief psychotic disorder [n(%)]	--	--	--	--	--	--
	Schizophreniform disorder [n(%)]	2(2.6)	1(2.4)	3(4.6)	0.60(6)	--	--
	Schizophrenia	67(85.9)	35(83.3)	52(80.0)	0.88(6)	--	--
	PANSS Positive	20.5(4.7)	20.5(5.6)	20.6(5.6)	0.01(2,180)	--	--
	PANSS Negative	22.5(6.0)	22.6(6.1)	20.8(6.5)	1.70(2,179)	--	--
	PANSS Total	84.3(13.6)	83.2(14.7)	80.6(14.4)	1.22(2,178)	--	--
	IDSC Sum	10.0(9.9)	22.4(10.9)	19.3(10.2)	22.69(2,163)	2.03E-09	0.22
	YMRS Sum	3.3(6.1)	6.2(6.8)	7.3(8.1)	6.11(2,179)	2.72E-03	0.06
	CGI	4.8(0.7)	5.2(0.6)	5.2(0.6)	9.25(2,181)	1.50E-04	0.09
	GAF	46.8(13.0)	38.9(8.0)	39.5(8.3)	11.61(2,180)	1.80E-05	0.11
	WHO-QoL Global	11.7(3.6)	11.3(3.3)	12.3(3.3)	0.86(2,164)	--	--
	Verbal IQ (MWTB)	24.6(5.6)	24.7(5.9)	25.1(5.7)	0.08(2,109)	--	--

© 2020 American Medical Association. All rights reserved.

eTable 16. Günzburg exclusion site comparison table

Site frequencies N(%) for the analysis without cases from the Günzburg site.

Site	SG1	SG2	SG3	SG4	SG5	FChi	P	DPhi
Augsburg	18(7.1)	2(3.8)	1(1.0)	14(5.5)	1(2.4)	6.34(10)	1.75E-01	0.1
Bad Zwischenahn	16(6.3)	5(9.4)	13(13.4)	16(6.3)	5(11.9)	6.88(10)	1.43E-01	0.1
Bochum	14(5.6)	6(11.3)	21(21.6)	21(8.3)	16(38.1)	51.57(10)	1.70E-10	0.27
Bremen Ost	6(2.4)	3(5.7)	1(1.0)	16(6.3)	0(0.0)	10.10(10)	3.87E-02	0.12
Eschwege	7(2.8)	0(0.0)	0(0.0)	0(0.0)	0(0.0)	12.51(10)	1.39E-02	0.13
Goettingen	8(3.2)	2(3.8)	0(0.0)	1(0.4)	0(0.0)	10.32(10)	3.54E-02	0.12
Graz	47(18.7)	11(20.8)	16(16.5)	27(10.6)	0(0.0)	15.70(10)	3.44E-03	0.15
Guenzburg	0(0.0)	0(0.0)	0(0.0)	0(0.0)	0(0.0)	NaN(5)	NaN	NaN
Hildesheim	9(3.6)	0(0.0)	1(1.0)	6(2.4)	1(2.4)	3.50(10)	4.77E-01	0.07
LMU Muenchen	16(6.3)	6(11.3)	8(8.2)	43(16.9)	9(21.4)	18.61(10)	9.36E-04	0.16
Liebenburg	4(1.6)	1(1.9)	2(2.1)	2(0.8)	0(0.0)	1.83(10)	7.67E-01	0.05
Lueneburg	10(4.0)	2(3.8)	2(2.1)	19(7.5)	0(0.0)	8.32(10)	8.06E-02	0.11
Muenster	2(0.8)	0(0.0)	1(1.0)	0(0.0)	0(0.0)	3.11(10)	5.40E-01	0.07
Osnabruck	17(6.7)	1(1.9)	5(5.2)	14(5.5)	1(2.4)	2.92(10)	5.72E-01	0.06
Rotenburg/Wamme	14(5.6)	1(1.9)	0(0.0)	13(5.1)	0(0.0)	8.80(10)	6.63E-02	0.11
Tiefenbrunn	1(0.4)	0(0.0)	1(1.0)	2(0.8)	0(0.0)	1.25(10)	8.70E-01	0.04
UMG Goettingen	60(23.8)	11(20.8)	21(21.6)	57(22.4)	8(19.0)	0.67(10)	9.55E-01	0.03
Wilhelmshaven	3(1.2)	2(3.8)	4(4.1)	3(1.2)	1(2.4)	5.10(10)	2.77E-01	0.09

© 2020 American Medical Association. All rights reserved.

eTable 17. Günzburg site replacement site comparison table

Site frequencies N(%) for the analysis with Günzburg excluded and replaced by cases from the replication cohort.

Variable	SG1	SG2	SG3	SG4	SG5	FChi	P	DPhi
Augsburg	2(1.7)	2(3.1)	20(7.1)	22(8.9)	3(5.6)	8.53(10)	7.40E-02	0.11
Bad Zwischenahn	13(10.9)	7(10.8)	15(5.4)	15(6.1)	6(11.1)	6.72(10)	1.52E-01	0.09
Bochum	30(25.2)	22(33.8)	18(6.4)	26(10.5)	8(14.8)	50.65(10)	2.64E-10	0.26
Bremen Ost	1(0.8)	1(1.5)	6(2.1)	15(6.1)	3(5.6)	10.55(10)	3.21E-02	0.12
Eschwege	0(0.0)	0(0.0)	7(2.5)	0(0.0)	0(0.0)	12.24(10)	1.57E-02	0.13
Goettingen	0(0.0)	0(0.0)	8(2.9)	1(0.4)	2(3.7)	10.48(10)	3.31E-02	0.12
Graz	19(16.0)	0(0.0)	52(18.6)	24(9.7)	7(13.0)	20.16(10)	4.63E-04	0.16
Guenzburg	0(0.0)	0(0.0)	0(0.0)	0(0.0)	0(0.0)	NaN(5)	NaN	NaN
Hildesheim	1(0.8)	2(3.1)	10(3.6)	5(2.0)	0(0.0)	4.56(10)	3.35E-01	0.08
LMU Muenchen	9(7.6)	13(20.0)	19(6.8)	43(17.4)	6(11.1)	20.56(10)	3.87E-04	0.16
Liebenburg	2(1.7)	0(0.0)	4(1.4)	2(0.8)	1(1.9)	1.68(10)	7.93E-01	0.05
Lueneburg	2(1.7)	0(0.0)	11(3.9)	19(7.7)	1(1.9)	12.65(10)	1.31E-02	0.13
Muenster	2(1.7)	0(0.0)	3(1.1)	0(0.0)	0(0.0)	5.09(10)	2.78E-01	0.08
Osnabruck	5(4.2)	3(4.6)	19(6.8)	11(4.5)	0(0.0)	5.09(10)	2.79E-01	0.08
Rotenburg/Wamme	1(0.8)	0(0.0)	15(5.4)	12(4.9)	3(5.6)	7.86(10)	9.70E-02	0.1
Tiefenbrunn	1(0.8)	0(0.0)	1(0.4)	2(0.8)	0(0.0)	1.39(10)	8.45E-01	0.04
Tuebingen	2(1.7)	2(3.1)	0(0.0)	0(0.0)	0(0.0)	14.27(10)	6.47E-03	0.14
UMG Goettingen	25(21.0)	12(18.5)	69(24.6)	47(19.0)	12(22.2)	2.91(10)	5.72E-01	0.06
Wilhelmshaven	4(3.4)	1(1.5)	3(1.1)	3(1.2)	2(3.7)	4.28(10)	3.69E-01	0.07

© 2020 American Medical Association. All rights reserved.

eTable 18. Supervised learning classification of subgroups using cognitive variables

Supervised learning classification of subgroups using cognitive variables in the discovery sample.

SG	TP	FP	TN	FN	Sp(%)	Se(%)	BA(%)	AUC	PPV	NPV	DOR
1	26	55	458	226	89.28	10.32	49.80	0.50	32.10	66.96	0.93
2	1	34	687	43	95.28	2.27	48.78	0.49	2.86	94.11	0.23
3	3	14	620	128	97.79	2.29	50.04	0.50	17.65	82.89	1.08
4	116	164	349	136	68.03	46.03	57.03	0.57	41.43	71.96	2.07
5	71	281	398	15	58.62	82.56	70.59	0.71	20.17	96.37	3.98

Notes: SG, subgroup; TP, true positive; FP, false positive; TN, true negative; FN, false negative; Sp(%), specificity percentage correct; Se(%), sensitivity percentage correct; BA, balanced accuracy; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; DOR, diagnostic odds ratio.

© 2020 American Medical Association. All rights reserved.

eTable 19. Total number of participants across time points for each subgroup

Numbers of participants across subgroups and visits for each of the variables analysed using mixed models.

	Subgroups	Total N	Visit 1	Visit 2	Visit 3	Visit 4	Replication	Total N	Visit 1	Visit 2	Visit 3	Visit 4
GAF	1	252	247	182	150	133		138	134	70	47	26
	2	44	44	26	17	16		37	34	20	7	5
	3	131	130	79	60	52		129	128	72	29	19
	4	252	251	167	147	137		90	82	52	25	17
	5	86	86	40	32	31		64	61	25	19	19
	Total	**765**	**758(99%)**	**494(65%)**	**406(53%)**	**369(48%)**		**458**	**439(95%)**	**239(52%)**	**127(28%)**	**86(19%)**
WHO QoL	1	252	242	177	146	130		138	116	65	47	24
	2	44	43	25	17	15		37	22	18	7	4
	3	131	120	77	57	49		129	111	66	28	18
	4	252	236	160	140	130		90	66	53	22	16
	5	86	80	31	27	26		64	15	14	13	13
	Total	**765**	**721(94%)**	**470(61%)**	**387(51%)**	**350(46%)**		**458**	**330(72%)**	**216(47%)**	**117(26%)**	**75(16%)**
IDSC	1	252	212	157	133	117		138	122	67	42	24
	2	44	41	24	16	13		37	28	19	7	5
	3	131	112	74	54	46		129	121	62	27	18
	4	252	220	160	128	126		90	74	46	23	16
	5	86	80	37	31	27		64	54	24	18	15
	Total	**765**	**665(87%)**	**452(59%)**	**362(47%)**	**329(43%)**		**458**	**399(87%)**	**218(48%)**	**117(26%)**	**78(17%)**
PANSS	1	252	234	177	147	125		138	125	67	43	25
	2	44	41	26	17	14		37	32	19	7	5
	3	131	123	77	56	50		129	117	69	26	16
	4	252	238	159	139	131		90	79	49	23	16
	5	86	86	40	30	29		64	62	25	19	16
	Total	**765**	**722(94%)**	**479(63%)**	**389(51%)**	**349(46%)**		**458**	**415(91%)**	**229(50%)**	**118(26%)**	**78(17%)**

© 2020 American Medical Association. All rights reserved.

eTable 20. Mixed-model analyses controlling for missing data

Mixed model analyses of illness course controlling for the amount of missing data over visits for each participant (N.Missing).

	Term	Mean Sq	df	F	P-value	Sig
PANSS						
	N.Missing	13.741	1008.2	0.162	0.687	
	Group	5211.070	551.84	61.551	<2.20E-16	***
	Time (linear)	5191.925	1225.48	61.325	1.04E-14	***
	Time (quadratic)	2063.291	1276.66	24.371	9.00E-07	***
	N.Missing x Group	171.667	998.23	2.028	0.088	.
	Time x Group	1393.877	1227.94	16.464	3.79E-13	***
	$Time^2$ x Group	689.939	1270.71	8.149	1.73E-06	***
IDS-C_{30}						
	N.Missing	4.105	1012.24	0.096	0.757	
	Group	1365.980	560.88	31.860	<2.20E-16	***
	Time (linear)	296.966	1200.31	6.926	0.009	**
	Time (quadratic)	47.405	1257.56	1.106	0.293	
	N.Missing x Group	69.454	993.74	1.620	0.167	
	Time x Group	667.891	1217.54	15.578	1.95E-12	***
	$Time^2$ x Group	195.069	1259.7	4.550	0.001	**
GAF						
	N.Missing	45.711	1027.55	0.554	0.457	
	Group	1313.784	594.05	15.917	2.19E-12	***
	Time (linear)	584.136	1299.58	7.077	0.008	**
	Time (quadratic)	3311.396	1340.44	40.119	3.26E-10	***
	N.Missing x Group	70.973	1018.98	0.860	0.488	
	Time x Group	371.841	1304.06	4.505	0.001	**
	$Time^2$ x Group	450.875	1338.23	5.462	0.0002	***
WHOQoL-BREF						
	N.Missing	15.034	1000.23	2.863	0.091	.
	Group	130.453	568.25	24.841	<2.20E-16	***
	Time (linear)	40.623	1250.79	7.735	0.005	**

© 2020 American Medical Association. All rights reserved.

	Term	Mean Sq	df	F	P-value	Sig
	Time (quadratic)	26.057	1291.57	4.962	0.026	*
	N.Missing x Group	3.702	992.15	0.705	0.589	
	Time x Group	31.852	1259.87	6.065	7.81E-05	***
	$Time^2$ x Group	20.761	1288.12	3.953	0.003	**

Notes: PANSS, positive and negative symptom total score; IDSC, inventory of depressive symptomatology total score; YMRS, young mania rating scale total score; GAF, global assessment of functioning; WHOQoL, WHO Quality of Life Questionnaire global score; Mean Sq, mean square error; df, degrees of freedom; F, F statistic. Significance codes: <0.001, '***'; <0.01, '**'; <0.05, '*' ; <0.1, '.'.

© 2020 American Medical Association. All rights reserved.

eTable 21. Post hoc analysis of mixed models controlling for missing data

Post-hoc analysis of mixed model quadratic trends of the model controlling for missing values corrected for multiple comparisons using the false-discovery rate (FDR).

	Contrast	Estimate	SE	df	T	p (FDR)	Sig.
PANSS							
	1 - 2	-119.363	134.843	1309.274	-0.885	0.38	
	1 - 3	-296.446	84.736	1291.793	-3.498	0.001	**
	1 - 4	67.225	65.955	1288.272	1.019	0.38	
	1 - 5	-403.342	107.718	1303.541	-3.744	< 0.001	***
	2 - 3	-177.083	145.414	1308.503	-1.218	0.32	
	2 - 4	186.588	135.333	1310.252	1.379	0.28	
	2 - 5	-283.979	159.900	1310.957	-1.776	0.15	
	3 - 4	363.671	85.515	1294.543	4.253	< 0.001	***
	3 - 5	-106.895	120.689	1303.589	-0.886	0.38	
	4 - 5	-470.566	108.332	1305.128	-4.344	< 0.001	***
IDS-C_{30}							
	1 - 2	80.299	99.661	1221.836	0.806	0.61	
	1 - 3	-230.800	63.122	1256.060	-3.656	0.001	**
	1 - 4	11.866	48.880	1222.342	0.243	0.81	
	1 - 5	-47.919	79.552	1244.443	-0.602	0.61	
	2 - 3	-311.098	107.267	1232.128	-2.900	0.013	*
	2 - 4	-68.432	99.555	1220.202	-0.687	0.61	
	2 - 5	-128.217	117.687	1230.884	-1.089	0.55	
	3 - 4	242.666	62.953	1252.117	3.855	0.001	**
	3 - 5	182.881	88.895	1255.008	2.057	0.100	.
	4 - 5	-59.785	79.418	1241.934	-0.753	0.61	
GAF							
	1 - 2	-26.912	129.476	1341.281	-0.208	0.84	
	1 - 3	141.379	82.034	1345.838	1.723	0.14	
	1 - 4	-90.067	63.490	1334.744	-1.419	0.22	
	1 - 5	356.794	105.112	1362.514	3.394	0.004	**
	2 - 3	168.291	139.778	1344.132	1.204	0.29	
	2 - 4	-63.155	129.766	1341.201	-0.487	0.70	
	2 - 5	383.706	154.457	1352.148	2.484	0.033	*

© 2020 American Medical Association. All rights reserved.

	Contrast	Estimate	SE	df	T	p (FDR)	Sig.
	3 - 4	-231.446	82.491	1345.589	-2.806	0.017	*
	3 - 5	215.415	117.569	1362.295	1.832	0.13	
	4 - 5	446.861	105.470	1362.247	4.237	< 0.001	***
WHOQoL-BREF							
	1 - 2	48.368	33.127	1270.932	1.460	0.36	
	1 - 3	77.483	21.125	1288.526	3.668	0.003	**
	1 - 4	6.841	16.330	1272.595	0.419	0.84	
	1 - 5	3.031	28.904	1327.984	0.105	0.92	
	2 - 3	29.116	35.860	1277.569	0.812	0.60	
	2 - 4	-41.526	33.261	1271.965	-1.248	0.42	
	2 - 5	-45.337	40.927	1299.670	-1.108	0.45	
	3 - 4	-70.642	21.335	1290.709	-3.311	0.005	**
	3 - 5	-74.453	31.999	1325.873	-2.327	0.067	.
	4 - 5	-3.810	29.057	1328.757	-0.131	0.92	

Notes: PANSS, positive and negative symptom total score; IDSC, inventory of depressive symptomatology total score; YMRS, young mania rating scale total score; GAF, global assessment of functioning; WHOQoL, WHO Quality of Life Questionnaire global score. Significance codes: <0.001, '***'; <0.01, '**'; <0.05, '*' ; <0.1, '.'.

© 2020 American Medical Association. All rights reserved.

eReferences.

1. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bull* 1987;13(2):261-276.
2. Rush AJ, Carmody T, Reimitz PE. The inventory of depressive symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. *International journal of methods in psychiatric research* 2000;9:45-59.
3. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry* Nov 1978;133:429-435.
4. Lehrl S, Triebig G, Fischer B. Multiple choice vocabulary test MWT as a valid and short test to estimate premorbid intelligence. *Acta neurologica Scandinavica* May 1995;91(5):335-345.
5. Association AP. *Diagnostic and statistical manual of mental disorders, 4th ed.* US: Washington, DC: American Psychiatric Association; 2002.
6. Angermeyer MC, Kilian R, Matschinger H. *WHOQOL100 und WHOQOL-BREF. Handbuch für die deutschsprachigen Versionen der WHO Instrumente zur Erfassung von Lebensqualität*. Göttingen: Hogrefe; 2000.
7. Rammstedt B, John OP. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J Res Personal* 2007;41:203-2012.
8. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* Jun 2001;17(6):520-525.
9. Brat DJ, Verhaak RG, Aldape KD, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England journal of medicine* Jun 25 2015;372(26):2481-2498.
10. Chen F, Zhang Y, Senbabaoglu Y, et al. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell reports* Mar 15 2016;14(10):2476-2489.
11. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* Aug 14 2014;158(4):929-944.
12. Network CGAR. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* Sep 11 2014;513(7517):202-209.
13. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 2004;101(12):4164-4169.
14. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 2003;52(1):91-118.
15. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* Jun 15 2007;23(12):1495-1502.
16. Li Y, Ngom A. The non-negative matrix factorization toolbox for biological data mining. *Source Code for Biology and Medicine* 2013;8(1):10.
17. Senbabaoglu Y, Michailidis G, Li JZ. A reassessment of consenuss clustering for class discovery. *Sci Rep* 2014;4:6207.
18. Koutsouleris N, Meisenzahl E, Borgwardt S, et al. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain* 2015;138(7):2059-2073.

© 2020 American Medical Association. All rights reserved.

19. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986;1(1):54-77.

20. McIntosh AR, Chau W, Protzner AB. Spatiotemporal analysis of event-related fMRI data using partial least squares. *Neuroimage* Oct 2004;23(2):764-775.

21. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* Sep 2007;81(3):559-575.

22. Budde M, Anderson-Schmidt H, Gade K, et al. A longitudinal approach to biological psychiatric research: the PsyCourse study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2018:in press.

23. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS genetics* Dec 2006;2(12):e190.

24. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* Jan 2013;10(1):5-6.

25. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* Jun 2009;5(6):e1000529.

26. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* Jul 23 2018;50(8):1112-1121.

27. Pardinas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* Mar 2018;50(3):381-389.

28. Stahl EA, Breen G, Forstner AJ, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics* May 2019;51(5):793-803.

29. Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* May 2018;50(5):668-681.

30. Ruderfer DM, Fanous AH, Ripke S, et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* Sep 2014;19(9):1017-1024.

31. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* Oct 1 2015;526(7571):68-74.

32. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015;67(1):1-48.

33. Halekoh U, Højsgaard S. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software* 2014;59(9):1-30.

34. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001;29(4):1165-1188.

35. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 2013;4:133-142.

36. Johnson PCD. Extension of Nakagawa & Schielzeth's R2glmm to random slopes models. *Methods in Ecology and Evolution* 2014;5(9).

37. Jaeger BC, Edwards LJ, Das K, Sen PK. An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics* 2017;44(6):1086-1105.

© 2020 American Medical Association. All rights reserved.

38. Hastie T, Tibshirani R. Classification by pairwise coupling. Paper presented at: Advances in neural information processing systems, 1998.
39. Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the positive and negative syndrome scale for schizophrenics. *Psychiatry research* Jan 1988;23(1):99-110.
40. McGuffin P, Farmer A, Harvey I. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Arch Gen Psychiatry* Aug 1991;48(8):764-770.

© 2020 American Medical Association. All rights reserved.