**Antonucci, et al. – Supplementary Information**

1. **Sample Determination**

This study was conducted on 896 patients with schizophrenia recruited within the multicentric consortium “Italian Network for Research on Psychoses (NIRP) (Galderisi et al., 2014). Within this consortium, patients living in the community and seen in outpatient units of 26 Italian university psychiatric clinics and/or mental health departments have been recruited. Inclusion criteria were a diagnosis of schizophrenia according to DSM-IV, confirmed with the Structured Clinical Interview for DSM-IV—Patient version (SCID-I-P), an age between 18 and 66 years, and, for patients undergoing drug treatment, pharmacological stability since at least 3 months. Exclusion criteria were: a history of head trauma with loss of consciousness; a history of moderate to severe mental retardation or of neurological diseases; a history of alcohol and/or substance abuse in the last 6 months; current pregnancy or lactation; inability to provide informed consent; treatment modifications and/or hospitalization due to symptom exacerbation in the last 3 months. During NIRP recruitment procedures, 1691 patients were screened and 1180 considered eligible; of these, 202 refused to participate and 57 dropped out before completing the procedures, leaving 921 included. We excluded 25 subjects as they did not fill out the Resilience Scale for Adults (see section 1.2 and Supplementary Information 2), and therefore our final sample was composed of 896 patients.

1. **The Resilience Scale for Adults (RSA)**

The resilience level of each patient included in this study was assessed via the Resilience Scale for Adults (RSA) (Bonfiglio, Renati, Hjemdal, & Friborg, 2016). This is a self-administered instrument including 33 items, based on a seven-point Likert Scale in which each item has a positive and negative attribute at each end of a scale continuum. Items are organized into six domains: perception of self (six items), perception of the future (four items), structured style (four items), social competence (six items), family cohesion (six items), and social resources (seven items). However, for analysis purposes, we only used the RSA total score as a global indicator of the level of resilience, with higher (i.e., above median) scores reflecting greater resilience abilities. The RSA shows good psychometric properties and high construct validity(Bonfiglio et al., 2016).

1. **Assessment**
	1. **Anamnestic information**

To evaluate and quantify context-related factors, an *ad hoc* questionnaire was developed to collect information about gender, age, marital status, schooling, housing, substance use, socio-economic status, alcohol use and abuse, and clinical course of the disease. The socio-economic status was obtained from the education level and type of work of each parent. The former was measured through a 7-point scale rating as 1 the elementary school level, and as 7 the post-degree/specialization courses level. The type of work was instead ranked on 9 levels spanning from laborer to high-level managerial position. The Hollingshead index(Hollingshead & Redlich, 1954) was calculated as the average of the socio-economic status of the two parents.

* 1. **Premorbid Adjustment Scale-PAS**

To assess the level of functioning before the onset of psychosis, the Premorbid Adjustment Scale (PAS) (Shapiro et al., 2009) was used. It evaluates the degree of achievement of developmental goals throughout childhood, adolescence, and adulthood. Based on these life periods, the scale is subdivided into four subscales, reflecting the level of functioning during childhood up to 11 years old, early adolescence to age 15, late adolescence to age 18, and then adulthood. For each of these periods, different areas are investigated. For childhood and adolescence, the items focus on sociability and social withdrawal, peer relationships, scholastic performance, adaptation to school, and ability to form socio-sexual relationships. Concerning the adult period, the attention shifts primarily on the social relationships, while the General Scale provides a global view, containing items estimating the highest level of functioning achieved, spanning from educational achievements to social relationships and level of interest in and enjoyment of major life activities (work, family, etc.).

* 1. **St. Hans Rating Scale-SHRS**

The level of discomfort generated by antipsychotics’ extrapyramidal side effects was taken into consideration and evaluated via the St. Hans Rating Scale (SHRS) (Gerlach et al., 1993). This is a multidimensional scale that consists of four main components. The first component evaluates the severity of hyperkinesia, both active and passive, in different body districts; the second one focuses on parkinsonism and its main manifestations, while the third one allows to acknowledge if the dystonia is present or not. The last one assesses the level of psychic (subjective) and motor (objective) akathisia. The items in each domain are rated from 0 (not present) to 6 (severe) and global scores are then calculated for hyperkinesia, parkinsonism, and the global extra-pyramidal symptomatology.

* 1. **Internalized Stigma of Mental Illness-ISMI**

The Internalized Stigma of Mental Illness (ISMI) (Ritsher, Otilingam, & Grajales, 2003) was used to measure the self-inflicted stigma and self-rejection. This scale is composed of 29 items organized into five subscales, assessing alienation, stereotype endorsement, discrimination experience, social withdrawal, and stigma resistance. For each item, patients are asked to give their opinion through a four-point Likert scale from 1 (strongly disagree) to 4 (strongly agree). The total score is calculated from the sum of all the items in each subscale. The results obtained are therefore directly proportional to the level of internalized stigmatization.

* 1. **Brief COPE-Coping Orientation to Problems Experienced**

To measure how individuals cope with stressful life events, the Brief-COPE, a shortened version of the original 60-items COPE scale was used. This is a 28-items self-report questionnaire organized into fourteen scales of two items each(Carver, Scheier, & Weintraub, 1989). Each scale reflects a coping strategy. The questionnaire allows researchers to determine the patients’ primary coping style as either problem-oriented or emotion-oriented. The difference between these two styles is that in the former, the person is oriented to face adversities acting on the source of stress and relying on problem-solving abilities, while in the latter the aim is to reduce the emotional distress caused by such stressors.

* 1. **Self-Esteem Rating Scale-SERS**

The Self-Esteem Rating Scale (SERS) (Lecomte, Corbiere, & Laisne, 2006) was employed to investigate the level of self-esteem. This instrument consists of two scales, positive and negative self-esteem, in which the 40 items are split. Each item is scored on a 7-point Likert scale with 20 of them being scored positively and 20 negatively. The SERS considers multiple aspects of self-evaluation such as overall self-worth, social competence, problem-solving ability, intellectual ability, self-competence, and worth compared to others.

* 1. **Recovery Style Questionnaire-RSQ**

Another domain evaluated is the method of recovery from psychosis through the Recovery Style Questionnaire (RSQ) (Gruber et al., 2018). This is a self-administered tool composed of 39 statements, divided into 13 concept scales, with which the participant is asked to agree or disagree. Choosing between two different types of answers for each item, the patient manifests expresses its predisposition to recover with an “integration” or “sealing over” style. The “integrators” are those patients that express interest and curiosity toward their condition and want to frame it into a coherent perspective. The “sealers”, on the other hand, are patients who want simply to forget their psychotic experience, considering it an unhappy parenthesis.

* 1. **Service Engagement Scale-SES**

Service Engagement Scale (SES) (Tait, Birchwood, & Trower, 2002) was the instrument used to assess the relationship of the individual with the community mental health services. It includes 14 items rated on a four-point Likert scale, from 0 meaning “not at all or rarely” to 3 corresponding to “most of the time”. The items are split into 4 sub-scales: availability, collaboration, help-seeking, and treatment adherence, whose scores get summed to obtain the final one. In this study, we used the global score, with higher results reflecting higher difficulty in engaging with services.

* 1. **Social Network Questionnaire**

The type and quality of social support were explored via the Social Network Questionnaire (SNQ)(Ribe, Salamero, Perez-Testor, Valero, & Garcia, 2015). This tool consists of 15 items grouped into four subscales, investigating social contacts, practical, affective, and partner support. Higher scores reflect broader social networks and better interaction of the patients with such networks.

* 1. **Calgary Depression Scale for Adults-CDSS**

Depressive symptoms were assessed using the Calgary Depression Scale for Schizophrenia (CDSS) (Schennach-Wolff et al., 2011). It is a nine-items structured interview in which each item is graded according to a 4-point Likert scale, going from 0 meaning absent to 3 meaning severe. It consists of eight structured questions and a ninth observational item that depends on the observation of the patient during the interview. It is aimed at assessing the level of depressive symptoms, independently from positive and negative symptoms of schizophrenia.

* 1. **Specific Level of Functioning Scale-SLOF**

The Specific Level of Functioning Scale (SLOF) (Mucci et al., 2014) was employed to evaluate the real-life functioning of the patients. This scale included 43 items and is based either on caregiver reports, or on direct observation of patients' behavior and functioning in several domains, combined with information referred by the caregiver. No patient in this study self-administered the SLOF scale. Each item is rated on a five-point scale and the total score ranges from 43 to 215, which indicates the level of assistance the participant needs to perform a given activity. Higher scores indicate better functioning. This scale is validated in Italian, it has adequate reliability, construct validity, internal consistency, and inter-rater reliability(Mucci et al., 2014). The SLOF scale allows to measure six different domains across five subscales, namely: a) physical efficiency; b) skills in self-care; c) interpersonal relationships; d) social acceptability; e) community activities (e.g., purchases, use of public transport); f) work skills.

* 1. **Pervasive Developmental Disorder Assessment - PDD**

The PDD Behavior Inventory (PDD) is a rating scale aimed at measuring, also retrospectively, the presence of a Pervasive Developmental Disorder (PDD; autism, Asperger disorder, PDD-NOS, or childhood disintegrative disorder) (Cohen, Schmidt-Lackner, Romanczyk, & Sudhalter, 2003). The scale is filled by the caregivers.

* 1. **UCSD Performance-Based Skill Assessment, version B- UPSA-B**

The UPSA-B is a role-play-based measure of daily life functioning for patients with schizophrenia. The role-play tasks aim at assessing the level of individual skills in performing everyday community tasks (e.g., counting money, simulating to buy an item and paying it, filling a bill form, calling a public service asking for an information, calling the doctor to postpone an appointment). A total score, as well as two subscores reflecting the level of financial and communication skills, can be obtained. The UPSA-B is widely used because of its short administration time and of its good psychometric properties(Mausbach et al., 2008; Mausbach et al., 2011; Mausbach et al., 2010).

* 1. **MATRICS Consensus Cognitive Battery-MCCB**

Neurocognitive functions were evaluated through the MATRICS Consensus Cognitive Battery (MCCB). This includes 10 tests to evaluate impairments in seven cognitive domains: processing speed, attention/vigilance, working memory, verbal learning, visual learning, reasoning and problem solving, and social cognition (Keefe et al., 2011). The latter, social cognition, was evaluated using the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) managing emotion section and included in MCBB, integrated by The Awareness of Social Inference Test (TASIT) (Westerhof-Evers, Visser-Keizer, McDonald, & Spikman, 2014) and the Facial Emotion Identification Test (FEIT) (Erol, Mete, Sonmez, & Unal, 2010). For correlation purposes, we calculated age, gender and education corrected Z-scores for each MCCB domain, as well as for the overall scores.

* 1. **Positive and Negative Syndrome Scale-PANSS**

The severity of symptomatology was measured via the Positive and Negative Syndrome Scale (PANSS) (Peralta & Cuesta, 1994). This is made up of 30 items grouped into scales for Positive Symptoms, Negative Symptoms, and General Psychopathology. However, after the development and validation of a 5-factors model(Wallwork, Fortgang, Hashimoto, Weinberger, & Dickinson, 2012), other sub-domain scores can be calculated, as is the case of the “disorganization” factor in this study. Even the “positive symptoms” factor included here was calculated using the afore-mentioned model. Therefore, for our analysis, we considered the scores from positive and negative symptoms, general psychopathology, and disorganization, and a total score obtained summing all sub-domain scores, but the latter.

The individual anamnestic factors were then combined with the 12 scales previously described (i.e., all except for MCCB and PANSS) and their subscales, to obtain a set of 85 variables to be employed in the machine learning analysis. Tables S1 reports all the variables, from each of the assessments, which entered the algorithm.

Table S1. *Variables feeding the machine learning algorithm. Abbreviations: PAS = Premorbid Adjustment Scale; PDD = Pervasive Developmental Disorder; SHRS = Saint Hans Rating Scale; SERS = Self Esteem Rating Scale; SNQ = Social Network Questionnaire; CDSS = Calgary Depression Scale for Adults; RSQ = Recovery Style Questionnaire; SES = Service Engagement Scale; UPSA = UCSD Performance-Based Skill Assessment; ISMI = Internalized Stigma of Mental Illness; SLOF = Specific Level of Functioning.*

|  |  |  |
| --- | --- | --- |
| Variable  | Assessment | Variable Type |
| Global Educational Attainment | Anamnestic Information | Categorical |
| Age of onset (any psychopathology) | Anamnestic Information | Continuous |
| Age of onset (psychosis only) | Anamnestic Information | Continuous |
| Age at first relapse | Anamnestic Information | Continuous |
| Remission after the first episode | Anamnestic Information | Binary |
| Further psychotic episodes | Anamnestic Information | Binary |
| Number of therapy changes | Anamnestic Information | Continuous |
| Number of hospitalizations | Anamnestic Information | Continuous |
| Alcohol abuse | Anamnestic Information | Binary |
| Illicit drugs abuse | Anamnestic Information | Binary |
| Prescription drugs abuse | Anamnestic Information | Binary |
| Substance abuse | Anamnestic Information  | Binary |
| Attempted suicide | Anamnestic Information | Binary |
| Hollingshead index | Anamnestic Information | Continuous |

|  |  |  |
| --- | --- | --- |
| Academic adjustment in childhood | PAS | Continuous |
| Academic adjustment in early adolescence | PAS | Continuous |
| Academic adjustment in late adolescence | PAS | Continuous |
| Social adjustment in childhood | PAS | Continuous |
| Social adjustment early adolescence | PAS | Continuous |
| Social adjustment late adolescence | PAS | Continuous |
| Social adjustment in adulthood | PAS | Continuous |

|  |  |  |
| --- | --- | --- |
| Subjective akathisia | SHRS | Continuous |
| Objective akathisia | SHRS | Continuous |
| Dystonia | SHRS | Continuous |
| Total parkinsonism | SHRS | Continuous |
| Global parkinsonism | SHRS | Continuous |
| Total passive dyskinesia | SHRS | Continuous |
| Total active dyskinesia | SHRS | Continuous |
| Global passive dyskinesia | SHRS | Continuous |
| Global active dyskinesia | SHRS | Continuous |
| Mean dyskinesia | SHRS | Continuous |
| Total extra-pyramidal symptoms | SHRS | Continuous |
| SHRS total, standardized | SHRS | Continuous |

|  |  |  |
| --- | --- | --- |
| Alienation | ISMI | Continuous |
| Stereotype Endorsement | ISMI | Continuous |
| Discrimination Experienced | ISMI | Continuous |
| Social Withdrawal | ISMI | Continuous |
| Stigma Resistance | ISMI | Continuous |
| Total ISMI score | ISMI | Continuous |
| Total ISMI score 2 | ISMI | Continuous |

|  |  |  |
| --- | --- | --- |
| Retrospective Assessment of a Pervasive Developmental Disorder | PDD | Continuous |

|  |  |  |
| --- | --- | --- |
| Positive reframing | COPE | Continuous |
| Self-distraction | COPE | Continuous |
| Venting | COPE | Continuous |
| Use of informational support | COPE | Continuous |
| Active coping | COPE | Continuous |
| Denial | COPE | Continuous |
| Religion | COPE | Continuous |
| Humor | COPE | Continuous |
| Behavioral disengagement | COPE | Continuous |
| Emotional support | COPE | Continuous |
| Substance abuse | COPE | Continuous |
| Acceptance | COPE | Continuous |
| Planning | COPE | Continuous |
| Self-blame | COPE | Continuous |

|  |  |  |
| --- | --- | --- |
| Total score | SERS | Continuous |

|  |  |  |
| --- | --- | --- |
| Continuity | RSQ | Continuous |
| Ownership | RSQ | Continuous |
| Responsibility | RSQ | Continuous |
| Curiosity | RSQ | Continuous |
| Education | RSQ | Continuous |
| Help seeking | RSQ | Continuous |
| Blame | RSQ | Continuous |
| Cause | RSQ | Continuous |
| Optimism | RSQ | Continuous |
| Impact | RSQ | Continuous |
| Fear | RSQ | Continuous |
| Liking | RSQ | Continuous |
| Satisfaction | RSQ | Continuous |
| Total score | RSQ | Continuous |

|  |  |  |
| --- | --- | --- |
| Total score | SES | Continuous |

|  |  |  |
| --- | --- | --- |
| Social Contact | SNQ | Continuous |
| Practical Support | SNQ | Continuous |
| Affective Support | SNQ | Continuous |
| Partner support | SNQ | Continuous |

|  |  |  |
| --- | --- | --- |
| Communication task | UPSA-B | Continuous |
| Financial task | UPSA-B | Continuous |
| Total score | UPSA-B | Continuous |

|  |  |  |
| --- | --- | --- |
| Total score | CDSS | Continuous |

|  |  |  |
| --- | --- | --- |
| Physical functioning | SLOF | Continuous |
| Personal care skills | SLOF | Continuous |
| Interpersonal relationships | SLOF | Continuous |
| Social acceptability | SLOF | Continuous |
| Activities of community living | SLOF | Continuous |
| Work skills | SLOF | Continuous |

1. **Machine Learning pipeline**

To allow for unbiased estimation of the model’s generalizability and prevent information leaking between subjects used for training the models and subjects used for validating decisions (Ruschhaupt, Huber, Poustka, & Mansmann, 2004), we built a double cycle, cross-validation (CV) framework (Koutsouleris et al., 2021; Koutsouleris et al., 2016) in which we mixed repeated-nested CV (10 repetitions, 10 folds) at the inner CV level, and leave-site-out (17 sites) at the outer CV level. That is, at the outer CV level, we iteratively held back every study site as validation sample, while the rest of the data entered the inner CV cycle, where cases were again iteratively assigned to training data and test site samples used to identify optimally predictive hyperparameter combinations. This was done in order to obtain a geographic validation of our decisions.

Our NeuroMiner machine learning preprocessing pipeline consisted of the following steps:

1. Features were corrected for age and gender effects. Specifically, we removed the variance associated with age and gender from the feature scores within each inner and outer CV fold through partial correlations.
2. As many machine learning algorithms are sensitive to scale differences between features, we scaled each variable to a [0, 1] range of to remove these effects from each training sample matrix. The scaling parameters were then applied to the inner and outer CV cycles.
3. Scaled data entered a k-Nearest Neighbor-(k-NN) imputation step to fill the missing values in the data (Antonucci et al., 2020). For each missing value of a given subject, we identified a subset of cases in the training data that provided values for the given variable and that had values in all other variables which were non-empty in given subject (source data).
Subjects in the source subset were sorted according to their similarity with the target subject using the Euclidean distance. Then, we computed the median of the given variable in the 7 nearest neighbors of given subjects and filled the missing value with this median value. This process was repeated until all missing values were filled with the respectively computed nearest-neighbor medians. We always used the original, non-imputed training matrix for the imputation.

Preprocessed data entered a greedy forward search wrapper(Saeys, Inza, & Larranaga, 2007) which allows identifying the most parsimonious subset of variables within the variable pool, thus providing maximum classification performance with the smallest amount of discriminative features. More specifically, within every classifier, the wrapper algorithm used a Support Vector Machine model (SVM, (Noble, 2006)) to evaluate the discriminative value of each variable within the modality, then extracted the most predictive variable and reiterated over the remaining variable pool to select the 2nd best performing variable, which was added to the first one. This process was re-iterated until the optimal variable subspace had been identified. We stopped the variable search when the top 80% of the variables had been extracted by the wrapper, thus allowing to identify a clinically applicable set of top-performing variables for decision purposes. The wrapper added single features up to the 80% of the total and then tested models. We ranked variables with the mean of training and test data performance, instead of using only the test data.

The preprocessed and trained model was applied to the outer CV cycle (CV2) by preprocessing the best discriminative variables, and determining each validation individual’s class (HR vs. LR) through majority voting across all ensemble models. In other words, in each variable evaluation step in the CV1, the SVM algorithm modeled linear relationships between features and classification labels (higher vs lower resilience). In the linear kernel space, the SVM optimized a hyperplane that maximized separability between most “higher resilience-like” and most “lower resilience-like” subjects (i.e., the Support Vectors). Based on the trained hyperplane, the algorithm then decided subjects' classification of the inner CV1 cycle by projecting its data into the learned kernel space and measuring their geometric distance to the decision boundary. This resulted in a decision value and a predicted classification label per participant.

The wrapper-based feature selection was carried out for each CV1 training and test sample and then repeated for every combination of the SVM parameters C (misclassification cost) within a grid defined by the ranges C = [0.0125 – 16, 8 parameters]. Because of our nested cross-validation framework, we created an ensemble of 17 models (one per site) for each CV2 partition („CV1 ensemble“). We were therefore able to establish a final out-of-training class membership decision for a given individual by combining all CV1 ensembles into a larger CV2 ensemble, in which the given individual had not served for model training and optimization at the CV1 level. This ensemble generation procedure has been repeatedly described in our previous work(Antonucci et al., 2020) and is a feature of the model generation and validation process implemented in NeuroMiner.

To better understand which variables might inform HR and LR classes at the single-subject level, we checked which features were the most reliable. Reliability for each feature is defined in terms of a Cross-Validation Ratio (CVR = mean(w) / standard error(w)) (Koutsouleris et al., 2018). In this formula, w represents the normalized individual weights from SVM models generated in the repeated nested CV scheme. Normalization is performed using the Euclidean norm of w, defined as s= w/||w||2 (Koutsouleris et al., 2018). Higher CVR values represent a high probability for a given variable of being selected in the CV scheme as highly discriminative for classification purposes.

To assign statistical significance to the observed classification, we employed permutation (Golland & Fischl, 2003). We performed 1000 random permutations of the outcome labels. For each permutation, we retrained all linear SVM models in the repeated nested CV experiment using the respective feature subsets obtained from the observed-label analyses. For each permutation, we accumulated the decisions of the random models into a permuted ensemble prediction for each outer cycle subject. Thus, we built a null distribution of out-of-training classification performance (BAC). Finally, we calculated the significance of the observed out-of-training BAC as the number of events where the permuted out-of-training BAC was higher or equal to the observed BAC divided by the number of permutations performed. The significance of the model was determined at α=0.05.

1. **Sanity Checks: association with pharmacological treatment and additional random reshuffling**

We performed some sanity checks in order to exclude (i) that our machine learning algorithm was associated with the type of pharmacological treatment in use by patients, and (ii) that despite the robust double cycle, nested, leave-site-out CV implemented, our algorithm could carry any latent site effect, thus showing poor stability. To carry out the first sanity check, we conducted a 3x2 chi-square (χ2) test between individual classification rates and the type of medication in used (i.e., typical antipsychotics, atypical, both, or none). A classification rate equal to 0 means that the observed and the algorithmic label of a given individual are equal (i.e., the individual has been correctly seen by the machine). We performed the test three times: on the whole sample, only on HR patients, and only on LR patients. A classification rate equal to 1 means that the observed and the algorithmic label of a given individual are unequal (i.e., the individual has been unseen by the machine). Of note, quantitative information regarding pharmacological treatment (i.e., chlorpromazine equivalent scores) were not available in the NIRP dataset. The 4x2 chi-square tests were not significant (whole sample: χ2 =5.42, p=0.66; HR patients: χ2=4.45, p=011; LR patients: χ2=4.56, p=0.10), thus suggesting that the type of pharmacological treatment did not affect the machine learning performance.

To explicitly test for the stability of our machine learning solution, we generated three new random reshufflings of our original NIRP datasets (N=896), leading to three new subjects and sites distributions:

* Reshuffling 1: 595 patients across 17 sites in discovery, 301 patients across 9 sites in validation:

*Table S2 = Characteristics of the Discovery and Validation cohorts randomly generated after Reshuffling 1. Abbreviations: SD = Standard Deviation; M/F = Male/Female.*

|  |  |  |  |
| --- | --- | --- | --- |
| DISCOVERY SAMPLE, Reshuff 1 | No. | Age-median (SD) | Gender (M/F) |
| Whole sample | 595 | 40.35 (10.63) | 177/418 |
| Higher Resilience | 284 | 39.72 (10.13) | 93/191 |
| Lower Resilience | 311 | 40.85 (11.06) | 84/227 |

|  |  |  |  |
| --- | --- | --- | --- |
| VALIDATION SAMPLE, Reshuff 1 | No. | Age-median (SD) | Gender (M/F) |
| Whole sample | 301 | 40.11 (10.73) | 92/209 |
| Higher Resilience | 158 | 39.7 (10.57) | 49/109 |
| Lower Resilience | 143 | 40.56 (10.92) | 43/100 |

* Reshuffling 2: 598 patients across 18 sites in discovery, 298 patients across 11 sites in validation:

*Table S3 = Characteristics of the Discovery and Validation cohorts randomly generated after Reshuffling 2. Abbreviations: SD = Standard Deviation; M/F = Male/Female.*

|  |  |  |  |
| --- | --- | --- | --- |
| DISCOVERY SAMPLE, Reshuff 2 | N | Mean age (SD)  | Gender (M/F) |
| Whole sample | 598 | 39.43(10.72) | 184/414 |
| Higher Resilience | 294 | 39.27(10.31) | 95/199 |
| Lower Resilience | 304 | 39.56(11.12) | 89/215 |

|  |  |  |  |
| --- | --- | --- | --- |
| VALIDATION SAMPLE, Reshuff 2 | N | Mean age (SD)  | Gender (M/F) |
| Whole sample | 298 | 41.95(10.35) | 85/213 |
| Higher Resilience | 148 | 40.72(10.17) | 47/101 |
| Lower Resilience | 150 | 43.18(10.39) | 38/112 |

* Reshuffling 3: 597 patients across 26 sites in discovery, 299 patients across 26 sites in validation:

*Table S4 = Characteristics of the Discovery and Validation cohorts randomly generated after Reshuffling 3. Abbreviations: SD = Standard Deviation; M/F = Male/Female.*

|  |  |  |  |
| --- | --- | --- | --- |
| DISCOVERY SAMPLE | No. | Age-median (SD) | Gender (M/F) |
| Whole sample | 597 | 40.28 (10.65) | 179/418 |
| Higher Resilience | 297 | 39.95 (10.4) | 98/199 |
| Lower Resilience | 300 | 40.6 (10.9) | 81/219 |
| VALIDATION SAMPLE | No. | Age-median (SD) | Gender (M/F) |
| Whole sample | 299 | 40.24 (10.69) | 90/209 |
| Higher Resilience | 145 | 39.36 (10.04) | 44/101 |
| Lower Resilience | 154 | 41.06 (11.24) | 46/108 |

As highlighted by these distributions, in reshuffling 2 and 3 we let the random function assign a patient to the discovery or to the validation sampling without respecting the rule of assigning its entire recruitment center either to the discovery, or to the validation sample. This was done on purpose, to explicitly test that such rule of separation did not affect our machine learning findings.

After each reshuffling, we re-performed the machine learning pipeline described in Manuscript section 2.3 and in Supplementary Information section 4 in each reshuffling independently, without altering the described pipeline. Table S5 shows that the discovery and the validation performance of the algorithm stays stable across reshuffling and that performances of the algorithm on the reshuffling solutions largely overlap the performance of the algorithm described in the main text. Therefore, our machine learning solution for discriminating between higher and lower resilience patients shows very good stability and is free of any latent site or sampling effect.

*Table S5: Machine learning higher vs. lower resilience classification performance in the original discovery and validation division, and across the three reshufflings. Abbreviations: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives; ACC = Accuracy; SENS = Sensitivity; SPEC = Specificity; FPR = False Positives Rate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; AUC = Area Under The Curve; BAC = Balanced Accuracy.*

**

**References:**

Antonucci, L. A., Pergola, G., Pigoni, A., Dwyer, D., Kambeitz-Ilankovic, L., Penzel, N., . . . Bertolino, A. (2020). A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects. *Biological Psychiatry, 87*(8), 697-707. doi:10.1016/j.biopsych.2019.11.007

Bonfiglio, N. S., Renati, R., Hjemdal, O., & Friborg, O. (2016). The resilience scale for adults in italy: A validation study comparing clinical substance abusers with a nonclinical sample. *Psychology of Addictive Behaviors, 30*(4), 509-515. doi:10.1037/adb0000176

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology, 56*(2), 267-283. doi:10.1037//0022-3514.56.2.267

Cohen, I. L., Schmidt-Lackner, S., Romanczyk, R., & Sudhalter, V. (2003). The PDD Behavior Inventory: a rating scale for assessing response to intervention in children with pervasive developmental disorder. *Journal of Autism and Developmental Disorders, 33*(1), 31-45. doi:10.1023/a:1022226403878

Erol, A., Mete, L., Sonmez, I., & Unal, E. K. (2010). Facial emotion recognition in patients with schizophrenia and their siblings. *Nordic Journal of Psychiatry, 64*(1), 63-67. doi:10.3109/08039480903511399

Galderisi, S., Rossi, A., Rocca, P., Bertolino, A., Mucci, A., Bucci, P., . . . Italian Network For Research on, P. (2014). The influence of illness-related variables, personal resources and context-related factors on real-life functioning of people with schizophrenia. *World Psychiatry, 13*(3), 275-287. doi:10.1002/wps.20167

Gerlach, J., Korsgaard, S., Clemmesen, P., Lauersen, A. M., Magelund, G., Noring, U., . . . Casey, D. E. (1993). The St. Hans Rating Scale for extrapyramidal syndromes: reliability and validity. *Acta Psychiatrica Scandinavica, 87*(4), 244-252. doi:10.1111/j.1600-0447.1993.tb03366.x

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. *Information Processing in Medical Imaging, 18*, 330-341. doi:10.1007/978-3-540-45087-0\_28

Gruber, M., Rumpold, T., Schrank, B., Sibitz, I., Otzelberger, B., Jahn, R., . . . Unger, A. (2018). Recover recovery style from psychosis: a psychometric evaluation of the German version of the Recovery Style Questionnaire (RSQ). *Epidemiology and Psychiatric Sciences, 29*, e4. doi:10.1017/S2045796018000471

Hollingshead, A. B., & Redlich, F. C. (1954). Schizophrenia and social structure. *American Journal of Psychiatry, 110*(9), 695-701. doi:10.1176/ajp.110.9.695

Keefe, R. S., Fox, K. H., Harvey, P. D., Cucchiaro, J., Siu, C., & Loebel, A. (2011). Characteristics of the MATRICS Consensus Cognitive Battery in a 29-site antipsychotic schizophrenia clinical trial. *Schizophrenia Research, 125*(2-3), 161-168. doi:10.1016/j.schres.2010.09.015

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., . . . Consortium, P. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry, 78*(2), 195-209. doi:10.1001/jamapsychiatry.2020.3604

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., . . . Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry, 3*(10), 935-946. doi:10.1016/S2215-0366(16)30171-7

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., . . . Consortium, P. (2018). Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry, 75*(11), 1156-1172. doi:10.1001/jamapsychiatry.2018.2165

Lecomte, T., Corbiere, M., & Laisne, F. (2006). Investigating self-esteem in individuals with schizophrenia: relevance of the Self-Esteem Rating Scale-Short Form. *Psychiatry Research, 143*(1), 99-108. doi:10.1016/j.psychres.2005.08.019

Mausbach, B. T., Bowie, C. R., Harvey, P. D., Twamley, E. W., Goldman, S. R., Jeste, D. V., & Patterson, T. L. (2008). Usefulness of the UCSD performance-based skills assessment (UPSA) for predicting residential independence in patients with chronic schizophrenia. *Journal of Psychiatric Research, 42*(4), 320-327. doi:10.1016/j.jpsychires.2006.12.008

Mausbach, B. T., Depp, C. A., Bowie, C. R., Harvey, P. D., McGrath, J. A., Thronquist, M. H., . . . Patterson, T. L. (2011). Sensitivity and specificity of the UCSD Performance-based Skills Assessment (UPSA-B) for identifying functional milestones in schizophrenia. *Schizophrenia Research, 132*(2-3), 165-170. doi:10.1016/j.schres.2011.07.022

Mausbach, B. T., Harvey, P. D., Pulver, A. E., Depp, C. A., Wolyniec, P. S., Thornquist, M. H., . . . Patterson, T. L. (2010). Relationship of the Brief UCSD Performance-based Skills Assessment (UPSA-B) to multiple indicators of functioning in people with schizophrenia and bipolar disorder. *Bipolar Disorders, 12*(1), 45-55. doi:10.1111/j.1399-5618.2009.00787.x

Mucci, A., Rucci, P., Rocca, P., Bucci, P., Gibertoni, D., Merlotti, E., . . . Italian Network for Research on, P. (2014). The Specific Level of Functioning Scale: construct validity, internal consistency and factor structure in a large Italian sample of people with schizophrenia living in the community. *Schizophrenia Research, 159*(1), 144-150. doi:10.1016/j.schres.2014.07.044

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565-1567. doi:10.1038/nbt1206-1565

Peralta, V., & Cuesta, M. J. (1994). Psychometric properties of the positive and negative syndrome scale (PANSS) in schizophrenia. *Psychiatry Research, 53*(1), 31-40. doi:10.1016/0165-1781(94)90093-0

Ribe, J. M., Salamero, M., Perez-Testor, C., Valero, A., & Garcia, M. C. (2015). Spanish version of the social network questionnaire in schizophrenia patients' caregivers. *Schizophrenia Research, 164*(1-3), 273-274. doi:10.1016/j.schres.2015.03.008

Ritsher, J. B., Otilingam, P. G., & Grajales, M. (2003). Internalized stigma of mental illness: psychometric properties of a new measure. *Psychiatry Research, 121*(1), 31-49. doi:10.1016/j.psychres.2003.08.008

Ruschhaupt, M., Huber, W., Poustka, A., & Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology, 3*, Article37. doi:10.2202/1544-6115.1078

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517. doi:10.1093/bioinformatics/btm344

Schennach-Wolff, R., Obermeier, M., Seemuller, F., Jager, M., Messer, T., Laux, G., . . . Riedel, M. (2011). Evaluating depressive symptoms and their impact on outcome in schizophrenia applying the Calgary Depression Scale. *Acta Psychiatrica Scandinavica, 123*(3), 228-238. doi:10.1111/j.1600-0447.2010.01608.x

Shapiro, D. I., Marenco, S., Spoor, E. H., Egan, M. F., Weinberger, D. R., & Gold, J. M. (2009). The Premorbid Adjustment Scale as a measure of developmental compromise in patients with schizophrenia and their healthy siblings. *Schizophrenia Research, 112*(1-3), 136-142. doi:10.1016/j.schres.2009.04.007

Tait, L., Birchwood, M., & Trower, P. (2002). A new scale (SES) to measure engagement with community mental health services. *Journal of Mental Health, 11*(2), 191-198. doi:10.1080/09638230020023570-2

Wallwork, R. S., Fortgang, R., Hashimoto, R., Weinberger, D. R., & Dickinson, D. (2012). Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. *Schizophrenia Research, 137*(1-3), 246-250. doi:10.1016/j.schres.2012.01.031

Westerhof-Evers, H. J., Visser-Keizer, A. C., McDonald, S., & Spikman, J. M. (2014). Performance of healthy subjects on an ecologically valid test for social cognition: the short, Dutch Version of The Awareness of Social Inference Test (TASIT). *Journal of Clinical and Experimental Neuropsychology, 36*(10), 1031-1041. doi:10.1080/13803395.2014.966661