Leveraging Stacked Classifiers for Multi-task Executive Function in Schizophrenia Yields Diagnostic and Prognostic Insights

Tongyi Zhang[1†], Xin Zhao[1†*], B.T. Thomas Yeo[2,3,4,5,6], Xiaoning Huo[7], Simon B. Eickhoff[8,9], and Ji Chen[8,10,11*]

[1]School of Psychology, Northwest Normal University, Lanzhou, China
[2]Centre for Sleep and Cognition & Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[4]N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore
[5]Integrative Sciences & Engineering Programme, National University of Singapore, Singapore
[6]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA
[7]The Third People's Hospital of Lanzhou, Lanzhou, China
[8]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany
[9]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[10]Center for Brain Health and Brain Technology, Global Institute of Future Technology, Institute of Psychology and Behavioral Science,
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai, China 200240
[11]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China

[*]Corresponding authors:
E-mail: psyzhaoxin@nwnu.edu.cn; chen.ji@sjtu.edu.cn
[†]These authors contributed equally to this work.

Abstract

Cognitive impairment is a central characteristic of schizophrenia. Executive functioning (EF) impairments are often seen in mental disorders, particularly schizophrenia, where they relate to adverse outcomes. As a heterogeneous construct, how specifically each dimension of EF to characterize the diagnostic and prognostic aspects of schizophrenia remains opaque. We used classification models with a stacking approach on systematically measured EFs to discriminate 195 patients with schizophrenia from healthy individuals. Baseline EF measurements were moreover employed to predict symptomatically remitted or non-remitted prognostic subgroups. EF feature importance was determined at the group-level and the ensuing individual importance scores were associated with four symptom dimensions. EF assessments of inhibitory control (interference and response inhibitions), followed by working memory, evidently predicted schizophrenia diagnosis (area under the curve [AUC]=0.87) and remission status (AUC=0.81). The models highlighted the importance of interference inhibition or working memory updating in accurately identifying individuals with schizophrenia or those in remission. These identified patients had high-level negative symptoms at baseline and those who remitted showed milder cognitive symptoms at follow-up, without differences in baseline EF or symptom severity compared to non-remitted patients. Our work indicates that impairments in specific EF dimensions in schizophrenia are differentially linked to individual symptom-load and prognostic outcomes. Thus, assessments and models based on EF may be a promising tool that can aid in the clinical evaluation of this disorder.

Introduction

Schizophrenia is a serious mental health condition that can severely impair an individual's functioning and quality of life. Individuals with schizophrenia present a wide range of symptoms of varying severity. These symptoms feature hallucinatory and delusional experiences termed "positive symptoms" as well as negative symptoms which manifest as atypical emotional and social behaviors. Distinct from negative symptoms, though related, cognitive symptoms involve the impairment of mental functions related to memory, attention, and executive tasks. This greatly affects the ability to live independently given the difficult to treat these symptoms using the currently available antipsychotic medications[1]. The cognitive symptoms in schizophrenia are a core aspect of psychopathology; they are considered a trait marker that emerges in the prodromal phase and persists throughout the illness[2], unlike those manifested in affective psychotic disorders or drug-induced psychosis where cognitive deficits are epiphenomenal. However, cognitive performance in patients with schizophrenia is heterogeneous, and can vary from virtually unaffected to severely impaired[3,4]. While there is mixed evidence regarding the impairments in some cognitive domains, executive dysfunction is pervasively abnormal in schizophrenia. Previous studies have consistently indicated that mild-to-severe deficits in processes are related with executive functions (EFs)[5].

EF represents a series of higher-order cognitive processes that involve impulse control and behavior orchestration[6]. Deficits in this domain can hinder goal-directed activity and contribute to aggression, violence, and poor compliance to medication in patients with schizophrenia, leading to worse clinical outcomes[7]. In the field of psychiatry, prognostic prediction remains a significant challenge in research and clinical practice, though it is crucial for early assessment and intervention. Previous studies employing baseline neuroimaging, genetic, or clinical data only approached chance-level accuracy in most cases[8,9]. This emphasized the lack of reliable markers for tracking the disease trajectory[10,11]. EF deficits emerge in the early stages (e.g., ultra-high risk, first episode) of schizophrenia[12,13] and have been related to disease progression, symptom severity, and recovery of social and occupational skills[14]. Hence, they might be potential markers for tracking the clinical courses and prognostic statuses of schizophrenia.

A broad range of EF impairments has been associated with this disorder, including increased difficulties in inhibiting automatic responses and switching to new ones, reduced cognitive flexibility, and disturbances in the maintenance and updating of goal-related or rule-based information in working memory. These actually align well with the three-factor model of EF proposed by Miyake et al.[15], which characterizes 1) interference inhibition and response inhibition; 2) cognitive flexibility and switching; and (3) working memory updating and maintenance. This three-dimensional representation of EF functions robustly captures individual variation in EF subcomponents across a broad spectrum of age groups and clinical cohorts, including patients with psychiatric disorders[16,17]. These dimensions of EF differ in concepts and neurobiological substrates, highlighting the need to consider and assess these dimensions, while studying their correlations with diagnostic and prognostic aspects of schizophrenia. Such finer characterization of EF functions may assist in understanding different psychopathological processes in schizophrenia (e.g., disorganization symptoms), which manifest as difficulties in the goal-directed sequencing of thoughts and behaviors[18,19]. These symptoms are linked to the cognitive dimension in our recently introduced four-dimensional representation (positive, negative, cognitive, and affective) of schizophrenia psychopathology, as generalizable across populations and clinical settings[20]. Disturbances in EF have likewise been implicated in negative and positive symptoms. Firstly, failure in effectively monitoring volitional behaviors and inhibiting false inference in predictive processing would have consequences on positive symptoms (e.g., hallucinations and delusions)[21,22]. Secondly, cognitive rigidity hampers adjustment of thoughts and actions for environmental volatility[23]. Thirdly, impaired working memory updating and maintenance has been linked to poor abstract thinking[24], which can increase the risk of developing negative symptoms (e.g., apathy and diminished expressive behavior) commonly observed in schizophrenia[23,25]. However, some dimensions of EF may be more severely affected than others in schizophrenia[26]. Currently, it remains unclearly the abnormalities in which EF dimensions characterize schizophrenia and would play a role in providing prognostic information.

Assessment of the underlying processes of the dimensions of EF is a challenge.

Currently available tools (e.g., Cambridge Neuropsychological Test Automated Battery[27] and the National Institutes of Health Toolbox[28]) do not feature a comprehensive assessment to cover the various dimensions of EF. Furthermore, assessments included in available neuropsychological batteries (e.g., Delis–Kaplan Executive Function System) are paper tests rather than experiments. Consequently, they do not offer a trial-by-trial based dynamic quantification, e.g., the reaction time in a sequential task. Trial-by-trial responses help detect subtle cognitive impairments in schizophrenia, including EF[29,30]. In addition, the cognitive symptom items routinely used in clinical practice, including EF rating such as in the Positive and Negative Syndrome Scale (PANSS), are retrospective based on information collected from patient interviews or contributions by relatives. Comparatively, task paradigms drawn from the cognitive psychology literature provide objective trial-by-trial tests that facilitate measuring particular cognitive functions with likely improved sensitivity and specificity[16,31]. Such a tailored assessment strategy would be ideal for investigating the diagnostic and prognostic value of EF dimensions in schizophrenia by establishing classification models. Previous psychiatric machine learning studies mainly considered single algorithms, such as support vector machine (SVM) or random forest (RF), comparing their respective accuracies[32]. Methods such as stacking, a mainstay multi-view learning approach, may be another strategy for improving model performance[33].

In this study, we systematically applied six well-established behavioral paradigms to assess individual baseline functions along the three EF dimensions (i.e., inhibitory control, working memory maintenance and updating, and cognitive flexibility) to determine their consistency in characterizing patients with schizophrenia and their prognostic statuses at follow-up. This is tested via establishing diagnostic and prognostic classification models using machine learning methods: SVM, RF, Adaptive Boosting (AdaBoost), and their stacked model with a stringent nested cross-validation (CV) and independent testing. The importance of EF feature contributions to classification models was determined via the SHapley Additive exPlanations (SHAP) approach, which facilitates the identification of important features at both the group- and individual-levels. This approach enables a link between feature importance and individual psychopathology (Fig. 1).

Results

Participant identification, screening, and follow-up

In this hospital-based study, we initially identified 270 individuals with schizophrenia (International Classification of Diseases, tenth edition [ICD-10], additional screening with the Structured Clinical Interview for DSM-IV axis I Disorders) aged 18–65 years, from a total of 580 inpatients in the psychiatric department of the Third People's Hospital of Lanzhou (Lanzhou, China) (Fig. 2). These 270 patients with schizophrenia were clinically stable (change in total PANSS score<20% with a particular type of antipsychotics at a maintenance dosage within the last 6 weeks; details in Supplementary Table S1). Furthermore, 75 of the 270 patients with schizophrenia were excluded due to illiteracy (*N*=35) or refusal to participate (*N*=40). Finally, 195 patients were included. The study was approved by the ethics committees of Northwest Normal University and the Third People's Hospital of Lanzhou (Lanzhou, China). They underwent evaluations that included the electronic medical records, the Positive and Negative Syndrome Scale (PANSS), six EF behavioral tasks, and the fluid intelligence (Raven's Progressive Matrices). To enable comparative analyses, 169 demographically matched healthy control participants who were free of a history of mental illness or brain injury were recruited and underwent the same assessments, except the PANSS.

Among the 195 patients with schizophrenia, 86 participants completed follow-up assessments (PANSS and EF tests) in 4–6 weeks after the initial evaluation. The loss of 109 participants at follow-up was primarily attributed to hospital discharge (*N*=98) or concerns regarding the potential impact of assessment results on their discharge timing (*N*=11) (Supplementary Table S2). Fifty-eight participants at follow-up identified as being in remission; who had scores ≤ 3 on key PANSS items (P1, P2, P3, N1, N4, N6, G5, G9; the Remission in Schizophrenia Working Group [RSWG] criteria[34]). The other 28 participants were not in remission (Supplementary Table S2). Alternatively, defining the remission status by a reduction in PANSS total score at follow-up assessment relatively to baseline showed that[35–37]: 1) a 25% reduction (68 remitted vs. 18 non-remitted); 2) a 35% reduction (48 remitted vs. 38 non-remitted); 3) a 50% reduction (16 remitted vs. 70 non-remitted)

(Supplementary Table S2). In addition, prognostic statuses may also be reflective in a change of subtype membership from baseline to the end of follow-up, particularly surrounds the negative symptom subtype. This is because many patients tend to experience increased negative symptoms and diminished positive symptoms due to standard antipsychotic treatments and related factors at follow-up, while others feature primary (stable) negative symptoms[38] which relate to poorer clinical outcomes[39]. By using a subtyping system (*http://webtools.inm7.de/sczDCTS/)* from our previous work, patients with schizophrenia were assigned based on their symptom patterns as predominantly negative, positive, or ambiguous cases[20]. Of the 86 patients, 64 non-negative subtype patients (50 ambiguous and 14 positive) at baseline had a negative subtype assignment at follow-up assessment. Thirteen patients with schizophrenia maintained their negative subtype membership over time.

EF dimensions

EF dimensions are consistently and differentially affected in schizophrenia

Six behavioral paradigms (e.g., Zhao et al., 2023)[40], were administered to assess the three EF dimensions (inhibitory control, working memory maintenance and updating, and cognitive flexibility; Fig. 3A), based on 14 measurements: 1) The inhibitory control dimension: four measurements for the interference control function based on the *Stroop* task, three measurements for the response inhibition function based on the *Go/No-Go* task; 2) The working memory dimension: two measurements for working memory updating function based on the *running memory* task, three measurements for numeric working memory maintenance capacity based on the *Corsi block* test and the *digit span backward* task; 3) The cognitive flexibility dimension: two measurements (switch cost and mixing cost) in the *number-letter switching* task. Besides a conceptual formulation of the 14 task measurements according to the three-dimensional representation of EF, we supplemented five composite scores. This included an *Inhibition* composite score, an abbreviated version for representing general EF functions, and three cross-dimensional composite scores (*Inhibition/Switching, Inhibition/Working memory updating and Switching/Working memory updating*). Among these measurements, all reaction times,switching cost, and three composite scors (abbreviated

general EFs, *Inhibition/Switching, Inhibition/Updating*) are the higher the worse EF performance, while the remaining the higher the better.

To evaluate whether schizophrenia differentially affected the three-dimensional EF measurements and the five EF composite scores, we performed a mixed-model analysis of covariance (ANCOVA). This has revealed a significant two-way interaction between group (schizophrenia vs. HC) and EF measurements (p<0.001). Follow-up one-way ANCOVAs revealed that, except for 2 measurements which assess working memory maintenance based on the *Corsi block* test, the accuracy in No-Go trails, the difference of reaction time between the congruent and the incongruent condition trials (i.e., the interference effect) in the Stroop task, the mixing cost in the switching task and the two composite scores of *Inhibition/Working memory updating* and *Inhibition*, other measurements and composite scores were all significantly different between patients with schziohrenia and healthy controls (all p<0.05, false discovery rate [FDR] corrected) (Table 2; Fig. 3B). These differed EF performance cover all of the three EF dimensions with the reaction times for Go trails (response inhibition) in the *Go/No-Go* task presenting the largest effect size (η^2= 0.160) (Fig. 3C, Table 2).

Remitted patients show improved interference inhibition at follow-up, without difference in any EF dimension at baseline, compared with non-remitted patients

Further analyses were conducted to examine differences in baseline EF dimensions between the remission and non-remission groups (RSWG criteria); there were no significant between-group differences observed on any EF dimension (Supplementary Table S3). When comparing respective changes in these dimensions from baseline to follow-up (Supplementary Table S5), significant differences were only observed in the remission group. Specifically, on the Stroop task, the remission group demonstrated shorter reaction times under neutral, congruent, and incongruent conditions compared with baseline (i.e., better interference inhibition ability; all p<0.05).

Regarding clinical outcomes defined by subtype transmission, there were no significant differences in baseline EF measurements among the three subgroups (i.e., positive, negative, and ambiguous) of patients with schizophrenia (all p>0.05, FDR corrected). In patients with

more prominent secondary negative symptoms (i.e., baseline non-negative subtype with transition to a negative subtype), the Stroop task revealed significantly reduced reaction times at follow-up compared with baseline for incongruent stimuli (p=0.022), congruent stimuli (p=0.007), and neutral stimuli (p=0.037). Additionally, the mixing cost in the switching task was significantly lower at follow-up (p=0.048). However, patients of the stable negative subtype during follow-up did not show significant differences between baseline and follow-up across all EF measurements. There were no significant differences observed in baseline EF measurements between the secondary negative group and the primary negative group (all p> 0.05).

Psychopathology

Psychopathology dimensions specifically correlated with different EF measurements

Previous studies showed that EF, and specifically their dimensional measurements, could be associated with different aspects of symptomatology[41]. We probed the potential association of multifaceted aspects of psychopathology with different EF functions at the baseline assessment using Pearson correlation analysis (Supplementary Fig. S4). Using the four-dimensional structure of the PANSS[20], we observed significant reductions in severity across negative, positive, cognitive, and affective symptoms in patients with schizophrenia at follow-up (p<0.001, Supplementary Table S6). Among the 86 follow-up patients, the ability to inhibit conflict, as reflected by reaction times to incongruent (r=0.212, p=0.050) and congruent stimuli (r=0.246, p=0.022) in the *Stroop* task, were significant correlated with baseline positive symptom scores. Patient capacity to maintain and shift mental sets, quantified by switch cost (r=0.298; p=0.005) and mixing cost (r=0.265; p=0.014) in the *number-letter switching* task, was significantly correlated with baseline cognitive and positive symptom scores, respectively, within the follow-up subset but not in the overall patient sample. For comparison, correlation analyses were repeated using PANSS three-original subscale scores; these yielded similar results, except for correlations with *Stroop* metrics (Supplementary Fig. S5).

EF dimensions show consistent and distinct impairments in schizophrenia and offer prognostic insights via machine learning classification models

Using a multivariate approach, classification models can be employed to identify feature variables (e.g., EF measurements) which can reliably differentiate target cases (i.e., patients with schizophrenia) from reference cases (i.e., HC). Additionally, they can provide insights into future categories (e.g., prognostic status) based on baseline assessments. Three methods (i.e., RF, SVM, and AdaBoost), and their stacked assembler, were used to construct classification models. The original data was repeatedly split into discovery and test sets, with each discovery set nested for hyperparameters tuning and model validation as a CV design. Next, the ensuing best model was applied to the test set from each repeat to obtain out-of-sample performance. The whole procedure was repeated for 100 tests, resulting in 100 hold-out, test sets. This approach has been demonstrated to effectively gauge generalization while balancing practical acquisitions of clinical sample data[42–44].

1) Diagnostic classification

For our classification experiments, two feature sets were used: 1) 19 EF assessments, which reflect three EF dimensions measured by six behavioral paradigms (Table 2); and 2) 32 features, which added 13 sociodemographic variables to the 19 EF variables in feature set 1 (Table 2). Performance metrics were assessed on the 100 test sets (Supplementary Table S7). For the feature set relying on only EF assessments (i.e., feature set 1), we aimed for a model to classify new patients, regardless of sociodemographics. The highest out-of-sample classification was achieved by the stacking model (area under the curve [AUC]=0.87). The feature set 2, which also included sociodemographic variables, was aimed at incorporating information from routine clinical interviews that indicates disease susceptibility[45]. With the addition of sociodemographic variables, improved model performance was observed (stacking model AUC=0.91). In addition, repeating the entire CV process on the EF assessments after controlling for the effects of sociodemographic variables largely maintained the performance (stacking model AUC=0.80) (Fig. 4A).

2) Prognostic classification

For prognostic classification, we applied the same models (RF, SVM, AdaBoost, and their stacking, as described in the diagnostic classification above) on two feature sets, to assess their performance in discriminating remission status: (1) 19 baseline EF variables; (2) 32 features (the 19 baseline EF variables plus 13 sociodemographic variables). The stacking model achieved the highest classification accuracy based on the EF assessments-only feature set (AUC=0.81), and the performance was identical to that observed when the EF plus sociodemographic features set was used.

Additionally, to test the influence of demographic variables and medication on our models, we conducted several control analyses (Fig. 5A). By regressing out the effects of sociodemographic variables on EF assessments, we found that the model performance was decreased to AUC=0.65, indicating a poorer setup (Fig. 5A). However, there was no significant difference in any of the sociodemographic variables adjusted in our classification models between the remitted and non-remitted patient subgroups (all $p>0.05$) (Supplementary Table S3). Additionally, controlling for medication effects using an olanzapine (OZP) equivalent dosage (which did not differ significantly between remitted and non-remitted patients; $p=0.15$) diminished the prognostic classification accuracy to an AUC of 0.72. By investigating only those patients ($N=46$) treated with a commonly effective OZP-equivalent dosage of 10–20 mg/day in clinical practice produced a similar prognostic classification performance (AUC=0.82) to that recorded in the group of followed up patients ($N=86$).

We supplemented classifications to discriminate prognostic subgroups defined by the reduction in re-assessed total PANSS score. The results showed a poorer discriminative power, with AUCs of 0.74, 0.58, and 0.73 (Fig. 5A; Supplementary Table S8) for a 25%, 35%, and 50% reduction in the PANSS score, respectively. Alternatively, we established a classification model for the clinical outcomes of patients based on subtype-membership transition to distinguish between 1) baseline non-negative subtype with a transition to negative subtype (secondary) and 2) stable negative subtype during follow-up (primary). A promising classification performance was revealed (AUC=0.81).

Moreover, two sensitive analyses were performed to take the attribution condition in our

study into consideration. The demographic variables, symptoms, OZP-equivalent dosage, and EF measurements did not differ significantly between patients who were followed up and those who were not (all $p>0.05$, Supplementary Table S3). Furthermore, by treating the attribution patients as best cases (i.e., all remitted), the classification accuracy decreased slightly to an AUC of 0.75, while treating the attribution patients as worst cases (none remitted) yielded a further decease (AUC=0.68).

Feature importance and association with individual psychopathology

SHAP analysis was performed on the best-performing classifiers trained on the 19 EF dimension assessments (for the feature sets including sociodemographic variables please refer to Supplementary Table S9)[46]. The goal was to determine the directional contribution of EF dimensions for classification informed by decision path (i.e., better or worse performance in an EF dimension drives the model to assign a schizophrenia or HC label). Besides group-level determination of important dimensions, the Shapley value for each EF feature in the best-performing classifiers was calculated for each individual[47,48], facilitating a link to the expression level of patients along several psychopathological dimensions.

1) *Group-level feature importance and decision path*

Absolute Shapley values derived from the best-performing classifier (highest AUC) identified in the 100 test sets were used to rank each EF dimension to indicate its importance in discriminating patients with schizophrenia, and those remitted at follow-up (Figs. 4B, 5B). Including only the 19 EF dimensions scores as the feature set in both diagnostic and prognostic classifications, the inhibition control dimension—comprised of response and interference inhibitions—ranked highest in both classifying schizophrenia group participants and their follow-up remission status. Within the inhibition control dimension, important features were from the *Go/No-Go task*, which assesses response inhibition (Go trial accuracy and reaction time), and the Stroop task, which assesses interference inhibition (reaction time for neutral stimuli in the diagnostic model, and reaction time for incongruent stimuli in the prognostic model). Adding sociodemographic variables to the diagnostic and prognostic classifiers generally replicated inhibition control as the strongest contributing dimension

(Supplementary Figs S6A, S7A).

Next, group-level decision path analysis of each EF dimension identified that poor performance on any of these drove the model to correctly classify individuals as patients with schizophrenia (Fig. 4C). However, worse performance on the inhibition control dimension increased the likelihood of non-remission status classification (Fig. 5C). These results were replicated in additional models in which sociodemographic variables were included in diagnostic and prognostic classifiers (Supplementary Figs. S6B, S7B).

2) *Individual-level decision path and association with psychopathology*

Decision path analysis was likewise conducted at the individual level, to determine the relative performance of EF dimensions (as identified in group averages) for correctly assigning individuals. Among correctly identified participants, patients with schizophrenia scored below the averages of both HC and schizophrenia groups on at least one EF dimension (Fig. 4D). As expected for the prognostic classification model, remitted status for most participants in the schizophrenia group was correctly assigned based on higher baseline inhibitory control dimension (including both interference control and response inhibition) performance compared with the averages of both remitted and non-remitted participants. Specifically, remitted participants showed shorter reaction times to the incongruent condition in the *Stroop* task, and higher accuracy in the response to the Go trials during the *Go/No-Go* task. However, a few patients presented both worse performance in inhibitory control and higher baseline abilities in other dimensions such as working memory updating or shifting (Fig. 5D; Supplementary Fig. S9). These findings were replicated when sociodemographic variables were included in diagnostic classification models (Supplementary Materials).

Pearson correlation analysis was further performed on Shapley values for each EF feature and scores on the four symptom dimensions, across the overall schizophrenia group (N=195) and follow-up subset (N=86). Results showed that individual Shapley values of the interference inhibition function, as assessed by a difference in reaction time between the congruent and the incongruent trials in the *Stroop* task, that promoted a correct assignment of cases versus HC were significantly associated with the negative symptoms (r=0.439, p=0.042, FDR corrected) for individuals with schizophrenia. The inference inhibition function, though

assessed by a different behavioral metric, was similarly identified as a factor of top importance at the group-level for the diagnostic model described above (Fig. 6A). After including the 13 sociodemographic variables, there was no significant correlation found.

For prognostic classification, the importance of working memory updating (assessed in the *running memory* task; r=0.618, p=0.023, FDR corrected) and maintenance (assessed in the *Corsi block* test; r=0.597, p= 0.031, FDR corrected) in the model that accurately assigned remitted patients was correlated with low-level cognitive symptoms at follow-up. Additionally, working memory updating function contributing to the accurate assignment of remitted patients was associated with more severe re-assessed negative symptoms (r=0.596, p=0.031, FDR correction). Following the inclusion of sociodemographic variables in the model, the significantly associated EF variables were changed (Fig. 6C).

Using PANSS three-subscale scores in correlation analyses did not reveal significant correlation with the importance scores of any EF features of the dimensions identified in the diagnostic models. Significant correlation patterns among the importance scores of EF dimension features in prognostic models were a subset of those reported (Supplementary Fig. S10) when using the four dimensions of psychopathology, as described above.

Discussion

This study is the first to investigate the classification power of comprehensively three EF dimensions (i.e., inhibitory control, working memory maintenance and updating, cognitive flexibility) for discriminating both patients with schizophrenia from HC and determine their remission status at follow-up. Importantly, our SHAP approach parsed the relative importance of each feature in these classification tasks, at the group and individual level. Collectively, we found that EF dimensions consistently and differentially characterize schizophrenia and are informative regarding the prognostic status, though certain dimensions are more closely linked to the disease trait and related psychopathology.

The four primary findings are as follows. Firstly, EF assessments could be used to both classify patients with schizophrenia (AUC=0.87) and to identify those with remitted status

(AUC=0.81). Importantly, there is no significant difference in baseline EF and symptom severity between the two prognostic subgroups. Secondly, inhibition control was the most strongly contributing dimension to patient classification of both schizophrenia and remission outcome. Thirdly, at the individual level, correctly identified patients presented below-average performance on at least one EF dimension. However, except for a few patients with correctly assigned remission status who had worse performance in inhibitory control, others featured higher baseline performance in this dimension compared with the averages of both remitted and non-remitted patients. Finally, the EF dimension interference inhibition, which is important in promoting the correct classification of patients with schizophrenia by the model, was significantly associated with patient negative symptoms. The model importance of working memory for accurate remission assignment covaried with low-level follow-up cognitive symptoms.

The paradigms measuring EF and its dimensions are readily available through software platforms, such as *E*-Prime and Matlab-based Psychtoolbox. Thus, they can be implemented in clinical practice. The present findings may encourage the use of a feasible, low-cost, and effective approach to schizophrenia diagnosis and psychopathology evaluation.

Schizophrenia, and its remission status, are classified by EF dimensions

Previous machine learning studies employing neuropsychological test batteries (e.g., the Cambridge neuropsychological test automated battery, the Wechsler adult intelligence scale, and the brief assessment of cognition in schizophrenia) to differentiate patients with schizophrenia from HC have yielded accuracy rates <70%[8,49]. Neuroimaging-derived assessments are an alternative and broadly attempted approach[50]. While showing promise for improving classification performance[51,52], this strategy is associated with other challenges[53] including heterogeneous data acquisition, high dimensionality due to large numbers of voxels or measures, and limited applicability in low-income countries and regions, as illustrated by our recent meta-analysis of global psychiatric neuroimaging data[54]. Another direction is systematic modeling based on readily attenable data with improved objectivity and reliability, namely behavioral EF tasks, for improved clinical translation[10,55]. By carefully assessing three EF dimensions via six tasks, and thus 19 variables, our stacking model achieved

reasonably high diagnostic accuracy. Furthermore, the inclusion of sociodemographic characteristics increased the AUC to 0.91. This is broadly consistent with previous findings of an association between sociodemographics and disease susceptibility, clinical course, and symptom expression in schizophrenia[45]. In clinical practice, early and accurate prediction of remission outcomes holds significant implications for effective treatments. Nevertheless, this remains difficult, with recent models based on baseline neuroimaging, genetics, and clinical factors often producing accuracy rates marginally better than the chance level (i.e., 50%). Comparatively, our prognosis classification model, incorporating only baseline EF assessments, demonstrated improved performance in denoting remission status among patients with schizophrenia at 4–6-week follow-up (AUC=0.81). We moreover tested the classification models by two methods for defining remission, i.e., based on a reduction in re-assessed PANSS scores calculated with specific items or all items, to assess the robustness of our findings to the definition of treatment response[36]. The discriminative power of the classification model was higher when specific items, compared with all items, of the PANSS were involved in defining prognostic statuses. This evidence implied specificities in mapping EF dimensions and symptom recovery in patients with schizophrenia.

Classification-important EF dimensions and associations with individual psychopathology

Leveraging the SHAP framework, we conducted feature importance analysis on our classification models to identify the relative contributions of each EF dimension variable.

Diagnostic classification

At the group level, we found that the inhibitory control dimension (i.e., interference inhibition and response inhibition) ranked highest in importance for classifying participants in the schizophrenia group. This is consistent with previous works showing abnormal alterations in the temporal and spatial characteristics of inhibition-related brain responses and behaviors in schizophrenia[56,57]. Moreover, our decision path plot pertaining to all correctly classified individuals indicated that those who performed poorly on any EF dimension tended to be classified in the schizophrenia group. This emphasizes the general EF impairments

among patients with schizophrenia, consistent with a previous meta-analysis showing that EF deficits within this patient population cover broad dimensions[26]. Considering individual variation in this context through decision path analysis for each participant, we noted that a few accurately identified patients with schizophrenia performed slightly above average on either of the three EF dimensions (e.g., inhibition, updating, or shifting). This aligns well with data showing that cognitive performance, including EF functions, in schizophrenia can vary from mild deficiency[58,59] to severely impaired[60,61], and connects the neuropsychological and neurobiological heterogeneity systematically observed among these patients[62,63].

By extending the SHAP framework to individual-level analyses, we further linked the importance of each EF feature in diagnostic classifier—quantified by Shapley values—to the landscape of individual psychopathology. Notably, we found Shapley values from the dimension identified as important at the group-level—interference inhibition—to be valuable within the model correctly classifying patients with schizophrenia, and significantly associated with a patient's negative symptom expression. Consistent with our finding, earlier research showed that individuals with more severe negative symptoms tend to have diminished inhibitory control, as measured by the *Stroop* task[64]. This deficiency, particularly within interference inhibition, might play a role in the manifestation of negative symptoms, which may be understood through the lens of target-speech recognition deficit in schizophrenia[64]. Specifically, the interference inhibitory function assists individuals in accurately segregating target speech from noisy background environments in which multiple speakers are talking simultaneously[65]. An impairment in this function would thus be connected to the disorganized speech information processes in schizophrenia, including the inability to either inhibit unrelated speech signals or capture desired speech signals[65]. Such impairment has been correlated with the severity of negative symptoms, including poverty of speech and hypobulia[66].

Prognostic classification

Interestingly, while inhibitory control contributed to diagnostic classification, it was also the top predictor of remission outcome at follow-up. Our decision path analysis showed that patients who were correctly classified as remitted generally (despite some exceptions)

showed good baseline performance on inhibitory control tasks. Previous work has consistently demonstrated an association between EF and long-term post-treatment remission outcomes in patients with schizophrenia[67]. Specifically, patients with higher EF performance are more likely to remit relative to those with lower EF performance[68,69], especially on the inhibition control dimension[70]. A possible interpretation is that patients with better inhibitory control are more adherent to therapeutic plans, including pharmacological interventions and lifestyle modifications[71,72]. Alternatively, because inhibitory control assessments are closely related to specific clinical manifestations in patients with schizophrenia[73–75], better performance on this EF dimension along with milder symptom expressions among such patients, implies higher chance of remission[68]. Our results showed significant post-treatment improvement in interference inhibition in the remission subgroup, but not in the non-remission subgroup (Supplementary Materials), corroborating a relationship between inhibitory performance and remission in schizophrenia. Interestingly, we did not find significant difference in any baseline EF assessment and symptom dimension score between remitted and non-remitted patients. This points to a dissociation between symptoms of schizophrenia and the construct of EF; nonetheless, it highlights the role of EF dimensions in predicting remission rather than merely acting as a marker of illness severity.

A few patients who were correctly classified as having symptom remission exhibited poor inhibitory control performance; nevertheless, they demonstrated better baseline abilities in other dimensions, such as working memory updating or shifting. Furthermore, we revealed that the importance of working memory updating and maintenance contributions to the prognostic model in classifying remission status covaried with poor patient cognition at follow-up. Supporting this observation, previous studies have shown that working memory (updating and maintenance) performance is superior among stably remitted patients with schizophrenia versus non-remitted patients[68]. Disrupted working memory in patients with schizophrenia has been linked to cognitive disorganization and poorer performance on tasks requiring abstract thinking[24,76,77], which is similarly assessed within our cognitive symptom dimension based on PANSS.

Limitations and considerations

First, applied machine learning tends to rely on multiple datasets from independent medical centers for extensive tests. However, such concerns are moderated by our use of multiple random-splits to set aside a test 'lock box' in each repeat, while performing nested CVs on the remaining sample, as well-established previously[50,53]. This strategy effectively gauges the out-of-sample generalization performance, while balancing practical clinical data collection issues[42,43,78,79]. Nevertheless, future multisite and population-level EF studies may help expand the applicability. Second, patients with schizophrenia in our study had been treated with antipsychotics, reflecting typical clinical practice. In our sample, the OZP equivalent dosage did not show a significant correlation with most symptom scores (except for the affective symptom dimension) or EF measurements (except for one variable). Moreover, there was no significant difference in baseline OZP dosage between patients who were in remission and those who were not at follow-up. Nevertheless, the exact dosage of antipsychotic medication could have an impact on individual prognostic status[80]. As expected, the classification accuracy decreased when adjusting for individual variations in OZP equivalent dosage. However, a control analysis that only included patients who received the suggested starting dose for OZP (e.g., 10–20 mg/day)—a dose commonly effective in individuals with schizophrenia[81]—remained the classification performance as in our main experiments. Notwithstanding, future research involving drug-naïve patients at baseline and continuous assessments of EF function along with detailed records of medication usage over time would help establish the causal relationships between antipsychotic effects, prognostic statuses, and specific EF dimensions. Third, the attrition rate in our patient sample was similar to those reported previously.[82–84] This rate pertains to the representativeness of the findings derived from the patients who continued in the study, though we did not observe significant differences in all of the baseline characteristics between the followed up and drop-out patients (Supplementary Table S3). Furthermore, classification models established based on the two extreme conditions (i.e., treating the attribution patients as best-case [all remitted] or worst-case [non-remitted] scenarios) showed decreased prognostic discrimination accuracy. This was in line with the previous notion[85,86] that the worst case

analysis would lead to underestimated results[87], though these might not reflect the true potential attrition bias. Future research may also incorporate outpatients to develop models representing the broader spectrum of patient populations, but managing the attrition rates remains a challenge.

To conclude, here we tested the classification power of six well-established behavioral paradigms, which assess three EF dimensions, for discriminating patients with schizophrenia from HC at baseline, as well as the remission status at follow-up. Results from robust validation and testing revealed promising performance and, thus, a consistent impairment in dimensions of EF to characterize individuals with schizophrenia and provide important prognostic information. Furthermore, different EF dimensions characterized diagnosis and prognosis to varying extents. Inhibitory control and working memory were identified as the most important factors for accurate classification of schizophrenia and remission status. Additionally, the classification strength of these EF dimension features was associated with specific psychopathologies. Thus, our research presents evidence that certain dimensions of EF are reliably compromised in individuals with schizophrenia. These deficits correlate with the severity of symptoms and can predict future outcomes. Hence, these measures may serve as valuable aids in the clinical assessment of this disorder.

Methods

Ethics approval and consent

This study was conducted in full compliance with the ethical guidelines and approved protocols of the Ethics Committee at the Third People's Hospital of Lanzhou City and Northwest Normal University (Lanzhou, China). Before participation, all participants involved in the study were provided with comprehensive information regarding the study aims, procedures, potential risks, and benefits. The study adhered to the principles outlined in the Declaration of Helsinki, and informed consent was provided by all participants.

Data collection

Participant recruitment, clinical characterization, and definition of prognostic status

The study was conducted from March to August, 2023. Participants were 195 individuals who had been diagnosed with schizophrenia (of those, 86 were further assessed after a 4–6-week interval), and 169 healthy individuals (HC group) (Table 1). Participants in the schizophrenia group had received inpatient treatment at the Third People's Hospital in Lanzhou City within the past 2 years. Diagnoses were reached by two resident psychiatrists using the ICD-10 diagnostic criteria for schizophrenia (F20.9). The participants were further screened with the Structured Clinical Interview for DSM-IV axis I Disorders. The patients were in a stable condition and receiving consistent treatment, with no changes in medication expected during the study. Inclusion criteria were age 18–65 years and the ability to communicate effectively, complete experimental tasks, and voluntarily sign the informed consent form. Individuals with a severe physical disease, visual abnormality, or adverse drug reactions were excluded from the study (details for the inclusion and exclusion criteria are listed in Supplementary Table S1). Participants in the HC group, recruited through offline promotions and online advertisements, were matched with those in the schizophrenia group for age, sex, education level, and socioeconomic status. All HC group participants were physically healthy and did not have a history of mental illness or brain injury.

Symptom severity in each patient with schizophrenia was evaluated using the PANSS[87]. Scores for four symptom dimensions (i.e., positive, negative, affective, and cognitive factors) were derived for each patient via the Dimensions and Clustering Tool for Schizophrenia Symptomatology (DCTS; *http://webtools.inm7.de/sczDCTS/*). These dimensions have been previously identified as stable and generalizable across populations, regions, and clinical settings[20]. Higher scores denote more severe symptoms within each dimension. Patients were further categorized into the positive subtype, negative subtype, or ambiguous cases lying in-between these two subtypes based on their DCTS-derived symptom dimensional scores and membership values. We employed a heuristic membership degree of 0.6 as the cutoff value for the ambiguous cases subgroup given that these participants were not clearly assigned to any of the two more differentiated negative-positive subtypes.

For the 86 followed up patients with schizophrenia, remission versus non-remission prognostic statuses were determined based on the RSWG criteria[34]. Remission is defined by scores ≤ 3 on key PANSS items (P1, P2, P3, N1, N4, N6, G5, G9). Considering the current lack of consensus on a definition of clinical outcomes in schizophrenia, we utilized an alternative definition. This definition was based on the reduction of the total PANSS score, and three remitted or non-remitted conditions were specified by: 1) a 25% reduction; 2) a 35% reduction; and 3) a 50% reduction[36]. Clinical outcomes of patients were moreover defined based on subtype-membership transition: 1) baseline non-negative subtype with transition to a negative subtype, and 2) stable negative subtype during follow-up. This definition helps in identifying patients that experience primary (and stable) negative symptoms or secondary symptoms due to antipsychotic treatments and related factors[38,39].

Assessments

Sociodemographic and electronic records

The standard 60-item Raven's Progressive Matrices test was used to assess fluid intelligence[88]. Family socioeconomic status was assessed using a family financial status questionnaire[89]. Thereafter, we collected detailed clinical information from the electronic medical records of patients (e.g., disorder onset, number of episodes, age at diagnosis, duration of illness, types and dosages of antipsychotic drugs, family medical history).

EF

This study was based on the influential model subdividing EF into three core dimensions: inhibitory control; working memory (updating and maintenance); and cognitive flexibility/shifting[15,90]. Working memory updating is the process of continuously replacing old information with new in working memory, according to current task requirements. Working memory span/maintenance is the ability to maintain and process information over a period of time, often directly linked to short-term memory capacity[40]. Inhibitory control involves the ability to suppress dominant responses and adapt to a changing environment, minimizing the impact of irrelevant information on ongoing information processing[15]. Therefore, inhibition is also divided into two dimensions: interference inhibition (or

interference control) and response inhibition (or behavioral inhibition)[91]. Cognitive flexibility is considered a single dimension, characterizing the ability to flexibly switch between different tasks and modes of thought[92].

According to their complexity, we selected six behavioral tasks to measure these EF dimensions (Fig. 3)[40]: 1) *number running memory updating* task was used to examine working memory updating[89]; 2) *digit span backward* task was used to measure working memory maintenance (span)[40]; 3) *Corsi block* test was used to measure working memory maintenance (span), which more comprehensively assesses maintenance in the spatial dimension[93]; 4) *Stroop* task was used to measure interference inhibition[94]; 5) *Go/No-Go* task was used to measure response inhibition[95]; and 6) *number switching* task was used to measure shifting[92]. All behavioral tasks were performed using *E*-Prime 3.0 software (Psychology Software Tools, Inc., Pittsburgh, PA, USA). Accuracy and reaction time on each task can be weighted to derive 14 comprehensive assessment indicators from these six behavioral tests.

Inhibitory control

The four measurements for assessing the interference control function included the reaction times for the incongruent, congruent, and neutral stimuli, and the difference of reaction time between the congruent and the incongruent condition trials (i.e., the interference effect) in the *Stroop* task. Three measurements for assessing the response inhibition function included the reaction time for the "Go" stimuli and the accuracy for the Go and No-Go stimuli in the *Go/No-Go* task.

Working memory

Two measurements for assessing working memory updating function included the proportion of digits correctly recalled and placed in the correct sequence at two different speeds of presentation (1,750 ms and 750 ms per digit) in the *running memory* task. Three measurements for assessing the numeric working memory maintenance capacity included the length of the last correctly repeated sequence, the count of sequences correctly repeated until the conclusion of the test (i.e., the total number of successful trials) from the *Corsi block* test, and the maximal number of digits accurately recalled in the reverse order of the *digit span*

backward task.

Cognitive flexibility

Two measurements which included the difference in reaction time between the switch and the non-switch trials [switch cost], as well as the difference in the reaction time between the non-switch and the single-task trials [mixing cost]) measured in the *number-letter switching* task.

Besides a conceptual formulation of the 14 task measurements according to the three-dimensional representation of EF, we supplemented five composite scores calculated based on these measured variables[96]. Firstly, an inhibitory composite score was calculated by averaging: 1) the difference in reaction time between the congruent and the incongruent condition trials in the *Stroop* task; and 2) the accuracy for the No-Go stimuli in the *Go/No-Go* task as in previous studies[97,98]. The purpose of this approach was to denote the combined response and inference inhibitory functions. Furthermore, this inhibitory composite score was aggregated with the assessments in the *running memory* task (the proportion of digits correctly recalled and placed in the correct sequence at the speed of 1,750 ms per digit) and the *number-letter switching* task (switch cost) to form an abbreviated version for representing general EF functions. Such abbreviation is in compliance with the previous notion on a single-condition indicator that these trails require greater executive control demands[99]. Considering that EF functions interplay across conceptual constructs during cognitive engagement (e.g., problem-solving) for processing particular behaviors[100], we additionally created three cross-dimensional EF composite scores by collapsing the cardinal three dimensions of EF, which the inhibitory composite score was similarly used: 1) Inhibition and Updating composite; 2) Inhibition and Switching composite; and 3) Switching and Updating composite. The measurements used to assess working memory updating and switching were employed to calculate the abbreviated version of the EF composite score. These composite measures would provide additional insights into the dimensional and cross-dimensional contributing features to diagnostic and prognostic clssifications[100]. Consequently, 19 EF-related indicators were used as input features for the machine learning models (Fig. 3).

Behavioral task and clinical scale analyses

We performed a mixed-model ANCOVA with group (schizophrenia vs. healthy control) as a between-subjects factor and all EF measurements as well as composite scores as within-subjects factors, while controlling for sociodemographic variables, to examine whether schizophrenia has differentially affected EF performance. The mixed-model ANCOVA was followed by multiple one-way ANCOVAs to examine group differences in each EF measure. also controlling for Sociodemographic variables were likewise controlled and a Benjamini-Hochberg (BH) false discovery rate (FDR) correction was applied to adjust for multiple comparisons. Eta squared effect sizes were calculated to describe the magnitude of effect sizes, with certain values interpreted as small ($\eta^2 \leq 0.01$), medium ($0.01 < \eta^2 \leq 0.06$), or large ($0.06 < \eta^2 \leq 0.14$)[101]. Two-sample t tests were used to compare clinical symptoms, demographic characteristics, and EF between the remission and non-remission subgroups. Chi-squared tests were used to compare categorical variables between these groups. We used paired-samples *t*-tests to compare clinical symptoms between baseline and follow-up (after 4–6 weeks of treatment) for the subset of followed participants. Finally, we used Pearson and Spearman correlation analyses to examine the relationships among clinical symptoms and EF, with the former applied to continuous variables and the latter applied to categorical variables.

Classification modeling procedure

Features and models

Participants were categorized into schizophrenia and HC groups. Two feature sets were tested for diagnostic classification accuracy: 1) 19 baseline EF assessments, subsuming the three EF dimensions measured by six behavioral paradigms (Table 2); and 2) 32 features—the 19 baseline EF measures plus 13 routinely attainable sociodemographic variables. For prognostic classification, the same two feature sets were tested to distinguish patient remission status (remitted vs. non-remitted) after 4–6 weeks of antipsychotic treatment. SVM, RF, and AdaBoost, which are widely used in psychiatric machine-learning research[102], were used for classification tasks, along with a synthesized stacking model of the three. Stacking models are a multi-view approach integrating classification weight estimates

from single classifiers to improve ultimate performance[103].

Complementary investigations

Control analysis

Several control analyses were performed, in which we controlled the effects of 1) 13 demographic characteristics; 2) both the 13 demographic variables and the OZP equivalent dosage; 3) only the OZP equivalent dosage on the baseline EF assessments using regression approaches[104] when establishing the diagnostic and prognostic classification models. The continuous variables in the demographic dataset included age, years of education, body mass index, socioeconomic status, and fluid intelligence measured by Raven's Progressive Matrices test. The categorical variables consisted of sex, ethnicity, residence, employment status, only-child status, marital status, smoking history, and drinking history. One-hot encoding was applied to the categorical variables, converting them into binary vectors. Furthermore, we sought to mitigate the potential impact of antipsychotic drug dosage and evaluate the robustness of our classification results. Thus, we established the prognostic classification models with only those patients receiving a clinically standard, commonly effective OZP equivalent dosage of 10–20 mg/day.

Sensitivity analysis

In our research, the majority of dropouts were due to hospital discharge. Therefore, we conducted a best-case sensitivity analysis, assuming that all participants who were lost to follow-up had positive treatment outcomes (remission). Additionally, we performed a worst-case analysis, assuming that all participants who were lost to follow-up had the least favorable treatment outcomes (no remission). These data help establish the potentially extreme scenarios due to attrition with respect to our main analyses with an observed prognostic status. Sensitivity analyses, such as best-worst (assuming all participants lost to follow-up in one group [referred as group 1] have had a beneficial outcome and all those with missing outcomes in the other group [group 2] have had a harmful outcome) and worst-best case (assuming that all participants lost to follow-up in group 1 have had a harmful outcome; and that all those lost to follow-up in group 2 have had a beneficial outcome), are utilized in

medical and psychiatric research to evaluate the reliability of results in relation to participants who do not complete the study[105,106].

Machine learning design

Machine learning and CV were implemented using Python (version 3.10.11) and the scikit-learn package (version 1.3.0). Specifically, the original data were first preprocessed to accommodate missing values, outliers, and class imbalance issues (see Supplementary Materials for details). Thereafter, we randomly split the preprocessed data into a discovery dataset with 80% of the overall schizophrenia and HC groups ('training set') and a 'lock-box' test dataset with the remaining 20% of these samples ('test set') to determine out-of-sample classification performance(Fig. 1)[4,42–44,107]. The random split procedure was stratified for the outcome variable (diagnostic label or remission status), ensuring a balanced representation of labels in each dataset[42,43]. Using the discovery dataset, we performed a nested CV loop[108] (also termed double CV), which differentiates two CV roles to avoid 'circularity' introduced by overfitting when the same sample subset is used for both hyperparameter tuning and model validation[53]. Specifically, within a nested CV loop, the inner CV (k= 3), encompassing 80% of the discovery sample, operates all data-dependent decisions while determining optimal hyperparameters. The outer CV (k= 5) is subsequently utilized for parameter assessment and model selection[78]. For optimal hyperparameter selection, we used the Optuna optimization technique (version 3.5.0) and selection based on the achievable AUC of candidate hyperparameters within the validation sets of the inner loop[109]. The AUC metric, representing the degree of separability, is widely used to evaluate model performance. In this investigation, it indicated the ability of the model to distinguish between the schizophrenia and HC groups, and between the patient remission and non-remission groups. Hyperparameters with the highest average performance over the 5 × 3 nested CV were used to train a model on the entire discovery sample without further modification; next, they were tested using the independent 'test set' sample[50]. In addition to AUC, sensitivity, specificity, and balanced accuracy performance metrics were assessed[110]. To avoid potential bias from random splitting, the aforementioned machine learning procedure was repeated 100 times[111,112].

Feature importance analysis

To evaluate the contributions of EF features to our classification models, we assigned an importance score (i.e., Shapley value) to each feature[47]. Specifically, we used the SHAP library (version 0.39.0) model-agnostic SHAP KernelExplainer approach, which is generally used to estimate Shapley values for prediction models[113]. SHAP KernelExplainer employs a Monte Carlo approach to randomly sample feature combinations based on input predictors. Initially, it estimates the importance of these combinations with varying features present in model predictions. Subsequently, individual Shapley values are calculated to denote the contribution of each feature to the target prediction based on a weighted linear regression model[48].

After obtaining the individual-level Shapley values for each feature, we computed the mean absolute Shapley value (i.e., feature importance score) across all individuals (i.e., the group-level Shapley values) where larger Shapley values indicate stronger importance of this feature to the classification model. The group-level Shapley values were next used to depict a group-level decision path for each feature. Essentially, among those participants in the schizophrenia group who were correctly classified, performance by each EF dimension assessment was averaged across both the HC and schizophrenia groups for the diagnostic models, and across the remission and non-remission patient groups for the prognostic models. The averaged performance of each EF feature was then normalized and compared with a positive (or negative) value; higher (or lower) performance by an EF feature drove the model to more accurately classify true cases. Using individual Shapley values, we also plotted the decision path for each individual who was correctly classified to complement the group-level results; this was performed because individuals with schizophrenia have heterogeneous expressions across EF dimensions[60]. Finally, based on the individual Shapley values, we tested the extent to which the importance of the contribution of an EF feature to a classification was linked to individual psychopathology. This was conducted using Pearson correlation analysis based on individual scores (entire patient group N=195, diagnostic model; subset sample N= 86, prognostic model) along the four symptom dimensions which were assessed at baseline and follow-up, and the difference between the two. The FDR approach

was used for the correction of multiple comparisons to statistically rule out potential false-positive associations[114].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 32260207 [to Xin Zhao], No. 82371506 [to Ji Chen] & No. 82201658 [to Ji Chen]), the STI2030-Major Projects (No. 2022ZD0214000 [to Ji Chen]), and the National Key R&D Program of China (No. 2021YFC2502200 [to Ji Chen]). BTTY is supported by the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC CTG-IIT (CTGIIT23jan-0001), NMRC STaR (STaR20nov-0003), Singapore Ministry of Health (MOH) Centre Grant (CG21APR1009), the Temasek Foundation (TF2223-IMH-01), and the United States National Institutes of Health (R01MH120080 & R01MH133334). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NMRC, MOH or Temasek Foundation.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

Information for the main sample used in the present study have been included in the Supplementary Materials. The raw data of our used sample are protected and are not publicly available due to data privacy. These data can be accessed upon reasonable request to the corresponding author (X. Z.). Derived data supporting the findings of this study are available from the corresponding authors (X. Z. or J. C.) upon request.

Code availability statement

Scripts to run the main analyses have been made publicly available and can be accessed at https://doi.org/10.6084/m9.figshare.26086594.v1.

References

1. McCutcheon, R. A., Keefe, R. S. E. & McGuire, P. K. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. *Mol Psychiatry* 1–17 (2023).
2. Heilbronner, U., Samara, M., Leucht, S., Falkai, P. & Schulze, T. G. The Longitudinal Course of Schizophrenia Across the Lifespan: Clinical, Cognitive, and Neurobiological Aspects. *Harvard Review of Psychiatry* **24**, 118 (2016).
3. Vita, A. *et al.* European Psychiatric Association guidance on assessment of cognitive impairment in schizophrenia. *European Psychiatry* **65**, e58 (2022).
4. Lim, K. *et al.* Cognitive Subtyping in Schizophrenia: A Latent Profile Analysis. *Schizophrenia Bulletin* **47**, 712–721 (2021).
5. Barch, D. M. The Cognitive Neuroscience of Schizophrenia. *Annu. Rev. Clin. Psychol.* **1**, 321–353 (2005).
6. Diamond, A. Executive Functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
7. Kerns, J. G., Nuechterlein, K. H., Braver, T. S. & Barch, D. M. Executive Functioning Component Mechanisms and Schizophrenia. *Biological Psychiatry* **64**, 26–33 (2008).
8. Ambrosen, K. S. *et al.* A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naive schizophrenia patients based on multimodal neuropsychiatric data. *Translational Psychiatry* **10**, 276

(2020).

9. Soldatos, R. F. *et al.* Prediction of Early Symptom Remission in Two Independent Samples of First-Episode Psychosis Patients Using Machine Learning. *Schizophrenia Bulletin* **48**, 122–133 (2022).
10. Coutts, F., Koutsouleris, N. & McGuire, P. Psychotic disorders as a framework for precision psychiatry. *Nat Rev Neurol* (2023) doi:10.1038/s41582-023-00779-1.
11. McGorry, P. *et al.* Biomarkers and clinical staging in psychiatry. *World Psychiatry* **13**, 211–223 (2014).
12. Sheffield, J. M., Karcher, N. R. & Barch, D. M. Cognitive Deficits in Psychotic Disorders: A Lifespan Perspective. *Neuropsychol Rev* **28**, 509–533 (2018).
13. Freedman, D. & Brown, A. S. The developmental course of executive functioning in schizophrenia. *International Journal of Developmental Neuroscience* **29**, 237–243 (2011).
14. Green, M. F., Kern, R. S. & Heaton, R. K. Longitudinal studies of cognition and functional outcome in schizophrenia: implications for MATRICS. *Schizophrenia research* **72**, 41–51 (2004).
15. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
16. Friedman, N. P. & Robbins, T. W. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacol.* **47**, 72–89 (2022).
17. Karr, J. E. *et al.* The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin* **144**, 1147–1185 (2018).
18. Kerns, J. G. & Berenbaum, H. Cognitive impairments associated with formal thought disorder in people with schizophrenia. *Journal of Abnormal Psychology* **111**, 211–224 (2002).
19. Kerns, J. G. & Berenbaum, H. The relationship between formal thought disorder and executive functioning component processes. *Journal of Abnormal Psychology* **112**, 339 (2003).
20. Chen, J. *et al.* Neurobiological Divergence of the Positive and Negative Schizophrenia Subtypes Identified on a New Factor Structure of Psychopathology Using Non-negative Factorization: An International Machine Learning Study. *Biological Psychiatry* **87**, 282–293 (2020).
21. Yang, H. *et al.* Diminished self-monitoring in hallucinations–Aberrant anterior insula connectivity differentiates auditory hallucinations in schizophrenia from subjective tinnitus. *Asian Journal of Psychiatry* **52**, 102056 (2020).
22. Raveendran, V. & Kumari, V. Clinical, cognitive and neural correlates of self-monitoring deficits in schizophrenia: an update. *Acta Neuropsychiatrica* **19**, 27–37 (2007).
23. Raffard, S. *et al.* Working Memory Deficit as a Risk Factor for Severe Apathy in Schizophrenia: A 1-Year Longitudinal Study. *Schizophrenia Bulletin* **42**, 642–651 (2016).
24. Glahn, D. C., Cannon, T. D., Gur, R. E., Ragland, J. D. & Gur, R. C. Working memory

constrains abstraction in schizophrenia. *Biological Psychiatry* **47**, 34–42 (2000).

25. Hager, O. M. *et al.* Reward-dependent modulation of working memory is associated with negative symptoms in schizophrenia. *Schizophrenia Research* **168**, 238–244 (2015).
26. Thai, M. L., Andreassen, A. K. & Bliksted, V. A meta-analysis of executive dysfunction in patients with schizophrenia: Different degree of impairment in the ecological subdomains of the Behavioural Assessment of the Dysexecutive Syndrome. *Psychiatry Research* **272**, 230–236 (2019).
27. De Luca, C. R. *et al.* Normative data from the CANTAB. I: development of executive function over the lifespan. *Journal of clinical and experimental neuropsychology* **25**, 242–254 (2003).
28. Weintraub, S. *et al.* Cognition assessment using the NIH Toolbox. *Neurology* **80**, S54–S64 (2013).
29. Chidharom, M., Krieg, J. & Bonnefond, A. Impaired Frontal Midline Theta During Periods of High Reaction Time Variability in Schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **6**, 429–438 (2021).
30. Waltz, J. A., Frank, M. J., Wiecki, T. V. & Gold, J. M. Altered probabilistic learning and response biases in schizophrenia: Behavioral evidence and neurocomputational modeling. *Neuropsychology* **25**, 86–97 (2011).
31. Snyder, H. R., Miyake, A. & Hankin, B. L. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. *Front. Psychol.* **6**, (2015).
32. Dhakal, P., Tayara, H. & Chong, K. T. An ensemble of stacking classifiers for improved prediction of miRNA-mRNA interactions. *Computers in Biology and Medicine*.
33. Pavlyshenko, B. Using stacking approaches for machine learning models. in *2018 IEEE second international conference on data stream mining & processing (DSMP)* 255–258 (IEEE, 2018).
34. Andreasen, N. C. *et al.* Remission in Schizophrenia: Proposed Criteria and Rationale for Consensus. *AJP* **162**, 441–449 (2005).
35. Beitinger, R., Lin, J., Kissling, W. & Leucht, S. Comparative remission rates of schizophrenic patients using various remission criteria. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **32**, 1643–1651 (2008).
36. Chekroud, A. M. *et al.* Illusory generalizability of clinical prediction models. *Science* **383**, 164–167 (2024).
37. Schennach-Wolff, R. *et al.* An early improvement threshold to predict response and remission in first-episode schizophrenia. *The British Journal of Psychiatry* **196**, 460–466 (2010).
38. Kirschner, M., Aleman, A. & Kaiser, S. Secondary negative symptoms — A review of mechanisms, assessment and treatment. *Schizophrenia Research* **186**, 29–38 (2017).
39. Veerman, S. R. T., Schulte, P. F. J. & de Haan, L. Treatment for Negative Symptoms in

Schizophrenia: A Comprehensive Review. *Drugs* **77**, 1423–1459 (2017).

40. Zhao, X., Zhang, W., Tong, D. & Maes, J. H. R. Creative thinking and executive functions: Associations and training effects in adolescents. *Psychology of Aesthetics, Creativity, and the Arts* **17**, 79–90 (2023).
41. Leucht, S., Davis, J. M., Engel, R. R., Kissling, W. & Kane, J. M. Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. *Acta Psychiatrica Scandinavica* **119**, 7–14 (2009).
42. Schultebraucks, K. *et al.* Assessment of early neurocognitive functioning increases the accuracy of predicting chronic PTSD risk. *Mol Psychiatry* **27**, 2247–2254 (2022).
43. Webb, C. A. *et al.* Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology* **88**, 25–38 (2020).
44. Perochon, S. *et al.* Early detection of autism using digital behavioral phenotyping. *Nat Med* 1–9 (2023) doi:10.1038/s41591-023-02574-3.
45. van Os, J., Kenis, G. & Rutten, B. P. F. The environment and schizophrenia. *Nature* **468**, 203–212 (2010).
46. Habets, P. C. *et al.* Multimodal Data Integration Advances Longitudinal Prediction of the Naturalistic Course of Depression and Reveals a Multimodal Signature of Remission During 2-Year Follow-up. *Biological Psychiatry* (2023).
47. Shah, R. V. *et al.* Personalized machine learning of depressed mood using wearables. *Transl Psychiatry* **11**, 1–18 (2021).
48. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).
49. Ebdrup, B. H. *et al.* Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naive schizophrenia patients. *Psychological Medicine* **49**, 2754–2763 (2019).
50. Chen, J., Patil, K. R., Yeo, B. T. T. & Eickhoff, S. B. Leveraging Machine Learning for Gaining Neurobiological and Nosological Insights in Psychiatric Research. *Biological Psychiatry* **93**, 18–28 (2023).
51. Rutherford, S. The promise of machine learning for psychiatry. *Biological Psychiatry* **88**, e53–e55 (2020).
52. Walter, M. *et al.* Translational machine learning for psychiatric neuroimaging. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **91**, 113–121 (2019).
53. Wu, J., Li, J., Eickhoff, S. B., Scheinost, D. & Genon, S. The challenges and prospects of brain-based prediction of behaviour. *Nat Hum Behav* **7**, 1255–1264 (2023).
54. Chen, Z. *et al.* Sampling inequalities affect generalization of neuroimaging-based diagnostic classifiers in psychiatry. *BMC Med* **21**, 241 (2023).
55. Myszczynska, M. A. *et al.* Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* **16**, 440–456 (2020).

56. Hughes, M. E., Fulham, W. R., Johnston, P. J. & Michie, P. T. Stop-signal response inhibition in schizophrenia: Behavioural, event-related potential and functional neuroimaging data. *Biological Psychology* **89**, 220–231 (2012).
57. Kaladjian, A. *et al.* Blunted activation in right ventrolateral prefrontal cortex during motor response inhibition in schizophrenia. *Schizophrenia research* **97**, 184–193 (2007).
58. Wexler, B. E. *et al.* Neuropsychological near normality and brain structure abnormality in schizophrenia. *American Journal of Psychiatry* **166**, 189–195 (2009).
59. Raffard, S. & Bayard, S. Understanding the executive functioning heterogeneity in schizophrenia. *Brain and Cognition* **79**, 60–69 (2012).
60. Snitz, B. E., MacDonald III, A. W. & Carter, C. S. Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. (2006).
61. Chan, R. C. K., Chen, E. Y. H., Cheung, E. F. C., Chen, R. Y. L. & Cheung, H. K. The components of executive functioning in a cohort of patients with chronic schizophrenia: A multiple single-case study design. *Schizophrenia Research* **81**, 173–189 (2006).
62. Carruthers, S. P., Van Rheenen, T. E., Gurvich, C., Sumner, P. J. & Rossell, S. L. Characterising the structure of cognitive heterogeneity in schizophrenia spectrum disorders. A systematic review and narrative synthesis. *Neuroscience & Biobehavioral Reviews* **107**, 252–278 (2019).
63. Brugger, S. P. & Howes, O. D. Heterogeneity and Homogeneity of Regional Brain Structure in Schizophrenia: A Meta-analysis. *JAMA Psychiatry* **74**, 1104–1111 (2017).
64. Donohoe, G., Corvin, A. & Robertson, I. Evidence that specific executive functions predict symptom variance among schizophrenia patients with a predominantly negative symptom profile. *Cognitive Neuropsychiatry* **11**, 13–32 (2006).
65. Wu, C., Wang, C. & Li, L. Speech-on-speech masking and psychotic symptoms in schizophrenia. *Schizophrenia Research: Cognition* **12**, 37–39 (2018).
66. Zheng, Y. *et al.* Negative-emotion-induced reduction in speech-in-noise recognition is associated with source-monitoring deficits and psychiatric symptoms in mandarin-speaking patients with schizophrenia. *Comprehensive Psychiatry* **124**, 152395 (2023).
67. Johansson, M., Hjärthag, F. & Helldin, L. Cognitive markers related to long-term remission status in Schizophrenia Spectrum Disorders. *Psychiatry Research* **289**, 113035 (2020).
68. Helldin, L., Kane, J. M., Karilampi, U., Norlander, T. & Archer, T. Remission and cognitive ability in a cohort of patients with schizophrenia. *Journal of Psychiatric Research* **40**, 738–745 (2006).
69. Rabanea-Souza, T. *et al.* Neuropsychological correlates of remission in chronic schizophrenia subjects: The role of general and task-specific executive processes. *Schizophrenia Research: Cognition* **3**, 39–46 (2016).
70. Yun, D. Y. *et al.* Impairments in executive functioning in patients with remitted and non-remitted schizophrenia. *Progress in Neuro-Psychopharmacology and Biological*

Psychiatry **35**, 1148–1154 (2011).

71. Chuang, S. P., Wu, J. Y. W., Wang, C. S., Liu, C. H. & Pan, L. H. Self concepts, health locus of control and cognitive functioning associated with health-promoting lifestyles in schizophrenia. *Comprehensive Psychiatry* **70**, 82–89 (2016).
72. Na, E. *et al.* Relationships among medication adherence, insight, and neurocognition in chronic schizophrenia. *Psychiatry and clinical neurosciences* **69**, 298–304 (2015).
73. Thakkar, K. N., Schall, J. D., Boucher, L., Logan, G. D. & Park, S. Response inhibition and response monitoring in a saccadic countermanding task in schizophrenia. *Biological psychiatry* **69**, 55–62 (2011).
74. Waters, F. A. V., Badcock, J. C., Maybery, M. T. & Michie, P. T. Inhibition in schizophrenia: association with auditory hallucinations. *Schizophrenia Research* **62**, 275–280 (2003).
75. Ettinger, U. *et al.* Response inhibition and interference control: Effects of schizophrenia, genetic risk, and schizotypy. *Journal of Neuropsychology* **12**, 484–510 (2018).
76. Cameron, A. M. *et al.* Working memory correlates of three symptom clusters in schizophrenia. *Psychiatry Research* **110**, 49–61 (2002).
77. Perlstein, W. M., Carter, C. S., Noll, D. C. & Cohen, J. D. Relation of Prefrontal Cortex Dysfunction to Working Memory and Symptoms in Schizophrenia. *AJP* **158**, 1105–1113 (2001).
78. Sasse, L. *et al.* On Leakage in Machine Learning Pipelines. Preprint at http://arxiv.org/abs/2311.04179 (2023).
79. Youssef, A. *et al.* External validation of AI models in health should be replaced with recurring local validation. *Nature Medicine* **29**, 2686–2687 (2023).
80. Leucht, S. *et al.* Examination of Dosing of Antipsychotic Drugs for Relapse Prevention in Patients With Stable Schizophrenia: A Meta-analysis. *JAMA Psychiatry* **78**, 1238–1248 (2021).
81. Kane, J. M., Leucht, S., Carpenter, D. & Docherty, J. P. The expert consensus guideline series. Optimizing pharmacologic treatment of psychotic disorders. Introduction: methods, commentary, and summary. *The Journal of clinical psychiatry* **64**, 5–19 (2003).
82. Hofer, A. *et al.* Why do indiviuals with schizophrenia drop out of observational clinical trials? *Psychiatry Research* **256**, 1–5 (2017).
83. Martin, J. L. R. *et al.* Meta-analysis of drop-out rates in randomised clinical trials, comparing typical and atypical antipsychotics in the treatment of schizophrenia: Submitted to European Psychiatry. A poster of this work was presented at the WPA International Congress in Florence (November 10–13, 2004). Poster session PO1: Psychotic Disorders. *Eur. psychiatr.* **21**, 11–20 (2006).
84. Johnsen, E. *et al.* The serum level of C-reactive protein (CRP) is associated with cognitive performance in acute phase psychosis. *BMC Psychiatry* **16**, 60 (2016).
85. Cebula, M. *et al.* Resin Infiltration of Non-Cavitated Proximal Caries Lesions in Primary

and Permanent Teeth: A Systematic Review and Scenario Analysis of Randomized Controlled Trials. *Journal of Clinical Medicine* **12**, 727 (2023).

86. Sharma, M., Patel, R. K., Garg, A., SanTan, R. & Acharya, U. R. Automated detection of schizophrenia using deep learning: a review for the last decade. *Physiological Measurement* **44**, 03TR01 (2023).
87. Mavridis, D., Chaimani, A., Efthimiou, O., Leucht, S. & Salanti, G. Addressing missing outcome data in meta-analysis. *BMJ Ment Health* **17**, 85–89 (2014).
88. Raven, J. The Raven's progressive matrices: change and stability over culture and time. *Cognitive psychology* **41**, 1–48 (2000).
89. Zhao, X., Wang, Y. & Maes, J. H. R. The effect of working memory capacity and training on intertemporal decision making in children from low-socioeconomic-status families. *Journal of Experimental Child Psychology* **216**, 105347 (2022).
90. Miyake, A. *et al.* The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology* **41**, 49–100 (2000).
91. Nigg, J. T. On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin* **126**, 220–246 (2000).
92. Kray, J., Li, K. Z. & Lindenberger, U. Age-related changes in task-switching components: The role of task uncertainty. *Brain and cognition* **49**, 363–381 (2002).
93. Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J. & de Haan, E. H. F. The Corsi Block-Tapping Task: Standardization and Normative Data. *Applied Neuropsychology* **7**, 252–258 (2000).
94. Stroop, J. R. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* **18**, 643–662 (1935).
95. Gomez, P., Ratcliff, R. & Perea, M. A model of the go/no-go task. *Journal of Experimental Psychology: General* **136**, 389 (2007).
96. Willoughby, M. T. & Blair, C. B. Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment* **28**, 319–330 (2016).
97. Cepeda, N. J., Blackwell, K. A. & Munakata, Y. Speed isn't everything: complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science* **16**, 269–286 (2013).
98. Gardiner, E., Hutchison, S. M., Müller, U., Kerns, K. A. & Iarocci, G. Assessment of executive function in young children with and without ASD using parent ratings and computerized tasks of executive function. *The Clinical Neuropsychologist* **31**, 1283–1305 (2017).
99. Yangüez, M., Bediou, B., Chanal, J. & Bavelier, D. In search of better practice in executive functions assessment: Methodological issues and potential solutions. *Psychological Review* **131**, 402–430 (2024).

100. Ronold, E. H., Schmid, M. T., Oedegaard, K. J. & Hammar, Å. A Longitudinal 5-Year Follow-Up Study of Cognitive Function After First Episode Major Depressive Disorder: Exploring State, Scar and Trait Effects. *Front. Psychiatry* **11**, (2020).
101. Fritz, C. O., Morris, P. E. & Richler, J. J. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* **141**, 2 (2012).
102. Verma, S. *et al.* Machine learning techniques for the Schizophrenia diagnosis: a comprehensive review and future research directions. *J Ambient Intell Human Comput* **14**, 4795–4807 (2023).
103. Pernía-Espinoza, A., Fernandez-Ceniceros, J., Antonanzas, J., Urraca, R. & Martinez-de-Pison, F. J. Stacking ensemble with parsimonious base models to improve generalization capability in the characterization of steel bolted components. *Applied Soft Computing* **70**, 737–750 (2018).
104. Dobson, A. J. & Barnett, A. G. *An Introduction to Generalized Linear Models*. (CRC press, 2018).
105. Akl, E. A. *et al.* Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ* **344**, e2809 (2012).
106. Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* **17**, 162 (2017).
107. Martini, G. *et al.* Machine learning can guide food security efforts when primary data are not available. *Nat Food* **3**, 716–728 (2022).
108. Varoquaux, G. *et al.* Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* **145**, 166–179 (2017).
109. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperp. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (2019). doi:10.1145/3292500.3330701.
110. Koutsouleris, N. *et al.* Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry* **3**, 935–946 (2016).
111. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* **11**, 2079–2107 (2010).
112. Lee, L.-H. *et al.* Evaluating the performance of machine learning models for automatic diagnosis of patients with schizophrenia based on a single site dataset of 440 participants. *European Psychiatry* **65**, e1 (2021).
113. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. in *Joint European conference on machine learning and knowledge discovery in databases* 417–431 (Springer, 2020).

114. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

1149 **Table 1 | Participant demographics and clinical characteristics**

Characteristic		Schizophrenia group (*N*= 195)	Healthy control group (*N*= 169)	*p*-value	Remission group at baseline (*N*= 58)	Non-remission group at baseline (*N*= 28)	*p*-value
Demographic							
	Age	35.35 ± 9.35	37.69 ± 13.71	0.055	34.03 ± 8.69	34.54 ± 9.75	0.810
	Sex			0.264			0.192
	Male	114 (58.5%)	88 (52.1%)		39 (67.2%)	14 (50%)	
	Female	81 (41.5%)	81 (47.9%)		19 (32.8%)	14 (50%)	
	Ethnicity, Han	173 (88.7%)	152 (89.9%)	0.837	52 (89.7%)	24 (85.7%)	0.861
	Education, years	11.12 ± 4.58	10.90 ± 3.94	0.615	10.52 ± 3.98	11.07 ± 4.24	0.555
	BMI	23.21 ± 3.73	23.95 ± 4.32	0.494	23.60 ± 3.13	22.65 ± 3.78	0.217
	Residence, urban	114 (58.5%)	98 (58.0%)	1.000	27 (24.6%)	16 (57.1%)	0.490
	SES	23.21 ± 7.32	23.88 ± 5.80	0.338	22.72 ± 7.25	23.43 ± 8.02	0.684
	RPM	32.56 ± 11.50	39.52 ± 9.63	**< 0.001**	34.84 ± 11.17	33.04 ± 12.49	0.500
	Employed, yes	53 (27.2%)	111 (65.7%)	**< 0.001**	44 (75.9%)	21 (75.0%)	.
	Only child, yes	63 (32.3%)	24 (14.2%)	**< 0.001**	42 (72.4%)	14 (50.0%)	0.072
	Marital status			**< 0.001**			0.205
	Unmarried	115 (59.0%)	54 (32.0%)		34 (58.6%)	19 (67.9%)	
	Married	48 (24.6%)	108 (63.9%)		18 (31.0%)	4 (14.3%)	
	Divorced	31 (15.9%)	7 (4.1%)		6 (10.3%)	5 (17.9%)	
	Widowed	1 (0.5%)	0 (0.0%)		0 (0.0%)	0 (0.0%)	
	Smoking history			0.079			0.731
	Never	119 (61.0%)	122 (72.2%)		36 (62.1%)	19 (67.9%)	
	1–3 years	19 (9.7%)	11 (6.5%)		15 (8.6%)	3 (10.7%)	
	>3 years	57 (29.2%)	36 (21.3%)		17 (29.3%)	6 (21.4%)	

	Alcohol consumption history					
	Never	133 (68.2%)	99 (58.6%)	0.118	42 (72.4%)	18 (64.3%)	0.729
	Occasionally	54 (27.7%)	64 (40.3%)		14 (24.1%)	9 (32.1%)	
	Regularly	8 (4.1%)	6 (3.8%)		2 (3.4%)	1 (3.6%)	
Clinical							
Electronic medication records							
	Age at onset	27.43 ± 8.63	.	.	27.02 ± 8.02	27.25 ± 9.39	0.906
	Duration of disorder	9.74 ± 7.17	.	.	9.09 ± 5.98	10.48 ± 7.51	0.395
	Frequency of episodes	5.98 ± 4.27			6.17 ± 3.84	6.32 ± 4.05	0.871
	First episode, yes	10 (5.13%)	.	.	1 (1.72%)	1 (3.57%)	0.984
	Family medical history, yes	33 (16.92%)	.	.	49 (84.5%)	23 (82.1%)	.
	Dose equivalent to olanzapine (mg/day)	14.48 ± 6.41	.	.	14.30 ± 5.46	16.65 ± 7.66	0.154
	Type of antipsychotic medication						0.829
	First generation	9 (5%)	.	.	2 (3.4%)	2 (7.1%)	
	Second generation	186 (95%)	.	.	56 (96.6%)	26 (92.9%)	
Clinical scale							
	3 PANSS subscales						
	PANNS-Negative	21.42 ± 6.48	.	.	21.59 ± 6.59	21.79 ± 5.99	0.893
	PANNS-Positive	22.01 ± 4.46	.	.	21.64 ± 4.06	21.18 ± 5.99	0.676
	PANNS-General	40.3 ± 6.81	.	.	39.67 ± 7.39	40.39 ± 6.20	0.657
	PANNS-Total	83.79 ± 13.85	.	.	82.90 ± 13.97	83.36 ± 15.33	0.890
	4 dimensions of PANSS						
	Negative factor	8.22 ± 2.5	.	.	8.24 ± 2.70	8.71 ± 2.11	0.425
	Positive factor	6.24 ± 1.74	.	.	5.99 ± 1.66	6.23 ± 2.07	0.560
	Affective factor	5.65 ± 0.86	.	.	5.67 ± 0.93	5.57 ± 0.78	0.619
	Cognitive factor	9.78 ± 1.72	.	.	9.66 ± 1.82	9.76 ± 1.61	0.823

Note: Data are presented as the mean ± standard deviation or n (%). The *p*-values in bold face indicate statistically significant differences ($p < 0.05$). **Calculation:** BMI is calculated as weight (kg) divided by height squared (m²). Remission status (remission or non-remission) was determined based on the RSWG remission criteria. **Abbreviations:** BMI, body mass index; SES, socioeconomic status; RPM, Raven's Progressive Matrices. PANSS, Positive and Negative Syndrome Scale; RSWG, Remission in Schizophrenia Working Group.

1151 **Table 2 | Comparisons of executive function dimensions between groups**

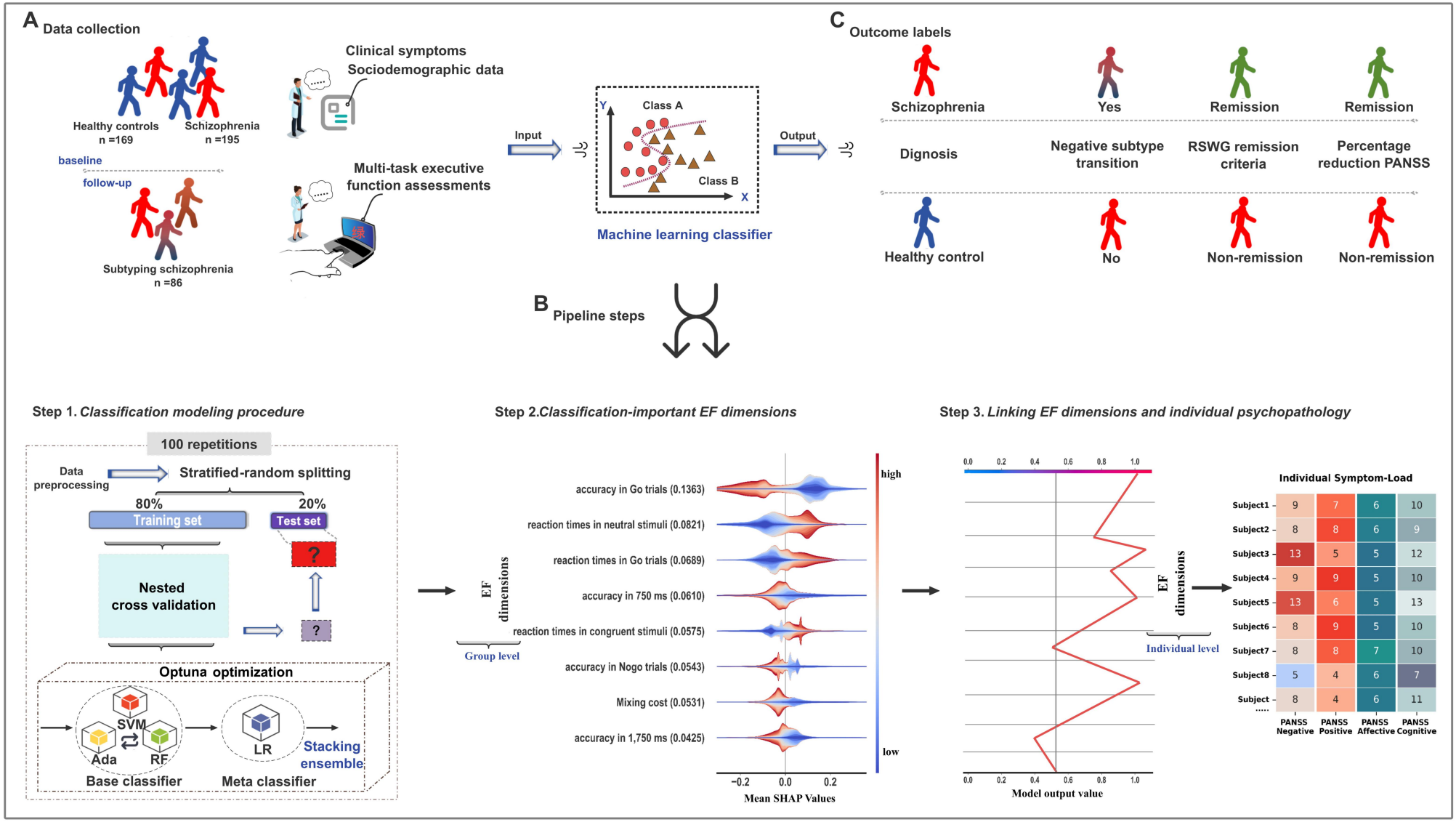
Dimension /Subdimension	Measurement	Diagnostic classification (N=364)				Prognosis classification (N=86)		
		Schizophrenia group (*N*= 195)	Healthy control group (*N*= 169)	*p*-value	η^2	Baseline (*N*= 86)	4–6 weeks treatment (*N*= 86)	*p*-value
Dimension I: Inhibition								
Interference inhibition (*Stroop task*)	Reaction times in neutral stimuli	727.27 ± 143.34	611.64 ± 114.15	**0.002**	0.040	719.80 ± 133.13	677.10 ± 124.85	0.258
	Reaction times in congruent stimuli	704.26 ± 132.24	604.55 ± 118.40	**< 0.001**	0.080	723.00 ± 133.13	677.56 ± 121.80	**0.022**
	Reaction times in incongruent stimuli	751.79 ± 148.77	663.61 ± 149.02	**< 0.001**	0.090	764.28 ± 152.17	721.14 ± 140.05	**0.049**
	Stroop interference effect	–23.01 ± 73.67	–7.09 ± 44.24	0.217	0.010	3.20 ± 145.73	–8.53 ± 63.52	0.616
Response inhibition (*Go/No-Go task*)	Accuracy in Go trials	0.83 ± 0.14	0.94 ± 0.08	**< 0.001**	0.120	0.82 ± 0.15	0.82 ± 0.15	0.968
	Reaction times in Go trials	486.47 ± 70.21	426.96 ± 49.87	**< 0.001**	0.160	489.80 ± 71.90	498.17 ± 74.78	0.507
	Accuracy in No-Go trials	0.83 ± 0.12	0.88 ± 0.11	0.166	0.010	0.84 ± 0.13	0.84 ± 0.14	0.840
Dimension II: Working memory updating and maintenance								
Updating (*Running memory task*)	Accuracy in 1,750 ms	0.60 ± 0.25	0.74 ± 0.21	**< 0.001**	0.060	0.56 ± 0.26	0.52 ± 0.28	0.395
	Accuracy in 750 ms	0.51 ± 0.29	0.66 ± 0.22	**0.001**	0.050	0.49 ± 0.29	0.44 ± 0.27	0.258
Maintenance (*Digit span backward*;	Span in digit span backward task	5.07 ± 1.51	6.06 ± 2.11	**< 0.001**	0.090	5.05 ± 1.65	5.31 ± 1.76	0.308
Corsi block test)	Span in Corsi block test	4.50 ± 1.00	4.76 ± 1.27	0.101	0.010	4.56 ± 0.99	4.84 ± 1.13	0.258
	Accuracy in Corsi block test	0.53 ± 0.20	0.55 ± 0.21	0.480	0.000	0.54 ± 0.21	0.56 ± 0.20	0.616
Dimension III: Cognitive flexibility/Switching								

Switching (*Number switching task*)	Switching cost	304.80 ± 303.40	198.95 ± 238.37	**0.001**	0.030	292.36 ± 300.00	277.50 ± 286.88	0.840
	Mixing cost	62.43 ± 189.87	112.30 ± 149.69	0.278	0.010	70.90 ± 213.90	14.32 ± 213.43	0.258
Composite scores								
.	Inhibition composite	–11.09 ± 36.83	–3.10 ± 22.12	0.217	0.010	2.02 ± 72.87	–3.85 ± 31.75	0.484
	Executive function composite	98.10 ± 102.30	65.53 ± 80.06	**0.001**	0.030	98.31 ± 103.95	91.39 ± 94.78	0.624
	Inhibition and updating composite	–5.24 ± 18.43	–1.18 ± 11.05	0.217	0.010	1.29 ± 36.43	–1.67 ± 15.89	0.481
	Inhibition and switching composite	146.86 ± 153.44	97.92 ± 120.07	**0.001**	0.030	147.19 ± 155.91	136.82 ± 142.15	0.625
	Updating and switching composite	152.70 ± 151.71	99.84 ± 119.20	**0.001**	0.030	1146.46 ± 150.01	139.01 ± 143.46	0.729

Note: The *p*-values in bold face indicate statistically significant differences ($p<0.05$). The *p*-values were adjusted using the false discovery rate (FDR) with Benjamini–Hochberg procedure. η^2, eta-squared effect size. The interference effect is quantified as the difference in reaction time between incongruent and congruent trials in the *Stroop* task.

1152

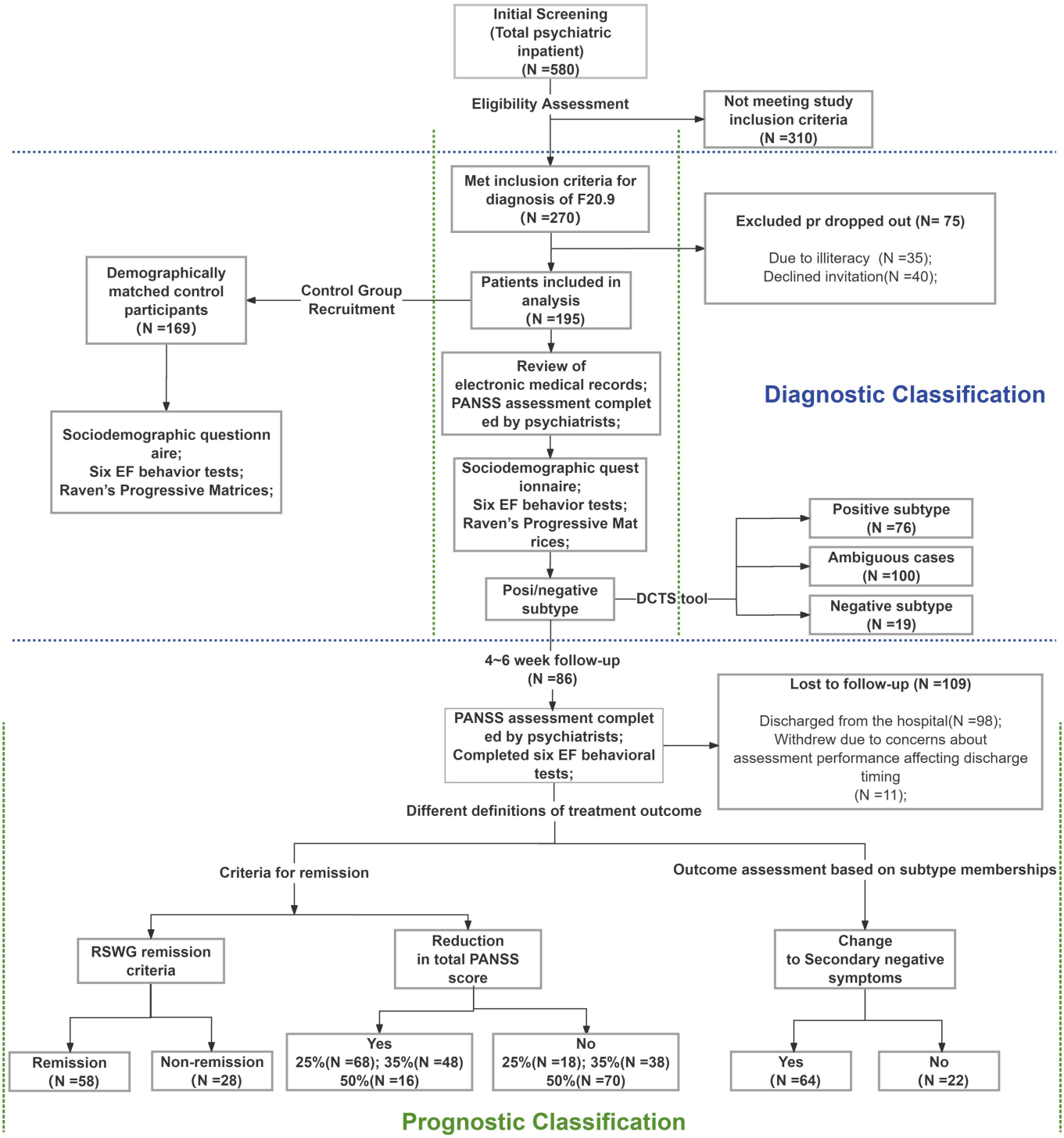
Fig. 1 | Study overview.



A). The diagnostic model included the recruitment of patients with schizophrenia, further stratified into positive and negative symptom subtypes, and healthy controls. All participants underwent a comprehensive battery of tests and assessments tailored to their respective groups. The prognostic model was developed from a cohort of 86 patients with schizophrenia who completed a standard treatment regimen within 4–6 weeks of hospitalization. At follow-up, patients were evaluated using the RSWG criteria as the primary outcome measure. Additionally, three widely accepted definitions in the field were incorporated to comprehensively classify treatment response (i.e., 25%, 35%, and 50% symptom reduction thresholds), as well as changes in positive and negative symptom subtypes. **B).** Data underwent preprocessing, followed by a stratified random division into an 80% discovery dataset and a 20% test dataset, balanced for diagnostic and remission outcomes. The discovery set was subjected to nested cross-validation, with model performance assessed on the test set using various metrics. To reduce splitting variance, this procedure was repeated 100 times. **C).** SHAP values assigned to executive function features indicated their importance in model predictions, with mean absolute values reflecting overall impact. Individual Shapley values highlighted feature influence on correct classifications, which were also assessed for their relationships to psychopathology measures. FDR was used to control for false-positives in multiple comparisons. Correlational analysis between individual-level Shapley values for each feature and individual psychopathology.

Abbreviations: Ada, AdaBoost; FDR, false discovery rate; RF, random forest; RSWG, Remission in Schizophrenia Working Group; SHAP, SHapley Additive exPlanations; SVM, support vector machine. EF, executive function.

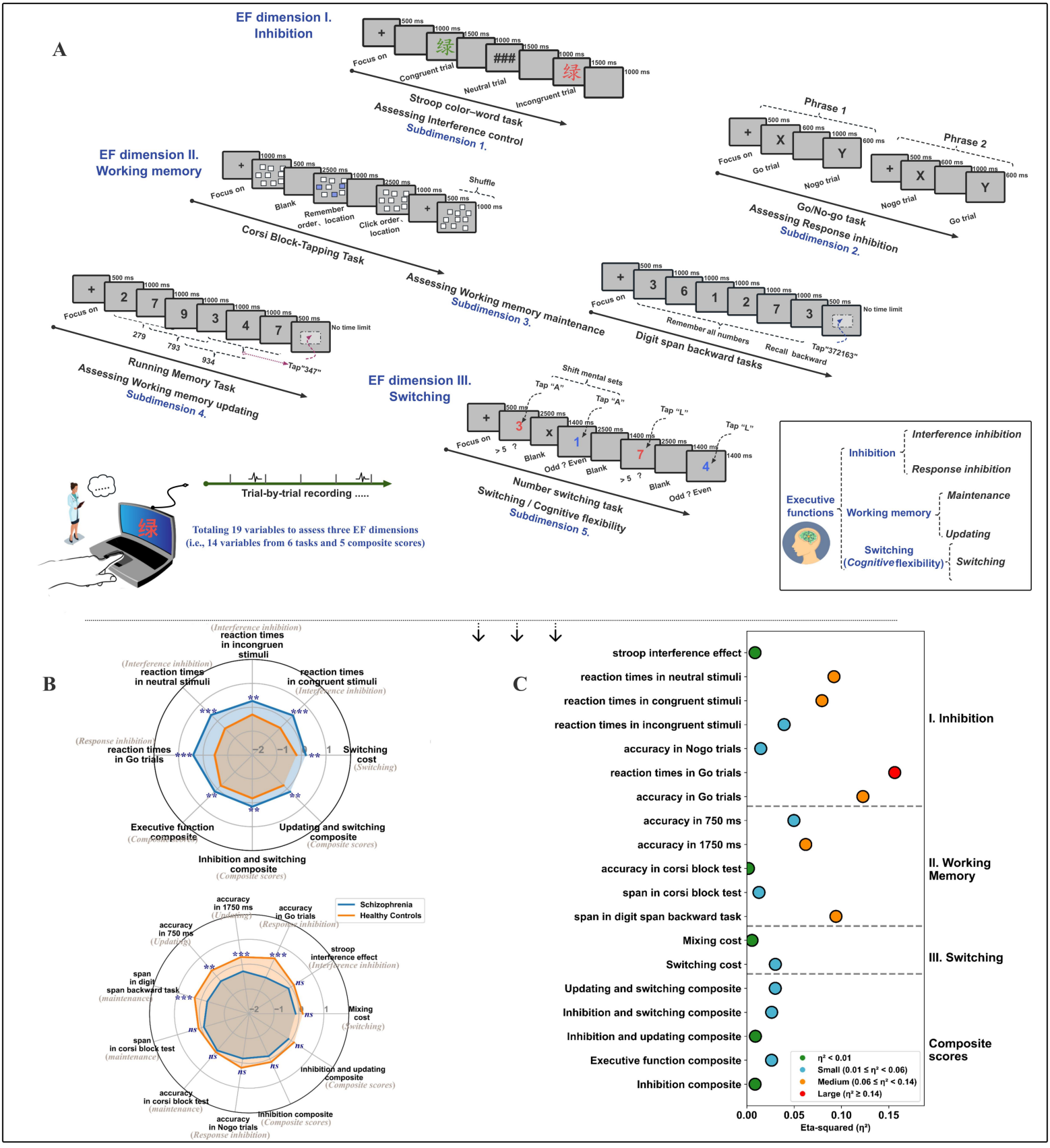
Fig. 2 | Flowchart of patient identification



Flow chart depicting the inclusion and exclusion criteria for participants.

Abbreviations: DCTS, Dimensions and Clustering Tool for Schizophrenia Symptomatology; EF, executive function; PANSS, Positive and Negative Syndrome Scale; RSWG, Remission in Schizophrenia Working Group.

1172 **Fig. 3 | Multi-task executive function dimension assessments**



1173

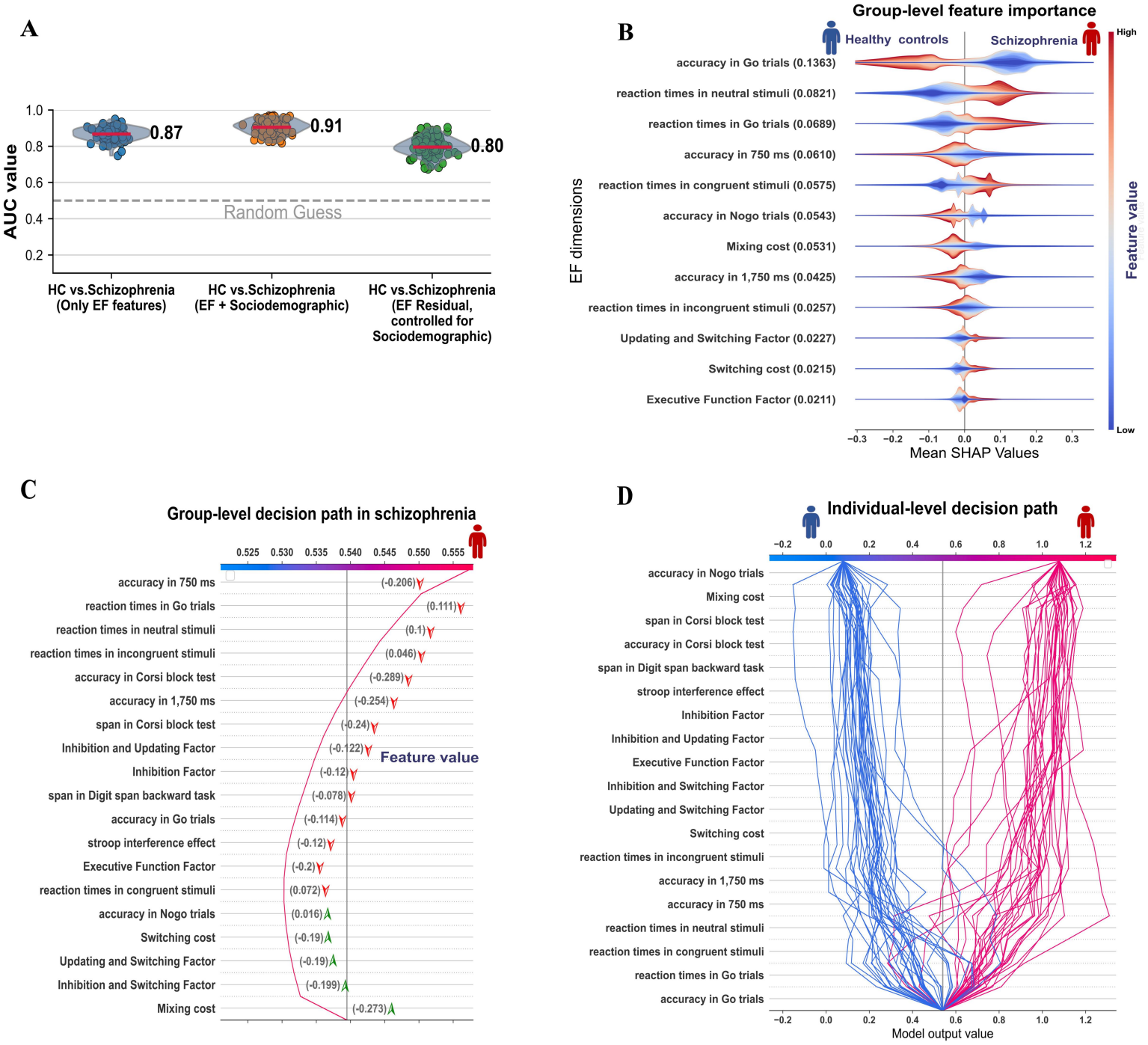
1174 **A).** The six behavioral paradigms (e.g., Zhao et al., 2023)[40] used to assess the three EF

dimensions (i.e., inhibitory control, working memory maintenance and updating, and cognitive flexibility) based on 14 measurements. **B)** The radar plot was constructed based on the Z-scores of the 14 EF measurements and the 5 composite scores, with annotated *p*-values (FDR corrected) resulted from one-way ANCOVAs following a mixed model ANCOVA. **C**). Effect sizes are colored as small ($0.01 \leq \eta^2 < 0.06$), medium ($0.06 \leq \eta^2 < 0.14$), or large ($\eta^2 \geq 0.14$) based on the guidelines proposed by Cohen (1988).

Abbreviations: EF, executive function; FDR, false discovery rate; *ns*, not significant.

Note: *n*s: Not significant difference. *: $p < 0.05$. **: $p < 0.01$. ***: $p < 0.001$.

1184 **Fig. 4 | Model metrics and feature importance for diagnostic models**



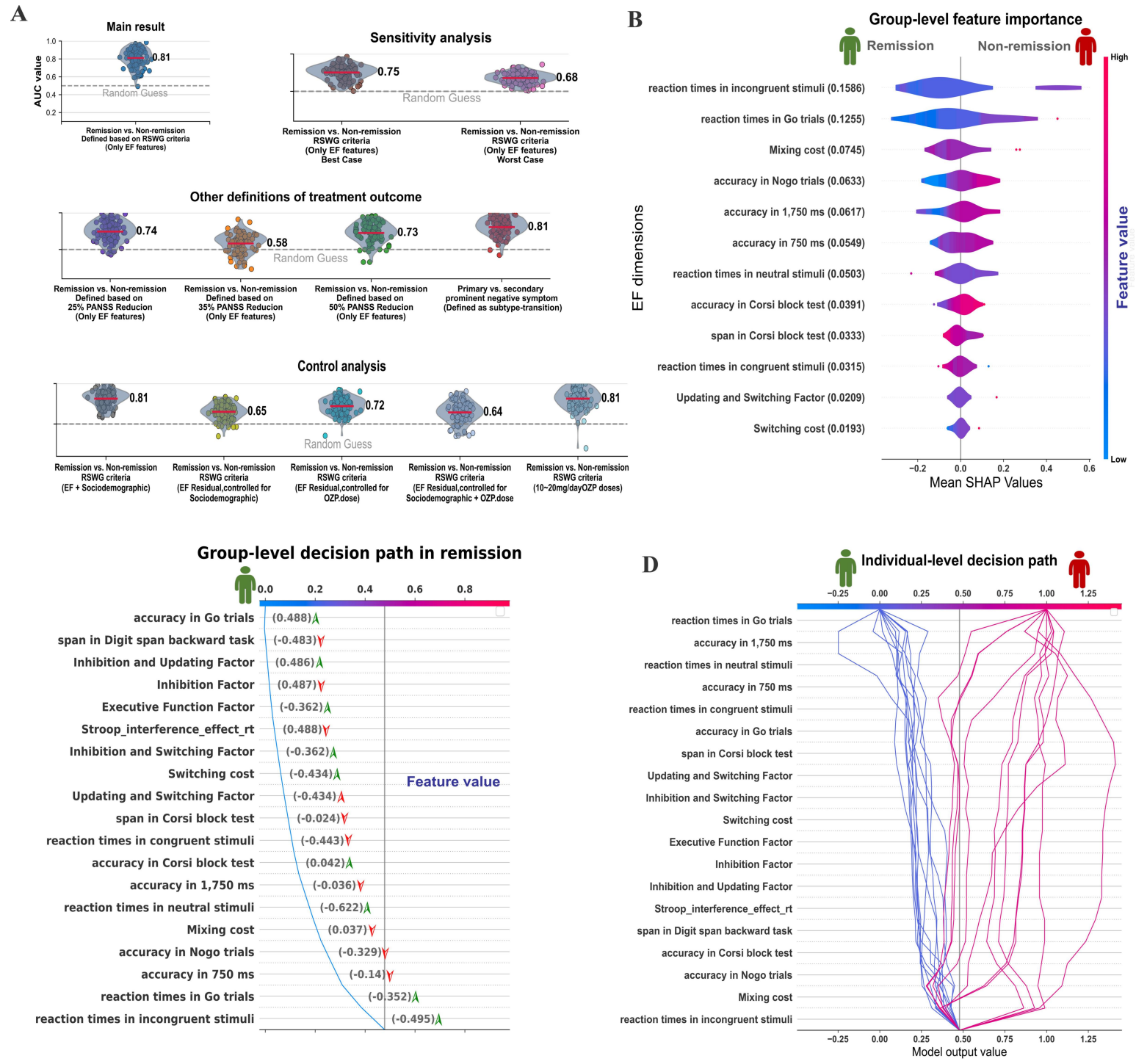
1185

1186 **A).** Violin plots show the values of area under the curve (AUC) for discriminating patients
1187 with schizophrenia from healthy participants by diagnostic classification models. Each point
1188 within the violin plots represents the AUC value derived from the hold-out test data of each
1189 random split procedure (repeated 100 times) in our machine learning design. The red line
1190 within each violin plot denotes the mean. **B).** The group-level feature importance plot ranks
1191 EF features on the y-axis by their absolute average Shapley value across individuals,
1192 representing their overall importance in the ability of the model to distinguish between
1193 patients and healthy participants. The original feature weights for each EF variable were color

coded, with blue color denoting a negative weight value and red color denoting a positive weight value. Values along the x-axis indicate a positive or negative effect of an EF feature on classifying an individual, with a negative and positive value promoting the model towards a classification of "healthy" and “schizophrenia”, respectively. Collectively, these findings indicated that higher accuracy in the Go trials was associated with a higher likelihood of classifying an individual as healthy. **C).** The group-level decision path, generated based on all correctly classified schizophrenia samples. EF features are ranked from upper to bottom based on their group-level importance present in b), the color bar denotes the impact of an EF feature on model’s classification towards “healthy” (blue) or “schizophrenia” (red), the red curve shows the values for each EF feature coded in the color bar. The numbers in parentheses represent the z-score standardized original measurement values of each EF feature by the averages of this EF feature across the healthy and the patient groups in the model test samples (a negative number to the right of the perpendicular line denotes the measurement of an EF feature that is below the average in patients with schizophrenia). The red or green arrows adjacent to the parentheses indicate the higher or lower values within the parentheses that are associated with better or worse EF functions, respectively. This is because some measurements from the EF tasks indicate better performance with higher values, while others are more favorable with lower values. **D).** The decision path for each individual in the test sets of the classification modeling iterations, that for each individual, how each of these important EF features have promoted the diagnostic model to classify this individual as a healthy participant or patient with schizophrenia, given the expression level (task measurements) of this individual in each of the EF features. The blue and red lines indicate accurate classifications as healthy participants and schizophrenia patients, respectively.

Abbreviations: EF, executive function; HC, healthy control.

1218 **Fig. 5 | Model metrics and feature importance for prognostic models**



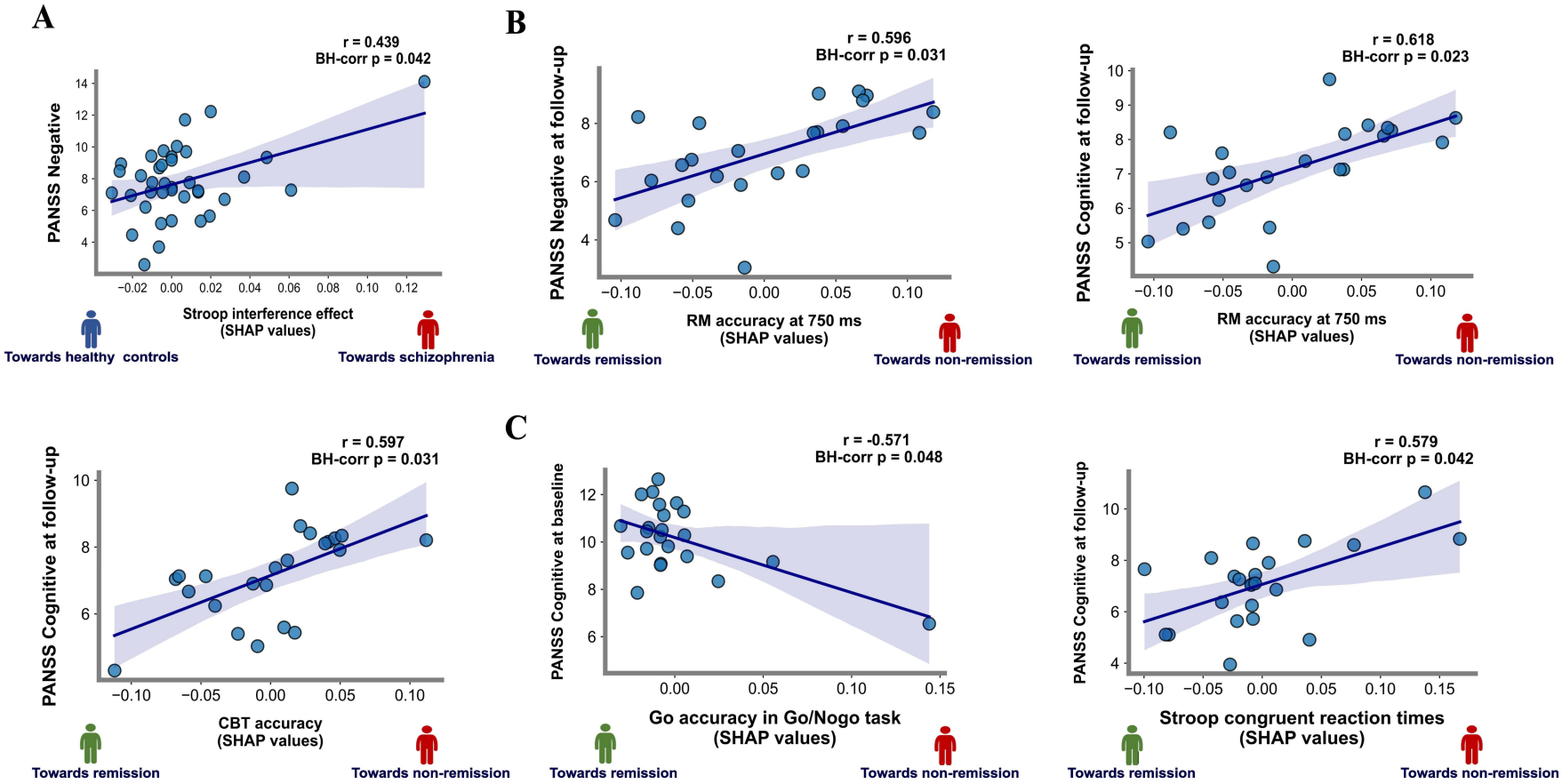
1219

1220 **A).** Violin plots show the values of area under the curve (AUC) for discriminating remitted
1221 patients with schizophrenia from non-remitted patients by prognostic classification models.
1222 Each point within the violin plots represents the AUC value derived from the hold-out test
1223 data of each random split procedure (repeated 100 times) in our machine learning design. The
1224 red line within each violin plot denotes the mean. **B).** The group-level feature importance plot
1225 ranks EF features on the y-axis by their absolute average Shapley value across individuals,
1226 representing their overall importance in the model's distinction between remitted and

non-remitted patients (i.e., greater separation of the violin-like plots towards the extremes denotes higher importance). The original feature weights for each EF variable are color-coded, with blue indicating a negative weight value and red indicating a positive weight value. Values along the x-axis indicate a positive or negative effect of an EF feature on classifying an individual, with a negative and positive value promoting the model towards a classification of "remission" and “non-remission”, respectively. Collectively, these data suggest that longer reaction times in incongruent stimuli are associated with a higher likelihood of classification as non-remitted patient. **C).** The group-level decision path, generated based on all correctly classified remitted schizophrenia patients. EF features are ranked from upper to bottom based on their group-level importance present in b), the color bar denotes the impact of an EF feature on model’s classification towards “remission” (blue) or “non-remission” (red), and the blue curve shows the values for each EF feature coded in the color bar. The numbers in parentheses represent the z-score standardized original measurement values of each EF feature by the averages of this EF feature across the remitted and non-remitted patient subgroups in the model test samples (a negative number to the left of the perpendicular line denotes the measurement of an EF feature that is below the average in remitted patients). The red or green arrows adjacent to the parentheses indicate the higher or lower values within the parentheses that are associated with better or worse EF functions, respectively, as some measurements from the EF tasks indicate better performance with higher values, while others are more favorable with lower values. **D).** The decision path for each individual in the test sets of the classification modeling iterations, illustrating how each of these important EF features influenced the prognostic model to classify each individual as remitted or non-remitted, given their performance (task measurements) in each of the EF features. The blue and red lines represent correct classifications as remitted and non-remitted patients, respectively.

Abbreviations: EF, executive function; OZP, olanzapine; PANSS, Positive and Negative Syndrome Scale; RSWG, Remission in Schizophrenia Working Group; SHAP, SHapley Additive exPlanations.

Fig. 6 | Correlation between the importance of an executive function feature and individual psychopathology along four symptom dimensions



A). Correlation in the diagnostic model using executive function features only. **B).** Correlation in the prognostic model using executive function features only. **C).** Correlation in the prognostic model using executive function features and sociodemographic features.

Abbreviations: BH, Benjamini–Hochberg correction; CBT, Corsi block test, assessing numeric working memory maintenance capacity; Go/No-Go task, assessing response inhibition function; PANSS, Positive and Negative Syndrome Scale; RM, running memory task, assessing working memory updating capability; SHAP, SHapley Additive exPlanations; Stroop, Stroop task, assessing interference control function.