**Schizophrenia Bulletin** vol. 52 no. 2 pp. 1–20, 2026
https://doi.org/10.1093/schbul/sbaf218

**REGULAR ARTICLE**

# 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]; Ji Chen[8,10,11,*]**

[1]School of Psychology, Northwest Normal University, Lanzhou 730070, China; [2]Centre for Sleep and Cognition & Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117549, Singapore; [3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore; [4]N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore 117456, Singapore; [5]Integrative Sciences & Engineering Programme, National University of Singapore, Singapore 119077, Singapore; [6]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA 02129, United States; [7]The Third People's Hospital of Lanzhou, Lanzhou 730050, China; [8]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich 52425, Germany; [9]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany; [10]Center for Brain Health and Brain Technology, Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China; [11]School of Psychology, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed: Xin Zhao, School of Psychology, Northwest Normal University, Lanzhou 730070, China. (psyzhaoxin@nwnu.edu.cn) and Ji Chen, Center for Brain Health and Brain Technology, Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: (chen.ji@sjtu.edu.cn)

T. Zhang and X. Zhao contributed equally to this work.

***Background:*** **Executive function (EF) impairments are often seen in mental disorders, particularly schizophrenia (SZ), where they relate to adverse outcomes. As a heterogeneous construct, how specifically each dimension of EF to characterize the diagnostic and prognostic aspects of SZ remains opaque.**

***Study Design:*** **We used classification models with a stacking approach on systematically measured EFs using 6 tasks to discriminate 195 patients with SZ 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 4 symptom dimensions.**

***Study Results:*** **The models highlighted the importance of inhibitory control (interference and response inhibitions) or working memory (WM) in accurately identifying individuals with SZ (area under the curve [AUC] = 0.87) or those in remission (AUC = 0.81). Patients who are correctly classified, in the association with the contribution of interference inhibition function to our diagnostic classifier, present more severe baseline negative symptoms compared to those who are more likely to be misclassified. Also, linked to the function of WM updating, patients who are successfully classified as remitted display milder cognitive symptoms at follow-up. Remitted patients do not differ significantly from non-remitted cases in baseline EF assessments or overall symptom severity.**

***Conclusions:*** **Our work indicates that impairments in specific EF dimensions in SZ are differentially linked to individual symptom-load and prognostic outcomes. Thus, assessments and models based on EF may be promising in the clinical evaluation of this disorder.**

**Key words:** schizophrenia; executive functions; machine learning; prognostic prediction.

## Introduction

Cognitive symptoms, a core aspect of schizophrenia (SZ), emerge in the prodromal phase and persist throughout the illness[1]—unlike in affective psychotic disorders or drug-induced psychosis, where cognitive deficits are epiphenomenal. Symptoms involve impaired mental functions related to memory, attention, and executive tasks.[2] Executive function (EF) represents higher-order cognitive processes involving impulse control and behavior orchestration.[3] EF deficits hinder goal-directed activity and contribute to poor medication compliance, worsening clinical outcomes.[4]

In psychiatry, prognostic prediction remains challenging. Assessments using baseline neuroimaging, genetic, or clinical data often approach chance-level accuracy.[5,6] EF deficits—related to disease progression, symptom severity, and loss of social and occupational skills[7]—represent potential clinical and prognostic

markers for SZ and align well with the 3-factor model of EF of Miyake et al.,[8] characterizing (1) interference inhibition and response inhibition; (2) cognitive flexibility (CF) and switching; and (3) working memory (WM) updating and maintenance.[9] These EF dimensions differ in concept and neurobiological substrate, highlighting the need to assess them separately when correlating them with diagnostic and prognostic aspects of SZ. Furthermore, the inhibitory and WM dimension contain sub-functions—specifically, interference inhibition and response inhibition fall under the inhibition dimension, while the WM dimension contains updating (dynamic manipulation) and maintenance (short-term storage) functions. The inhibition and WM constructs each comprise at least 2 distinct sub-functions,[10,11] which can be differentially affected in SZ.[12,13] Such finer characterization of EFs may assist in understanding specific psychopathological processes in SZ, for example, disorganization symptoms.[14,15] EF disturbances have been implicated in negative and positive symptoms. Failure in effectively monitoring volitional behaviors and inhibiting false inference in predictive processing would give rise to positive symptoms (eg, hallucinations).[16,17] Impaired WM updating and maintenance is linked to poor abstract thinking,[18] which increases the risk of developing negative symptoms (eg, apathy) common in SZ.[19,20] However, some EF dimensions may be more severely affected than others,[21] leaving unknown which EF dimension deficits best characterize SZ and offer prognostic information.

Widely used tools, for example, the Cambridge Neuropsychological Test Automated Battery[22] and the National Institutes of Health Toolbox,[23] do not cover various EF dimensions. Assessments included in available neuropsychological batteries (eg, Delis–Kaplan EF System) do not offer trial-by-trial dynamic quantification that could detect subtle cognitive impairments, including in EF.[24,25] Furthermore, routinely used cognitive symptom items, such as in the Positive and Negative Syndrome Scale (PANSS), are retrospective, based on information collected from patient interviews or relatives' contributions. In contrast, task paradigms offer a tailored assessment strategy, with objective trial-by-trial tests that measure particular cognitive functions, with improved sensitivity and specificity.[9,26]

Here, 6 well-established behavioral paradigms were implemented to assess individual functions along 3 EF dimensions to determine their role in characterizing SZ patients at base-line and their prognostic statuses at follow-up (Figure 1). This is tested via establishing diagnostic and prognostic classification models using multiple machine learning methods, along with their stacked model, following a stringent nested cross-validation (CV) and independent testing. The importance of EF feature contributions to classification models was determined at both the group- and individual-levels, which the latter further facilitates establishing a link between feature importance and individual psychopathology.

## Methods

### *Participant Recruitment, Clinical Characterization, and Definition of Prognostic Status*

We recruited 195 SZ aged $35.35 \pm 9.35$ years (81 females), and 169 age- and gender-matched healthy individuals (HC group) (Table 1). Eighty-six patients were successfully followed with reassessment after 4-6 weeks. Diagnoses were made by 2 resident psychiatrists using the ICD-10 diagnostic criteria for SZ. Participants provided written informed consent, and were screened with the Structured Clinical Interview for DSM-IV axis I Disorders. Patients were stable and receiving consistent treatment, with no medication changes expected during the study (cf. Table S1). This study was approved by the Ethics Committee at the Third People's Hospital of Lanzhou City and Northwest Normal University (Lanzhou, China).

Symptom severity was assessed using the PANSS.[27] Scores of 4 dimensions representing positive, negative, affective, and cognitive symptoms 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,[27] with higher scores denoting more severe symptoms. Patient subgroups, either showing prominent positive or negative symptoms, were also defined using this DCTS tool to explore potential interactions between symptom profiles, classification accuracy, and key EF impairments. Patients with an ambiguous symptom profile were excluded to improve specificity.

For the 86 followed-up patients, reassessments included item-wise PANSS and EF behaviors. Their remission versus non-remission prognostic statuses were determined based on the Remission in Schizophrenia Working Group (RSWG) criteria[28] defined by scores $\leq 3$ on specific PANSS items (P1, P2, P3, N1, N4, N6, G5, G9). Considering the lack of an established definition of clinical outcomes in SZ, we utilized an alternative definition based on different rates (25%, 35%, and 50%) of reduction in total PANSS score.[29] Prognostic statuses may also be reflected in a change of subtype membership from baseline to the end of follow-up, particularly surrounding the negative symptom subtype, which we differentiated: (1) baseline non-negative subtype (ie, positive subtype and the ambiguous class wherein the patients lack clear assignment) with transition to a negative subtype; and (2) stable negative subtype during follow-up. Negative symptoms can be primary (stable) or secondary to positive symptoms and antipsychotic treatments[30] which relate to poorer clinical outcomes.[31]

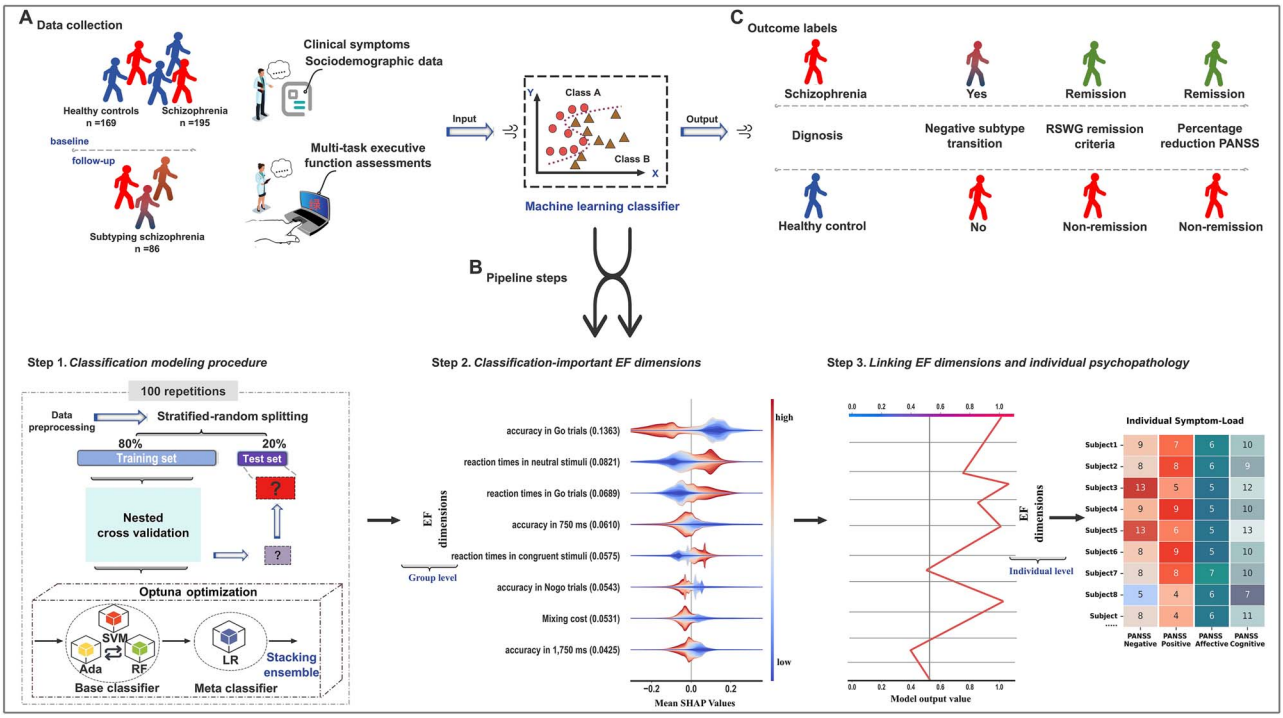


**Figure 1.** Study overview. (A) The diagnostic model included the recruitment of patients with schizophrenia, further stratified into positive and negative symptom subtypes, and healthy controls. All participants underwent a comprehensive battery of tests and assessments tailored to their respective 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, 3 widely accepted definitions in the field were incorporated to comprehensively classify treatment response (ie, 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.

### *Assessments*

*Executive Function.* We adopted a 3-dimensional definition of EF, specifying inhibitory control (IC) (interference inhibition and response inhibition), WM (updating and maintenance), and CF[8]. Six behavioral tasks were employed to measure these dimensions[32] (Figure 2A): (1) *number running memory updating* task (WM updating)[33]; (2) *digit span backward* task (WM maintenance)[32]; (3) *Corsi block* test (WM maintenance)[34]; (4) *Stroop* task (interference inhibition)[35]; (5) *Go/No-Go* task (response inhibition)[36]; and (6) *number switching* task (CF).[37] Accuracy and reaction time (RT) for these tasks were used to compute values for 14 assessment indicators, that is, features for model training. These indicators are effective in measuring different aspects of EF (cf. Table S2) to reflect the underlying latent construct,[38] and are sensitive to SZ-related deficits.[39]

*Inhibitory Control.* The 4 measurements for assessing the interference control function included the RT for the incongruent, congruent, and neutral stimuli, and the difference in RT between the congruent and the incongruent condition trials (ie, the interference effect) in the *Stroop* task. Three measurements for assessing the response inhibition function included the RT 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 WM updating function included the proportion of digits correctly recalled and placed in the correct sequence at 2 different speeds of presentation (1750 and 750 ms per digit) in the *running memory* task. Three measurements for assessing the numeric WM maintenance capacity included the length of the last correctly repeated sequence, the count of sequences correctly repeated until the conclusion of the test (ie, 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.

**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 | .055 | 34.03 ± 8.69 | 34.54 ± 9.75 | .810 |
| Sex | | | .264 | | | .157 |
| Male | 114 (58.5%) | 88 (52.1%) | | 39 (67.2%) | 14 (50.0%) | |
| Female | 81 (41.5%) | 81 (47.9%) | | 19 (32.8%) | 14 (50.0%) | |
| Ethnicity, Han | 173 (88.7%) | 152 (89.9%) | .837 | 52 (89.7%) | 24 (85.7%) | .465 |
| Education, years | 11.12 ± 4.58 | 10.90 ± 3.94 | .615 | 10.52 ± 3.98 | 11.07 ± 4.24 | .555 |
| BMI | 23.21 ± 3.73 | 23.95 ± 4.32 | .494 | 23.60 ± 3.13 | 22.65 ± 3.78 | .217 |
| Residence, urban | 114 (58.5%) | 98 (58.0%) | 1.000 | 27 (46.6%) | 16 (57.1%) | .255 |
| SES | 23.21 ± 7.32 | 23.88 ± 5.80 | .338 | 22.72 ± 7.25 | 23.43 ± 8.02 | .684 |
| RPM | 32.56 ± 11.50 | 39.52 ± 9.63 | **<.001** | 34.84 ± 11.17 | 33.04 ± 12.49 | .500 |
| Employed, yes | 53 (27.2%) | 111 (65.7%) | **<.001** | 15 (25.9%) | 5 (17.9%) | .587 |
| Only child, yes | 63 (32.3%) | 24 (14.2%) | **<.001** | 42 (72.4%) | 14 (50.0%) | .150 |
| Marital status | | | | | | |
| Unmarried | 115 (59.0%) | 54 (32.0%) | **<.001** | 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 | | | .079 | | | .508 |
| Never | 119 (61.0%) | 122 (72.2%) | | 34 (58.6%) | 20 (71.4%) | |
| 1-3 years | 19 (9.7%) | 11 (6.5%) | | 6 (10.3%) | 2 (7.1%) | |
| >3 years | 57 (29.2%) | 36 (21.3%) | | 18 (31.0%) | 6 (21.4%) | |
| Alcohol consumption history | | | | | | |
| Never | 133 (68.2%) | 99 (58.6%) | .118 | 42 (72.4%) | 18 (64.3%) | .968 |
| 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 | .906 |
| Duration of disorder | 9.74 ± 7.17 | . | . | 9.09 ± 5.98 | 10.48 ± 7.51 | .395 |
| Frequency of episodes | 5.98 ± 4.27 | | | 6.17 ± 3.84 | 6.32 ± 4.05 | .871 |
| First episode, yes | 10 (5.1%) | . | . | 1 (1.7%) | 1 (3.6%) | .548 |
| Family medical history, yes | 33 (16.9%) | . | . | 10 (17.2%) | 4 (14.3%) | .971 |
| Dose equivalent to olanzapine (mg/day) | 14.48 ± 6.41 | . | . | 14.30 ± 5.46 | 16.65 ± 7.66 | .154 |
| Type of antipsychotic medication | | | | | | .829 |
| First generation | 9 (4.6%) | . | . | 2 (3.4%) | 2 (7.1%) | |
| Second generation | 186 (95.4%) | . | . | 56 (96.6%) | 26 (92.9%) | |
| Clinical scale | | | | | | |
| 3 PANSS subscales | | | | | | |
| PANSS-Negative | 21.42 ± 6.48 | . | . | 21.59 ± 6.59 | 21.79 ± 5.99 | .893 |
| PANSS-Positive | 22.01 ± 4.46 | . | . | 21.64 ± 4.06 | 21.18 ± 5.99 | .676 |
| PANSS-General | 40.30 ± 6.81 | . | . | 39.67 ± 7.39 | 40.39 ± 6.20 | .657 |
| PANSS-Total | 83.79 ± 13.85 | . | . | 82.90 ± 13.97 | 83.36 ± 15.33 | .890 |
| 4 dimensions of PANSS | | | | | | |
| Negative factor | 8.22 ± 2.50 | . | . | 8.24 ± 2.70 | 8.71 ± 2.11 | .425 |
| Positive factor | 6.24 ± 1.74 | . | . | 5.99 ± 1.66 | 6.23 ± 2.07 | .560 |
| Affective factor | 5.65 ± 0.86 | . | . | 5.67 ± 0.93 | 5.57 ± 0.78 | .619 |
| Cognitive factor | 9.78 ± 1.72 | . | . | 9.66 ± 1.82 | 9.76 ± 1.61 | .823 |

Data are presented as the mean ± standard deviation or *n* (%). The *P*-values in bold face indicate statistically significant differences between groups ($P < .05$). Calculation: BMI is calculated as weight (kg) divided by height squared ($m^2$). 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.

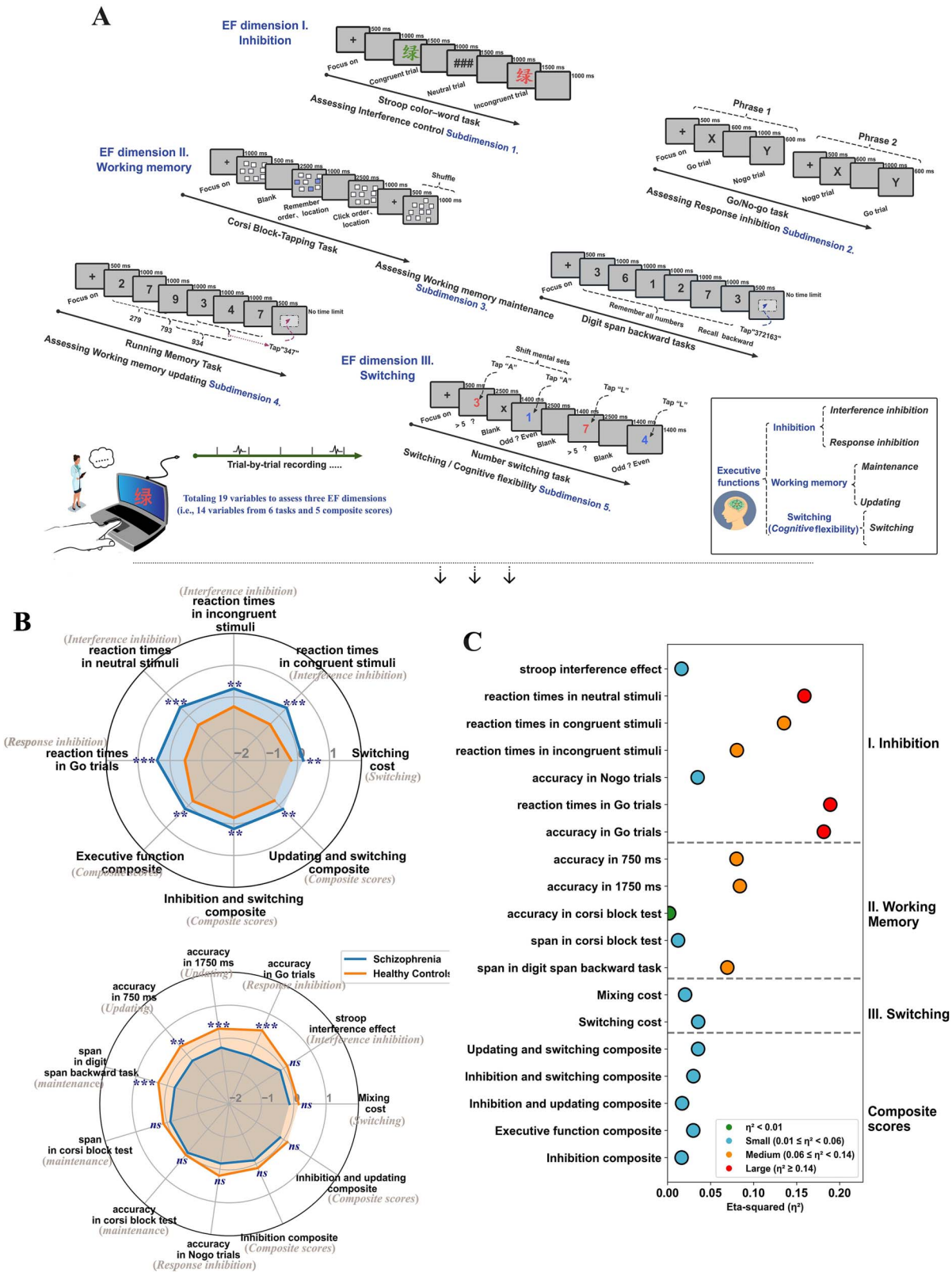


**Figure 2.** Multi-task executive function dimension assessments. (A) The 6 behavioral paradigms (eg, Zhao et al.[40]) used to assess the three EF dimensions (ie, inhibitory control, working memory maintenance and updating, and cognitive flexibility) based on 14 measurements. (B) The radar plot was 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. *ns*: Not significant difference. $^{*}P < .05$; $^{**}P < .01$; $^{***}P < .001$.

*Cognitive Flexibility.* Two measurements included the difference in RT between the switch and the non-switch trials [switch cost] as well as between the non-switch and the single-task trials [mixing cost] measured in the *number-letter switching* task.

Our measurements were in line with those commonly used in works involving *Miyake's* 3 theoretical factors.[8] For instance, compared with the "Heaton's Wisconsin Card Sorting Test (WCST)," the number-letter switching task we adopted provides a more process-pure measurement of CF[40] by reducing engagement of other functions such as WM and rule learning.[41] The CF domain is unidimensional, while the IC and WM domains comprised sub-functions. Therefore, separate tasks are required for comprehensive measurement, leading to a greater number of variables for characterizing these 2 domains than CF. To evaluate the construct validity of the variables included against the theoretical 3-factor structure, confirmatory factor analysis (CFA) was conducted.

Besides conceptual formulation of the 14 task measurements according to the 3-dimensional representation of EF, we supplemented 5 composite scores calculated based on these measured variables.[42] This is because real-world cognitive tasks often require multiple EFs to engage simultaneously for goal representation and maintenance.[43] It is possible that in SZ, the interaction between these functions is impaired,[44] leading to symptoms and poor clinical outcomes.[39,45] Consequently, composite scores that measure this interplay may offer a distinct predictive contribution beyond that of any single variable. An inhibitory composite score was calculated by averaging: (1) the difference in RT between the congruent and the incongruent condition trials; and (2) the accuracy for the No-Go stimuli in the *Go/No-Go* task[46,47] to denote the combined response and interference 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 1750 ms per digit) and the *number-letter switching* task (switch cost) to abbreviate general EF functions. Such abbreviation complies with the previous notion on a single-condition indicator that these trials require greater executive control demands.[38] We moreover created 3 cross-dimensional EF composite scores[48]: (1) Inhibition/Updating composite; (2) Inhibition/Switching composite; and (3) Switching/Updating composite.

### *Between-Group Comparisons in EF and Symptoms*

Multiple one-way ANCOVAs corrected using a false discovery rate (FDR) approach were followed by a mixed-model ANCOVA with group (SZ vs. healthy control) as a between-subjects factor and all EF variables as within-subjects factors, controlling for sociodemographic variables, to examine whether SZ differentially affected EF performance. Two-sample *t*-tests compared symptoms, demographics, and EF between remission and non-remission subgroups. Chi-squared tests compared categorical variables between these groups. Paired-samples *t*-tests compared symptoms between baseline and follow-up assessments; and Pearson and Spearman correlation analyses were performed to examine the relationships among symptoms and EF where appropriate.

### *Classification Modeling Procedure*

*Features and Models.* Two feature sets were tested for diagnostic classification (SZ vs. HC): (1) 19 baseline EF assessments, reflecting the 3 EF dimensions (Table 2); and (2) 32 features—the 19 baseline EF measures plus 13 sociodemographic variables (age, years of education, body mass index, socioeconomic status, and fluid intelligence as well as sex, ethnicity, residence, employment status, only-child status, marital status, smoking history, and drinking history). For prognostic classification (remission vs. non-remission) (Table S3), the same 2 feature sets (Table S4) were tested. Support vector machine (SVM), random forest (RF), and AdaBoost—widely used in psychiatric machine-learning research[49]—were used for classification tasks, along with a synthesized stacking model of the three.[50] After training the 3 base-models separately, they are provided examples from an intermediate holdout set and their predictions are collected. These predictions serve as the input features to a meta-model (stacked atop the base-models), which is tuned to the intermediate holdout set. Finally, the full system of models is used to make predictions about examples that neither the base-models nor the meta-models were trained on (the final holdout set).

*Machine Learning Design.* The original data were first preprocessed to accommodate missing values, outliers, and class imbalance issues (see Supplementary material for details). Then, data were 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 to avoid potential bias from random splitting,[51,52] which resulted in 100 hold-out, test sets. This approach effectively gauges generalization while balancing practical acquisitions of clinical sample data.[53,54] Specifically, in each repeat, we randomly split the preprocessed data into a discovery dataset with 80% of the overall SZ 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 (Figure 1).[53-56] The random split procedure was stratified for the outcome variable (diagnostic label or remission status), ensuring balanced representation of labels in each dataset.[42,43] Using

**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 | $\eta^2$ | 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 | **.002** | 0.040 | 719.80 ± 133.13 | 677.10 ± 124.85 | .258 |
| | Reaction times in congruent stimuli | 704.26 ± 132.24 | 604.55 ± 118.40 | **<.001** | 0.080 | 723.00 ± 133.13 | 677.56 ± 121.80 | **.022** |
| | Reaction times in incongruent stimuli | 751.79 ± 148.77 | 663.61 ± 149.02 | **<.001** | 0.090 | 764.28 ± 152.17 | 721.14 ± 140.05 | **.049** |
| | Stroop interference effect | −23.01 ± 73.67 | −7.09 ± 44.24 | .217 | 0.010 | 3.20 ± 145.73 | −8.53 ± 63.52 | .616 |
| **Response inhibition** (Go/No-Go task) | Accuracy in Go trials | 0.83 ± 0.14 | 0.94 ± 0.08 | **<.001** | 0.120 | 0.82 ± 0.15 | 0.82 ± 0.15 | .968 |
| | Reaction times in Go trials | 486.47 ± 70.21 | 426.96 ± 49.87 | **<.001** | 0.160 | 489.80 ± 71.90 | 498.17 ± 74.78 | .507 |
| | Accuracy in No-Go trials | 0.83 ± 0.12 | 0.88 ± 0.11 | .166 | 0.010 | 0.84 ± 0.13 | 0.84 ± 0.14 | .840 |
| **Dimension II: working memory updating and maintenance** | | | | | | | | |
| **Updating** (Running memory task) | Accuracy in 1750 ms | 0.60 ± 0.25 | 0.74 ± 0.21 | **<.001** | 0.060 | 0.56 ± 0.26 | 0.52 ± 0.28 | .395 |
| | Accuracy in 750 ms | 0.51 ± 0.29 | 0.66 ± 0.22 | **.001** | 0.050 | 0.49 ± 0.29 | 0.44 ± 0.27 | .258 |
| **Maintenance** (Digit span backward; Corsi block test) | Span in digit span backward task | 5.07 ± 1.51 | 6.06 ± 2.11 | **<.001** | 0.090 | 5.05 ± 1.65 | 5.31 ± 1.76 | .308 |
| | Span in Corsi block test | 4.50 ± 1.00 | 4.76 ± 1.27 | .101 | 0.010 | 4.56 ± 0.99 | 4.84 ± 1.13 | .258 |
| | Accuracy in Corsi block test | 0.53 ± 0.20 | 0.55 ± 0.21 | .480 | 0.002 | 0.54 ± 0.21 | 0.56 ± 0.20 | .616 |
| **Dimension III: cognitive flexibility/switching** | | | | | | | | |
| **Switching** (Number switching task) | Switching cost | 304.80 ± 303.40 | 198.95 ± 238.37 | **.001** | 0.030 | 292.36 ± 300.00 | 277.50 ± 286.88 | .840 |
| | Mixing cost | 62.43 ± 189.87 | 112.30 ± 149.69 | .278 | 0.010 | 70.90 ± 213.90 | 14.32 ± 213.43 | .258 |
| **Composite scores** | | | | | | | | |
| | Inhibition factor | −11.09 ± 36.83 | −3.10 ± 22.12 | .217 | 0.010 | 2.02 ± 72.87 | −3.85 ± 31.75 | .484 |
| | Executive function factor | 98.10 ± 102.30 | 65.53 ± 80.06 | **.001** | 0.030 | 98.31 ± 103.95 | 91.39 ± 94.78 | .624 |
| | Inhibition and updating factor | −5.24 ± 18.43 | −1.18 ± 11.05 | .217 | 0.010 | 1.29 ± 36.43 | −1.67 ± 15.89 | .481 |
| | Inhibition and switching factor | 146.86 ± 153.44 | 97.92 ± 120.07 | **.001** | 0.030 | 147.19 ± 155.91 | 136.82 ± 142.15 | .625 |
| | Updating and switching factor | 152.70 ± 151.71 | 99.84 ± 119.20 | **.001** | 0.030 | 146.46 ± 150.01 | 139.01 ± 143.46 | .729 |

The *P*-values were adjusted using the false discovery rate (FDR) with Benjamini–Hochberg procedure. $\eta^2$, eta-squared effect size. The interference effect is quantified as the difference in reaction time between incongruent and congruent trials in the *Stroop* task.

the discovery dataset, we performed a nested CV loop[57] (also termed double CV), which differentiates 2 CV roles to avoid "circularity" introduced by overfitting when the same sample subset is used for both hyperparameter tuning and model validation.[58] 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 in the validation sets of the inner loop.[59] The outer CV ($k = 5$) is subsequently utilized for out-of-sample validation and parameter assessment.[60] Hyperparameters with the highest average performance over the $5 \times 3$ 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.[61] Performance metrics were then assessed on the 100 test sets. In addition to AUC, sensitivity, specificity, and balanced accuracy performance metrics were assessed.[62]

*Feature Importance.* To evaluate the contributions of EF features to our classifications, we assigned an importance score (ie, Shapley value) to each feature.[47] Specifically, we used the SHAP library (version 0.39.0) model-agnostic SHAP KernelExplainer approach generally used to estimate Shapley values for prediction models.[63] After obtaining individual-level Shapley values for each feature based on a weighted linear regression model as detailed elsewhere,[64] we computed the mean absolute Shapley value (ie, feature importance score) across all individuals (ie, the group-level Shapley values) where larger Shapley values indicate stronger importance of this feature to the classification model. SHAP analysis was performed on the classifiers best-performing in the test sets with the 19 EF assessments.[65]

To determine the directional contribution of EF dimensions for classification, a decision path analysis was performed. Among correctly classified participants in the SZ group, performance by each EF dimension assessment was averaged across both the HC and SZ groups for the diagnostic models, and across the remission and non-remission patient groups for the prognostic models. The averaged performance was then used to normalize the original measurement values of each EF feature for correctly assigned patients, with a positive (or negative) value denoting higher (or lower) performance of an EF feature that has driven the model towards an accurate classification. Using individual Shapley values, the decision path was plotted for each correctly classified individual, because SZ have heterogeneous expressions across EF dimensions.[66] Finally, based on these values, we tested the extent to which the importance of the contribution of an EF feature to a classification was linked to individual psychopathology along the 4 symptom dimensions, using Pearson correlation analysis followed by a FDR approach for correcting multiple comparisons.[67]

### *Complementary Investigations*

*Control Analysis.* We controlled the effects of (1) 13 demographic characteristics; (2) the 13 demographic variables and the olanzapine (OZP) equivalent dosage; and (3) only the OZP equivalent dosage on the baseline EF assessments using regression approaches[68] when establishing the classification models. Furthermore, to mitigate the potential impact of antipsychotic drug dosage on prognostic classification, we included only patients receiving a clinically standard, commonly effective OZP equivalent dosage of 10-20 mg/day.

*Sensitivity Analysis.* We conducted a best- and worst-case sensitivity analysis to account for the effect of attrition, assuming that all participants who were lost to follow-up had positive or unfavorable treatment outcomes. This helped establish potentially extreme scenarios due to attrition with respect to our main analyses.[69,70]

*Subsampling and Subgroup Analyses.* To investigate whether an unequal number of variables across EF dimensions affected the model results and our identification of feature importance, we performed subsampling analysis with 100 repetitions to draw an equal number of individual indicators across dimensions from the original data when establishing classifiers (cf. Supplementary methods).

We stratified our diagnostic and prognostic models by sex, that is, repeating the stacking ensemble within the male or female patient subgroup independently, to assess the influence of sex on classification performance and contributing EF variables. We also performed diagnostic and prognostic classifications separately for each symptom subgroup.

## Results

### *EF Dimensions Are Consistently and Differentially Affected in Schizophrenia*

CFA confirmed adequate model fit of the included EF variables against the theoretical 3-factor construct (cf Table S5, Figure S1). All variables demonstrated acceptable standardized factor loadings (>0.30, per).[71-73] All latent factors showed good composite reliability (all>0.60) in the combined group of patients and controls.

ANCOVA revealed significant 2-way interaction between group (SZ vs. HC) and EF measurements ($P < .001$). Follow-up one-way ANCOVAs revealed that, except for 2 measurements which assess WM maintenance based on the *Corsi block* test, the accuracy in No-Go trials, the difference of RT between the congruent and the incongruent condition trials (ie, the interference effect) in the Stroop task, the mixing cost in the switching task and the 2 composite scores of *Inhibition/WM updating* and *Inhibition*, other measurements and composite scores

were all significantly different between SZ patients and healthy controls (all $P_{fdr} < .05$) (Table 2; Figure 2B). The differential EF performance covers all of 3 EF dimensions with the RT for Go trials (response inhibition) in the *Go/No-Go* task presenting the largest effect size ($\eta^2 = 0.160$) (Table 2; Figure 2C). Response inhibition showed a significant sex difference, with females demonstrating superior accuracy in No-Go trials ($0.86 \pm 0.10$ vs. $0.81 \pm 0.13$, $P_{fdr} = .039$) (Table S6 compares male and female patients).

### *Remitted Patients Show Improved Interference Inhibition at Follow-Up, without Difference in any EF Dimension at Baseline, Compared with Non-remitted Patients*

The differences in baseline EF dimensions between the remission (58 patients) and non-remission groups (28 patients; RSWG criteria) were analyzed. No significant between-group differences were observed on any EF dimension (Table S7). When comparing respective changes in these dimensions from baseline to follow-up (Table S8), significant differences were only observed in the remission group. Specifically, on the Stroop task, the remission group demonstrated shorter RT under neutral, congruent, and incongruent conditions compared with baseline (ie, better interference inhibition ability; all $P < .05$).

Regarding clinical outcomes defined by subtype transmission, there were no significant differences in baseline EF measurements among the 3 subgroups (ie, positive, negative, and ambiguous) of SZ patients (all $P_{fdr} > .05$). In patients with more prominent secondary negative symptoms (ie, baseline non-negative subtype with transition to a negative subtype), the Stroop task revealed significantly reduced RT at follow-up compared with baseline for incongruent stimuli ($P_{fdr} = .022$), congruent stimuli ($P_{fdr} = .007$), and neutral stimuli ($P_{fdr} = .037$). Additionally, the mixing cost in the switching task was significantly lower at follow-up ($P_{fdr} = .048$). However, 13 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 in baseline EF measurements between the secondary negative group (64 patients of non-negative subtype transition to negative subtype) and the stable negative group (all $P_{fdr} > .05$).

### *Psychopathology Dimensions Specifically Correlated with Different EF Measurements*

EF, and specifically their dimensional measurements, is associated with different aspects of symptomatology.[74] Using the 4-dimensional structure of the PANSS,[20] we observed significant reductions in severity across negative, positive, cognitive, and affective symptoms in SZ patients at follow-up ($P < .001$, Table S9). Among the 86 follow-up patients, the ability to inhibit conflict, as reflected by congruent stimuli ($r = 0.246$, $P = .022$) in the *Stroop* task, was positively correlated with baseline positive symptom scores (Figure S2). Patients' capacity to maintain and shift mental sets, quantified by switch cost and mixing cost in the *number-letter switching* task, was positively correlated with baseline cognitive ($r = 0.298$; $P = .005$) and positive symptom scores ($r = 0.265$; $P = .014$), respectively, within the follow-up subset but not in the overall patient sample. For comparison, correlation analyses were repeated using PANSS 3-original subscale scores; these yielded similar results, except for correlations with *Stroop* metrics (Figure S3).

### *EF Dimensions Show Consistent and Distinct Impairments in Schizophrenia and Offer Prognostic Insights Via Machine Learning Classification Models*

*Diagnostic Classification.* For the feature set relying on only EF assessments (ie, 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 ([AUC] = 0.87) (Figure 3A; Table S10). The feature set 2, which also included sociodemographic variables, was aimed at incorporating information from routine clinical interviews that indicates disease susceptibility.[75] With the addition of sociodemographic variables, improved model performance was observed (stacking model AUC = 0.91). Additionally, we regressed demographic variables (age, gender, and education) from the EF assessments and used the residuals as features for model training, removing the variance explained by demographic factors and maintaining model performance (stacking model AUC = 0.80).

*Prognostic Classification.* The stacking model achieved the highest classification accuracy based on the EF assessments-only feature set (AUC = 0.81) (Figure 4A), and the performance was identical to that observed when the EF plus sociodemographic features set was used.

By regressing out the effects of sociodemographic variables on EF assessments, we found that the model performance decreased to AUC = 0.65, indicating a poorer setup (Figure 4A). 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 $P > .05$) (Table S7). Additionally, controlling for medication effects using an OZP equivalent dosage (which did not differ significantly between remitted and non-remitted patients; $P = .15$) diminished the prognostic classification accuracy to an AUC of 0.72. Investigating only 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 ($n = 86$).

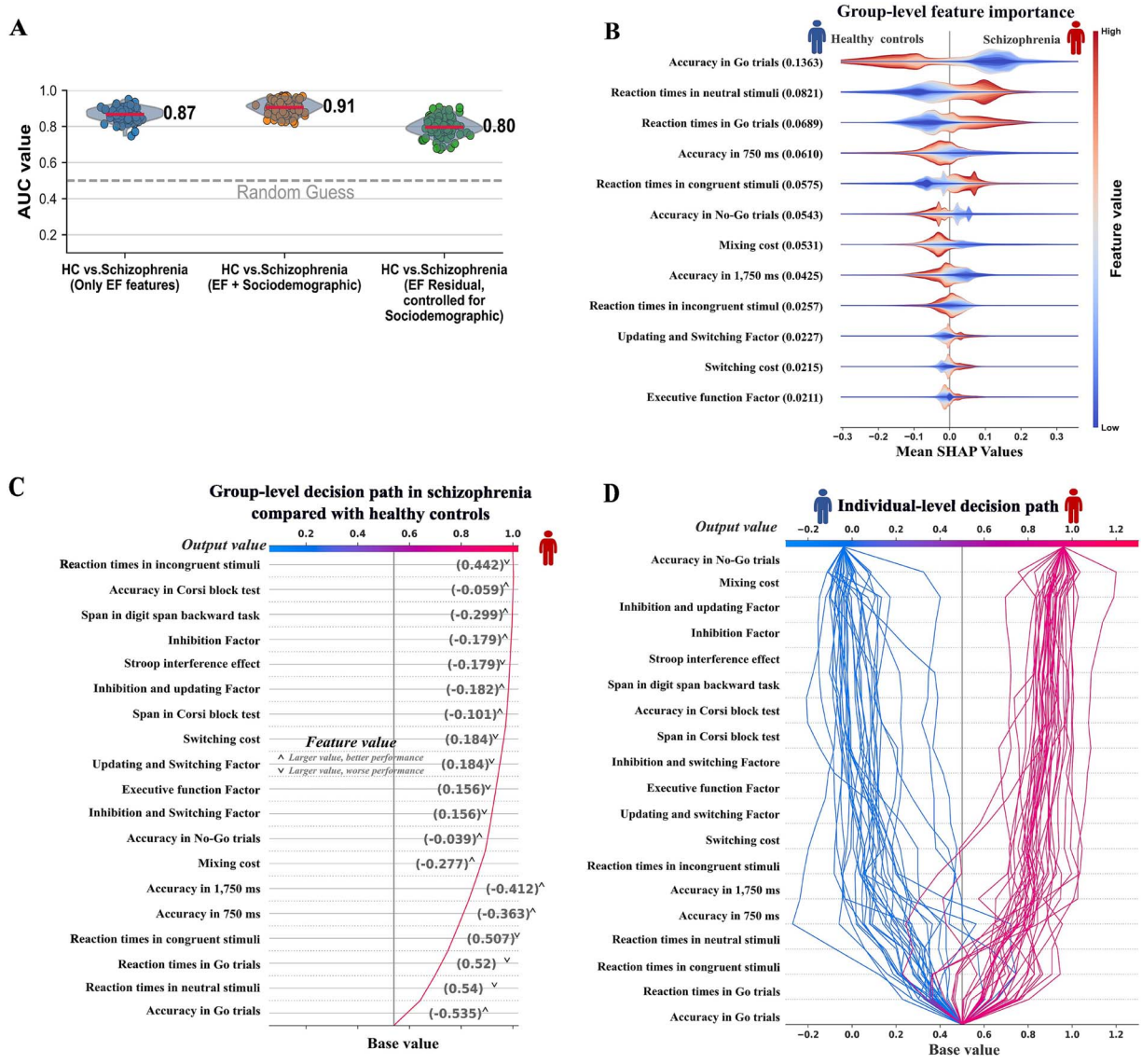


**Figure 3.** Performance and feature importance for diagnostic models. (A) Violin plots show the values of area under the curve (AUC) for discriminating patients with schizophrenia from healthy participants by diagnostic classification models. Each point within the violin plots represents the AUC value derived from the hold-out test data of each random split procedure (repeated 100 times) in our machine learning design. The horizontal line within each violin plot denotes the mean. (B) The group-level feature importance plot ranks EF features on the y-axis by their absolute average Shapley value across individuals, representing their overall importance in the ability of the model to distinguish between patients and healthy participants. The original feature weights for each EF variable were color coded. 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 samples. EF features are presented in an order with more closely related ones to have more similar decision paths. The color bar denotes the impact of an EF feature on model's classification towards "healthy" (0) or "schizophrenia" (1), the 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 denotes the measurement of an EF feature that is below the average in patients with schizophrenia). The ^ or ˇ markers 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 diagnostic classification to illustrate the model's decision-making process. The visualization begins from a uniform baseline at the bottom, corresponding to the model's average prediction, here 0.5 as we were performing binary classifications. Each curve from bottom to top denotes the Shapley values assigned to each EF feature per individual given the expression level this individual in these features associated with corresponding outcome. The terminal point of each trajectory at the top of the plot represents the model's final output score, which maps to a classified class of either 0 (healthy) or 1 (patient). Predictive path for an accurately classified individual was shown, demonstrating the individual-level reliability of the feature contributions that align with the observed diagnostics. Abbreviations: EF = executive function; HC = healthy control.

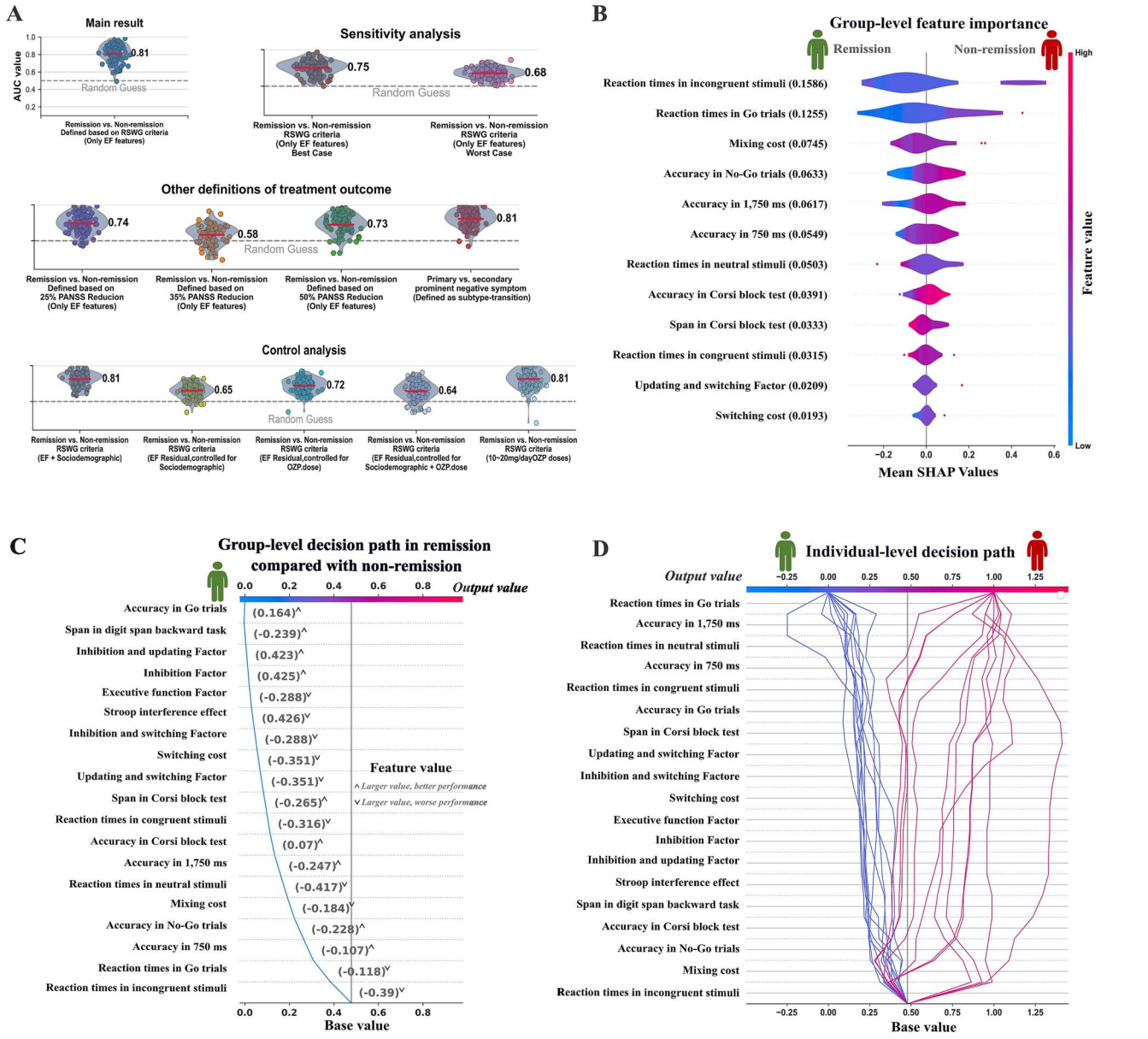


**Figure 4.** Performance and feature importance for prognostic models. (A) Violin plots show the values of area under the curve (AUC) for discriminating remitted patients with schizophrenia from non-remitted patients by prognostic classification models. Each point within the violin plots represents the AUC value derived from the hold-out test data of each random split procedure (repeated 100 times) in our machine learning design. The horizontal line within each violin plot denotes the mean. (B) The group-level feature importance plot ranks EF features on the y-axis by their absolute average Shapley value across individuals, representing their overall importance in the model's distinction between remitted and non-remitted patients (ie, greater separation of the violin-like plots towards the extremes denotes higher importance). The original feature weights for each EF variable are color-coded. 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 presented in an order with more closely related ones to have more similar decision paths. The color bar denotes the impact of an EF feature on model's classification towards "remission" (0) or "non-remission" (1), and the 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 denotes the measurement of an EF feature that is below the average in remitted patients). The ^ or ˅ markers 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 prognostic classification to illustrate the model's decision-making process. The visualization begins from a uniform baseline at the bottom, corresponding to the model's average prediction, here 0.5 as we were performing binary classifications. Each curve from bottom to top denotes the Shapley values assigned to each EF feature per individual given the expression

We supplemented classifications to discriminate prognostic subgroups defined by the reduction in re-assessed total PANSS score. The results showed poorer discriminative power, with AUCs of 0.74, 0.58, and 0.73 (Figure 4A; Table S11) for a 25% (68 remitted vs. 18 non-remitted), 35% (48 remitted vs. 38 non-remitted), and 50% (16 remitted vs. 70 non-remitted) reduction in the PANSS score, respectively. Alternatively, we established a classification model for the clinical outcomes of patients based on subtype-membership transition which yielded a promising performance (AUC = 0.81).

Moreover, 2 sensitive analyses were performed to take the attrition condition in our study into consideration. The demographic variables, symptoms, OZP-equivalent dosage, and EF measurements did not differ significantly between followed-up patients and those who were not (all $P > .05$, Table S7). Furthermore, by treating the patients lost to attrition as best cases (ie, all remitted), classification accuracy decreased slightly to an AUC of 0.75; treating them as worst cases (none remitted) yielded a further decrease (AUC = 0.68).

### *Feature Importance and Association with Individual Psychopathology*

*Group-Level Feature Importance and Decision Path.* Including only the 19 EF dimensions scores as the feature set in both diagnostic and prognostic classifications, the IC dimension—comprised of response and interference inhibitions—ranked highest in both classifying SZ group participants and their follow-up remission status by the Shapley approach (Figures 3B and 4B). Within the IC dimension, important features were from the *Go/No-Go task* and the Stroop task. Adding sociodemographic variables to the diagnostic and prognostic classifiers generally replicated IC as the strongest contributing dimension (Figures S4A and S5A).

Next, group-level decision path analysis of each EF dimension identified that poor performance on any of these drove the model to correctly classify SZ patients (Figure 3C). However, worse performance on the IC dimension increased the likelihood of non-remission status classification (Figure 4C). These results were replicated in additional models in which sociodemographic variables were included (Figures S4B and S5B).

*Individual-Level Decision Path and Association with Psychopathology.* Among correctly identified participants in diagnostic classification, SZ patients scored below the averages of both HC and SZ groups on at least one EF dimension (Figure 3D; Figure S6). For the prognostic classification model, remitted status for most participants in the SZ group was correctly assigned based on higher baseline IC dimension (including both interference control and response inhibition) compared with the averages of both remitted and non-remitted participants. Specifically, remitted participants showed shorter RT 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, several patients in remission presented worse baseline performance in WM than the group averages (Figure 4D; Figure S7). These findings were replicated when sociodemographic variables were included in diagnostic classification models (Figures S4 and S5).

Pearson correlation analysis was further performed on individual Shapley values for each EF feature and scores on the 4 symptom dimensions. The interference inhibition function, as assessed by a difference in RT 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 ($r = 0.439$, $P_{fdr} = .042$) for SZ (Figure 5A). After including the 13 sociodemographic variables, no significant correlation was found. For prognostic classification, the importance of WM updating (assessed in the *running memory* task; $r = 0.618$, $P_{fdr} = .023$) and maintenance (assessed in the *Corsi block* test; $r = 0.597$, $P_{fdr} = .031$) in the model that accurately assigned remitted patients was correlated with low-level cognitive symptoms at follow-up. Additionally, WM updating function contributing to the accurate assignment of remitted patients was associated with less severe re-assessed negative symptoms ($r = 0.596$, $P_{fdr} = .031$) (Figure 5B). Following inclusion of sociodemographic variables in the model, the significantly associated EF variables were different (Figure S8).

Using PANSS 3-subscale scores in correlation analyses did not reveal significant correlation with the importance scores of any EF features 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 above (Figure S9).

Equalizing variables across EF dimensions when classifying patients and their remission outcomes exhibited decreased performance with the number of features reducing from 19 to 11 (diagnostic model: AUC = 0.81, prognostic model: AUC = 0.70) (Figure S10). Therefore, all variables contributed to the classification, supporting

level this individual in these features associated with corresponding outcome. The terminal point of each trajectory at the top of the plot represents the model's final output score, which maps to a classified class of either 0 (non-remission) or 1 (remission). Predictive path for an accurately classified individual was shown, demonstrating the individual-level reliability of the feature contributions that align with the observed outcomes. Abbreviations: EF = executive function; OZP = olanzapine; PANSS = Positive and Negative Syndrome Scale; RSWG = Remission in Schizophrenia Working Group; SHAP = SHapley Additive exPlanations.

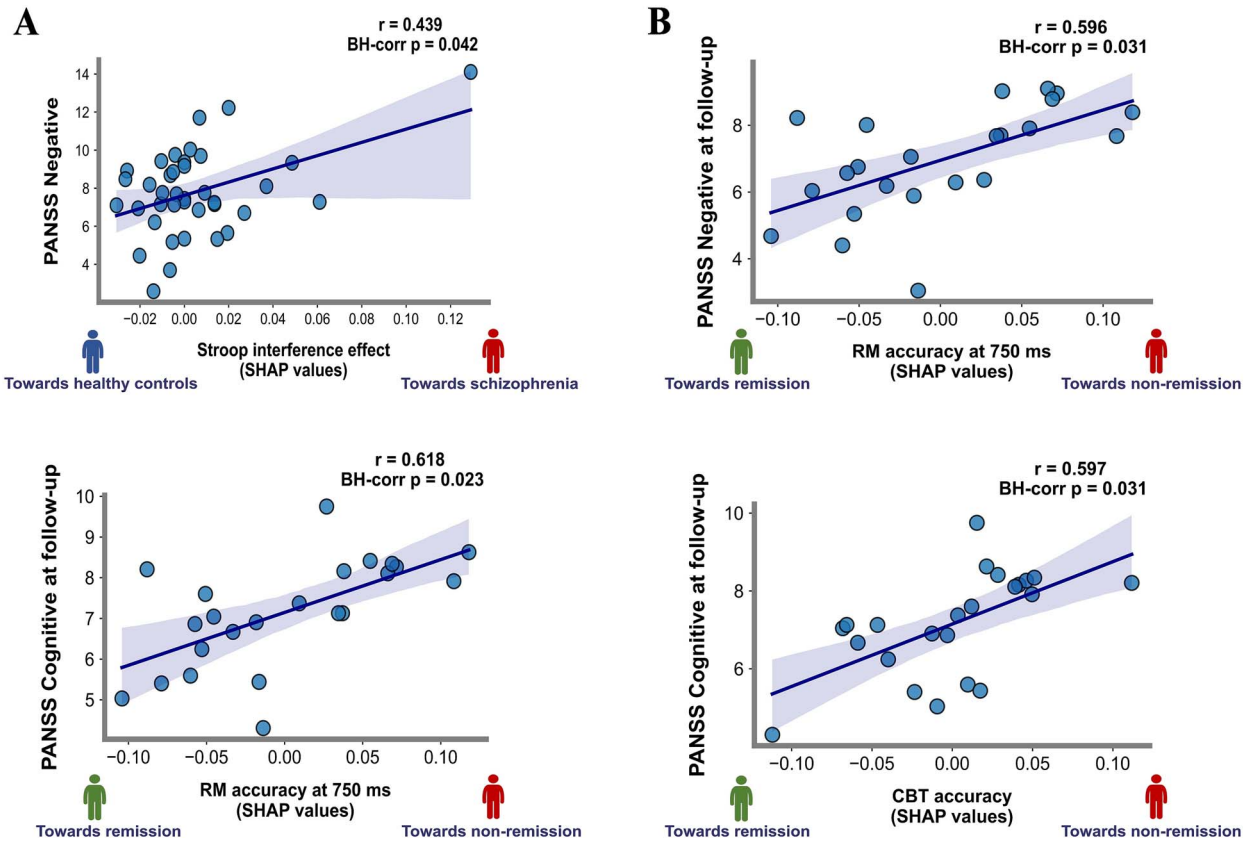


**Figure 5.** Correlation between the importance of an executive function feature and individual psychopathology along four symptom dimensions. (A) Correlation of model-derived EF feature importance with individual symptom scores in diagnostic classification. (B) Correlation of model-derived EF feature importance with individual symptom scores in prognostic classification. 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.

their inclusion. Feature ranking using SHAP replicated the main results.

Classification of patients versus controls in each sex subgroup demonstrated similar performance as in our main results (AUC = 0.86 for males, AUC = 0.84 for females) (Figure S11). However, prognostic classification of remission outcomes within female patients revealed an apparent decrease in AUC (0.73). Nevertheless, feature importance identified by SHAP replicated the IC dimension as the most important predictor contributed to correctly assign a diagnostic label and a remission status (Figure S11). Establishing models separately for patients dominant in positive or negative symptom profiles revealed comparable discriminative power in diagnostic classification (AUCs = 0.88 and 0.85 for the positive and negative subgroups respectively) to the main results, and the important features pointed to the inhibitory dimension as well. In contrast, in prognostic classification, modeling within either the positive (AUC = 0.77) or negative (AUC = 0.74) subgroup resulted in lower AUC, with WM maintenance identified as the top important EF assessment for the negative subgroup (Figure S12).

## Discussion

This study investigated the classification power of 3 EF dimensions for discriminating both SZ patients 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. Previous studies classifying patients using cognitive data have generally achieved accuracies below 78% (Table S12); those reaching up to 90% accuracy[76] typically include symptom data.[77] Our stacking model achieved reasonably high diagnostic accuracy, identifying specific EF dimensions and assessments informative for identifying patients. Furthermore, inclusion of sociodemographic characteristics increased the AUC to 0.91, consistent with previous findings.[75] At the group level, the IC dimension ranked highest in importance for classifying patients. Previous works showing abnormal temporal and spatial characteristics of inhibition-related brain responses and behaviors in SZ.[78,79] Moreover, our decision path plot indicated that those performing poorly on any EF dimension tended to be classified in the SZ group. This corroborates the general EF impairments among

these patients.[21] However, we noted from individual-level decision paths that the performance on either of 3 EF dimensions for a few accurately identified patients was relatively preserved. This aligns well with reports that cognitive performance, including EFs, in SZ can vary from mild deficiency[80,81] to severe impairment,[66,82] and connects the neuropsychological and neurobiological heterogeneity observed among these patients.[83] Our study addressed this heterogeneity from a dimensional perspective in consistent with the "unity and diversity" structure of EF,[8] which differs from approaches that use clustering to define patient subgroups based on EF profiles.[84] That is, instead of creating categories, we captured inter-individual variability by treating each patient as a unique profile of continuous EF variables. Our model was then trained on these individual-level variables—without imposing any conceptual groupings—to classify patients versus controls. Furthermore, in the aforementioned clustering study, the author implemented 3 types of models to instantiate the "three-component theory of flexible cognition," plus an inner-speech component (deep neural network), to fit the behavioral data obtained from healthy controls and patients with SZ while performing WCST. They found that specific model parameters successfully differentiated the 3 patient subgroups with varying degree of impairments in CF.[84] There are also other attempts[85,86] to probe the possible mechanisms that would be related to the executive functioning impairment in SZ using theory-driven computational models (cf. Table S12 for details). The group-level important EF dimension—interference inhibition—was significantly associated with negative symptoms. Consistent with this, individuals with more severe negative symptoms tend to have diminished IC.[1,2] This association, particularly for interference inhibition, may be understood as target-speech recognition deficits in SZ.[64] Specifically, impaired interference inhibition is connected to disorganized speech information processes in SZ, including the inability to inhibit unrelated speech signals or capture desired speech signals.[87] The putative excitatory/inhibitory imbalance in SZ[88] could lead to the linguistic disorganization and impoverishment characteristic of formal thought disorder (FTD). Impairments in EF, particularly IC, have been consistently associated with FTD in SZ.[88] At the neurobiological level, multiple brain regions are implicated,[89,90] involving deficits in distributed networks for sematic, executive, attentional and social functions.[91] Especially a compromised frontoparietal-striatal circuit underlies EF impairments with IC being tied to frontostriatal dysfunction and WM to prefrontal-parietal disturbances. The weakened frontoparietal connectivity and reduced frontostriatal influence in SZ may give rise to the excitatory/inhibitory imbalance linked with abnormal striatal dopamine levels.[92] These collectively suggest that the executive dysfunction in SZ would relate to an impaired synergy between cognitive and reward processes, hindering goal-oriented thought and behavior (cf. Supplementary discussion).

Prognostic prediction is even more challenging; one study using only cognitive data achieved a chance-level accuracy of 50.3%. With clinical data included, accuracies usually hover around 70%. Our model—trained using only EF assessments—achieved a balanced accuracy of 74% (AUC: 0.81). We tested the classifications by alternative definitions for remission, but the discriminative power of the classification model was higher when specific items, relative to all items, of the PANSS were involved in defining prognostic statuses. This implies specificities in mapping EF dimensions and symptom recovery in patients with SZ. Interestingly, IC was again appeared as the top predictor of remission outcome. Furthermore, patients correctly classified as remitted generally showed good baseline performance on IC tasks. Previous work consistently demonstrates an association between EF and long-term post-treatment outcomes.[93] Specifically, patients with higher EF performance are more likely to remit,[94,95] especially on the IC dimension,[96] likely because patients with better IC are more therapy-adherent.[97] Alternatively, because IC assessments are closely related to specific clinical manifestations in SZ,[13,98,99] better performance on this EF dimension along with milder symptoms among such patients implies higher remission likelihood.[93] Our results showed significant post-treatment improvement in interference inhibition in the remission subgroup, but not in the non-remission subgroup (Supplementary material), corroborating a relationship between inhibitory performance and remission. Interestingly, we found no significant difference in any baseline EF assessment and symptom dimension score between subgroups. While suggestive of dissociation between SZ symptoms and the EF construct, this highlights the role of EF dimensions in predicting remission rather than merely being a marker of illness severity. The absence of significant differences in any baseline EF variables between patients in remission and those who are not at follow up, despite the presence of successful classification of patients' remission outcomes with multivariate machine learning, implied that prediction of remission in SZ is contingent upon patterns of executive dysfunction rather than deficits in any single measure. In fact, there are non-linear interactions across EF variables and dimensions, with an inclusion of interaction terms to out-perform a regression model formulating only main effects, as illustrated (cf. Figure S13). Our SHAP method accounts for non-linear relationships and interactions among features[100] when decomposing each feature's specific contributions to a prediction model on the individual level.[101] Therefore, an EF variable can be identified as important by SHAP—because of its role in complex, individual-level patterns—while failing to show a significant average effect across the entire group in univariate analyses. Also, the calculation

of Shapley values is not affected by the imbalanced number of variables we included for measuring different EF dimensions. This property is particularly relevant as variables within the same dimension are often highly inter-correlated. These findings are consistent with previous empirical works.[65]

Furthermore, the importance of WM contribution to the prognostic model in classifying remission status covaried with poor patient cognitive symptoms at follow-up. WM performance is superior among stably remitted SZ patients versus non-remitted patients,[94] and worse performance in WM is linked to cognitive disorganization (the major symptom within the cognitive dimension we used in this study).[18,102,103] Sex differences exist in SZ.[104,105] Our results regarding between-sex comparisons on sociodemographic and EF variables resonate with findings that female patients typically exhibit later onset, higher social functioning, and milder symptoms.[106,107] Possibly relate to the influences of psychosocial factors and hormonal variations[54,108] on a greater heterogeneity in the illness trajectories for female patients, we observed lower accuracy when classifying the remission outcomes within this population group. However, IC was consistently identified as the important contribution dimension irrespective of sex subgroups.

Multiple prior works employing several machine learning algorithms to classify SZ have generated variable accuracies.[109,110] They have not synthesized the predictive power of single classifiers. Here, we implemented a stacking model to integrate the strengths of various algorithms operating different assumptions[111] to reduce generalization error as demonstrated in both healthy populations[112] and psychiatric disorders.[113] Our stacking model built upon 3 algorithms yielded slight performance improvement over the best-performing single model, RF. This suggests RF's assumptions were a good fit for our data. Consequently, the stacking model should have leveraged the characteristics of RF most heavily. The outcome highlights a practical tradeoff between model complexity (related to computational costs, regularizations[114] and number of hyperparameters [Table S13]) and performance gain. Thus, the choice to adopt such a model would benefit from a cost-effective analysis that considers computational resources, presumed complexity of the data, and clinical utility of the resulting performance enhancement. Nevertheless, the flexibility of stacking models to integrate heterogeneous data sources makes them a powerful and versatile tool for future medical applications.[115]

Several limitations exist. Our use of multiple random-splits to set aside a test "lock box" in each repeat, while performing nested CVs on the remaining sample,[63] effectively gauges the out-of-sample generalization performance, while balancing practical clinical data collection issues.[53,60,116] Nevertheless, future studies might consider incorporating additional sites, while performing careful data harmonization to address heterogeneity and ensure the validity of any predictions.[61] Owing to potential ethical restrictions and data privacy constraints, methods such as federated learning[117] can be leveraged without sharing individual raw data upon a uniform processing pipeline. Alternatively, there is a recent note about "recurring local validation" that protects against performance-disruptive data variability across populations, geographies, and facilities.[118] Second, present patients had been treated with antipsychotics, reflecting typical clinical practice, and the exact dosage of antipsychotic medication could have influenced individual prognostic status[119]. 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 (eg, 10-20 mg/day)—commonly effective in SZ[120]—retained the classification performance in our main experiments. Nevertheless, future research involving drug-naïve patients at baseline and continuous assessments of EF along with detailed records of medication usage over time would help establish causal relationships between antipsychotic effects, prognostic statuses, and specific EF dimensions. Third, classifications were poorer when patient attrition was accounted by assuming extreme outcomes (either all remitted or all non-remitted). Such analyses can underestimate results as reported,[118] though the true bias may differ. Although urged, managing patient attrition remains challenging. In addition, we focused on patients diagnosed with SZ (F20.9) according to DSM-IV criteria is to avoid introducing additional heterogeneity between categories. However, including the wider spectrum of patient populations with psychosis (based on the DSM-5) would allow capture of additional important transdiagnostic, dimensional information. Lastly, this study did not employ verbal tasks to assess the language deficits in SZ, for example, incoherent speech, which are usually subsumed under the broad umbrella of FTD.[91] While FTD is linked with executive dysfunction in SZ,[88] its pathophysiology is complex, including multiple cognitive functions beyond EF[89]; this leads to difficulties in dissecting the extent to which FTD severity is due to executive dysfunction. Formal language assessments may be integrated to explore how FTD interacts with executive dysfunction in identifying SZ and related prognostic outcomes.

To conclude, results from our models revealed promising performance and, importantly, different EF dimensions characterized diagnosis and prognosis to varying extents. IC and WM were the most important factors for accurate classification of SZ and remission outcomes. Together with the classification strength of these EF dimension features to associate with specific psychopathologies, our research thus provides measures that may serve as valuable aids in the clinical assessment of this disorder.

## Supplementary Material

Supplementary material is available at https://academic.oup.com/schizophreniabulletin.

## Funding

This work was supported by the National Natural Science Foundation of China (No. 32260207 [to X.Z.], No. 82201658 [to J.C.] & No. 82371506 [to J.C.]), the Shanghai Rising-Star Program (24QA2704700 [to J.C.]) and the STI2030-Major Projects (No. 2022ZD0214000 [to J.C.]) . B.T.T.Y. 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 OF-IRG (OFIRG24jan-0006; OFIRG24jul-0049), NMRC STaR (STaR20nov-0003), Singapore Ministry of Health (MOH) Centre Grant (CG21APR1009), The United States National Institutes of Health (R01MH133334 & 2R01MH120080), and The Singapore National Research Foundation (NRF) Investigatorship (NRFI10-2024-0014). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the funders.

## Data Availability

Information for the main sample used in the present study has been included in the Supplementary Materials. The raw data of the sample used in this study 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. Scripts to run the main analyses have been made publicly available and can be accessed at https://github.com/SCZ-EF-stacked-classifiers.

## References

1. Heilbronner U, Samara M, Leucht S, Falkai P, Schulze TG. The longitudinal course of schizophrenia across the lifespan: clinical, cognitive, and neurobiological aspects. *Harv Rev Psychiatry*. 2016;24:118-128. https://doi.org/10.1097/HRP.0000000000000092
2. McCutcheon RA, Keefe RSE, McGuire PK. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. *Mol Psychiatry*. 2023;28:1902-1918. https://doi.org/10.1038/s41380-023-01949-9
3. Diamond A. Executive functions. *Annu Rev Psychol*. 2013;64:135-168. https://doi.org/10.1146/annurev-psych-113011-143750
4. Kerns JG, Nuechterlein KH, Braver TS, Barch DM. Executive functioning component mechanisms and schizophrenia. *Biol Psychiatry*. 2008;64:26-33. https://doi.org/10.1016/j.biopsych.2008.04.027
5. Ambrosen KS, Skjerbaek MW, Foldager J, 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. *Transl Psychiatry*. 2020;10:276. https://doi.org/10.1038/s41398-020-00962-8
6. Soldatos RF, Cearns M, Nielsen MØ, et al. Prediction of early symptom remission in two independent samples of first-episode psychosis patients using machine learning. *Schizophr Bull*. 2022;48:122-133. https://doi.org/10.1093/schbul/sbab107
7. Green MF, Kern RS, Heaton RK. Longitudinal studies of cognition and functional outcome in schizophrenia: implications for MATRICS. *Schizophr Res*. 2004;72:41-51. https://doi.org/10.1016/j.schres.2004.09.009
8. Friedman NP, Miyake A. Unity and diversity of executive functions: individual differences as a window on cognitive structure. *Cortex*. 2017;86:186-204. https://doi.org/10.1016/j.cortex.2016.04.023
9. Friedman NP, Robbins TW. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacol*. 2022;47:72-89. https://doi.org/10.1038/s41386-021-01132-0
10. Gärtner A, Strobel A. Individual differences in inhibitory control: a latent variable analysis. *J Cogn*. 2021;4:17. https://doi.org/10.5334/joc.150
11. Frischkorn GT, von Bastian CC, Souza AS, Oberauer K. Individual differences in updating are not related to reasoning ability and working memory capacity. *J Exp Psychol Gen*. 2022;151:1341-1357. https://doi.org/10.1037/xge0001141
12. Friedman NP, Miyake A. The relations among inhibition and interference control functions: a latent-variable analysis. *J Exp Psychol Gen*. 2004;133:101-135. https://doi.org/10.1037/0096-3445.133.1.101
13. Ettinger U, Aichert DS, Wöstmann N, Dehning S, Riedel M, Kumari V. Response inhibition and interference control: effects of schizophrenia, genetic risk, and schizotypy. *J Neuropsychol*. 2018;12:484-510. https://doi.org/10.1111/jnp.12126
14. Kerns JG, Berenbaum H. Cognitive impairments associated with formal thought disorder in people with schizophrenia. *J Abnorm Psychol*. 2002;111:211-224. https://doi.org/10.1037/0021-843X.111.2.211
15. Kerns JG, Berenbaum H. The relationship between formal thought disorder and executive functioning component processes. *J Abnorm Psychol*. 2003;112:339-352. https://doi.org/10.1037/0021-843X.112.3.339
16. Yang H, Wang M, Wu F, Li Q, Zheng Y, Qin P. Diminished self-monitoring in hallucinations–aberrant anterior insula connectivity differentiates auditory hallucinations in schizophrenia from subjective tinnitus. *Asian J Psychiatr*. 2020;52:102056. https://doi.org/10.1016/j.ajp.2020.102056
17. Raveendran V, Kumari V. Clinical, cognitive and neural correlates of self-monitoring deficits in schizophrenia: an update. *Acta Neuropsychiatr*. 2007;19:27-37. https://doi.org/10.1111/j.1601-5215.2007.00151.x
18. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. *Biol Psychiatry*. 2000;47:34-42. https://doi.org/10.1016/S0006-3223(99)00187-0

19. Raffard S, Gutierrez LA, Yazbek H, et al. Working memory deficit as a risk factor for severe apathy in schizophrenia: a 1-year longitudinal study. *Schizophr Bull*. 2016;42:642-651. https://doi.org/10.1093/schbul/sbw002
20. Hager OM, Kirschner M, Bischof M, et al. Reward-dependent modulation of working memory is associated with negative symptoms in schizophrenia. *Schizophr Res*. 2015;168:238-244. https://doi.org/10.1016/j.schres.2015.08.024
21. Thai ML, Andreassen AK, 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 Res*. 2019;272:230-236. https://doi.org/10.1016/j.psychres.2018.12.088
22. De Luca CR, Wood SJ, Anderson V, et al. Normative data from the CANTAB. I: development of executive function over the lifespan. *J Clin Exp Neuropsychol*. 2003;25:242-254. https://doi.org/10.1076/jcen.25.2.242.13639
23. Weintraub S, Dikmen SS, Heaton RK, et al. Cognition assessment using the NIH toolbox. *Neurology.* 2013;80:S54-S64.
24. Chidharom M, Krieg J, Bonnefond A. Impaired frontal midline theta during periods of high reaction time variability in schizophrenia. *Biol Psychiatry: Cogn Neurosci Neuroimaging*. 2021;6:429-438. https://doi.org/10.1016/j.bpsc.2020.10.005
25. Waltz JA, Frank MJ, Wiecki TV, Gold JM. Altered probabilistic learning and response biases in schizophrenia: behavioral evidence and neurocomputational modeling. *Neuropsychology.* 2011;25:86-97. https://doi.org/10.1037/a0020882
26. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. *Front Psychol*. 2015;6:6. https://doi.org/10.3389/fpsyg.2015.00328
27. Chen J, Patil KR, Weis S, 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. *Biol Psychiatry*. 2020;87:282-293. https://doi.org/10.1016/j.biopsych.2019.08.031
28. Andreasen NC, Carpenter WT, Kane JM, Lasser RA, Marder SR, Weinberger DR. Remission in schizophrenia: proposed criteria and rationale for consensus. *Am J Psychiatry*. 2005;162:441-449. https://doi.org/10.1176/appi.ajp.162.3.441
29. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science.* 2024;383:164-167. https://doi.org/10.1126/science.adg8538
30. Kirschner M, Aleman A, Kaiser S. Secondary negative symptoms—a review of mechanisms, assessment and treatment. *Schizophr Res*. 2017;186:29-38. https://doi.org/10.1016/j.schres.2016.05.003
31. Veerman SRT, Schulte PFJ, de Haan L. Treatment for negative symptoms in schizophrenia: a comprehensive review. *Drugs.* 2017;77:1423-1459. https://doi.org/10.1007/s40265-017-0789-y
32. Zhao X, Zhang W, Tong D, Maes JHR. Creative thinking and executive functions: associations and training effects in adolescents. *Psychol Aesthet Creat Arts*. 2023;17:79-90. https://doi.org/10.1037/aca0000392
33. Zhao X, Wang Y, Maes JHR. The effect of working memory capacity and training on intertemporal decision making in children from low-socioeconomic-status families. *J Exp Child Psychol*. 2022;216:105347. https://doi.org/10.1016/j.jecp.2021.105347
34. Kessels RPC, van Zandvoort MJE, Postma A, Kappelle LJ, de Haan EHF. The corsi block-tapping task: standardization and normative data. *Appl Neuropsychol*. 2000;7:252-258. https://doi.org/10.1207/S15324826AN0704_8
35. Stroop JR. Studies of interference in serial verbal reactions. *J Exp Psychol*. 1935;18:643-662. https://doi.org/10.1037/h0054651
36. Gomez P, Ratcliff R, Perea M. A model of the go/no-go task. *J Exp Psychol Gen*. 2007;136:389-413. https://doi.org/10.1037/0096-3445.136.3.389
37. Kray J, Li KZ, Lindenberger U. Age-related changes in task-switching components: the role of task uncertainty. *Brain Cogn*. 2002;49:363-381. https://doi.org/10.1006/brcg.2001.1505
38. Yangüez M, Bediou B, Chanal J, Bavelier D. In search of better practice in executive functions assessment: methodological issues and potential solutions. *Psychol Rev*. 2024;131:402-430. https://doi.org/10.1037/rev0000434
39. Berberian AA, Gadelha A, Dias NM, et al. Component mechanisms of executive function in schizophrenia and their contribution to functional outcomes. *Braz J Psychiatry*. 2019;41:22-30. https://doi.org/10.1590/1516-4446-2018-0021
40. Feng X, Perceval GJ, Feng W, Feng C. High cognitive flexibility learners perform better in probabilistic rule learning. *Front Psychol*. 2020;11:11. https://doi.org/10.3389/fpsyg.2020.00415
41. Miles S, Howlett CA, Berryman C, Nedeljkovic M, Moseley GL, Phillipou A. Considerations for using the Wisconsin card sorting test to assess cognitive flexibility. *Behav Res Methods*. 2021;53:2083-2091. https://doi.org/10.3758/s13428-021-01551-3
42. Willoughby MT, Blair CB, The Family Life Project Investigators. Measuring executive function in early childhood: a case for formative measurement. *Psychol Assess*. 2016;28:319-330. https://doi.org/10.1037/pas0000152
43. Goghari VM, MacDonald AW. The neural basis of cognitive control: response selection and inhibition. *Brain Cogn*. 2009;71:72-83. https://doi.org/10.1016/j.bandc.2009.04.004
44. Eisenberg DP, Berman KF. Executive function, neural circuitry, and genetic mechanisms in schizophrenia. *Neuropsychopharmacol.* 2010;35:258-277. https://doi.org/10.1038/npp.2009.111
45. Barch DM, Ceaser A. Cognition in schizophrenia: core psychological and neural mechanisms. *Trends Cogn Sci*. 2012;16:27-34. https://doi.org/10.1016/j.tics.2011.11.015
46. Cepeda NJ, Blackwell KA, Munakata Y. Speed isn't everything: complex processing speed measures mask individual differences and developmental changes in executive control. *Dev Sci*. 2013;16:269-286. https://doi.org/10.1111/desc.12024
47. Gardiner E, Hutchison SM, Müller U, Kerns KA, Iarocci G. Assessment of executive function in young children with and without ASD using parent ratings and computerized tasks of executive function. *Clin Neuropsychol*. 2017;31:1283-1305. https://doi.org/10.1080/13854046.2017.1290139
48. Ronold EH, Schmid MT, Oedegaard KJ, 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*. 2020;11:11. https://doi.org/10.3389/fpsyt.2020.575867
49. Verma S, Goel T, Tanveer M, Ding W, Sharma R, Murugan R. Machine learning techniques for the schizophrenia diagnosis: a comprehensive review and future research directions. *J Ambient Intell Human Comput*. 2023;14:4795-4807. https://doi.org/10.1007/s12652-023-04536-6

50. Pernía-Espinoza A, Fernandez-Ceniceros J, Antonanzas J, Urraca R, Martinez-de-Pison FJ. Stacking ensemble with parsimonious base models to improve generalization capability in the characterization of steel bolted components. *Appl Soft Comput*. 2018;70:737-750. https://doi.org/10.1016/j.asoc.2018.06.005
51. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.
52. Lee LH, Chen CH, Chang WC, 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. *Eur Psychiatry*. 2021;65:e1. https://doi.org/10.1192/j.eurpsy.2021.2248
53. Schultebraucks K, Ben-Zion Z, Admon R, et al. Assessment of early neurocognitive functioning increases the accuracy of predicting chronic PTSD risk. *Mol Psychiatry*. 2022;27:2247-2254. https://doi.org/10.1038/s41380-022-01445-6
54. Perochon S, Di Martino JM, Carpenter KLH, et al. Early detection of autism using digital behavioral phenotyping. *Nat Med*. 2023;29:2489-2497. https://doi.org/10.1038/s41591-023-02574-3
55. Lim K, Smucny J, Barch DM, Lam M, Keefe RSE, Lee J. Cognitive subtyping in schizophrenia: a latent profile analysis. *Schizophr Bull*. 2021;47:712-721. https://doi.org/10.1093/schbul/sbaa157
56. Martini G, Bracci A, Riches L, et al. Machine learning can guide food security efforts when primary data are not available. *Nat Food*. 2022;3:716-728. https://doi.org/10.1038/s43016-022-00587-8
57. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 2017;145:166-179. https://doi.org/10.1016/j.neuroimage.2016.10.038
58. Wu J, Li J, Eickhoff SB, Scheinost D, Genon S. The challenges and prospects of brain-based prediction of behaviour. *Nat Hum Behav*. 2023;7:1255-1264. https://doi.org/10.1038/s41562-023-01670-1
59. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM; 2019:2623-2631. https://doi.org/10.1145/3292500.3330701
60. Sasse L, Nicolaisen-Sobesky E, Dukart J, Eickhoff SB, Patil KR. Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*. 2025;12:5. https://doi.org/10.1186/s40537-024-01059-7
61. Chen J, Patil KR, Yeo BTT, Eickhoff SB. Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. *Biol Psychiatry*. 2023;93:18-28. https://doi.org/10.1016/j.biopsych.2022.07.025
62. Koutsouleris N, Kahn RS, Chekroud AM, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry*. 2016;3:935-946. https://doi.org/10.1016/S2215-0366(16)30171-7
63. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning–a brief history, state-of-the-art and challenges. In: Koprinska I, Kamp M, Appice A, et al., eds. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 417-431. Springer, 2020.
64. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56-67. https://doi.org/10.1038/s42256-019-0138-9
65. Habets PC, Thomas RM, Milaneschi Y, 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. *Biol Psychiatry*. 2023;94:948-958. https://doi.org/10.1016/j.biopsych.2023.05.024
66. Snitz BE, MacDonald AW III, Carter CS. Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. *Schizophr Bull*. 2006;32:179-194. https://doi.org/10.1093/schbul/sbi048
67. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57:289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
68. Dobson AJ, Barnett AG. *An Introduction to Generalized Linear Models*. CRC Press, 2018.
69. Akl EA, Briel M, You JJ, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ*. 2012;344:e2809. https://doi.org/10.1136/bmj.e2809
70. Jakobsen JC, 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*. 2017;17:162. https://doi.org/10.1186/s12874-017-0442-1
71. Hassim SR, Arifin WN, Kueh YC, Yaacob NA. Confirmatory factor analysis of the Malay version of the smartphone addiction scale among medical students in Malaysia. *Int J Environ Res Public Health*. 2020;17:3820. https://doi.org/10.3390/ijerph17113820
72. DiStefano C. The impact of categorization with confirmatory factor analysis. *Struct Equ Model Multidiscip J*. 2002;9:327-346. https://doi.org/10.1207/S15328007SEM0903_2
73. Cossette S, Pepin J, Côté JK, De Courval FP. The multidimensionality of caring: a confirmatory factor analysis of the caring nurse–patient interaction short scale. *J Adv Nurs*. 2008;61:699-710. https://doi.org/10.1111/j.1365-2648.2007.04566.x
74. Leucht S, Davis JM, Engel RR, Kissling W, Kane JM. Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. *Acta Psychiatr Scand*. 2009;119:7-14. https://doi.org/10.1111/j.1600-0447.2008.01308.x
75. van Os J, Kenis G, Rutten BPF. The environment and schizophrenia. *Nature*. 2010;468:203-212. https://doi.org/10.1038/nature09563
76. Shen C, Popescu FC, Hahn E, Ta TTM, Dettling M, Neuhaus AH. Neurocognitive pattern analysis reveals classificatory hierarchy of attention deficits in schizophrenia. *Schizophr Bull*. 2014;40:878-885. https://doi.org/10.1093/schbul/sbt107
77. Walsh-Messinger J, Jiang H, Lee H, Rothman K, Ahn H, Malaspina D. Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning. *Psychiatry Res*. 2019;278:27-34. https://doi.org/10.1016/j.psychres.2019.03.048
78. Hughes ME, Fulham WR, Johnston PJ, Michie PT. Stop-signal response inhibition in schizophrenia: behavioural, event-related potential and functional neuroimaging data. *Biol Psychol*. 2012;89:220-231. https://doi.org/10.1016/j.biopsycho.2011.10.013

79. Kaladjian A, Jeanningros R, Azorin JM, Grimault S, Anton JL, Mazzola-Pomietto P. Blunted activation in right ventrolateral prefrontal cortex during motor response inhibition in schizophrenia. *Schizophr Res*. 2007;97:184-193. https://doi.org/10.1016/j.schres.2007.07.033
80. Wexler BE, Zhu H, Bell MD, et al. Neuropsychological near normality and brain structure abnormality in schizophrenia. *Am J Psychiatry*. 2009;166:189-195. https://doi.org/10.1176/appi.ajp.2008.08020258
81. Raffard S, Bayard S. Understanding the executive functioning heterogeneity in schizophrenia. *Brain Cogn*. 2012;79:60-69. https://doi.org/10.1016/j.bandc.2012.01.008
82. Chan RCK, Chen EYH, Cheung EFC, Chen RYL, Cheung HK. The components of executive functioning in a cohort of patients with chronic schizophrenia: a multiple single-case study design. *Schizophr Res*. 2006;81:173-189. https://doi.org/10.1016/j.schres.2005.08.011
83. Carruthers SP, Van Rheenen TE, Gurvich C, Sumner PJ, Rossell SL. Characterising the structure of cognitive heterogeneity in schizophrenia spectrum disorders. A systematic review and narrative synthesis. *Neurosci Biobehav Rev*. 2019;107:252-278. https://doi.org/10.1016/j.neubiorev.2019.09.006
84. Granato G, Costanzo R, Borghi A, et al. An experimental and computational investigation of executive functions and inner speech in schizophrenia spectrum disorders. *Sci Rep*. 2025;15:5185. https://doi.org/10.1038/s41598-025-89555-3
85. Berdia S, Metz JT. An artificial neural network stimulating performance of normal subjects and schizophrenics on the Wisconsin card sorting test. *Artif Intell Med*. 1998;13:123-138. https://doi.org/10.1016/S0933-3657(98)00007-4
86. Cella M, Bishara AJ, Medin E, Swan S, Reeder C, Wykes T. Identifying cognitive remediation change through computational modelling—effects on reinforcement learning in schizophrenia. *Schizophr Bull*. 2014;40:1422-1432. https://doi.org/10.1093/schbul/sbt152
87. Wu C, Wang C, Li L. Speech-on-speech masking and psychotic symptoms in schizophrenia. *Schizophr Res Cogn*. 2018;12:37-39. https://doi.org/10.1016/j.scog.2018.02.005
88. Chang X, Zhao W, Kang J, et al. Language abnormalities in schizophrenia: binding core symptoms through contemporary empirical evidence. *Schizophr.* 2022;8:95. https://doi.org/10.1038/s41537-022-00308-x
89. Chen J, Wensing T, Hoffstaedter F, et al. Neurobiological substrates of the positive formal thought disorder in schizophrenia revealed by seed connectome-based predictive modeling. *Neuroimage-Clinical.* 2021;30:102666. https://doi.org/10.1016/j.nicl.2021.102666
90. Paul T, See JW, Vijayakumar V, et al. Neurostructural changes in schizophrenia and treatment-resistance: A narrative review. *Psychoradiology*, 2024;4. https://doi.org/10.1093/psyrad/kkae01s
91. Palaniyappan L, Homan P, Alonso-Sanchez MF. Language network dysfunction and formal thought disorder in schizophrenia. *Schizophr Bull*. 2023;49:486-497. https://doi.org/10.1093/schbul/sbac159
92. Limongi R, Mackinley M, Dempster K, Khan AR, Gati JS, Palaniyappan L. Frontal–striatal connectivity and positive symptoms of schizophrenia: implications for the mechanistic basis of prefrontal rTMS. *Eur Arch Psychiatry Clin Neurosci*. 2021;271:3-15. https://doi.org/10.1007/s00406-020-01163-6
93. Johansson M, Hjärthag F, Helldin L. Cognitive markers related to long-term remission status in schizophrenia Spectrum disorders. *Psychiatry Res*. 2020;289:113035. https://doi.org/10.1016/j.psychres.2020.113035
94. Helldin L, Kane JM, Karilampi U, Norlander T, Archer T. Remission and cognitive ability in a cohort of patients with schizophrenia. *J Psychiatr Res*. 2006;40:738-745. https://doi.org/10.1016/j.jpsychires.2006.07.005
95. Rabanea-Souza T, Akiba HT, Berberian AA, Bressan RA, Dias ÁM, Lacerda ALT. Neuropsychological correlates of remission in chronic schizophrenia subjects: the role of general and task-specific executive processes. *Psychiatry Res*. 2016;3:39-46. https://doi.org/10.1016/j.scog.2015.12.001
96. Yun DY, Hwang SSH, Kim Y, Lee YH, Kim YS, Jung HY. Impairments in executive functioning in patients with remitted and non-remitted schizophrenia. *Prog Neuro-Psychopharmacol Biol Psychiatry*. 2011;35:1148-1154. https://doi.org/10.1016/j.pnpbp.2011.03.018
97. Na E, Yim SJ, Lee JN, et al. Relationships among medication adherence, insight, and neurocognition in chronic schizophrenia. *Psychiatry Clin Neurosci*. 2015;69:298-304. https://doi.org/10.1111/pcn.12272
98. Thakkar KN, Schall JD, Boucher L, Logan GD, Park S. Response inhibition and response monitoring in a saccadic countermanding task in schizophrenia. *Biol Psychiatry*. 2011;69:55-62. https://doi.org/10.1016/j.biopsych.2010.08.016
99. Waters FAV, Badcock JC, Maybery MT, Michie PT. Inhibition in schizophrenia: association with auditory hallucinations. *Schizophr Res*. 2003;62:275-280. https://doi.org/10.1016/S0920-9964(02)00358-4
100. Hatami F, Rahman MM, Nikparvar B, Thill JC. Non-linear associations between the urban built environment and commuting modal Split: a random Forest approach and SHAP evaluation. *IEEE Access*. 2023;11:12649-12662. https://doi.org/10.1109/ACCESS.2023.3241627
101. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17:e70056. https://doi.org/10.1111/cts.70056
102. Cameron AM, Oram J, Geffen GM, Kavanagh DJ, McGrath JJ, Geffen LB. Working memory correlates of three symptom clusters in schizophrenia. *Psychiatry Res*. 2002;110:49-61. https://doi.org/10.1016/S0165-1781(02)00036-7
103. Perlstein WM, Carter CS, Noll DC, Cohen JD. Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia. *AJP.* 2001;158:1105-1113. https://doi.org/10.1176/appi.ajp.158.7.1105
104. Amoretti S, Mezquida G, Verdolini N, et al. Negative symptoms and sex differences in first episode schizophrenia: what's their role in the functional outcome? A longitudinal study. *Span J Psychiatry Ment Health*. 2025;18:91-99. https://doi.org/10.1016/j.sjpmh.2023.04.001
105. Zhao N, Wang XH, Kang CY, et al. Sex differences in association between cognitive impairment and clinical correlates in Chinese patients with first-episode drug-naïve schizophrenia. *Ann General Psychiatry*. 2021;20:26. https://doi.org/10.1186/s12991-021-00347-1
106. Usall J, Suarez D, Haro JM. Gender differences in response to antipsychotic treatment in outpatients with schizophrenia. *Psychiatry Res*. 2007;153:225-231. https://doi.org/10.1016/j.psychres.2006.09.016
107. Ceskova E, Prikryl R, Libiger J, Svancara J, Jarkovsky J. Gender differences in the treatment of first-episode schizophrenia: results from the European first episode schizophrenia trial.

*Schizophr Res*. 2015;169:303-307. https://doi.org/10.1016/j.schres.2015.10.013

108. García de la Garza Á, Blanco C, Olfson M, Wall MM. Identification of suicide attempt risk factors in a national US survey using machine learning. *JAMA Psychiatry*. 2021;78:398. https://doi.org/10.1001/jamapsychiatry.2020.4165
109. de Miras J, Ibanez-Molina A, Soriano M, Iglesias-Parro S. Schizophrenia classification using machine learning on resting state EEG signal. *Biom Signal Process Control*. 2023;79:104233. https://doi.org/10.1016/j.bspc.2022.104233
110. Wang M, Zhao S-W, Wu D, et al. Transcriptomic and neuroimaging data integration enhances machine learning classification of schizophrenia. *Psychoradiology*, 2024;4:kkae005. https://doi.org/10.1093/psyrad/kkae005
111. Fast A, Jensen D. Why stacked models perform effective collective classification. In: *2008 Eighth IEEE International Conference on Data Mining*. Pisa, Italy: IEEE; 2008:785-790. https://doi.org/10.1109/ICDM.2008.126
112. Chen Z, Liu H, Zhang Y, et al. Identifying major depressive disorder among US adults living alone using stacked ensemble machine learning algorithms. *Front Public Health*. 2025;13:13. https://doi.org/10.3389/fpubh.2025.1472050
113. Daza A, Arroyo-Paz BJ, Apaza O, Pinto J. Stacking ensemble learning model for predict anxiety level in university students using balancing methods. *Inform Med Unlocked*. 2023;42:101340. https://doi.org/10.1016/j.imu.2023.101340
114. Eleftheriadis P, Leva S, Ogliari E. Bayesian Hyperparameter optimization of stacked bidirectional long short-term memory neural network for the state of charge estimation. *Sustain Energy Grids Netw*. 2023;36:101160. https://doi.org/10.1016/j.segan.2023.101160
115. Ghasemieh A, Lloyed A, Bahrami P, Vajar P, Kashef R. A novel machine learning model with stacking ensemble learner for predicting emergency readmission of heart-disease patients. *Decis Anal J*. 2023;7:100242. https://doi.org/10.1016/j.dajour.2023.100242
116. Webb CA, Cohen ZD, Beard C, Forgeard M, Peckham AD, Björgvinsson T. Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: a comparison of machine learning approaches. *J Consult Clin Psychol*. 2020;88:25-38. https://doi.org/10.1037/ccp0000451
117. Loftus TJ, Ruppert MM, Shickel B, et al. Federated learning for preserving data privacy in collaborative healthcare research. *Digit Health*. 2022;8:205520762211344. https://doi.org/10.1177/20552076221134455
118. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29:2686-2687. https://doi.org/10.1038/s41591-023-02540-z
119. Leucht S, Bauer S, Siafis S, et al. Examination of dosing of antipsychotic drugs for relapse prevention in patients with stable schizophrenia: a meta-analysis. *JAMA Psychiatry*. 2021;78:1238-1248. https://doi.org/10.1001/jamapsychiatry.2021.2130
120. Kane JM, Leucht S, Carpenter D, Docherty JP. The expert consensus guideline series. Optimizing pharmacologic treatment of psychotic disorders. Introduction: methods, commentary, and summary. *J Clin Psychiatry*. 2003;64:5-19.




*Schizophrenia Bulletin* vol. 52 no. 2 pp. 1–20, 2026
https://doi.org/10.1093/schbul/sbaf218
Regular Article