nature medicine

Article

https://doi.org/10.1038/s41591-025-03560-7

Prediction of mental health risk in adolescents

Received: 16 September 2024

Accepted: 3 February 2025

Published online: 05 March 2025

Check for updates

Elliot D. Hill[1,2] ✉, Pratik Kashyap[3], Elizabeth Raffanello[3], Yun Wang[4], Terrie E. Moffitt[5,6,7], Avshalom Caspi[5,6,7], Matthew Engelhard[1,2,8] & Jonathan Posner[3,8]

Prospective prediction of mental health risk in adolescence can facilitate early preventive interventions. Here, using psychosocial questionnaires and neuroimaging measures from over 11,000 children in the Adolescent Brain and Cognitive Development Study, we trained neural network models to stratify general psychopathology risk. The model trained on current symptoms accurately predicted which participants would convert into the highest psychiatric illness risk group in the following year (area under the receiver operating characteristic curve = 0.84). The model trained solely on potential etiologies or disease mechanisms achieved an area under the receiver operating characteristic curve of 0.75 without relying on the child's current symptom burden. Sleep disturbances emerged as the most influential predictor of high-risk status, surpassing adverse childhood experiences and family mental health history. Including neuroimaging measures did not enhance predictive performance. These findings suggest that artificial intelligence models trained on readily available psychosocial questionnaires can effectively predict future psychiatric risk while highlighting potential targets for intervention. This is a promising step toward artificial intelligence-based mental health screening for clinical decision support systems.

Since the onset of the COVID-19 pandemic, mental illness rates among youth have risen substantially in the United States[1] and globally[2], adding strain to already overburdened mental health systems[3]. A key challenge in enhancing the effectiveness of mental health services is identifying youth most vulnerable or at the highest risk for psychiatric illness. Accurately predicting which youth in the general population will develop psychiatric problems would enable efficient allocation of preventive resources. To address this challenge, we trained neural network models[4] on longitudinal psychosocial and neurobiological data to predict future mental health risk, thereby providing an efficient approach for predicting psychiatric illness risk over time and identifying key contributing factors.

Understanding the diverse psychosocial and neurobiological factors contributing to youth mental health problems remains challenging[5]. Mental health issues seldom stem from a single cause; multiple factors typically influence them, each contributing added risk. Moreover, conventional studies assessing psychiatric risk often rely on categorical frameworks such as the Diagnostic and Statistical Manual, which does not adequately address the high rates of comorbidity across psychiatric disorders. It may be more beneficial to characterize risk factors as contributing to psychopathology broadly rather than to specific psychiatric disorders. Addressing these challenges could lead to better predictors of psychiatric risk.

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. [2]Duke AI Health, Duke University School of Medicine, Durham, NC, USA. [3]Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC, USA. [4]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA. [5]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA. [6]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [7]PROMENTA Center, University of Oslo, Oslo, Norway. [8]These authors jointly supervised this work: Matthew Engelhard, Jonathan Posner. ✉e-mail: elliot.d.hill@duke.edu

Based on longitudinal cohort research across the lifespan, investigators have recently proposed an approach to characterizing psychiatric illness based on one underlying dimension, the General Factor of Psychopathology or 'p-factor'. The p-factor reflects shared variation across psychiatric disorders, much as the g-factor reflects shared variance across intelligence domains. Studies on the social and neurobiological mechanisms underlying youth mental health issues have used the p-factor to capture general psychopathology. For example, a recent study of the Adolescent Brain and Cognitive Development (ABCD) cohort associated higher p-factor scores with smaller global brain volume and surface area on magnetic resonance imaging (MRI)[6], and a second ABCD analysis associated the p-factor with connectivity in the default mode and dorsal attention networks[7].

While predictive models of psychiatric risk have been developed, and some have shown good accuracy, they have relied upon symptom burden as predictors[8,9]. Findings from these models typically can be summarized as pointing to current symptom burden as the key predictor of future burden. While useful for screening purposes, this focus on symptoms rather than underlying etiologies has not led to new prevention strategies. An illustrative counterexample is the Framingham Risk Score[10], which indexes cardiovascular risk based in part on presumed etiologies (for example, elevated lipids) rather than the disease or its symptoms, enabling lipid-lowering interventions as a prevention strategy. For psychiatric risk assessment to have a similar impact—not just by identifying risk but guiding prevention—it must also incorporate presumed etiologies or disease mechanisms. Our analyses include two approaches: a 'symptom-driven' model using current symptom burden to predict future symptoms and a 'mechanism-driven' model using predictors based on potential etiologies of mental illness. The symptom-driven approach provides a benchmark, or upper limit, of predictive accuracy, allowing us to evaluate the performance of the mechanism-driven approach in identifying etiological predictors of psychiatric risk.

Our study examined psychosocial and neurobiological factors associated with mental illness using data from the ongoing ABCD Study, which includes over 11,000 youth and multiple assessments of their psychosocial environment and brain development collected over 5 years. These data were used to train neural network models to predict from current and previous assessment data which youth would have higher p-factor scores (indicative of a higher level of general psychopathology) 1 year later. It is possible to use traditional modeling tools in this setting, for example, by extracting summary or factor scores for each measurement instrument and applying a mixed modeling approach[11]. However, this requires simplifying assumptions about how to (1) summarize measures, (2) represent measurement history and (3) model relationships between outcomes and latent factors. In contrast, neural networks make fewer assumptions about data distribution, allowing relationships between predictors and outcomes to be entirely learned from data. While this flexibility can be a liability in smaller datasets, it is an advantage here due to the ABCD Study's size and uncertainty about the validity of the above assumptions.

We aimed to achieve five goals. First, we aimed to accurately predict mental health risk using current and past measurements. Second, we assessed whether accuracy could be maintained with accessible measurements (for example, self- or parent-report questionnaires), avoiding expensive, difficult-to-access tests. Third, we examined whether predictions remained accurate when using measurements of mechanisms, etiologies or protective factors rather than symptoms. Fourth, we focused on predicting 'conversion', identifying children whose p-factor rose from below to above the 75th percentile (that is, the most at-risk group) 1 year later. Fifth, we examined which psychosocial and neurobiological factors most influenced model predictions.

Results

Cohort descriptive statistics

After applying our filtering criteria, the dataset contained 11,416 participants. The training, validation and test sets contained 9,132, 1,142 and 1,142 participants. The mean number of events per participant was 3.1. The baseline and 1–3-year follow-up events had 11,409, 7,954, 7,667 and 4,585 participants, respectively. The mean, minimum and maximum ages were 11.4 ± 1.3, 8.9 and 15.4 years, respectively. There were 5,485 (47.7%) female and 6,027 (52.4%) male participants. There were 242 (2.1%) Asian, 1,682 (14.6%) Black, 2,310 (20.1%) Hispanic, 1,209 (10.5%) Other and 6,069 (52.7%) White participants. The number of participants in all risk groups decreased with age, follow-up event and event year due to loss-to-follow-up (Table 1).

Across the baseline and 1–3-year follow-up events, the conversion rate to high-risk was 11%—that is, about 1 of 10 participants in the sample moved from a no-, low- or medium-risk group into the high-risk group at the subsequent event. The rate of high-risk persistence was 66%—that is, about 2 of 3 participants in the high-risk group at a given event remained in the high-risk group at the next event.

Hyperparameter tuning

Models incorporating temporal dependencies between measurements (recurrent neural network (RNN), long short-term memory (LSTM) and transformer) tended to have lower validation loss than simpler models (linear model (LM) and multilayer perceptron (MLP)) that assume independent measurements. The LSTM achieved the lowest validation loss but was comparable to the RNN. For the questionnaires model, the hyperparameters that gave the lowest validation loss of 1.21 were: architecture = LSTM, hidden dimension = 191, number of layers = 1, dropout = 0.182, number of epochs = 56, learning rate = 0.003, momentum = 0.9 and weight decay = 1.61×10^{-6} (Supplementary Table 1).

Model performance

For the mechanism-driven approaches, the models trained on the data-driven MRI predictor set did not converge, even when applying substantial dimensionality reduction and regularization (Supplementary Table 2). As a result, we focused on the smaller, theory-driven set of MRI features. However, even these did not improve model performance when added to the questionnaire predictor set (Supplementary Fig. 1). Given these findings, we prioritized the questionnaire model (mechanism-driven approach), which provides strong performance and scalability.

Among the symptom-driven approaches, models using Child Behavior Checklist (CBCL) scales performed well, maintaining high predictive accuracy while requiring fewer questions, reducing the burden on participants. To streamline the analysis, we present results from two primary models:

(1) A mechanism-driven model using only questionnaire data.
(2) A symptom-driven model using CBCL scales.

Evaluation metrics for additional approaches are available in Supplementary Table 3.

The symptom-driven CBCL scales model outperformed the mechanism-driven questionnaires model across all three high-risk scenarios (Fig. 1). The mechanism-driven model trained on questionnaires had an area under the receiver operating characteristic curve (AUROC) for conversion, persistence and agnostic (any previous risk group) of 0.75 ± 0.01, 0.69 ± 0.02 and 0.8 ± 0.01, respectively. The symptom-driven model trained on CBCL scales had an AUROC for conversion, persistence and agnostic of 0.84 ± 0.01, 0.78 ± 0.01 and 0.89 ± 0.01, respectively. The models performed better when predicting the extreme compared with intermediate-risk groups; predicting which youth would be in the no-risk and high-risk groups was easier than in the low-risk and moderate-risk groups (Table 2).

The mechanism-driven model's predictive performance did not vary substantially across demographic subgroups (Table 3). Conversely, the symptom-driven performance increased with age, event year and follow-up event but did not vary considerably across race, sex or area deprivation index (ADI) quartile (Supplementary Table 3).

https://doi.org/10.1038/s41591-025-03560-7

Table 1 | Distribution of participant measurements across mental health risk categories stratified by demographic and socioeconomic subgroups

Variable	Group	No-risk	Low-risk	Moderate-risk	High-risk	Total
ADI quartile	1	2,585 (27%)	2,477 (27%)	2,372 (26%)	2,001 (22%)	9,435 (26%)
ADI quartile	2	2,408 (25%)	2,400 (26%)	2,385 (26%)	2,116 (24%)	9,309 (25%)
ADI quartile	3	2,268 (24%)	2,219 (24%)	2,278 (25%)	2,266 (25%)	9,031 (25%)
ADI quartile	4	2,227 (23%)	2,039 (22%)	2,199 (24%)	2,621 (29%)	9,086 (25%)
Age	9	767 (8%)	807 (9%)	946 (10%)	910 (10%)	3,430 (9%)
Age	10	2,003 (21%)	2,145 (23%)	2,186 (24%)	2,084 (23%)	8,418 (23%)
Age	11	2,558 (27%)	2,553 (28%)	2,562 (28%)	2,529 (28%)	10,202 (28%)
Age	12	2,364 (25%)	2,056 (23%)	2,146 (23%)	2,111 (23%)	8,677 (24%)
Age	13	1,432 (15%)	1,287 (14%)	1,147 (12%)	1,119 (12%)	4,985 (14%)
Age	14	364 (4%)	287 (3%)	247 (3%)	251 (3%)	1,149 (3%)
Event year	2016	140 (1%)	161 (2%)	178 (2%)	175 (2%)	654 (2%)
Event year	2017	1,536 (16%)	1,703 (19%)	1,683 (18%)	1,688 (19%)	6,610 (18%)
Event year	2018	2,692 (28%)	2,752 (30%)	3,005 (33%)	2,759 (31%)	11,208 (30%)
Event year	2019	2,813 (30%)	2,583 (28%)	2,542 (28%)	2,559 (28%)	10,497 (28%)
Event year	2020	2,084 (22%)	1,788 (20%)	1,674 (18%)	1,687 (19%)	7,233 (20%)
Event year	2021	223 (2%)	148 (2%)	152 (2%)	133 (1%)	656 (2%)
Follow-up event	Baseline	2,629 (28%)	2,883 (32%)	3,009 (33%)	2,985 (33%)	11,506 (31%)
Follow-up event	1-year	2,809 (30%)	2,664 (29%)	2,772 (30%)	2,567 (29%)	10,812 (29%)
Follow-up event	2-year	2,688 (28%)	2,424 (27%)	2,412 (26%)	2,437 (27%)	9,961 (27%)
Follow-up event	3-year	1,362 (14%)	1,164 (13%)	1,041 (11%)	1,015 (11%)	4,582 (12%)
Race	Asian	328 (3%)	195 (2%)	174 (2%)	93 (1%)	790 (2%)
Race	Black	1,594 (17%)	1,171 (13%)	1,034 (11%)	1,174 (13%)	4,973 (13%)
Race	Hispanic	2,005 (21%)	1,758 (19%)	1,794 (19%)	1,763 (20%)	7,320 (20%)
Race	White	4,766 (50%)	5,133 (56%)	5,219 (57%)	4,823 (54%)	19,941 (54%)
Race	Other	795 (8%)	878 (10%)	1,013 (11%)	1,151 (13%)	3,837 (10%)
Sex	Female	4,643 (49%)	4,472 (49%)	4,420 (48%)	3,990 (44%)	17,525 (48%)
Sex	Male	4,845 (51%)	4,663 (51%)	4,814 (52%)	5,014 (56%)	19,336 (52%)

Risk categories—no-risk, low-risk, moderate-risk and high-risk—correspond to quartiles of the p-factor, with the high-risk group representing the highest quartile of general psychopathology. Subgroups are stratified by demographic variables such as ADI quartile (divided into quartiles from least to most deprived) and age (in years). Each cell provides the count of participants and the percentage of the total within the specified subgroup and risk category. The Total column shows the overall count and percentage of participants within each subgroup across all risk categories.

SHAP values

The Shapley additive explanations (SHAP) value analysis using the mechanism-driven (questionnaires-only) model indicated that sleep disturbances, prosocial behaviors, adverse childhood experiences (ACEs), family mental health history and family conflict had the largest impact on high-risk predictions. Notably, sleep disturbances had a disproportionately large impact, and parent questionnaires influenced model predictions more strongly than youth questionnaires (Fig. 2).

For most participants, more sleep disturbances increased predicted high-risk probability for the following year, although extreme levels of disturbance reduced predicted risk in a small subset. More ACEs, family conflict and family history of mental illness increased high-risk probability, whereas more prosocial behaviors and parental monitoring decreased it. Although older males had a slightly higher high-risk probability than younger females, the influence of sex and age was negligible compared with the other factors (Supplementary Fig. 2). A list of the SHAP value sums for each predictor in the questionnaire model is presented in Supplementary Table 4.

Model generalization

The questionnaire model trained with the propensity score weighted loss function performed similarly to the unweighted loss function when predicting high-risk conversion, persistence and agnostic groups (Supplementary Fig. 3). The magnitude and relative ordering of the absolute SHAP value sums remained the same after reweighting.

The questionnaire model developed using site-based training, validation and test splits maintained the same predictive performance as participant-based data splits when predicting high-risk conversion, persistence and agnostic groups (Supplementary Fig. 3).

Sensitivity analysis

The AUROCs for the within-event and across-event p-factor models were comparable for the mechanism-driven questionnaires and symptom-driven CBCL scales models (Supplementary Fig. 4).

Discussion

Our study used artificial intelligence (AI) to prospectively predict mental health risk in youth. To our knowledge, this is the largest AI-based longitudinal study of generalized adolescent mental health risk prediction that assesses the relative importance of diverse predictors in making these prognostications.

Several findings stand out. First, our model predicted adolescent mental health (via the p-factor) 1 year after an assessment with relatively high accuracy. It effectively predicted high-risk conversion—predicting

https://doi.org/10.1038/s41591-025-03560-7

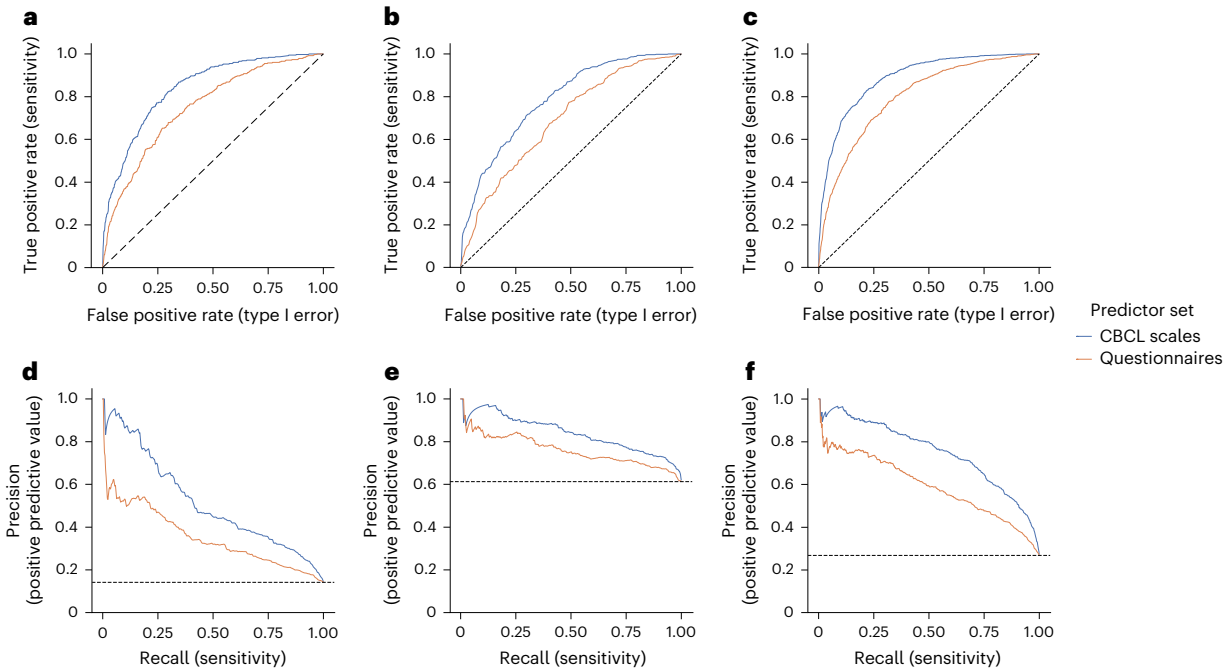

Fig. 1 | Model performance stratified by metric and high-risk groups. Receiver operating characteristic (ROC) and precision–recall (PR) curves comparing predictive model performance for mental health risk assessment for the test set (n = 1,142 participants). Each panel illustrates the performance of symptom-driven (CBCL scales) and mechanism-driven (Questionnaire) models. **a**, Conversion ROC curve: prediction of participants transitioning from no-, low- or medium-risk groups to the high-risk group. **b**, Persistence ROC curve: prediction of participants that remain in the high-risk group. **c**, Agnostic ROC curve: prediction of participants entering the high-risk group from any previous group. The dashed lines in **a**–**c** represent the baseline of random predictions. **d**, Conversion PR curve: positive predictive value for participants converting to the high-risk group. **e**, Persistence PR curve: positive predictive value for participants persisting in the high-risk group. **f**, Agnostic PR curve: positive predictive value for predicting high-risk status regardless of previous group. The curves highlight that while symptom-driven models (CBCL scales) outperform mechanism-driven models (Questionnaire), the sensitivity and precision of mechanism-driven models are strong. The dashed lines in **d**–**f** represent the outcome prevalence.

Table 2 | Predictive performance metrics for mental health risk stratified by risk scenarios and groups

Predictor set	Metric	Scenario	No-risk	Low-risk	Moderate-risk	High-risk
Questionnaires	AUROC	Conversion	0.73±0.01	0.56±0.01	0.63±0.01	0.75±0.01
Questionnaires	AUROC	Persistence	0.76±0.04	0.69±0.03	0.55±0.02	0.69±0.02
Questionnaires	AUROC	Agnostic	0.77±0.01	0.63±0.01	0.61±0.01	0.8±0.01
Questionnaires	AP	Conversion	0.56±0.02 (0.34)	0.33±0.01 (0.29)	0.35±0.02 (0.27)	0.36±0.02 (0.11)
Questionnaires	AP	Persistence	0.24±0.06 (0.03)	0.15±0.03 (0.05)	0.28±0.02 (0.26)	0.76±0.02 (0.66)
Questionnaires	AP	Agnostic	0.53±0.02 (0.26)	0.3±0.01 (0.22)	0.32±0.01 (0.26)	0.59±0.02 (0.26)
CBCL scales	AUROC	Conversion	0.82±0.01	0.64±0.01	0.72±0.01	0.84±0.01
CBCL scales	AUROC	Persistence	0.82±0.03	0.79±0.03	0.71±0.02	0.79±0.01
CBCL scales	AUROC	Agnostic	0.87±0.01	0.72±0.01	0.72±0.01	0.88±0.01
CBCL scales	AP	Conversion	0.68±0.02 (0.34)	0.38±0.02 (0.29)	0.42±0.02 (0.27)	0.52±0.03 (0.11)
CBCL scales	AP	Persistence	0.47±0.07 (0.03)	0.33±0.06 (0.05)	0.43±0.03 (0.26)	0.84±0.02 (0.66)
CBCL scales	AP	Agnostic	0.67±0.02 (0.26)	0.37±0.02 (0.22)	0.42±0.01 (0.26)	0.75±0.01 (0.26)

The table compares model performance for high-risk prediction across three scenarios—conversion, persistence and agnostic—and stratifies results by risk groups (no-risk, low-risk, moderate-risk and high-risk). Metrics reported include the AUROC and average precision (AP), denoted by mean±2s.e.m. For the AP metric, additional values in parentheses denote the prevalence of the high-risk scenario. The Predictor sets include the Questionnaires approach, reflecting the use of mechanism-driven predictors. Higher AUROC and AP values indicate better model performance in distinguishing participants likely to experience changes in their mental health risk status. (See Supplementary Glossary for additional explanation of terms).

which youth would move from lower-risk groups to the highest-risk group. Given the low prevalence of high-risk conversion (11%), the model performed well on this subset of participants (Fig. 1a,d). This is a promising step toward developing a clinical tool to preemptively identify at-risk patients who may require further evaluation to improve mental health outcomes.

Second, our model performed well with both the symptom-driven and mechanism-driven approaches, with the symptom-driven approach yielding slightly better results. This aligns with earlier research showing that current symptom burden is often a strong predictor of future burden. For example, the Chicago Adolescent Depression Risk Assessment achieves an AUROC of 0.8 for predicting

https://doi.org/10.1038/s41591-025-03560-7

Table 3 | Test set model evaluation metrics stratified by demographic subgroups

Variable	Group	No-risk	Low-risk	Moderate-risk	High-risk
Sex	Female	0.77±0.01	0.6±0.02	0.62±0.01	0.79±0.01
Sex	Male	0.78±0.01	0.65±0.01	0.6±0.01	0.8±0.01
Race	Asian	0.7±0.06	0.64±0.07	0.65±0.09	0.75±0.09
Race	Black	0.76±0.02	0.61±0.03	0.61±0.03	0.77±0.02
Race	Hispanic	0.77±0.02	0.59±0.03	0.62±0.02	0.79±0.02
Race	White	0.78±0.01	0.64±0.01	0.6±0.01	0.8±0.01
Race	Other	0.77±0.03	0.66±0.03	0.58±0.03	0.84±0.02
Age	9	0.77±0.03	0.57±0.04	0.59±0.03	0.79±0.03
Age	10	0.77±0.02	0.63±0.02	0.6±0.02	0.8±0.02
Age	11	0.73±0.02	0.61±0.02	0.59±0.02	0.79±0.02
Age	12	0.79±0.02	0.63±0.02	0.62±0.02	0.8±0.02
Age	13	0.82±0.02	0.68±0.03	0.6±0.03	0.82±0.02
Age	14	0.85±0.04	0.62±0.06	0.71±0.04	0.85±0.04
Follow-up event	Baseline	0.77±0.02	0.63±0.02	0.59±0.02	0.8±0.01
Follow-up event	1-year	0.76±0.02	0.62±0.02	0.59±0.02	0.79±0.01
Follow-up event	2-year	0.78±0.02	0.62±0.02	0.62±0.02	0.8±0.02
Follow-up event	3-year	0.8±0.03	0.67±0.03	0.65±0.03	0.81±0.02
ADI quartile	1	0.77±0.02	0.62±0.02	0.61±0.02	0.81±0.02
ADI quartile	2	0.78±0.02	0.66±0.02	0.61±0.02	0.81±0.02
ADI quartile	3	0.78±0.02	0.61±0.02	0.62±0.02	0.77±0.02
ADI quartile	4	0.76±0.02	0.6±0.02	0.58±0.02	0.8±0.02
Event year	2016	0.8±0.08	0.64±0.08	0.54±0.07	0.77±0.07
Event year	2017	0.77±0.02	0.63±0.03	0.62±0.02	0.79±0.02
Event year	2018	0.77±0.02	0.61±0.02	0.58±0.02	0.81±0.01
Event year	2019	0.77±0.02	0.62±0.02	0.61±0.02	0.79±0.02
Event year	2020	0.78±0.02	0.66±0.02	0.63±0.02	0.82±0.02
Event year	2021	0.76±0.06	0.58±0.11	0.61±0.08	0.75±0.07

Predictive performance metrics stratified by demographic subgroups for high-risk mental health prediction. The table reports AUROC values for each subgroup across risk categories (no-risk, low-risk, moderate-risk and high-risk). Subgroups include sex (female, male) and race/ethnicity (Asian, Black, Hispanic, white and other). AUROC values are presented as mean±2 s.e.m., reflecting the model's ability to distinguish participants within each demographic and risk category. Higher AUROC values indicate better model performance. The table demonstrates minimal variation in predictive performance across demographic subgroups, underscoring the model's robustness and potential fairness across diverse populations. See Supplementary Glossary for additional explanation of terms.

future depressive episodes using current mood, anxiety and affect regulation[8]. Relatedly, models predicting progression from at-risk to full-threshold schizophrenia[12] often rely on current symptoms and family history. While symptom-driven approaches may be useful for screening, they are susceptible to common methods bias, where predictors and outcomes come from the same source, potentially inflating associations.

In contrast, our mechanism-driven approach maintained strong predictive performance without relying on the child's current symptom load. Its predictions were within a general, nonreferred sample and included potential etiologies of illness, making it distinct from symptom-based approaches. By focusing on underlying mechanisms rather than symptom patterns, the mechanism-driven model offers the potential to identify actionable prevention targets.

Regarding potential prevention targets, sleep disturbances stood out as a robust predictor of high psychiatric illness risk. These were assessed with the 26-item Sleep Disturbance Scale for Children[13] (see Supplementary Table 5 for specific items and scoring). Their predictive influence surpassed variables such as ACEs, family mental health history and socioeconomic status. Approximately 29% of the ABCD cohort had sleep disturbance scores exceeding clinical thresholds[13], indicating that pathological sleep disturbance was common but not universal. Notably, the relationship between sleep disturbances and future mental illness risk was nonlinear; increased sleep disturbances raised risk up to a point, but extreme—and far less common—levels of sleep disturbances did not (Supplementary Fig. 2). This could suggest that while moderate sleep disturbances portend mental illness, severe sleep problems may instead relate to nonpsychiatric conditions. Importantly, our model's performance remained stable during the pandemic, a period of potentially heightened sleep disruption (Table 3). The stability of the model's performance during the COVID-19 pandemic supports its generalizability under varying real-world conditions.

The importance of sleep disturbance in predicting mental illness is consistent with earlier research. A meta-analysis[14] found sleep disturbances to be a precursor for several psychiatric disorders, including depression, anxiety and bipolar disorder. Other studies have associated sleep disturbances with subsequent substance use disorders, suicide attempts and suicide completion[15]. Sleep is vital for healthy neurodevelopment[16], and sleep disturbances, particularly in adolescence, are implicated as contributors to mental illness across diagnoses[17–19]. Although further research is needed, our finding that sleep disturbances were an influential predictor of high psychiatric

https://doi.org/10.1038/s41591-025-03560-7

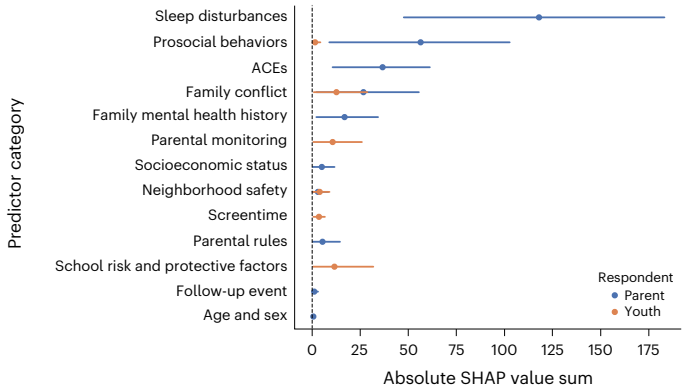

Fig. 2 | Predictor category SHAP analysis. SHAP values represent the influence of each predictor category on the model's predicted probability of a participant being classified as high-risk. Predictor categories are ranked by their absolute SHAP value sums, with higher values indicating a larger impact on model predictions. Absolute SHAP value sums are displayed as mean ± 95% confidence interval (n = 1,142 participants). The sleep disturbances category emerged as the most influential predictor, followed by prosocial behaviors, ACEs, family mental health history and family conflict. Each predictor category is further stratified by respondent type (youth versus parent), with parent-reported data generally showing more substantial influence on model predictions than youth-reported data. The questionnaire predictor responses can be found in Supplementary Table 4.

illness risk is potentially auspicious, as sleep disturbances are modifiable with evidence-based behavioral interventions[20]. Indeed, a recent clinical trial found that a sleep intervention effectively reduced symptoms across diagnoses[21].

Third, using the p-factor as an outcome supports the model's broad clinical applicability as a transdiagnostic indicator of mental illness. The model's predictions are not limited to a single diagnostic group. Additionally, the low variation in the model's predictive performance across demographic groups (Table 3) suggests resilience to demographic biases. The symptom-driven model's improved performance with participant age and event year suggests that the LSTM effectively utilized previous information, highlighting the benefit of longitudinal assessments for mental health prediction. In addition to its broad applicability, the p-factor also serves as a robust harbinger of future adverse psychiatric incidents. For example, a registry study of over 32,000 twins prospectively associated a general psychopathology factor (that is, the p-factor) with increased suicidal behavior, substance overdoses, criminality and new prescriptions for most psychiatric medications[22]. While our approach derived the p-factor from the multi-item CBCL, new approaches, including computerized adaptive testing, aim to enable faster p-factor assessments[23,24].

Fourth, adding MRI-derived neurobiological measures did not improve model performance beyond that of the questionnaires or CBCL models. This finding was consistent across both the theory-driven and data-driven approaches we employed and aligns with research showing marginal associations between brain measures and mental health[25]. This result underscores the model's potential clinical utility as its predictors can rely on accessible, low-cost instruments scalable to most clinical settings. An alternative explanation for the lower performance of MRI predictors could be that they had half as many measurements as the psychosocial questionnaires because MRI scans were only taken every other year in the ABCD Study. In contrast, the psychosocial predictors were measured every year. We used tabular MRI data for our experiments, primarily to compare them with the more scalable questionnaires; however, training computer vision models on ABCD's raw imaging data may offer a promising avenue for future research.

Fifth, our findings indicate that the model's performance was dominated by parent-reported data, with minimal contribution from youth-reported data (Fig. 2). With their broader life experience, parents may better contextualize behaviors and detect deviations from developmental norms. In contrast, youth may lack external reference points required to recognize atypical experiences. Future work could explore ways to integrate youth perspectives into predictive models better.

Applying propensity weights to account for sampling biases in the ABCD Study showed no change in the model's performance metrics and SHAP values, supporting its reliability when applied to a population more representative of the United States. Additionally, we conducted a site-specific analysis by excluding data from two ABCD sites during model training, using them exclusively for testing. This 'spatial distribution shift' evaluation yielded consistent results, suggesting the model's resilience to differences in assessment procedures, demographics and other site-specific factors. Together, these analyses provide strong evidence for the model's generalizability across diverse conditions. While our findings are promising, additional steps are needed to enhance the model's readiness for clinical deployment. Prospective evaluation in independent samples—such as real-world clinical settings where clinicians assess model predictions without influencing clinical decisions—will be critical[26]. Another priority is identifying a minimal set of questions that maintains high predictive performance while reducing patient burden. Sparsity-inducing loss functions (for example, L1 penalization[27]) could aid in optimizing this balance. Further performance improvements may be achieved using accessible, low-cost data sources such as medical records or behavioral assays. Lastly, while the ABCD Study sampled a diverse range of demographics across the United States, it is still possible that clinical populations are systematically different from the general US population. Thus, integrating methods that account for distribution shifts[28] into the modeling pipeline will be essential for improving clinical applicability. Future clinical applications could also include measures of adaptive functioning, or functional impairment, to augment mental health assessments.

This study demonstrates that AI models can accurately predict adolescent mental health prospectively using readily available questionnaires. Social environment and behavioral measures (particularly sleep disturbances) strongly influenced model-predicted psychiatric illness risk, whereas the impact of neurobiological measures on predicted risk was limited. Future research should validate these findings in clinical populations and assess whether AI-based early detection systems can help allocate mental health resources to patients who require preemptive intervention to improve their long-term outcomes. If validated, this approach could guide targeted preventive efforts, enhance early detection and ultimately contribute to more efficient mental health care delivery for adolescents.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-03560-7.

References

1. Xiao, Y., Brown, T. T., Snowden, L. R., Chow, J. C.-C. & Mann, J. J. COVID-19 policies, pandemic disruptions, and changes in child mental health and sleep in the United States. *JAMA Netw. Open* **6**, e232716 (2023).
2. Samji, H. et al. Review: Mental health impacts of the COVID-19 pandemic on children and youth – a systematic review. *Child Adolesc. Ment. Health* **27**, 173–189 (2022).
3. Kourgiantakis, T. et al. Navigating inequities in the delivery of youth mental health care during the COVID-19 pandemic: perspectives of youth, families, and service providers. *Can. J. Public Health* **113**, 806–816 (2022).

4. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
5. Posner, J. The role of precision medicine in child psychiatry: what can we expect and when? *J. Am. Acad. Child Adolesc. Psychiatry* **57**, 813 (2018).
6. Romer, A. L., Ren, B. & Pizzagalli, D. A. Brain structure relations with psychopathology trajectories in the ABCD Study. *J. Am. Acad. Child Adolesc. Psychiatry* **62**, 895–907 (2023).
7. Sripada, C. et al. Widespread attenuating changes in brain connectivity associated with the general factor of psychopathology in 9- and 10-year olds. *Transl. Psychiatry* **11**, 575 (2021).
8. Voorhees, B. W. V. et al. Predicting future risk of depressive episode in adolescents: the Chicago Adolescent Depression Risk Assessment (CADRA). *Ann. Fam. Med.* **6**, 503–511 (2008).
9. King, M. et al. Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD Study. *Arch. Gen. Psychiatry* **65**, 1368–1376 (2008).
10. Wilson, P. W. et al. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
11. Stroup, W. W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications* (CRC, 2012).
12. Shah, J. et al. Multivariate prediction of emerging psychosis in adolescents at high risk for schizophrenia. *Schizophr. Res.* **141**, 189–196 (2012).
13. Bruni, O. et al. The Sleep Disturbance Scale for Children (SDSC). Construction and validation of an instrument to evaluate sleep disturbances in childhood and adolescence. *J. Sleep Res.* **5**, 251–261 (1996).
14. Pigeon, W. R., Bishop, T. M. & Krueger, K. M. Insomnia as a precipitating factor in new onset mental illness: a systematic review of recent findings. *Curr. Psychiatry Rep.* **19**, 44 (2017).
15. Pigeon, W. R., Pinquart, M. & Conner, K. Meta-analysis of sleep disturbance and suicidal thoughts and behaviors. *J. Clin. Psychiatry* **73**, e1160–e1167 (2012).
16. Telzer, E. H., Goldenberg, D., Fuligni, A. J., Lieberman, M. D. & Gálvan, A. Sleep variability in adolescence is associated with altered brain development. *Dev. Cogn. Neurosci.* **14**, 16–22 (2015).
17. Uccella, S. et al. Sleep deprivation and insomnia in adolescence: implications for mental health. *Brain Sci.* **13**, 569 (2023).
18. Freeman, D., Sheaves, B., Waite, F., Harvey, A. G. & Harrison, P. J. Sleep disturbance and psychiatric disorders. *Lancet Psychiatry* **7**, 628–637 (2020).
19. Harvey, A. G., Murray, G., Chandler, R. A. & Soehner, A. Sleep disturbance as transdiagnostic: consideration of neurobiological mechanisms. *Clin. Psychol. Rev.* **31**, 225–235 (2011).
20. Lunsford-Avery, J. R., Bidopia, T., Jackson, L. & Sloan, J. S. Behavioral treatment of insomnia and sleep disturbances in school-aged children and adolescents. *Child Adolesc. Psychiatr. Clin. N. Am.* **30**, 101–116 (2021).
21. Harvey, A. G. et al. A randomized controlled trial of the Transdiagnostic Intervention for Sleep and Circadian Dysfunction (TranS-C) to improve serious mental illness outcomes in a community setting. *J. Consult. Clin. Psychol.* **89**, 537–550 (2021).
22. Pettersson, E., Larsson, H., D'Onofrio, B. M. & Lichtenstein, P. Associations between general and specific psychopathology factors and 10-year clinically relevant outcomes in adult Swedish twins and siblings. *JAMA Psychiatry* **80**, 728–737 (2023).
23. Moore, T. M. et al. Development of a computerized adaptive screening tool for overall psychopathology ('p'). *J. Psychiatr. Res.* **116**, 26–33 (2019).
24. Jones, J. D. et al. The general psychopathology 'p' factor in adolescence: multi-informant assessment and computerized adaptive testing. *Res. Child Adolesc. Psychopathol.* **52**, 1753–1764 (2024).
25. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
26. Antoniou, T. & Mamdani, M. Evaluation of machine learning solutions in medicine. *Can. Med. Assoc. J.* **193**, E1425–E1429 (2021).
27. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B* **58**, 267–288 (1996).
28. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset Shift in Machine Learning* (MIT, 2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

Methods

Ethics statement

Institutional review boards at each ABCD Study site approved the study procedures. Written consent was obtained from all parents and verbal assent was given by all children. The Duke Health Institutional Review Board approved analysis of ABDC data related to this paper.

Glossary of terms

For definitions of technical terms, refer to Supplementary Glossary.

Description of cohort

The ABCD Study collected data from a community sample of 11,880 children from 21 research sites across the United States. Recruitment for the ABCD Study used a multi-stage probability sampling method to capture the sociodemographic diversity of the US population (for additional details, see ref. 29). Post-sampling propensity weights were provided to improve the sample's representativeness, and we incorporated these weights in our generalization analyses[30]. All youth had in-person assessments once a year. Self-report and social assessments were taken yearly, whereas brain MRI scans were collected every 2 years. The baseline ABCD assessment was taken between 2017 and 2018 at ages 9–10, and four subsequent waves of assessment (1-year follow-up through 4-year follow-up; hereafter referred to as 'events') were included in our analysis. ABCD questionnaires were completed by either the youth or a parent. Our model predicted outcomes at each follow-up event based on predictors measured at all preceding events.

Questionnaires

To assess participants' social environment and behaviors, we included predictors from the following ABCD questionnaires in our analysis: CBCL, family conflict, neighborhood safety, problem monitoring, prosocial behaviors, school risk and protective factors, screentime, sleep disturbances, parental rules and family mental health history (see Supplementary Table 5 for predictors metadata, including the ABCD file, predictor names, questions and responses). In addition, we also included measures of demographics (age and sex assigned at birth), ACEs and socioeconomic status. We included age and sex even though our outcome (p-factor, see below) is adjusted for them because nonlinear interactions between these demographic variables and other predictors may not be accounted for in a linear adjustment. Because the ABCD Study did not include a questionnaire specific to ACEs (for example, Center for Disease Control–Kaiser ACEs Questionnaire[31]), we derived an ACEs score for each participant based on other questionnaires (see Supplementary Table 6 for details). For socioeconomic status, we used the ADI[32], the highest level of parental education and annual household income. Lastly, we included the time elapsed since the baseline measurement (in years) as a predictor.

MRI measures

We used two approaches for selecting MRI predictors: theory-driven and data-driven approaches.

In the theory-driven approach, we limited the MRI predictors to measures commonly associated with psychopathology[33], including transdiagnostic measures in youth[34]. These predictors included (1) functional connectivity estimates between the default mode, fronto-parietal and cingulo-opercular networks (based on ref. 35); (2) diffusion tensor imaging-derived average fractional anisotropy within 21 white matter tracts (see Supplementary Table 5 for specific tracts); and (3) mean activation (beta weights) from predefined regions of interest during the stop signal task[36], monetary incentive task[37] and emotional N-back task[38]. This approach prioritized computational efficiency and reduced model complexity while preserving the study's primary objectives.

To complement this targeted strategy, we conducted a data-driven analysis incorporating all available MRI measures from the ABCD Study, encompassing 55,207 predictors. This broader analysis allowed us to explore the potential utility of less commonly studied MRI features, ensuring no potentially relevant information was excluded.

p-factor

We performed factor analysis on the eight CBCL syndrome scales measured at all follow-up events to construct our outcome variable, the p-factor. The syndrome scales included anxious/depressed, withdrawn/depressed, somatic complaints, social problems, thought problems, attention problems, rule-breaking behavior and aggressive behavior. A previous analysis of p-factor in ABCD demonstrated that factor loadings and scores are robust to the choice of factor analysis model and whether individual items versus CBCL subscales were used[39]. Therefore, we used a single-factor exploratory factor analysis model to reduce the dimensionality of the CBCL subscales to a single factor—the p-factor—after which we binned the p-factor scores into quartiles. The model's task was to predict which quartile a participant's p-factor would be in at the next follow-up event. For example, we tested whether predictors (for example, questionnaires and MRI measures) at the baseline could predict a participant's p-factor quartile measured at the 1-year follow-up event. We term quartiles 1–4 as 'no-risk', 'low-risk', 'medium-risk' and 'high-risk', respectively. We defined 'conversion' as moving from the no-risk, low-risk or medium-risk groups into the high-risk group at the next follow-up event; 'persistence' as remaining in the high-risk group at the next follow-up event; and 'agnostic' as being in the high-risk group at the next follow-up event irrespective of group in the previous event.

Data preprocessing

To make a p-factor prediction, a participant needed at least one p-factor measurement available from an event after the predictors were measured. Therefore, we did not include outcomes (that is, p-factors) from the baseline event because no preceding data were available for this prediction. Similarly, we did not include predictors (questionnaires and MRI measures) from the 4-year follow-up event because there was no subsequent event in which the p-factor outcome could be assessed. Lastly, we excluded participants who had CBCL scales measured only at baseline.

From the questionnaires, we removed items that asked about the participant's preferred language, the number of missing/answered questions and summary scores (that is, measures aggregated from multiple questions). In addition, we excluded predictors with more than 25% of their values missing. We standardized our predictors to *z*-scores (0 mean and unit variance). To impute missing values, we forward-filled predictors within participants; for predictors without values at preceding events, we imputed the mean value of the predictor across all participants for that event. Lastly, we split our data into independent training, validation and test sets with proportions 8:1:1. The training set was used to learn the neural network parameters that best fit the data. The validation set was used to select optimal hyperparameters, and the test set was used to estimate out-of-sample model performance. To prevent data leakage (that is, when the model erroneously learns information from the test set during training), all learnable preprocessing steps (that is, *z*-score scaling and mean imputation) were fit only on the training set and then applied separately to the validation and test sets.

Model architecture

We trained neural network models to predict a participant's p-factor quartile at a measurement event from the predictors measured at all previous events. We tested five neural network model architectures in increasing order of complexity: LM, MLP[40], RNN[41], LSTM[42] and transformer[43]. The LM and MLP assume that future psychiatric risk (as measured by the p-factor) is conditionally independent of previous assessments, given information collected at the current

assessment. In contrast, the RNN, LSTM and transformer learn temporal dependencies between p-factor predictions and the full history of past measurements. Importantly, we used a unidirectional RNN and LSTM to prevent the model from erroneously learning from measurements not yet available at the current prediction time. For the same reason, we used a causal attention mask for the transformer.

We chose these sequence neural network models because they naturally handle repeated measurements and can efficiently learn high-dimensional, nonlinear relationships between predictors and outcomes. For the data-driven MRI predictor set, in addition to the architectures above, we used an additional autoencoder with linear encoder and decoder layers to reduce the dimensionality of the model input from 55,207 to 256 latent dimensions. Then, the learned latent encodings were used as the input for the five architectures listed above.

Training and hyperparameter tuning

The output dimension of all models was 4 (the number of predicted risk groups). We used multiclass cross-entropy (negative log-likelihood) as our loss function. We optimized the model's parameters using stochastic gradient descent[44] with Nesterov momentum[45]. We tuned the model architecture and optimization hyperparameters using 128 tuning trials[46]. The optimizer hyperparameter ranges for the number of epochs, initial learning rate, momentum and weight decay[47] were [10, 100], $[1 \times 10^{-4}]$, [0.9, 0.99] and $[1 \times 10^{-8}]$, respectively. The model hyperparameter ranges for the dropout rate[48], hidden dimension size and number of layers were [0.0, 0.75], [64, 256] and [1, 5]. For the transformer, we set the number of attention heads to 4. The five model architectures were treated as hyperparameters during tuning. Finally, the model hyperparameters that achieved the lowest cross-entropy loss on the validation set were selected for evaluation on the test set. We used a learning rate scheduler to reduce the learning rate by a factor of 0.1 when the validation loss did not decrease after two epochs.

Model performance metrics

All model performance metrics were generated using the test set to estimate out-of-sample model performance (that is, the model's performance on data not used to train it). To evaluate the model's predictive performance across all decision thresholds, we calculated the receiver operating characteristic curve (true positive rate versus false positive rate) and precision–recall curve (positive predictive value versus true positive rate) for each quartile. To summarize the model's performance, we calculated the AUROC and the average precision, that is, the area under the precision–recall curve using the model's predicted and observed outcomes for each quartile. All confidence intervals were generated by bootstrap resampling[49] with 1,000 resamples of the test set predictions.

Scenarios for model performance

To assess the potential clinical utility of the model, we reasoned that the most clinically useful feature of this predictive model would be its ability to predict conversion (that is, when a child's mental health changes from low-risk to high-risk). A high likelihood of conversion should affect treatment planning. Conversely, if a child already has severe psychiatric illness and the model predicts that this child will remain in the high-risk group in the following year, such a prediction has less clinical utility because the treatment plan is unlikely to change.

Toward this goal, we evaluated the model under three scenarios. The first scenario was conversion from no-, low- or moderate-risk to high-risk. Evaluating high-risk predictions for this scenario assesses the model's ability to identify which participants will convert from a healthier to a more severe psychiatric state. The second scenario was persistence (remaining in the high-risk group from one year to the next). The third scenario was agnostic, or any previous risk group—that is, evaluating the model's ability to predict being in the high-risk group, regardless of previous group status. Lastly, to measure how the model performance varied across demographic subgroups, we stratified the model performance metrics by sex, race/ethnicity, ADI quartile, age, follow-up event and event year.

We also examined model performance across six different approaches based on the predictor subsets used in each approach. The subsets included (1) questionnaires, (2) questionnaires and MRI measures, (3) previous p-factors, (4) CBCL scales, (5) questionnaires and CBCL scales and (6) questionnaires, MRI measures and CBCL scales. For simplicity, we refer to approaches (1) and (2) as 'mechanism-driven' and approaches (3–6) as 'symptom-driven'. Mechanism-driven approaches use distinct predictors and outcomes. In contrast, symptom-driven approaches involve predictors that overlap with the outcome measures, although they are measured at earlier time points (that is, autoregressive predictors). For example, the p-factor (our outcome) is derived from CBCL scales; therefore, any approach incorporating CBCL scales as predictors is considered symptom-driven.

Symptom-driven models generally provide high predictive accuracy but are less likely to uncover actionable insights for interventions. For instance, while current moderate illness might predict later severe illness, such a prediction does not reveal underlying causes or suggest treatments. Conversely, mechanism-driven approaches focus on predictors from domains such as family conflict and neighborhood safety—factors thought to influence or protect against mental illness. These domains are distinct from those that define the p-factor. Analogous to the Framingham Risk Score in cardiovascular health, mechanism-driven approaches emphasize identifying modifiable mechanisms or etiologies contributing to mental illness, making them better suited for guiding targeted interventions. As a baseline for symptom-driven models, we evaluated the naive case of simply using the previous p-factor quartile to predict the following year's p-factor quartile.

Predictor importance

We used SHAP[50] to evaluate predictor importance. SHAP values estimate the influence of a participant's predictor value on that participant's predicted risk probability. To derive a single measure of importance for each predictor, we summed the SHAP values across each participant and event. To determine which questionnaire had the largest relative influence on model predictions, we aggregated the SHAP values by taking the absolute value of the sum of all predictors within a given questionnaire (for example, SHAP values from each question in the family conflict questionnaire were summed). SHAP values must be estimated from a single model output; thus, we generated them using the predicted high-risk probability as we deemed it the most clinically relevant model output. We estimated SHAP values using expected gradients[51]. We generated 95% confidence intervals from 1,000 bootstrap resamples.

Model generalization

To explore the generalizability of our model, we employed two complementary analyses: (1) propensity score weighting and (2) site analysis. First, to address sampling biases and enhance the representativeness of our sample, we weighted the cross-entropy loss function using propensity scores provided by the ABCD Study[52]. These weights adjust for sampling biases, approximating a population more representative of the general US population. By re-running our model with these adjusted weights, we tested the robustness of our findings under conditions that better reflect broader US population characteristics.

Second, we performed a site-level analysis to evaluate the model's robustness to potential variability in assessment procedures and participant demographics across sites. Specifically, we excluded data from two ABCD sites during the model training and testing phases. These held-out sites were used solely for validation, enabling us to test model performance under 'spatial' shifts—situations where data

characteristics, such as questionnaire administration methods, order of administration or site-specific demographic distributions, differ. This approach allowed us to assess how well the model performs when applied to previously unseen contexts.

Sensitivity analysis

We tested whether estimating p-factors across versus within follow-up events affected model performance. When estimating factors across follow-up events, we assume the factor loadings do not change over time. In contrast, when we estimate factors within follow-up events, we assume that the corresponding factor loadings at different time points are statistically independent. These variations represent the two extremes of handling time in factor analysis, with more sophisticated modeling techniques likely falling somewhere in between.

Summary of methods

We trained neural network models on longitudinal psychosocial and MRI data from the ABCD cohort, which included 11,880 children aged 9–10 at baseline, to predict mental health risk 1 year later. Predictors comprised questionnaires capturing social environment, behaviors and brain imaging measures. Data preprocessing involved standardizing predictors, imputing missing values and splitting the data into stratified training, validation and test sets. The outcome variable, the p-factor, was derived through factor analysis of CBCL syndrome scales and categorized into risk quartiles. Model performance was evaluated across various scenarios, including predicting conversion to high-risk status. Predictor importance was analyzed using SHAP values. Sensitivity analyses assessed the impact of different factor models on p-factor estimation and the model's ability to generalize to different populations in the United States (see Supplementary Fig. 5 for an overview of the methods).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study were obtained from the Adolescent Brain Cognitive Development (ABCD) Study, a publicly available dataset managed by the National Institutes of Health (NIH). Researchers can request access to the ABCD Study data through the National Institute of Mental Health Data Archive (NDA) at https://nda.nih.gov/abcd. Access is restricted to qualified researchers affiliated with an institution, and approval is contingent on compliance with the NDA Data Use Certification, which is available at https://nda.nih.gov/ndapublicweb/Documents/NDA+Data+Access+Request+DUC+FINAL.pdf. Data can be used for scientific research purposes only and cannot be used for commercial or nonresearch applications. Requests for access are typically reviewed within 2–4 weeks.

Code availability

The code for all data processing, modeling and analysis can be found at https://github.com/Elliot-D-Hill/abcd (ref. 53).

References

29. Garavan, H. et al. Recruiting the ABCD sample: design considerations and procedures. *Dev. Cogn. Neurosci.* **32**, 16–22 (2018).
30. Karcher, N. R. & Barch, D. M. The ABCD Study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* **46**, 131–142 (2021).
31. Felitti, V. J. et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the Adverse Childhood Experiences (ACE) Study. *Am. J. Prev. Med.* **14**, 245–258 (1998).
32. Kind, A. J. et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Ann. Intern. Med.* **161**, 765–774 (2014).
33. Parlatini, V. et al. White matter alterations in attention-deficit/hyperactivity disorder (ADHD): a systematic review of 129 diffusion imaging studies with meta-analysis. *Mol. Psychiatry* **28**, 4098–4123 (2023).
34. Xia, J., Chen, N. & Qiu, A. Unraveling multimodal brain signatures: deciphering transdiagnostic dimensions of psychopathology in adolescents. *Adv. Intell. Syst.* **5**, 2300577 (2024).
35. Gordon, E. M. et al. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26**, 288–303 (2016).
36. Verbruggen, F. & Logan, G. D. Response inhibition in the stop-signal paradigm. *Trends Cogn. Sci.* **12**, 418–424 (2008).
37. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12**, 20–27 (2000).
38. Cohen, A., Conley, M., Dellarco, D. & Casey B. The impact of emotional cues on short-term and long-term memory during adolescence. [Abstract]. In *Proc. Soc. Neurosci.* (2016).
39. Clark, D. A. et al. The general factor of psychopathology in the Adolescent Brain Cognitive Development (ABCD) Study: a comparison of alternative modeling approaches. *Clin. Psychol. Sci.* **9**, 169–182 (2021).
40. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2**, 183–197 (1991).
41. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270 (2019).
42. Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks* (Springer, 2012).
43. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
44. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT2010* (eds Lechevallier, Y. & Saporta, G.) 177–186 (Physica-Verlag, 2010).
45. Nesterov, Y. E. A method for solving a convex programming problem with convergence rate O(1/k^2). *Dokl. Akad. Nauk*. 543–547 (1983).
46. Bergstra, J., Yamins, D. & Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. 30th Int. Conf. Mach. Learn* (eds Dasgupta, S. & McAllester, E.) 115–123 (PMLR, 2013).
47. Krogh, A. & Hertz, J. A simple weight decay can improve generalization. *Adv. Neural Inf. Process. Syst.* **4**, 950–957 (1991).
48. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
49. Efron, B. in *Breakthroughs in Statistics*, Vol. 2 (eds Kotz, S. & Johnson, N. L.) 569–593 (Springer, 1992).
50. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30**, 4765–4774 (2017).
51. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. Int. Conf. Mach. Learn* (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
52. Elliott, M. R. & Valliant, R. Inference for nonprobability samples. *Stat. Sci.* **32**, 249–264 (2017).
53. Hill, E. D. Elliot-D-Hill / abcd. *GitHub* https://github.com/Elliot-D-Hill/abcd (2025).

Acknowledgements

M.E. is supported by grant no. K01-MH127309 from the National Institute of Mental Health (NIMH). T.E.M. and A.C. are supported by grants no. R01-AG032282 and no. R01-AG069939 from the National Institute on

Aging (NIA) and grant no. MR/P005918/1 from the Medical Research Council. Health Data Science at Duke is supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant Award Number UL1 TR002553. The Duke AI Health Data Science Fellowship Program is supported by the above grant, the Duke Department of Biostatistics & Bioinformatics and Duke AI Health. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

E.D.H. was responsible for data curation, software, data analysis, interpretation and paper drafting. Y.W. and P.K. performed data curation and data analysis. E.R. performed data interpretation. A.C. and T.E.M. were responsible for interpretation and paper editing. M.E. performed data analysis, interpretation and paper drafting. J.P. performed data analysis, interpretation and paper drafting.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-03560-7.

Correspondence and requests for materials should be addressed to Elliot D. Hill.

Peer review information *Nature Medicine* thanks Nima Aghaeepour, Oliviero Bruni, Paul Moran and Lucina Uddin for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Elliot D. Hill

Last updated by author(s): Jan 29, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ABCD data collection details: https://www.sciencedirect.com/science/article/pii/S1878929317301214
Data analysis	The analysis code is provided here: https://github.com/Elliot-D-Hill/abcd

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data are from the Adolescent Brain and Cognitive Development (ABCD) study. Access to the raw data requires a data use agreement with NIH, as stupulated by the ABCD study's controlled access policies. Data can be requested here: https://nda.nih.gov/abcd/request-access

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We have reported participant sex as defined in the ABCD study, which is based on the question: "What sex was the child assigned
Reporting on race, ethnicity, or other socially relevant groupings	In Table 1 of our manuscript, we present the distribution of participants across mental health risk categories, stratified by demographic variables, including race/ethnicity.
Population characteristics	11,880 male and female children ages 9-16 years with races White, Black, Hispanic, Asian, and Other.
Recruitment	Multi-stage probability sampling method from 21 research sites across the United states.
Ethics oversight	The Duke Health Institutional Review Board approved analysis of ABDC data related to this manuscript.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data were collected data from a community sample of 11,880 children from 21 research sites across the United States.
Research sample	Representative sample of 11,880 male/female children ages 9-16 years. Race: White, Black, Hispanic, Asian, and Other.
Sampling strategy	Multi-stage probability sampling.
Data collection	Clinical, behavioral, demographic, and brain MRI measures collected.
Timing	ABCD assessments were taken between 2016 and 2018 at ages 9-10 with four subsequent yearly assessments.
Data exclusions	To our knowledge, the ABCD study has not publicly released detailed information on consent rates.
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	The ABCD study provided popensity scores to adjust for potiential non-representative sampling.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) — *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals — *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Reporting on sex — *Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall*

	numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks — *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes — *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication — *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication. — *For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*

Files in database submission — *Provide a list of all files available in the database submission.*

Genome browser session
(e.g. UCSC) — *Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates — *Describe the experimental replicates, specifying number, type and replicate agreement.*

Sequencing depth — *Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*

Antibodies — *Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Peak calling parameters — *Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality — *Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*

Software — *Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation — *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*

Instrument — *Identify the instrument used for data collection, specifying make and model number.*

Software — *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type: Brain MRI data were collected as part of the ABCD study

Design specifications: MRI data included structural MRI, task-based fMRI, resting fMRI, and DTI

Behavioral performance measures: Monetary Incentive Delay (MID) task, the Stop Signal task (SST) and an emotional version of the n-back task

Acquisition

Imaging type(s): MRI data included structural MRI, task-based fMRI, resting fMRI, and DTI

Field strength: 3T

Sequence & imaging parameters: ABCD imaging protocol is described here: https://abcdstudy.org/scientists/protocols-mri/

Area of acquisition: Whole brain

Diffusion MRI ☒ Used ☐ Not used

Parameters: *Specify # of directions, b-values, whether single shell or multi-shell, and if cardiac gating was used.*

Preprocessing

Preprocessing software: A description of preprocessing is provided here: https://pmc.ncbi.nlm.nih.gov/articles/PMC6981278/

Normalization: *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template: *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal: *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring: *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings: *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested: *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016): *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction: *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis: Neural network models (RNN, LSTM, and Transformer).