Using deep learning to predict internalizing problems from brain structure

Marlee M. Vandewouw[1,2], Bilal Syed[3], Noah Barnett[4], Alfredo Arias[5], Elizabeth Kelley[6,7], Jessica Jones[6,7], Muhammad Ayub[6,8], Alana Iaboni [1], Paul D. Arnold[9,10], Jennifer Crosbie[11,12], Russell J Schachar*[12,13], Margot J Taylor[11,14,15,16], Jason P. Lerch[11,17,18], Evdokia Anagnostou[1,11], Azadeh Kushki[1,2]

[1] Autism Research Centre, Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, ON, Canada
[2] Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
[3] Mouse Imaging Centre, The Hospital for Sick Children, Toronto, Canada
[4] Schulich School of Business, York University, Toronto, Canada
[5] Participant Advisory Committee, Province of Ontario Neurodevelopmental Network, Toronto, Canada
[6] Department of Psychiatry, Queen's University, Kingston, Canada
[7] Department of Psychology, Queen's University, Kingston, Canada
[8] Division of Psychiatry, University College London, London, United Kingdom
[9] The Mathison Centre for Mental Health & Education, Hotchkiss Brain Institute, University of Calgary, Calgary, Canada
[10] Departments pf Psychiatry and Metical Genetics. Cumming School of Medicine, University of Calgary, Calgary, Canada
[11] Program in Neurosciences & Mental Health, The Hospital for Sick Children, Toronto, Canada
[12] Department of Psychiatry, University of Toronto, Toronto, Canada
[13] Department of Psychiatry, The Hospital for Sick Children, Toronto, Canada
[14] Department of Diagnostic Imaging, The Hospital for Sick Children, Toronto, Canada
[15] Department of Psychology, University of Toronto, Toronto, Canada
[16] Department of Medical Imaging, University of Toronto, Toronto, Canada
[17] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom
[18] Department of Medical Biophysics, University of Toronto, Toronto, Canada
* Emiritus

Corresponding author:
Marlee M. Vandewouw
Autism Research Centre, Holland Bloorview Kids Rehabilitation Hospital
150 Kilgour Road, Toronto, ON M4G 1R8
Phone: 416-425-6220
Email: mvandewouw@hollandbloorview.ca

Abstract

Internalizing problems are associated with a wide range of adverse outcomes. While we have some understanding about risk factors (e.g., neurodevelopmental conditions), biological markers are not well understood. Here, we used deep learning to predict cross-sectional (*N*=14,523) and worsening longitudinal trajectories (*N*=10,540) of internalizing problems from measures of brain structure. A stratified cross-validation scheme was used, and performance was evaluated using the area under the receiving operating characteristic curve (AUC). The cross-sectional model performed well across the sample, reaching an AUC of 0.80 [95% CI: 0.71, 0.88]. For the longitudinal model, while performance was sub-optimal for predicting worsening trajectories in a sample of the general population (AUC=0.66 [0.65, 0.67]), good performance was reached across individuals with a neurodevelopmental condition (AUC=0.73 [0.70, 0.76]). Deep learning with features of brain structure is a promising avenue for biomarkers of internalizing problems, particularly for individuals who have a higher likelihood of experiencing difficulties.

Introduction

Mental health problems in children and youth are widespread, with 11% estimated to have a mental health disorder[1]. Internalizing problems, specifically, which include feelings of anxiety, depression, and social withdrawal, are associated with profoundly negative outcomes, including poorer quality of life and worse long-term social and economic outcomes[2,3]. This highlights the need for biological markers that help us understand who is most at risk for developing problems in this domain, taking a step towards improving our ability to deliver proactive care that can effectively minimize adverse outcomes. However, these markers have thus far remained elusive.

Neurobiological markers are a promising approach to bridging the gap between wide ranges of genetic and environmental variation and subsequent manifestations of internalizing problems, and brain structure is particularly appealing given its high reliability[4]. While there have been investigations into the structural correlates of internalizing problems, such as cortical thickness, cortical surface area, and cortical and subcortical volume, findings are mixed[5–10]. Thus far, this work has largely relied on methodological approaches that may be insufficient for characterizing the multi-dimensional and complex brain-behaviour associations related to mental health. For example, regression has typically been used, relying on manual feature selection, restricting brain-behaviour relationships to being linear, and modeling each brain region independently. This calls for complementary strategies which add depth to the existing work, advancing our understanding of the brain-behaviour associations of internalizing problems.

Here, we used deep learning to predict internalizing problems in youth from neurobiological measures. We drew from four multi-national, independently collected datasets of children and youth from Canada and the United States. We first used cross-sectional measures of brain structure to predict the presence of clinically significant internalizing problems (N=14,523), to provide insight into the underlying brain features beyond what can be revealed by traditional approaches. Next, we used longitudinal data (N=10,540) to examine whether brain structure can predict subsequent worsening trajectories of internalizing problems. The cohorts include both neurotypical (NT)

children and youth and those with neurodevelopmental (ND) conditions. Given the higher prevalence of internalizing problems in the ND compared to NT population[11], we also provide measures of model performance stratified by the presence or absence of any ND diagnosis alongside overall performance.

Methods

Participants

Data were extracted from four independent datasets: the Adolescent Brain Cognitive Development (ABCD[12]) study (Release 5.1), the Healthy Brain Network (HBN[13]; Release 10), the Human Connectome Project Development (HCP-D[14]) study (Release 2.0), and the Province of Ontario Neurodevelopmental (POND) network (exported January 2024). The ABCD and POND datasets contain longitudinal data, while HBN and HCP-D provide only cross-sectional data. The recruitment strategies, assent/consent procedures, institutional review board approvals, and data usage agreements for the datasets are provided in the **Supplementary material**. Participants were eligible for inclusion based on the availability of T1-weighted structural magnetic resonance images (MRIs) and the Child Behaviour Checklist (CBCL[15]), resulting in at least one datapoint from 14,950 participants (ABCD: 11,796, HBN: 1,959, HCP-D: 498, POND: 697). Flowcharts of the study samples are presented in **Supplemental Figure 1**. The Holland Bloorview Kids Rehabilitation Hospital's research ethics board approved the current study.

We acknowledge that individuals have different language preferences for referring to neurodevelopmental disorders/conditions and we honor such perspectives. Given that in this study we defined group membership according to the diagnostic criteria in the DSM-5, we have used this nomenclature in the manuscript, and in partnership with our family and youth advisories (represented by NB and AA).

Data

Biological sex, race/ethnicity, annual household income, and highest household education were collected as part of all four study protocols (see ***Sociodemographic assessments*** in the **Supplementary material**).

Internalizing problems were measured using the CBCL[15], a caregiver-reported assessment of problem behaviour in children and adolescents. The assessment derives syndromic scales corresponding to behavioural problems, including anxious/depressed, withdrawn/depressed, and somatic complaints, which are used to calculate the higher-order construct of internalizing problems. Raw scores were converted to age- and sex-normed standardized *T*-scores, with higher scores indicating greater problems. For the cross-sectional analysis, the provided threshold (*T*-score ≥64) was used to categorize individuals as having clinically significant internalizing problems. For the longitudinal analysis, the reliable change index (RCI[16]) was used to categorize individuals as having worsening (RCI>1.96) or non-worsening (RCI≤1.96) trajectories of internalizing problems (see the ***Reliable change index*** section in the **Supplementary material**).

T1-weighted images were acquired as part of all four study protocols (**Imaging protocols** in the **Supplementary material**). The FreeSurfer image analysis suite[17] was used to perform cortical reconstruction and volume segmentation of the structural MRIs (see the ***Imaging preprocessing*** section of the **Supplementary material**), which provided measures of cortical thickness, surface area, and volume, alongside volumes of subcortical structures, ventricles, white matter, brainstem, cerebellum, and whole-brain. Processed FreeSurfer outputs were provided by ABCD and HCP-D, while the pipeline was run in-house for POND and HBN. Datapoints with poor data quality were excluded from all subsequent analyses (see **Supplementary material**, section ***Quality control***).

After quality control, for the cross-sectional analysis, one datapoint per participant was selected, resulting in a sample size of 14,523 children and youth (ABCD: 11,632, HBN: 1,785, HCP-D: 498, POND: 608). For the longitudinal analysis, partial data were available for a third timepoint in the ABCD study, which was leveraged to maximize the total number of included datapoints, resulting in 10,532 datapoints (ABCD: 10,491, POND: 49) from 8,103 participants (ABCD: 8,062, POND: 49) were included. See the ***Datapoint selection*** section in the **Supplementary material** for further information on how datapoints were selected for both analyses.

Deep learning models

The primary goal of the cross-sectional deep learning model was to establish the features of brain structure that contribute to clinically significant internalizing problems. To do so, a binary classification model was built to predict internalizing problems from features of brain structure, and the contribution of each feature to the model's output was examined. The model was tuned, trained, and tested using data aggregated across all four datasets. The model architecture was a multilayer perceptron, consisting of an input layer, blocks of hidden layers, and an output layer (see ***Deep learning model architecture*** in the **Supplementary material** for further details). Input features included the measures of brain structure (*N*=241) along with age and sex, given their association with internalizing problems[18]; all features were *z*-scored prior to building the model. A nested 5-fold stratified cross-validation scheme was used to tune, train, and test the model (**Supplemental Figure 2**). For each iteration, the dataset was split into a training set (80%), which was used for hyperparameter tuning and model training, and 20% was withheld for testing purposes. Prediction scores (probabilities of the positive class) and feature importances (Shapley Additive exPlanations[19] (SHAP) values) were obtained for the test set. To ensure generalizability, this was performed across five folds.

The primary goal of the longitudinal deep learning model was to predict whether an individual would experience worsening internalizing problems from baseline measures of brain structure. The model architecture and tuning, training, and testing procedure was identical to the cross-sectional model, this time using baseline age, sex, internalizing problems, and between-timepoint age differences as features alongside the baseline measures of brain structure. Given the sample size imbalance between ABCD and POND, tuning, training, and testing were performed using the ABCD dataset, holding out POND as an external testing set.

Performance evaluation

The primary measure used to evaluate model performance was the area under the receiving operating characteristic (ROC) curve (AUC). The secondary measure was accuracy, computed as the prediction score cut-off that maximized the average recall of both classes. Outcome measures were computed across all samples, as well as

stratified by the presence or absence of any ND diagnosis. Feature importances (SHAP values) were used to measure each feature's contribution to the model's prediction.

Statistical analysis

AUCs and accuracies are reported across folds using the mean and 95% confidence intervals (CIs). Bias with respect to sociodemographic factors (sex, age, race/ethnicity, annual household income, and highest household education) was assessed by examining differences in AUC. Overall feature importances were computed across folds by taking the mean of the magnitudes of the SHAP values across all datapoints.

Results

Samples

Data from 14,523 children and youth between 5 – 21 years of age were included for the cross-sectional analysis; participant demographics are presented in **Table 1**. Across all four datasets, 1,967 [14%] participants had clinically significant internalizing problems, and internalizing problems were more prevalent in individuals with an ND diagnosis (31%) compared to those without (8%).

Table 1: Participant demographics for the cross-sectional analysis.

		***N* [%]**			
		ABCD	**HBN**	**HCP-D**	**POND**
N		11,632	1,785	608	498
ND diagnosis[1]		1,650 [14]	1,310 [74]	0 [0]	408 [67]
Internalizing problems[2]	Total sample	1,129 [10]	560 [31]	30 [6]	248 [41]
	No ND diagnosis	691 [7]	130 [28]	30 [6]	54 [27]
	ND diagnosis	435 [26]	428 [33]	–	194 [48]
Age range (years)		8 – 15	5 – 21	6 – 17	5 – 19
Median age (years) [IQR]		10.3 [9.5, 11.4]	10.1 [8.2, 13.1]	13.0 [10.2, 14.9]	11.7 [9.7, 14.4]
Sex	Male	6,051 [52]	1,143 [64]	229 [46]	430 [71]
	Female	5,581 [48]	642 [36]	269 [54]	178 [29]
Race and ethnicity[3]	Hispanic/Latino	2,368 [20]	301 [17]	40 [8]	23 [4]
	Non-Hispanic Asian	245 [2]	50 [3]	23 [5]	29 [5]
	Non-Hispanic Black	1,719 [15]	191 [11]	33 [7]	8 [1]
	Non-Hispanic White	6,077 [52]	750 [42]	308 [62]	293 [48]
	Multi-Racial/Other	1,222 [11]	275 [15]	89 [18]	86 [14]
	Unknown/Not reported	1 [0]	218 [12]	5 [1]	169 [28]
Annual household income[4]	<$50,000	3,002 [26]	287 [16]	61 [12]	81 [13]
	$50,000-$99,999	2,924 [25]	322 [18]	105 [21]	83 [14]
	≥$100,000	4,705 [40]	719 [40]	298 [60]	197 [32]
	Unknown/Not reported	1,001 [9]	457 [26]	34 [7]	247 [41]
Highest household education[5]	Below high school	429 [4]	45 [3]	6 [1]	2 [0]
	High school or equivalent	1,930 [17]	285 [16]	49 [10]	41 [7]
	Undergraduate	3,213 [28]	570 [32]	182 [37]	236 [39]
	Graduate	2,849 [24]	854 [48]	261 [52]	121 [20]
	Unknown/Not reported	3,211 [28]	31 [2]	0 [0]	208 [34]

ABCD: Adolescent Brain Cognitive Development study; HBN: Healthy Brain Network; HCP-D: Human Connectome Project Development study; POND: Province of Ontario Neurodevelopmental network; IQR: interquartile range; [1]According to the DSM-5 categories; data was not available for 519 (4%) of participants; [2]Determined using a clinically significant cutoff of ≥64; [3]Reported according to recommendations of the ABCD study[20]; [4]In local currency (CAD or USD); [5]High school or equivalent includes: high school graduate, General Education Development (GED) diploma or equivalent, less than one year of college credit/post-secondary education (or less than 10 classes), or one year or more of college credit without a degree; Graduate includes: master's degree, professional School degree, or doctoral degree

For the longitudinal analysis, 10,540 datapoints were included (**Table 2**), 2,029 [19%] of which had worsening changes in internalizing problems between the baseline and follow-up timepoints. The prevalence rates of worsening trajectories in individuals with (17%) and without (20%) an ND were similar.

Table 2: Participant demographics for the longitudinal analysis.

		***N* [%]**	
		ABCD	**POND**
N	Unique participants	8,062	49
	Datapoints	10,491	49
ND diagnosis[1]		1,522 [15]	32 [65]
Worsening internalizing problems[2]	Total sample	2,019 [19]	10 [20]
	No ND diagnosis	1,756 [20]	5 [29]
	ND diagnosis	259 [17]	5 [16]
Age range (years)	Baseline	8 – 13	7 – 15
	Follow-up	10 – 15	10 – 19
Median age (years) [IQR]	Baseline	10.3 [9.6, 10.9]	11.7 [9.8, 12.8]
	Follow-up	12.3 [11.6, 13.3]	15.2 [12.7, 16.8]
Sex	Male	5,584 [53]	35 [71]
	Female	4,907 [47]	14 [29]
Race and ethnicity[3]	Hispanic/Latino	2,038 [19]	1 [2]
	Non-Hispanic Asian	206 [2]	0 [0]
	Non-Hispanic Black	1,312 [13]	0 [0]
	Non-Hispanic White	5,850 [56]	33 [67]
	Multi-Racial/Other	1,085 [10]	9 [18]
	Unknown/Not reported	0 [0]	6 [12]
Annual household income[4]	<$50,000	2,616 [25]	10 [20]
	$50,000-$99,999	2,886 [28]	12 [24]
	≥$100,000	4,222 [32]	18 [37]
	Unknown/Not reported	767 [7]	9 [18]
Highest household education[5]	Below high school	355 [3]	0 [0]
	High school or equivalent	1,649 [16]	5 [10]
	Undergraduate	3,128 [30]	25 [51]
	Graduate	2,799 [27]	14 [29]
	Unknown/Not reported	2,560 [24]	5 [10]

ABCD: Adolescent Brain Cognitive Development study; POND: Province of Ontario Neurodevelopmental network; IQR: interquartile range; [1]Basline data, according to the DSM-5 categories; data was not available for 0.2% of samples; [2]Determined using the reliable change index (RCI); [3]Baseline data, reported according to recommendations of the ABCD study[20]; [4]Baseline data in local currency (CAD or USD); [5]Baseline data; High school or equivalent includes: high school graduate, General Education Development (GED) diploma or equivalent, less than one year of college credit/post-secondary education (or less than 10 classes), or one year or more of college credit without a degree; Graduate includes: master's degree, professional School degree, or doctoral degree

Predicting cross-sectional internalizing problems

The ROC curves across the entire sample and disaggregated by ND diagnosis are shown in **Figure 1A**. Across the cross-validation folds, the overall AUC was 0.80 (95% CI: [0.71, 0.88]), while the AUC for individuals with and without a diagnosis of an ND was 0.71 [0.60, 0.82] and 0.78 [0.68, 0.89], respectively. At the optimal prediction score cut-off (0.49 [0.46, 0.52]), an accuracy of 77% [68, 86] was achieved (ND: 62% [50, 75]; no ND: 82% [74, 90]).

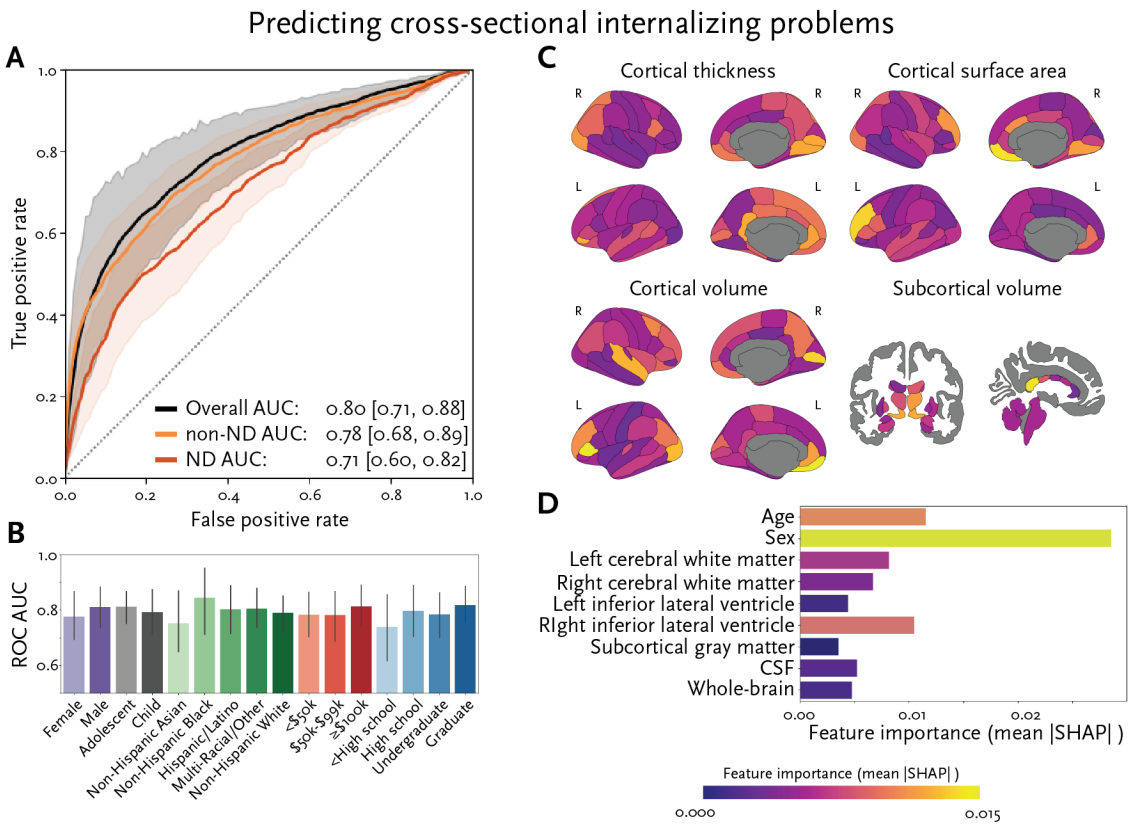


Figure 1: Performance and feature importances for the model predicting cross-sectional internalizing problems. (A) The mean [95% CI] receiving operating characteristic curve and corresponding area under the curve (AUC) across all data (black) and stratified by individuals with (red) and without (orange) an ND diagnosis. (B) AUCs in different sociodemographic groups (age, sex, race/ethnicity, annual household income, and highest household education level). (C) Feature importances (mean absolute SHAP values) for the cortical and subcortical features used in the prediction model. (D) Feature importances (mean absolute SHAP values) for the sociodemographic and global brain features.

AUCs were also compared amongst sociodemographic groups (age, sex, race/ethnicity, annual household income, and highest household education) to evaluate fairness (**Figure 1B**); no significant differences were observed (**Table 3**).

Table 3: Statistics comparing AUC amongst sociodemographic groups for the cross-sectional prediction model.

	F-statistic	*p*-value	η^2
ND	0.86	.380	.10
Sex	0.27	.615	.03
Age[1]	0.11	.754	.01
Race/ethnicity[2]	0.39	.813	.07
Annual household income[3]	0.14	.873	.02
Highest household education[4]	0.42	.743	.07

ND: neurodevelopmental condition; [1]Children (<13 years of age) versus adolescents (≥13 years of age); [2]Non-Hispanic Asian, Non-Hispanic Black, Hispanic/Latino, Multi-Racial/Other, Non-Hispanic White; [3]<$50,000, $50,000 - $99,999, ≥$100,000; [4]<High school, High school, Undergraduate, Graduate

Feature importances were calculated using the magnitude of SHAP values (**Figure 1C** and **D**). The most important features were sex (with females having a negative impact on prediction scores), the area and thickness of the right frontal pole, the areas of the left temporal pole, left rostral middle frontal, and right medial orbitofrontal gyri, and the volumes of the left medial orbitofrontal gyrus, left pars triangularis, right pericalcarine cortex, and posterior corpus callosum.

Predicting worsening trajectories of internalizing problems

The ROC curves across the entire sample and disaggregated by ND diagnosis are shown in **Figure 2A**. Across the cross-validation folds, the AUC was 0.66 (95% CI: [0.65, 0.67]) for ABCD, while the AUC for POND was 0.80 [0.78, 0.81]. Stratified by ND diagnosis across both datasets, the AUCs for individuals with and without an ND were 0.73 (95% CI: [0.70, 0.76]) and 0.65 (95% CI: [0.65, 0.67]), respectively. At the optimal prediction score cut-off (0.53 [0.50, 0.55]), an accuracy of 71% [65%, 77%] was achieved (ABCD: 63% [58, 69]; POND: 79% [76, 73]), and higher accuracy was achieved in those with (73% [70, 76]) an ND compared to those without (65% [64, 66]).

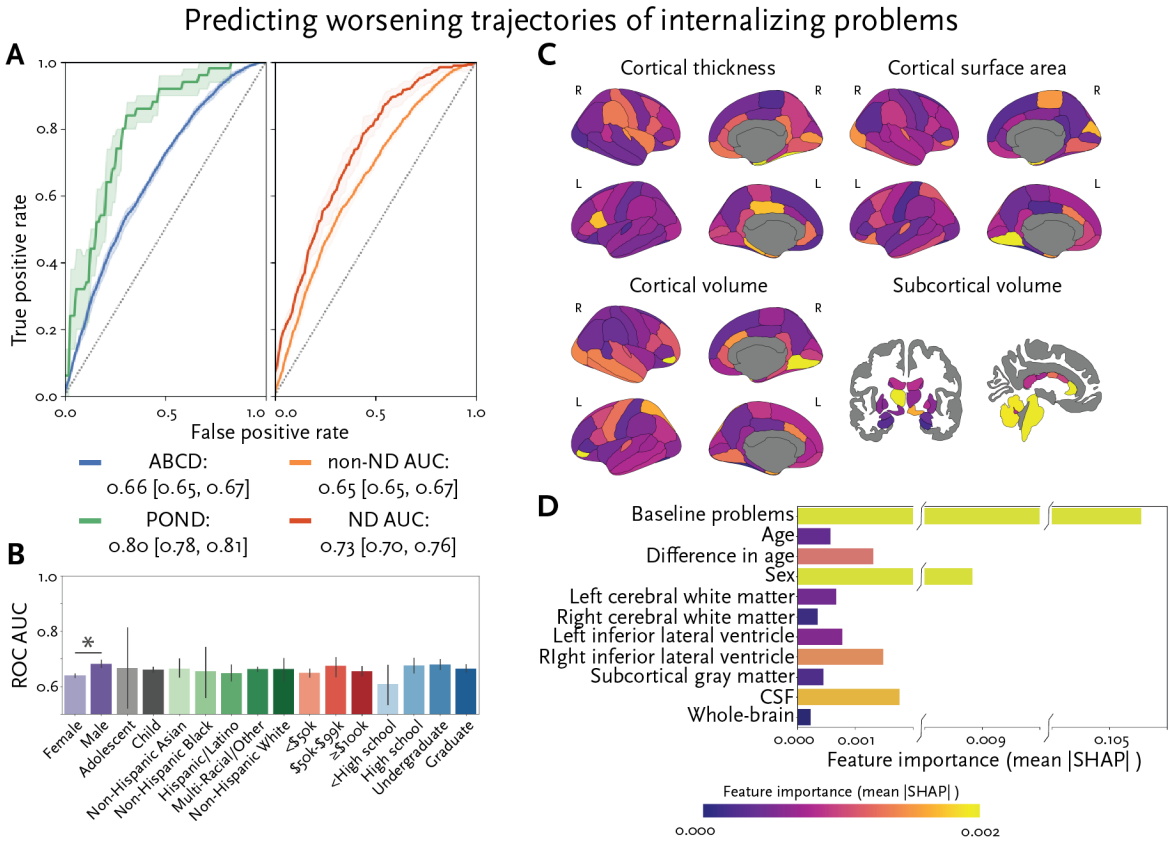


Figure 2: Performance and feature importances for the model predicting worsening trajectories of internalizing problems. (A) The mean [95% CI] receiving operating characteristic curve and corresponding area under the curve (AUC) across all data from ABCD (left; blue) and POND (left; green) and stratified by individuals with (right; red) and without (right; orange) an ND diagnosis. (B) AUCs in different sociodemographic groups (age, sex, race/ethnicity, annual household income, and highest household education level). (C) Feature importances (mean absolute SHAP values) for the cortical and subcortical features used in the prediction model. (D) Feature importances (mean absolute SHAP values) for the sociodemographic and global brain features.

AUCs were also compared amongst sociodemographic groups (age, sex, race/ethnicity, annual household income, and highest household education) to evaluate fairness (**Figure 2B**); a significant difference in sex was observed ($F(1,8)=24.9$, $p=.001$, $\eta^2=.76$), with males having higher AUC compared to females (**Table 4**).

Table 4: Statistics comparing AUC amongst sociodemographic groups for the longitudinal prediction model.

	F-statistic	*p*-value	η^2
ND	26.7	.001	.77
Sex	24.9	.001	.76
Age[1]	0.01	.943	.00
Race/ethnicity[2]	0.07	.990	.01
Annual household income[3]	1.03	.387	.15
Highest household education[4]	2.10	.141	.28

ND: neurodevelopmental condition; [1]Children (<13 years of age) versus adolescents (≥13 years of age); [2]Non-Hispanic Asian, Non-Hispanic Black, Hispanic/Latino, Multi-Racial/Other, Non-Hispanic White; [3]<$50,000, $50,000 - $99,999, ≥$100,000; [4]<High school, High school, Undergraduate, Graduate

Feature importances were calculated using the magnitude of SHAP values (**Figure 2C** and **D**). The most important features were the baseline score of internalizing problems (with fewer problems having a positive impact on prediction scores), sex (with females having a positive impact), the thicknesses of the right entorhinal cortex and fusiform, the areas of the right temporal pole and left lingual gyrus, and the volumes of the left thalamus, right lingual gyrus, anterior corpus callosum, and brainstem.

Discussion

This study examined the utility of using deep learning to predict mental health problems from measures of brain structure. Cross-sectionally, good performance (mean AUC=0.80) was achieved across a large sample (*N*=14,523) of children and youth, providing insight into the neurobiology underlying internalizing problems beyond what could be revealed by traditional regression. Longitudinally (*N*=10,540), performance was sub-optimal for predicting worsening trajectories of internalizing problems in children and youth from the general population (mean AUC=0.66). However, the model performed well for an external testing set comprised primarily of youth with an ND (mean AUC=0.80), and when only considering individuals across the whole sample with an ND diagnosis (mean AUC=0.73). These results indicate that deep learning with measures of neurobiology could potentially improve our ability to detect which individuals, especially which neurodivergent individuals, are at risk of developing mental health difficulties, allowing for timely interventions to help mitigate the rising burden on youth, families, and healthcare systems around the world.

Deep learning was able to predict internalizing problems from cross-sectional measures of neurobiology with good performance (mean AUC=0.80). While cross-

206 sectional prediction does not have direct clinical relevance, it does allow for the
207 comparison of the utility of using symptom measures, rather than diagnostic labels, as
208 prediction targets. As the popularity of artificial intelligence has grown, there has been
209 substantial work on its application to predicting ND conditions, such as autism and
210 attention-deficit/hyperactivity disorder, from measures of brain structure. However,
211 despite having distinct behaviour-based diagnostic criteria, these ND conditions are
212 highly heterogeneous in neurobiology and phenotype, and overlap with each other, and
213 even typical development [21,22]. This has posed two main challenges for the existing
214 literature. First, for clinical translation, the observed heterogeneity necessitates larger
215 sample sizes to capture the full spectrum of variability. Yet, most studies have used
216 small sample sizes[23]. While the diagnostic prediction accuracies of these underpowered
217 studies have been high when predicting ND diagnosis versus typical development,
218 accuracy was shown to decrease with sample size, with the best powered studies only
219 achieving an accuracy of just over 60%[23]. Second, considering the heterogeneity and
220 overlap, there is evidence that the diagnostic labels do not align with the underlying
221 neurobiology [21,22], which fundamentally limits work attempting to predict diagnosis from
222 neurobiology. Here, we have instead predicted mental health symptoms experienced
223 across neurotypical and neurodivergent children and youth from measures of brain
224 structure. Across this sample, we have achieved an accuracy of 77%, a substantial
225 improvement over the existing literature, supporting the use of transdiagnostic symptom
226 measures as prediction targets.

227 Comparatively, the deep learning model trained to predict worsening trajectories
228 of internalizing problems did not perform well for a sample of children and youth
229 representative of the general population (ABCD; mean AUC=0.66). However, when
230 applied to a small withheld external testing set comprised primarily of children with an
231 ND diagnosis (POND; *N*=49), performance improved considerably (mean AUC=0.80).
232 While this external sample was small, this provides some indication that the model
233 holds value for predicting worsening symptoms in individuals at high-risk of mental
234 health symptoms. When examining performance across the individuals with an ND
235 diagnosis across the entire sample, performance improved considerably (mean
236 AUC=0.73). This improvement cannot be explained by an increased representation of

worsening changes in the ND group, as prevalence rates were comparable in individuals with (17%) and without (20%) an ND diagnosis. We hypothesize that this suggests that biological differences associated with NDs more reliably predict subsequent mental health difficulties, potentially suggesting that mental health challenges may be part of the constellation of biological atypicality seen in NDs. In addition, there may be more heterogeneity in the biological correlates of mental health trajectories in NDs; while feature variability can make initial learning challenging for a deep learning model, it can eventually lead to more general and robust performance.

The cross-sectional model is also able to provide insight into the importance of individual neurobiological features for predicting internalizing problems. Nearly all of the most important features resided in the prefrontal cortex and included the medial orbitofrontal cortex, along with the temporal poles and corpus callosum. The role of these brain structures in internalizing disorders has been consistently documented[24–26], reflecting their roles in emotion processing, regulation, and awareness. However, while the deep learning model relied on features already known to be associated with internalizing problems to make predictions, we have provided further insight into the multivariate complexities in these associations. For example, the importance of the lateral prefrontal cortex was specific to its surface area and volume, not thickness, and the patterns of importances throughout the brain were largely hemisphere dependent. Traditional approaches, like regression, typically search for linear brain-behaviour associations within each brain region and structural modality in isolation. The capability of deep learning models to learn from high-dimensional data allows these models to capture intricate patterns and relationships across many variables that might be challenging for traditional statistical methods. This helps advance our understanding of the neural correlates of mental health problems, which can be used to generate hypotheses for future research.

There was some overlap between features important for the cross-sectional and worsening models, for example the temporal pole and corpus callosum; however, the patterns of feature importances were strikingly different between the two models. This finding suggests that there are neuroanatomical patterns that influence internalizing

problems over time, and these patterns are distinct from those that arise from the presence of clinically significant problems. This finding suggests that individuals who experience worsening internalizing problems have neuroanatomical differences that predate changes in mental health, and these patterns are different from those that are present in association with the onset of clinically significant problems. In addition to brain features, baseline levels of internalizing problems were also important for predicting worsening trajectories of internalizing problems. Caution should be used in interpreting this finding through a clinical lens, given the limited range of the CBCL and the use of the reliable change index to identify worsening trajectories. Thus, this likely reflects the fact that individuals with very few baseline problems are more likely to exhibit worsening trajectories than individuals with problems already at the upper limit of the scale.

Sex was also one of the most import predictive features in both the cross-sectional and worsening models. In the cross-sectional model, the feature importances indicated that being male positively impacted the predicted probability of having internalizing problems, while being female had a negative impact. While internalizing problems have classically been reported as increased in females compared to males[27], more recent work has reported no sex differences in both neurotypical and neurodivergent populations[28]. Considering this work, we hypothesize that positive predictive impact of males in the cross-sectional model may reflect the male preponderance in neurodivergence[18], which is in turn associated with increased internalizing problems[11]. On the other hand, in the longitudinal model, being female positively impacted the probability of predicted worsening internalizing problems, while being male had a negative impact. This is in line with evidence showing that both neurotypical and neurodivergent females are more likely to exhibit worsening longitudinal trajectories of internalizing problems across adolescence[29]. Together, the present study supports the importance of sex as an influencing factor on internalizing problems. Importantly, this study only considered biological sex, and not gender. There is some evidence of differences in internalizing problems between gender identities (e.g., [30]), although work in this domain has been limited; there is also evidence that sex differences in internalizing problems may be caused, in part, by gender socialization,

rather than biological sex[31]. While considering biological sex as a feature alongside neurobiology in the deep learning models, rather than regressing out its effects, is a step towards addressing its contribution to mental health; future work should consider measures of gender alongside measures of biological sex.

The present study provides data demonstrating that AI approaches using biological markers can be useful both at the cross-sectional level to uncover complex relationships between biology and phenotype that are not typically amenable to traditional statistical methods, as well as for predicting emerging clinically important phenotypes longitudinally. As such it is not an exhaustive search for brain-behaviour relationships during development in mental health. We only considered measures of brain structure and limited sociodemographic measures (age and sex) as predictors of cross-sectional and longitudinal internalizing problems. Other neurobiological measures, such as brain function and structural connectivity, alongside genetic, environmental, and broader sociodemographic variables, likely also hold predictive value. We also only examined internalizing problems as a prediction target, which are just one aspect of mental health; examining the utility of predicting sub-scores of internalizing problems (e.g., anxious/depressed, withdrawn/depressed, and somatic complaints), externalizing problems, or both problem domains, is an important next step. Furthermore, we measured internalizing problems using a caregiver report, which have been associated with response biases. Other limitations include that, due to the use of the ABCD dataset, the bulk of the sample's age is in the early adolescent years; this is particularly true for the longitudinal analysis, where two of the four datasets (HBN, HCP-D) cannot be used. Future work should include more younger children and older youth to ensure generalizability across development. We also did not consider the onset of the COVID-19 pandemic.

In conclusion, this study demonstrated that deep learning could be used to predict internalizing problems cross-sectionally, and performance exceeded comparable studies that aimed to predict ND diagnosis. We also used baseline measures of brain structure to predict worsening trajectories of internalizing problems, showing that acceptable performance could be reached for children and youth with an ND. This

suggests that neurobiology may be a promising avenue to pursue in the search for a biomarker for mental health problems in individuals who are most at risk, which can be used to deliver proactive care that ultimately prevents adverse outcomes. The study also provides further evidence for the promise of responsible AI methods for advancing personalized approaches to health care.

Data availability

Access to the data used in this study is controlled by each of the datasets (ABCD, HBN, HCP-D, and POND). Details on the data usage agreements can be found in the **Supplementary material**.

Acknowledgements

Funding for the current study was provided by the Canadian Institutes of Health Research and New Frontiers in Research Fund.

ABCD: Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development^{SM} (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from DOI: 10.15154/z563-zd24. DOIs can be found at https://nda.nih.gov/abcd.

HBN: This manuscript was prepared using a limited access dataset obtained from the Child Mind Institute Biobank, the Healthy Brain Network (HBN). This manuscript reflects the views of the authors and does not necessarily reflect the opinions or views of the Child Mind Institute.

HCP-D: Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number U01MH109589 and by funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. The HCP-Development 2.0 Release data used in this report came from DOI: 10.15154/1520708.

POND: This research was conducted with the support of the Ontario Brain Institute (POND, PIs: Anagnostou/Lerch), an independent non-profit corporation, funded partially by the Ontario government. The opinions, results and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred.

Author contributions

Conceptualization: MMV, AK; Data curation: MMV, BS; Formal analysis: MMV; Funding acquisition: MMV, JPL, EA, AK; Investigation: MMV; Methodology: MMV; Project administration: AK; Resources: EK, JJ, MA, PDA, JC, RJS, MJT, JPL, EA, AK; Software: MMV; Supervision: AK; Validation: MMV; Visualization: MMV; Writing – original draft: MMV; Writing – review & editing: MMV, BS, NB, AA, EK, JJ, MA, AI, PDA, JC, RJS, MJT, JPL, EA, AK

Competing interests

AK has a patent for holly™ (formerly Anxiety Meter) with royalties paid from Awake Labs. AK has received consulting fees from DNAStack and Shaftesbury. EA has received grants from Roche and Anavex, served as a consultant to Roche, Quadrant Therapeutics, Ono, and Impel Pharmaceuticals, has received in-kind support from AMO Pharma and CRA-Simons Foundation, received royalties from APPI and Springer, received an editorial honorarium from Wiley, and has a patent for holly™ (formerly Anxiety Meter). The remaining authors have no potential conflicts of interest to report.

Supplementary material

Recruitment strategies

The ABCD study is a population-based cohort study that aimed to recruit and follow 11,500 children that were representative of the population of the United States; full details on recruitment are described elsewhere[32]. Briefly, children between 9-10 years of age were recruited through public and private elementary schools to 21 nationally distributed sites, whose catchment areas contained one in every five eligible children in the US. To minimize bias and ensure a representative sample, school selection was informed by sociodemographic factors and recruitment was dynamically monitored. After their initial recruitment, participants were asked to return every two years for imaging and other assessments. Children were excluded for a lack of English fluency, neuroimaging contraindications, a major neurological disorder, <28 weeks gestational age, <1,200g birth weight, >1 month hospitalization due to birth complications, uncorrected sensory impairments, a current diagnosis of schizophrenia, autism, intellectual disability, or alcohol/substance use disorder, a history of traumatic brain injury, or an unwillingness to complete the testing protocol[33,34].

The objective of HBN was to create a large transdiagnostic dataset that captures a broad range of clinical developmental psychopathology[13]. A community-referred recruitment strategy was used, where advertisements that targeted families with concerns about their child's psychiatric well-being were distributed to community members, educators, and care providers in the New York City area of the United States. Children and youth were eligible for HBN if they were between 5 – 21 years of age, were fluent in English, could provide verbal assent, and whose parent/caregiver could provide informed consent. Exclusion criteria included the presence of a debilitating neurological disorder that prevented participation in the full protocol, injury or disease-related acute encephalopathy, a known neurodegenerative disorder, uncorrected sensory impairments, a diagnosis of schizophrenia, schizoaffective disorder, or bipolar disorder given in the past six months, currently untreated suicidality or homicidality with onset within the past three months, a history of substance dependence that requires chemical replacement therapy, and intoxication at any study visit.

418 The HCP-D study is a sub-study of the Human Connectome Project, which aims
419 to characterize the development of brain connectivity[14]. Children and adolescents
420 between 5 – 21 years of age were recruited from four cities in the United States
421 (Boston, Los Angeles, Minneapolis, and St. Louis). Participants were excluded based on
422 insufficient English fluency, premature birth, a lifetime history of serious medical or
423 endocrine conditions, treatment for a serious medical problem, treatment for more than
424 one year by a specialist for certain disorders, head injury, hospitalization for at least two
425 days for certain conditions, receiving special education services at school, and
426 contraindications for neuroimaging.

427 The objective of the POND study is to better understand the biological
428 underpinnings of neurodevelopmental conditions and consists of children and
429 adolescents with and without these conditions. Participants are recruited from clinical
430 research sites in four cities on Ontario, Canada, with three institutions collecting
431 neuroimaging data. POND contains a longitudinal component, where participants were
432 asked to return at least two years after their first visit. Participants were recruited if they
433 had a clinical diagnosis of a neurodevelopmental condition, including autism, attention-
434 deficit/hyperactivity disorder, and obsessive-compulsive disorder or typically developing
435 controls. Typically developing participants were recruited using flyers posted in the
436 community and word-of-mouth and were excluded on the basis of a diagnosed
437 neurodevelopmental, psychiatry, or neurological condition, a first-degree family member
438 with a diagnosed neurodevelopmental condition, or premature birth. All participants
439 were screened for contraindications for neuroimaging and sufficient English for consent
440 and the testing protocols.

441 ***Assent/consent procedures***

442 For the ABCD study, parent(s)/caregiver(s) provided written informed consent for
443 their child's participation, while written assent was obtained from the participant[35]. For
444 HBN, participants 18 or older provided written informed consent, while those under 18
445 provided verbal assent and written informed consent was obtained by a
446 parent/caregiver[13]. For the HCP-D study, parent(s)/caregiver(s) provided written
447 informed consent for their child's participation. For the POND study, written informed

448 consent was obtained from participants who had the capacity to do so; otherwise,
449 written consent was obtained from a parent(s)/caregiver(s), and verbal assent was
450 obtained from the participant, as per institutional ethics board SOPs.

451 ***Institutional review board approvals***

452 The majority of ABCD's 21 sites used a centralized Institutional Review Board at
453 the University of California, San Diego which reviewed and approved the study's
454 protocol, with some sites having used a local Institutional Review Board[36]. The study
455 protocol for HBN was reviewed and approved by the Chesapeake Institutional Review
456 Board. For the HCP-D study, a centralized Institutional Review Board at Washington
457 University, St. Louis reviewed and approved the study's protocol. For the POND study,
458 each site's Institutional Review Board approved the study protocol.

459 ***Data usage agreements***

460 Access to the ABCD and HCP-D data was obtained under the NIMH Data
461 Archive Data Use Agreement #15619 and #18618 for the Adolescent Brain Cognitive
462 Development (ABCD)/Connectome Coordination Facility (CCF) permission group.
463 Access to the HBN phenotypic data was granted after approval of a data usage
464 agreement awarded to A. Kushki; an agreement was not required to access the imaging
465 data. For the POND data, access is available as the study team reflected in the
466 authorship line is also part of the POND study team, according to internal procedures.

467 ***Sociodemographic assessments***

468 Biological sex was collected as part of all four study protocols by asking
469 parent(s)/caregiver(s) to specify what sex the child was assigned at birth. For ABCD,
470 options included male, female, intersex male, and intersex female. For HCP-D, male,
471 female and other were options. For POND and HBN, only male and female were
472 options.

473 Race and ethnicity data was collected as part of all study protocols. The ABCD
474 study provides the data pre-categorized into five recommended distinct racial/ethnic
475 categories (Hispanic/Latino, non-Hispanic Asian, non-Hispanic Black, non-Hispanic
476 White, and Multi-Racial/Other)[20], after asking the parent(s)/caregiver(s) to specify what
477 race they consider their child to be. The recommended procedure was followed to

categorize the raw race/ethnicity data from the remaining studies (described below) into these five categories.

In HBN, parent(s)/caregiver(s) are asked to select one category for their child's ethnicity (not Hispanic or Latino, Hispanic or Latino, decline to specify, or unknown) and race (White/Caucasian, Black/African American, Hispanic, Asian, Indian, Native American Indian, American Indian/Alaskan Native, Native Hawaiian/Other Pacific Islander, two or more races, or other race). In HCP-D, parent(s)/caregiver(s) or participants were asked to specify their ethnicity (Hispanic or Latino, Not Hispanic or Latino, or unknown/not reported), and asked what race they considered their child/themselves to be (American Indian/Alaska Native, Asian, Hawaiian or Pacific Islander, Black or African American, White, more than one race, and/or unknown or not reported); multiple categories were mapped to the "more than one race" category. In POND, parent(s)/caregiver(s) were asked to select all racial/ethnic groups that applied to their child (Aboriginal, Arab, Black, Chinese, East Asian, Filipino, Japanese, Jewish, Korean, Latin American/Hispanic, South Asian, Southeast Asian, West Asian, White, other, don't know, prefer not to answer).

Annual household income and highest household education were used as measures of socioeconomic status. For ABCD, HBN, and POND, parent(s)/caregiver(s) were asked to select their annual household income from several provided income bands. For HCP-D, either the participant or their parent(s)/caregiver(s) provided the numerical value for their annual household income. This raw data was used to categorize participants into low (<$50,000), medium ($50,000-$99,999) and high (≥$100,000) income households, along with unknown/not reported. Note that although POND is a Canadian study, income levels were not adjusted for the exchange rate. The highest education level achieved by one or two parent(s)/caregiver(s), where appropriate, was collected in all four datasets by asking responders to select from a provided list. This raw data was used to categorize the education levels for each parent/caregiver into several ordinal categories: below high school, high school or equivalent, Associate/Bachelor's degree, and unknown/not reported. High school or equivalent includes high school graduate, general education development (GED)

diploma or equivalent, and any college education that did not result in a degree. Graduate degree includes degrees from master's, doctoral, and professional school educations. When data from two caregivers was available, the highest level of education was selected.

ND diagnoses were also available for ABCD, HBN, and POND; for HCP-D, participants were excluded from the study based on the presence of a clinical diagnosis. For the purposes of this study, we defined an ND according to the diagnoses in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[37]: intellectual disabilities, communication disorders, autism spectrum disorder, attention-deficit/hyperactivity disorder, specific learning disorder, motor disorders, and other neurodevelopmental disorders. For ABCD, parents and youth completed a computerized version of the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS) for DSM-5 (K-SADS-COMP). Across both parent and youth reports, if the K-SADS indicated a diagnosis in any of the neurodevelopmental conditions (past, present, or in partial remission), the participant was labeled as having a neurodevelopmental condition. In HBN, after all assessments are complete, DSM-5 diagnoses that reach a consensus across clinicians are assigned; these diagnoses were used to categorize the participants as having a neurodevelopmental condition. In POND, diagnostic assessments and expert clinical judgement (e.g., [38–41]) are used to assign a label of typical development or assign a primary diagnosis of a neurodevelopmental and related conditions (e.g., obsessive-compulsive disorder), which was used to determine the presence of a neurodevelopmental condition.

Reliable change index

The reliable change index (RCI [16]) was used to quantify whether a change in internalizing problems between two timepoints was larger than what could have been solely attributed to measurement error. The index is computed by dividing the difference between the participant's problem scores by the standard error associated with the assessment. Mathematically:

$$\mathrm{RCI} = \frac{(x_2 - x_1)}{S_E}, \text{ where } S_E = SD\sqrt{1-r}$$

537 Here, x_1 and x_2 are the scores for the two timepoints, and SE_E is the standard error of
538 the score, which is computed using the scores standard deviation (SD=10) and
539 reliability (r=0.90[15]). This formula produces a z-score, and thus RCIs exceeding ±1.96
540 indicate a statistically significant, or reliable, change at the 5% threshold for
541 significance. This cutoff was used to categorize participants as having worsening
542 (RCI>1.96), improving (RCI<-1.96), or stable (-1.96≤RCI≤1.96) trajectories of
543 internalizing problems.

544 ***Imaging protocols***

545 T1-weighted images were obtained as part of the imaging protocols for all four
546 studies. For ABCD, data was collected from the 21 sites using three types of 3T
547 scanners, whose acquisition parameters are presented in **Supplemental Table 1**. HBN
548 collected data from four sites using one 1.5T and three 3T scanners, two of which used
549 the same acquisition parameter (**Supplemental Table 2**). All four sites that collected
550 data for HCP-D used the same Siemens 3T Prisma scanner, and the acquisition
551 parameters are presented in **Supplemental Table 3**. For the POND study, data was
552 collected using two 3T Siemens scanners at three imaging sites (**Supplemental Table**
553 **4**).

Supplemental Table 1: T1-weighted acquisition parameters for the ABCD study.

	Siemens Prisma	**Philips Achieve dStream Ingenia**	**GE Discovery**
Matrix	256×256	256×256	256×256
Slices	176	225	208
FOV	256×256	256×256	256×256
% FOV phase	100	93.75	100
Resolution (mm)	1.0×1.0×1.0	1.0×1.0×1.0	1.0×1.0×1.0
TR (ms)	2500	6.31	2500
TE (ms)	2.88	2.90	2.00
TI (ms)	1060	1060	1060
Flip angle (º)	8	8	8
Parallel imaging	2×	1.5×2.2	2×
Acquisition time (min:sec)	7:12	5:38	6:09

Supplemental Table 2: T1-weighted acquisition parameters for the HBN study.

	Siemens Avanto	Siemens Trio/Prisma
Matrix	256×256	320×320
Slices	176	224
FOV	256×256	256×256
% FOV phase	100	100
Resolution (mm)	1.0×1.0×1.0	0.8×0.8×0.8
TR (ms)	2730	2500
TE (ms)	1.64	3.15
TI (ms)	1000	1060
Flip angle (º)	7	8
Parallel imaging	2×	2×
Acquisition time (min:sec)	6:32	7:19

Supplemental Table 3: T1-weighted acquisition parameters for the HCP-D study.

	Siemens Prisma
Matrix	320×320
Slices	208
FOV	256×256
% FOV phase	93.8
Resolution (mm)	0.8×0.8×0.8
TR (ms)	2500
TE (ms)	1.81/3.6/5.39/7.18
TI (ms)	1000
Flip angle (º)	8
Parallel imaging	2
Acquisition time (min:sec)	8:22

Supplemental Table 4: T1-weighted acquisition parameters for the POND study.

	Siemens Prisma	Siemens Trio
Matrix	300×320	240×256
Slices	192	192
FOV	240×256	240×256
% FOV phase	93.75	93.75
Resolution (mm)	0.8×0.8×0.8	1.0×1.0×1.0
TR (ms)	1870	2300
TE (ms)	3.14	2.96
TI (ms)	945	945
Flip angle (º)	9	9
Parallel imaging	2×	2×
Acquisition time (min:sec)	5:01	5:01

554 ***Imaging preprocessing***

555 For all datasets, the FreeSurfer image analysis suite[17] was used to perform
556 cortical reconstruction and volume segmentation of the structural MRIs. Measures of
557 cortical thickness, cortical surface area, and cortical volume of the 68 region Desikan-
558 Killiany parcellation[42] were extracted as features from the FreeSurfer outputs, alongside
559 the volumes provided by the subcortical segmentation[43], which includes the subcortical
560 structures, ventricles, white matter, brainstem, cerebellum, and whole-brain. For the
561 ABCD study, the data was tabulated by the ABCD Data Analysis and Informatics Center

using FreeSurfer version 5.3 or version 6.0, which was modified to account for intensity scaling and inhomogeneity correction that were already applied to the T1-weighted images[44]. For HCP-D, the Lifespan HCP Release 2.0 includes data processed using the HCP Pipelines (version 4.0.0[45]) which includes a modified version of FreeSurfer (version 6.0), from which tabulated data was extracted. Data from POND and HBN was processed in-house through a FreeSurfer (version 5.3) pipeline. We include the FreeSurfer boilerplate text below for completeness.

Cortical reconstruction and volumetric segmentation was performed with the Freesurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). The technical details of these procedures are described in prior publications[43,46–58]. Briefly, this processing includes motion correction and averaging[57] of multiple volumetric T1 weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure [56], automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures[43,50] (including hippocampus, amygdala, caudate, putamen, ventricles) intensity normalization[59], tessellation of the gray matter white matter boundary, automated topology correction[49,60], and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class[46–48]. Once the cortical models are complete, a number of deformable procedures can be performed for further data processing and analysis including surface inflation[51], registration to a spherical atlas which is based on individual cortical folding patterns to match cortical geometry across subjects[52], parcellation of the cerebral cortex into units with respect to gyral and sulcal structure[42,53], and creation of a variety of surface based data including maps of curvature and sulcal depth. This method uses both intensity and continuity information from the entire three dimensional MR volume in segmentation and deformation procedures to produce representations of cortical thickness, calculated as the closest distance from the gray/white boundary to the gray/CSF boundary at each vertex on the tessellated surface[48]. The maps are created using spatial intensity gradients across tissue classes and are therefore not simply reliant on absolute signal intensity. The

maps produced are not restricted to the voxel resolution of the original data thus are capable of detecting submillimeter differences between groups. Procedures for the measurement of cortical thickness have been validated against histological analysis [61] and manual measurements[62,63]. FreeSurfer morphometric procedures have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths[54,58].

The distribution of cortical morphological features can vary between the site/scanner from which it is collected[64]. However, we also observed effects of site/scanner on the CBCL internalizing scores. For example, for the cross-sectional analysis, a significant effect of scanner was observed for ABCD ($H(31)=219.3$, $p<.001$, $\eta^2=0.02$) and POND ($H(4)=36.4$, $p<.001$, $\eta^2=0.06$), likely reflecting the geographic diversity and site-specific recruitment strategies, respectively. While techniques such as ComBat harmonization can be used to adjust for site effects while preserving covariate effects such as the RCIs[64], the algorithm can perform poorly when there is high collinearity between site and these covariates[65]. Thus, we chose not to correct for multisite effects.

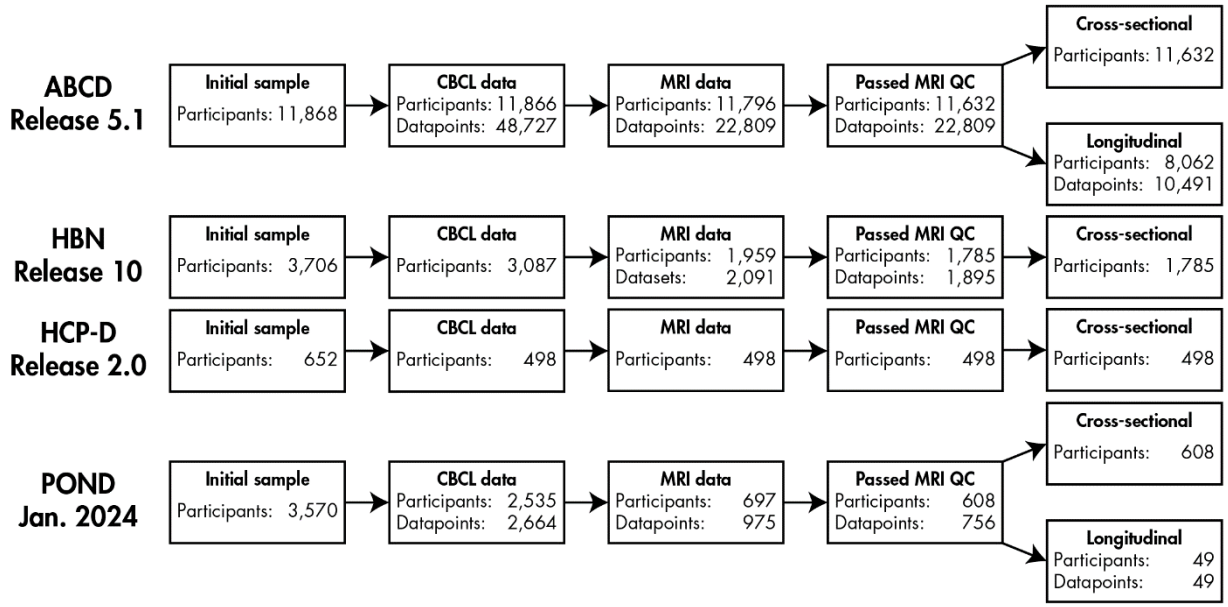
Quality control

The ABCD study provided measures relating to data quality[44]. Manual quality control was performed by trained technicians prior to the FreeSurfer pipeline to visually identify images with unacceptable severe artifacts or irregularities. After FreeSurfer, technicians also reviewed the quality of the extracted surfaces, and provided an overall quality control score which indicated whether the surfaces were recommended for inclusion or exclusion. Only participants who passed both aspects of quality control were retained for analyses. For HBN, T1-weighted images were manually inspected in-house for motion-related artifacts using a three-point scale, and those with severe motion artifacts were excluded. For HCP-D, only images rated highly (“good” or “excellent”) on a four-point scale by technicians are processed through the HCP preprocessing pipeline; thus, all FreeSurfer data available passed quality control. Data from the POND study was also processed through the CIVET image-processing

pipeline[66], which provides an automated index of quality control that was used to exclude participants for the analyses.

Datapoint selection

A flowchart of the study sample is presented in **Supplemental Figure 1**. The number of datapoints is the total number of unique pairs of CBCL assessments and MRI scans across all participants. While only one CBCL assessment was acquired per session in all four studies, some POND and HBN participants had multiple structural images acquired during a single scanning session. For both the cross-sectional and longitudinal analyses, if multiple structural images were eligible for the final sample, one was selected at random.



Supplemental Figure 1: Flowchart of the study samples. For the cross-sectional analysis, one datapoint (CBCL assessment and structural image) was selected for each participant. The ABCD and POND datasets provided longitudinal data, from which pairs of datapoints were selected for the longitudinal analysis.
ABCD: Adolescent Brain Cognitive Development study; HBN: Healthy Brain Network; HCP-D: Human Connectome Project Development study; POND: Province of Ontario Neurodevelopmental network; CBCL: Child Behaviour Checklist; MRI: magnetic resonance imaging; QC: quality control

For the cross-sectional analysis, a single session was selected from the longitudinal data available after quality control from ABCD and POND. For ABCD, given the higher proportion of participants with data from the initial recruitment session (9 – 11 years of age), the datapoint that gave the most uniform distribution across age was selected (per a Kolmogorov-Smirnov test). For POND, to minimize between-dataset

differences in age, the session whose age best matched the distribution of the ABCD sample was selected.

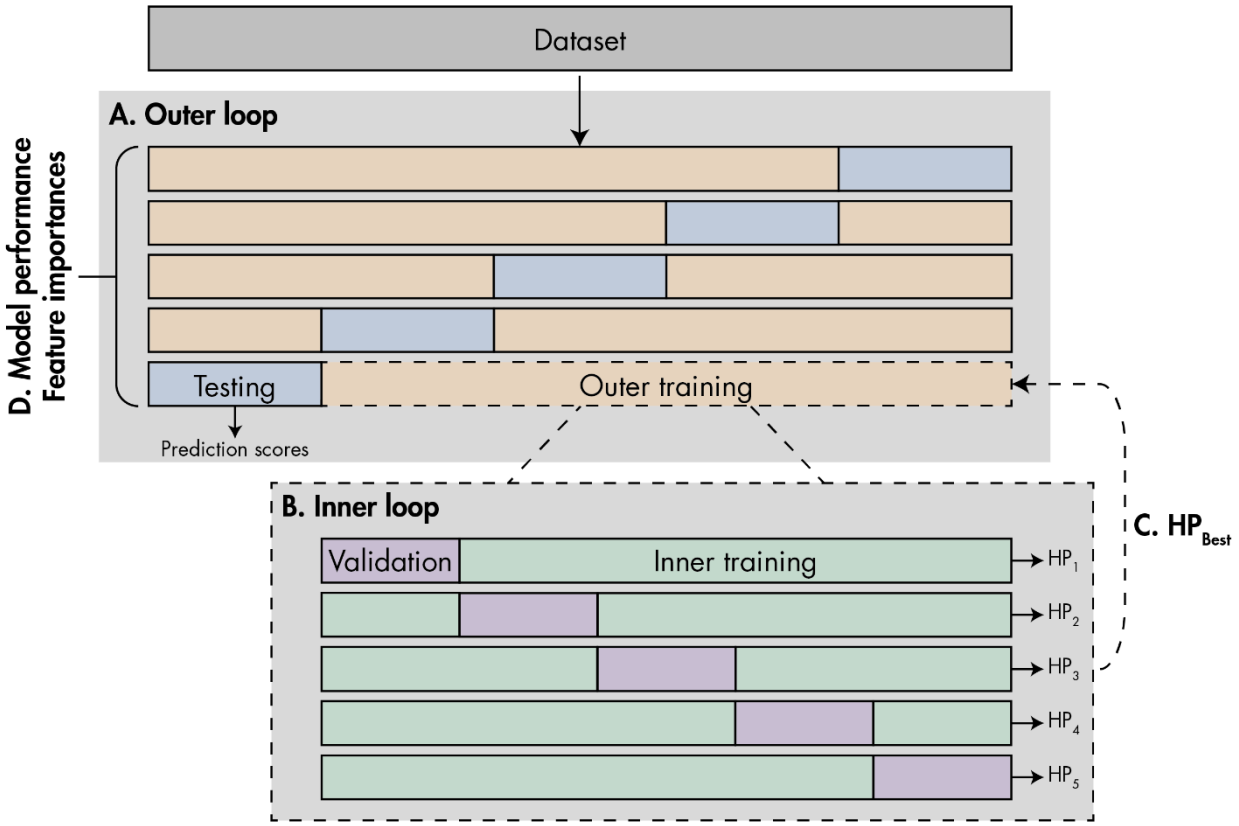
For the longitudinal analysis, partial data was available for a third timepoint in the ABCD study. Thus, to maximize the total number of included datapoints, measures of brain structure from the initial assessment were used to predict internalizing problems at the first follow-up, and measures of brain structure from the first follow-up were used to predict internalizing problems at the second follow-up. For POND, if multiple longitudinal timepoints were available, the timepoints that best overlapped with the age range of the ABCD sample were selected.

Deep learning model architecture

The model was implemented in Python (version 3.11) using the Keras framework (version 2.13.1;[67]) with TensorFlow (version 2.13.1;[68]) as the backend engine.

The model architecture was a multilayer perceptron, consisting of an input layer, blocks of hidden layers, and an output layer. Each block of hidden layers in the deep learning model consisted of a dense layer with rectified linear unit (RELU) activation, mitigating overfitting by applying L1 and L2 regularization to the layer's kernel followed by a dropout layer. Sigmoid activation was used in the output layer to predict the probability of the positive class. The models were trained to minimize binary cross-entropy loss, using stochastic gradient descent with an adaptive moment estimation (ADAM) optimizer. Due to imbalances between the classes, the models were trained using a balanced batch generator, which over-samples the minority class to ensure balanced data within each training batch[69].

A nested 5-fold stratified cross-validation scheme was used to tune, train, and test the model, with an inner loop for hyperparameter tuning and an outer loop for training and testing (**Supplemental Figure 2**). For the cross-sectional analysis, this scheme was performed using data aggregated across all four datasets. For the longitudinal analysis, given the imbalance between the POND and HBN datasets, the scheme was performed using ABCD, withholding POND as an external testing set.



Supplemental Figure 2: The nested cross-validation scheme used to train and test the deep learning models. The outer loop (A) uses 5-fold stratified cross-validation to split the data into testing (blue) and training (orange) sets. The inner loop (B) consists of a nested 5-fold stratified cross validation, where the outer training set is further split into validation (purple) and inner training (green) sets. Each validation fold is used to tune the model's hyperparameters (HP), and the best performing hyperparameter set (C; HP_{Best}) is selected to re-train the model using the entire outer training set, and prediction scores are calculated for the testing set. Measures of model performance and feature importances are then averaged across the outer folds (D).

The hyperparameter tuning was used to optimize the learning rate, batch size, and number of hidden blocks, as well as the number of neurons, regularization penalties, and dropout rate for each hidden block. Search criteria are presented in **Supplemental Table 5**. Given the class imbalance, the tuning was performed with the objective of maximizing the area under the precision-recall curve (AUC) of the validation set, using early stopping with a patience of 10 epochs and a maximum of 100 to prevent overfitting.

Supplemental Table 5: Search criteria for hyperparameter tuning.

		Values		
Learning rate		[0.1, 0.01, 0.001, 0.0001]		
Batch size		[8, 16, 32, 64, 128, 256, 512, 1024]		
		Minimum	**Maximum**	**Step size**
Learning rate		$1e^{-10}$	0.1	Log sampling
Number of hidden blocks		1	3	1
Per hidden block	**Number of neurons**	1	$N_{features}$	1
	Dropout rate	0.1	0.9	0.1
	L1 regularization penalty	$1e^{-10}$	1	Log sampling
	L2 regularization penalty	$1e^{-10}$	1	Log sampling

References

1. Kieling, C. *et al.* Worldwide prevalence and disability from mental disorders across childhood and adolescence: Evidence from the global burden of disease study. *JAMA Psychiatry* **81**, 347–356 (2024).

2. Vergunst, F. *et al.* Association of childhood externalizing, internalizing, and comorbid symptoms with long-term economic and social outcomes. *JAMA Netw Open* **6**, e2249568–e2249568 (2023).

3. Stevanovic, D. Impact of emotional and behavioral symptoms on quality of life in children and adolescents. *Quality of Life Research* **22**, 333–337 (2013).

4. Cannon, T. D. *et al.* Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis. *Hum Brain Mapp* **35**, 2424–2434 (2014).

5. 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).

6. Lipschutz, R. *et al.* Smaller hippocampal volume is associated with anxiety symptoms in high-risk Black youth. *Journal of Mood & Anxiety Disorders* **7**, 100065 (2024).

7. Andre, Q. R., Geeraert, B. L. & Lebel, C. Brain structure and internalizing and externalizing behavior in typically developing children and adolescents. *Brain Struct Funct* **225**, 1369–1378 (2020).

8. Muetzel, R. L. *et al.* Tracking brain development and dimensional psychiatric symptoms in children: A longitudinal population-based neuroimaging study. *American Journal of Psychiatry* **175**, 54–62 (2018).

9. Jarvers, I. *et al.* The relationship between adolescents' externalizing and internalizing symptoms and brain development over a period of three years. *Neuroimage Clin* **36**, 103195 (2022).

10. Bos, M. G. N., Peters, S., Van De Kamp, F. C., Crone, E. A. & Tamnes, C. K. Emerging depression in adolescence coincides with accelerated frontal cortical thinning. *Journal of Child Psychology and Psychiatry* **59**, 994–1002 (2018).

11. Lai, M. C. *et al.* Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *Lancet Psychiatry* **6**, 819–829 (2019).

12. Volkow, N. D. *et al.* The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cogn Neurosci* **32**, 4–7 (2018).

13. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci Data* **4**, 170181 (2017).

14. Somerville, L. H. *et al.* The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage* **183**, 456–468 (2018).

15. Achenbach, T. M. & Rescorla, L. A. *Manual for the ASEBA School-Age Forms & Profiles: An Integrated System of Multi-Informant Assessment*. (ASEBA, 2001).

16. Jacobson, N. S. & Truax, P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* **59**, 12–19 (1991).

17. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).

18. Rapee, R. M. *et al.* Adolescent development and risk for the onset of social-emotional disorders: A review and conceptual model. *Behaviour Research and Therapy* **123**, 103501 (2019).

19. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* **30**, (2017).

20. Cardenas-Iniguez, C. & Gonzalez, M. R. Recommendations for the responsible use and communication of race and ethnicity in neuroimaging research. *Nat Neurosci* **27**, 615–628 (2024).

21. Vandewouw, M. M. *et al.* Identifying replicable subgroups in neurodevelopmental conditions using resting-state functional magnetic resonance imaging data. *JAMA Netw Open* **6**, e232066–e232066 (2023).

22. Kushki, A. *et al.* Examining overlap and homogeneity in ASD, ADHD, and OCD: a data-driven, diagnosis-agnostic approach. *Transl Psychiatry* **9**, (2019).

23. Bahathiq, R. A., Banjar, H., Bamaga, A. K. & Jarraya, S. K. Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: Promising but challenging. *Front Neuroinform* **16**, 949926 (2022).

24. Schmaal, L. *et al.* Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry* **22**, 900–909 (2016).

25. Serra-Blasco, M. *et al.* Structural brain correlates in major depression, anxiety disorders and post-traumatic stress disorder: A voxel-based morphometry meta-analysis. *Neurosci Biobehav Rev* **129**, 269–281 (2021).

26. Wu, G. R. & Baeken, C. Normative modeling analysis reveals corpus callosum volume changes in early and mid-to-late first episode major depression. *J Affect Disord* **340**, 10–16 (2023).

27. Zahn-Waxler, C., Klimes-Dougan, B. & Slattery, M. J. Internalizing problems of childhood and adolescence: Prospects, pitfalls, and progress in understanding the development of anxiety and depression. *Dev Psychopathol* **12**, 443–466 (2000).

28. Mayes, S. D., Castagna, P. J. & Waschbusch, D. A. Sex differences in externalizing and internalizing symptoms in ADHD, autism, and general population samples. *J Psychopathol Behav Assess* **42**, 519–526 (2020).

29. Horwitz, E. *et al.* Sex differences in the course of autistic and co-occurring psychopathological symptoms in adolescents with and without autism spectrum disorder. *Autism* **27**, 1716–1729 (2023).

30. Herrmann, L. *et al.* Binary and non-binary fender identities, internalizing problems, and treatment wishes among adolescents referred to a gender identity clinic in Germany. *Arch Sex Behav* **53**, 91–106 (2024).

31. Farhane-Medina, N. Z., Luque, B., Tabernero, C. & Castillo-Mayén, R. Factors associated with gender and sex differences in anxiety prevalence and comorbidity: A systematic review. *Sci Prog* **105**, (2022).

32. Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures. *Dev Cogn Neurosci* **32**, 16–22 (2018).

33. Chaarani, B. *et al.* Baseline brain function in the preadolescents of the ABCD Study. *Nat Neurosci* **24**, 1176–1186 (2021).

34. Lisdahl, K. M. *et al.* Substance use patterns in 9-10 year olds: Baseline findings from the adolescent brain cognitive development (ABCD) study. *Drug Alcohol Depend* **227**, 108946 (2021).

35. Clark, D. B. *et al.* Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Dev Cogn Neurosci* **32**, 143–154 (2018).

36. Auchter, A. M. *et al.* A description of the ABCD organizational structure and communication framework. *Dev Cogn Neurosci* **32**, 8–15 (2018).

37. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Association, 2013).

38. Lord, C. *et al. Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1–4*. (Western Psychological Services, 2012).

39. Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* **24**, 659–685 (1994).

40. Kaufman, J. *et al.* Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial reliability and validity data. *J Am Acad Child Adolesc Psychiatry* **36**, 980–988 (1997).

41. Ickowicz, A. *et al.* The Parent Interview for Child Symptoms: A situation-specific clinical research interview for attention-deficit hyperactivity and related disorders. *The Canadian Journal of Psychiatry* **51**, 325–328 (2006).

42. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

43. Fischl, B. *et al.* Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).

44. Hagler, D. J. *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development study. *Neuroimage* **202**, 116091 (2019).

45. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).

46. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *Neuroimage* **9**, 179–194 (1999).

47. Dale, A. M. & Sereno, M. I. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *J Cogn Neurosci* **5**, 162–176 (1993).

48. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci U S A* **97**, 11050–11055 (2000).

49. Fischl, B., Liu, A. & Dale, A. M. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging* **20**, 70–80 (2001).

50. Fischl, B. *et al.* Sequence-independent segmentation of magnetic resonance images. *Neuroimage* **23**, S69–S84 (2004).

51. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *Neuroimage* **9**, 195–207 (1999).

52. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp* **8**, 272–284 (1999).

53. Fischl, B. *et al.* Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex* **14**, 11–22 (2004).

54. Han, X. *et al.* Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32**, 180–194 (2006).

55. Jovicich, J. *et al.* Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *Neuroimage* **30**, 436–443 (2006).

56. Segonne, F. *et al.* A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060–1075 (2004).

57. Reuter, M., Rosas, H. D. & Fischl, B. Highly accurate inverse consistent registration: A robust approach. *Neuroimage* **53**, 1181–1196 (2010).

58. Reuter, M., Schmansky, N. J., Rosas, H. D. & Fischl, B. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. *Neuroimage* **61**, 1402–1418 (2012).

59. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* **17**, 87–97 (1998).

60. Segonne, F., Pacheco, J. & Fischl, B. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans Med Imaging* **26**, 518–529 (2007).

61. Rosas, H. D. *et al.* Regional and progressive thinning of the cortical ribbon in Huntington's disease. *Neurology* **58**, 695–701 (2002).

62. Kuperberg, G. R. *et al.* Regionally localized thinning of the cerebral cortex in schizophrenia. *Arch Gen Psychiatry* **60**, 878–888 (2003).

63. Salat, D. *et al.* Thinning of the cerebral cortex in aging. *Cerebral Cortex* **14**, 721–730 (2004).

64. Fortin, J. P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120 (2018).

65. Bayer, J. M. M. *et al.* Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Front Neurol* **13**, 923988 (2022).

66. Ad-Dab'bagh, Y. *et al.* The CIVET image-processing environment: A fully automated comprehensive pipeline for anatomical neuroimaging research. in *Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping* (NeuroImage, Florence, Italy, 2006).

67. Chollet, F. Keras. *GitHub repository* Preprint at https://github.com/fchollet/keras (2015).

68. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. (2015).

69. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* **18**, 1–5 (2017).