Title?

Table of Contents

[Introduction 4](#_Toc196220334)

[Obsessive-Compulsive Disorder 4](#_Toc196220335)

[Neurobiology of OCD 4](#_Toc196220336)

[Structural Brain Abnormalities 7](#_Toc196220337)

[Informant Discrepancies 8](#_Toc196220338)

[Statistical Learning 9](#_Toc196220339)

[Learning methods 10](#_Toc196220340)

[Linear Model 10](#_Toc196220341)

[Decision Trees 10](#_Toc196220342)

[Boosting 11](#_Toc196220343)

[The present study 13](#_Toc196220344)

[Methods 15](#_Toc196220345)

[Data Source and Collection Procedures 15](#_Toc196220346)

[Data acquisition 15](#_Toc196220347)

[Structural MRI 15](#_Toc196220348)

[Preprocessing sMRI 16](#_Toc196220349)

[Brain Segmentation 17](#_Toc196220350)

[Regions of Interest 17](#_Toc196220351)

[Demographics and categorical diagnosis 18](#_Toc196220352)

[Self and Parental Reports of dimensional psychopathology 18](#_Toc196220353)

[Parent-Reported Child Behavior Checklist 19](#_Toc196220354)

[Self-Reported Brief Problem Monitor 19](#_Toc196220355)

[Sample 20](#_Toc196220356)

[Modelling approach 23](#_Toc196220357)

[Modeling Framework 23](#_Toc196220358)

[Initial Regression Approach 24](#_Toc196220359)

[Redefining the Task as Multiclass Classification 24](#_Toc196220360)

[Addressing Class Imbalance with Cost-Sensitive Learning 24](#_Toc196220361)

[Hyperparameter Tuning and Validation 25](#_Toc196220362)

[Threshold Calibration 25](#_Toc196220363)

[Evaluation Metrics 26](#_Toc196220364)

[Permutation Based Significance Testing 28](#_Toc196220365)

[Results 29](#_Toc196220366)

[Distribution of Outcome Variables 29](#_Toc196220367)

[Model Parameter Optimization 29](#_Toc196220368)

[Threshold Calibration and Class-Specific Performance 32](#_Toc196220369)

[Discussion 38](#_Toc196220370)

[References 39](#_Toc196220371)

[Appendix 54](#_Toc196220372)

Introduction

Obsessive-Compulsive Disorder

Obsessive-compulsive disorder (OCD) is recognized as a prevalent and persistent neuropsychiatric condition, impacting an estimated 2% to 3% of individuals worldwide (de Mathis et al., 2013). Almost half of adult patients with OCD report symptom onset during childhood or adolescence, with prevalence rates in these age groups ranging from 1% to 4% (Nazeer et al., 2020). OCD is a clinically and etiologically highly heterogeneous disorder, characterized by various overlapping symptom dimensions (Bragdon & Coles, 2017). Obsessive-Compulsive Disorder (OCD) is defined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) as the presence of obsessions, compulsions, or both. Obsessions are recurrent, intrusive, and unwanted thoughts, urges, or images that typically cause significant anxiety or distress, and which the individual attempts to ignore or neutralize. Compulsions are repetitive behaviors or mental acts performed in response to obsessions or rigid rules, aimed at reducing distress or preventing a feared outcome, though they are not realistically connected to the outcome or are clearly excessive. These symptoms are time-consuming or cause clinically significant distress or impairment in functioning (American Psychiatric Association, 2013). These symptoms cause significant distress or impairment in social, occupational, or other important areas of functioning.

OCD is unique among mental illnesses in that it exhibits both externalizing and internalizing symptom domains (Guzick et al., 2019). Internalizing features, such as anxiety and obsessive thinking, typically manifest as avoidance and withdrawal, while externalizing characteristics; like compulsive or ritualized behavior, are more overt and behaviorally disruptive (Achenbach, 2009; Shephard et al., 2021). Conceptualizing OCD through this dual framework offers a more nuanced understanding of its clinical presentation and has important implications for treatment planning. This framework is therapeutically beneficial and supported by empirical research (Kessler et al., 2011; Slade & Watson, 2006).

Neurobiology of OCD

Recent research increasingly conceptualizes OCD as a neurodevelopmental condition characterized by structural abnormalities in brain regions critical for emotional regulation, cognitive control, and habit formation (Graybiel & Rauch, 2000; Pagliaccio et al., 2021; van den Heuvel et al., 2016). The cortico-striato-thalamo-cortical (CSTC) model provides the foundational neurobiological framework for OCD, emphasizing dysfunction within circuits connecting the orbitofrontal cortex (OFC), anterior cingulate cortex (ACC), thalamus, and basal ganglia structures, particularly the caudate nucleus (Graybiel & Rauch, 2000). More recent models propose a broader neurocircuit-based taxonomy involving five partially overlapping networks, each linked to specific symptom dimensions: the fronto-limbic, sensorimotor, ventral cognitive, ventral affective, and dorsal cognitive circuits (Shephard et al., 2021).

In adolescents, structural deviations appear particularly pronounced and developmentally specific. Structural magnetic resonance imaging (sMRI) studies have identified consistent alterations. Large-scale meta-analyses from the ENIGMA-OCD working group have demonstrated widespread cortical and subcortical structural differences in pediatric OCD compared to healthy controls​enigma (van den Heuvel et al., 2022). In adolescents, sMRI findings include larger thalamic volume and smaller cortical thickness in prefrontal and parietal regions  (van den Heuvel et al., 2022). These abnormalities are thought to reflect disrupted neurodevelopmental processes involving fronto-striatal pathways. Importantly, medication status has emerged as a significant moderating factor in sMRI studies. Youth with OCD who were receiving selective serotonin reuptake inhibitors (SSRIs) showed different brain structural profiles compared to unmedicated patients. Specifically, unmedicated youth exhibited larger thalamic volumes relative to controls, whereas this enlargement was not observed among medicated individuals, suggesting that SSRIs may normalize or modulate structural brain alterations​enigma. Similarly, Wu et al. (2022) found that medicated and unmedicated adolescents differed in the magnitude and pattern of subcortical abnormalities​ (Wu et al., 2022). These findings underscore the need to account for medication exposure when interpreting structural brain data in pediatric OCD.Importantly, adolescence is a period of major brain restructuring, particularly within the prefrontal cortex and striatal regions. Normative developmental processes, such as synaptic pruning and myelination, contribute to age-related reductions in gray matter volume and increased efficiency of neural circuits (Casey et al., 2008). Deviations from these developmental trajectories have been proposed as a neurobiological basis for emerging psychiatric symptoms, including the onset and escalation of OCD during adolescence.

Subjects are excluded from the study if they are currently taking common antidepressant or antipsychotic prescription medications because these drugs can significantly affect brain structure.

Structural brain abnormalities have been particularly associated with internalizing symptom domains such as anxiety, fear, and intrusive thoughts—key features of OCD that often dominate the youth clinical presentation. Reduced medial prefrontal and orbitofrontal cortical thickness, combined with abnormal amygdala and thalamic volumes, have been linked to higher internalizing symptom severity across studies (Milad et al., 2013; Wang et al., 2022). Given the role of these regions in emotional regulation and cognitive control, structural alterations may underlie both symptom expression and treatment response. However, accurately capturing symptom severity during adolescence is complicated by informant discrepancies. Youth self-reports and parent reports frequently diverge, particularly for internalizing symptoms that are less externally observable. These differences introduce variability into efforts to map brain-behavior relationships and highlight the importance of evaluating how structural brain features predict symptoms as reported by different informants.

Informant Discrepancies

Traditionally, clinicians have depended on parents to provide comprehensive information about how an illness and its treatment affect their children. This reliance stems from the perception that children may not possess the cognitive and linguistic skills required to understand and respond to surveys accurately (Vygotsky, 1978). Consider the case of Liam, a 12-year-old having battled severe OCD for several years. After starting therapy, he was showing signs of improvement. According to Liam, he felt he was making excellent progress. He reduced his handwashing rituals from every hour to three times a day and started joining some family meals. He was also beginning to meet his friends for short walks around the neighborhood. However, his parents observed a different reality. While Liam had made some progress, he often became trapped in lengthy rituals that caused him significant distress. He had yet to return to school full-time, attending only partial days if he went at all. Though he started venturing out with friends, it was only to familiar, controlled environments. His parents continued to monitor his progress closely, supporting him in his journey while remaining aware of the continuous obstacles that his OCD presented. This vignette demonstrates the importance of recognizing that the child's perspective is distinct but equally valid. The insights provided by Liam can differ significantly from those of his mother, highlighting the potential discrepancies in information regardless of whether the goal is clinical assessment or research.

The phenomenon of informant discrepancy has been recognized for nearly 70 years, dating back to Lapouse and Monk's work in 1958. Achenbach, McConaughy, and Howell (1987) conducted a seminal analysis of 119 studies investigating these informant inconsistencies. Their key findings included: (a) reports of the same behavior by different informants generally show low to moderate agreement; (b) the reports of two informants observing children in the same setting are more similar than those of two informants observing children in different settings; (c) there is greater agreement between informants' reports for younger children compared to older ones; and (d) reports of externalizing behaviors like aggression show higher consistency than those of internalizing behaviors such as anxiety. They concluded by stating, "Different informants are needed for different situations. . . there is no royal road or preeminent gold standard for phenomena that are inevitably affected by assessment procedures and other situational variables" (p. 227–228). Consequently, the primary objectives of the informant discrepancies research summarized by Achenbach et al. (1987) were to outline the extent of informant discrepancies, identify the informant pairs (e.g., parent and child, teacher and parent) with the greatest discrepancies, and pinpoint the behavioral domains where these discrepancies were most pronounced. A prominent finding indicated discrepancies and varying accuracy in symptom reporting, with no clear consensus. This is especially important because obsessions (an internalizing symptom) are a distinguishing symptom of OCD.

The issue of informant discrepancies is particularly pertinent when interpreting study findings in the field of developmental psychopathology. A significant portion of the evidence about prevalence rates of psychological disorders, classification of diagnosis, and effectiveness of interventions for children is derived from reports by multiple informants (Weisz et al., 2005). For instance, the prevalence rates of conduct and oppositional defiant disorders in community samples can vary significantly depending on whether assessments are based solely on parental or teacher reports or if both sources are considered simultaneously. According to Offord et al. (1996), these rates range from 1.6% to 10.2%. More recent research by Munkvold et al. (2009) also noted substantial variability in oppositional defiant disorder prevalence, although the reported rates were lower, ranging from 0.2% to 2.6%. Furthermore, depending on the informant, it is typical to find inconsistent results from controlled studies that evaluate psychological therapies (De Los Reyes & Kazdin, 2005).

While the use of multiple informants in mental health assessment is thought to enhance our understanding of the psychological functioning of children, particularly in the infant population, we are still in the process of discovering how to effectively utilize this wealth of information (Reyes, 2013).

Statistical Learning

Statistical Learning Theory (SLT) is the basis for many modern machine learning algorithms, emphasizing generalization by extracting patterns from data for accurate predictions (Luxburg & Schoelkopf, 2008). SLT is central to supervised learning, which uses labeled data to train models that relate predictor variables to a response variable, focusing on prediction accuracy rather than causal understanding (Shmueli, 2011). Predictive modeling prioritizes accuracy, using probabilistic or supervised learning models, and employs loss functions to measure prediction discrepancies and refine models for real-world use. In the context of clinical neuroscience and neuroimaging, the most common machine learning paradigm learns to associate brain imaging data with specific classifications such as diagnostic groups. This approach can help identify key predictive features that differentiate between these categories (Enrico et al., 2021).

Learning methods

Linear Model

Linear regressions provide a simple technique for analyzing data by assuming a linear relationship between input variables (X) and an output variable (Y) and typically estimates parameters using the least squares method. While effective for straightforward linear relationships, they are limited with complex data, which has led to advanced adaptations such as generalized linear models and support vector machines, offering broader applicability and improved modeling techniques for diverse patterns.

Decision Trees

A diagram of a family

AI-generated content may be incorrect.

***Figure 2.*** *Example of a decision tree used for predicting severity scores based on individual and socioeconomic factors. The root node begins with the condition Age ≥ 16. If this condition is met, the model predicts a severity score of 4. If the condition is not met, the decision process continues down to the next node. Each non-leaf node represents a decision based on a feature threshold, while the leaf nodes indicate the predicted severity score (Y). This hierarchical structure illustrates how different combinations of age, income, and education contribute to the final prediction.*

Decision trees provide a significant advancement to linear models by effectively handling non-linear relationships and interactions between variables. Tree models operate under the assumption that the interaction between the response variable and the predictors can be modeled through locally constant fits (Breiman, 2017). Unlike the simplicity of linear models, decision trees dynamically segment the feature space through recursive binary splits, adeptly addressing both classification and regression tasks (James et al., 2021). In classification tasks, this is achieved by creating subsets with a dominant output class, whereas in regression, it involves reducing the variability of target values within each subset. As illustrated in Figure 2. each node of the tree serves as a decision point, directing data further down branches or reaching leaf nodes where predictions are determined by metrics such as class majority or mean values. Thus, when used for regression the aim is to split the data into subsets that minimize the resulting mean squared error, mean absolute error, or the variance of the target variable within these subsets (Ryan, 2025). Although they are effective in capturing complex patterns, they are also prone to overfitting, which occurs when the model captures noise rather than the underlying pattern. Thus, it crucial to balance bias and variance for accuracy and generalizability.

Boosting

A diagram of a tree

AI-generated content may be incorrect.

***Figure 3.*** *Illustration of the boosting process used in ensemble learning methods. Training error decreases over successive iterations as additional decision trees are added. Initially, a single tree is trained, followed by subsequent trees that correct the errors of the previous ones. The ensemble prediction is formed by summing the outputs of multiple trees, leading to improved accuracy and reduced training error over time.*

Boosting, as illustrated in Figure 3, is an ensemble method used to enhance predictive accuracy (Schapire & Freund, 2012). An ensemble combines multiple models to make more accurate predictions than a single model. Boosting iteratively adds simpler models, like decision trees, that improve the overall fit by addressing errors from previous models. Figure 3 demonstrates how training error decreases as more trees are added, with the ensemble prediction formed by summing tree outputs for improved accuracy. Gradient Boosting, which operates similarly to gradient descent, optimizes individual trees to reduce error and minimizes the ensemble's collective error by correcting residuals with each added tree (Friedman, 2001). This approach allows diverse loss functions for error minimization, enhancing the alignment of predictions with true outcomes (Ryan, 2025). Adjusting observation weights helps the models focus on challenging cases, correcting errors while maintaining flexibility. XGBoost, or extreme gradient boosting, is a particularly efficient algorithm for fitting boosting models (Ren et al., 2019).

Tuning Parameters

When using XGBoost for regression tasks, several key parameters can be fine-tuned to optimize model performance (XGBoost Developers, 2022). It's important to note that this overview is not exhaustive of all tuning parameters available in XGBoost; rather, it is a short summary of some of the most impactful parameters to consider when optimizing regression models. The learning rate determines how quickly the model learns patterns, with smaller values allowing for more cautious learning to reduce overconfidence and potential overfitting. The maximum depth sets how complex each decision tree is by limiting the number of splits, with deeper trees capturing more intricate patterns but risking overfitting. “Minimum child weight” sets the amount of data required in a leaf node before further splitting, promoting simpler, less complex trees. For each tree, only a random subset of the predictor variables is made available to the model, and the size of this subset is a tuning parameter denoted colsample\_bytree in XGBoost. Furthermore, a threshold on the magnitude of improvement required for making a new split is controlled by the parameter “gamma”. Finally, only a random subset of the training data is used when fitting each tree, and the size of this subset is controlled by the parameter “subsample”. Together, these parameters help balance the model's ability to learn complex patterns with its ability to generalize well to new, unseen data.

In conclusion, the application of machine learning techniques, particularly XGBoost, offers a powerful methodology for modeling complex interactions between variables to predict an outcome of interest. This approach provides a robust framework for integrating diverse data types, such as neuroimaging, behavioral, and demographic variables. By using the collective strengths of multiple models, boosted ensembles surpass the predictive capabilities of single decision trees and linear models, effectively managing intricate relationships in high-dimensional data. This integration of computational tools with clinical insights holds promise for refining diagnostic criteria and enhancing personalized intervention strategies for OCD, paving the way for more precise and effective treatment approaches.

The present study

This thesis examines the relationship between structural brain features and OCD symptomatology in adolescents, with particular emphasis on the role of informant discrepancies in symptom reporting. In the field of developmental psychopathology, informant differences are a well-documented methodological and clinical challenge, as much of the diagnostic and epidemiological data are derived from subjective reports provided by both youth and their caregivers. These discrepancies can substantially influence diagnostic outcomes, perceived symptom severity, and treatment decisions. OCD presents a particularly informative case for this investigation, as it encompasses both internalizing (e.g., obsessions) and externalizing (e.g., compulsions) symptom dimensions, and clinical diagnosis during adolescence is often based primarily on parent-reported information.

The present study seeks to elucidate brain–behavior relationships by evaluating the extent to which sMRI data can predict OCD symptom severity as reported by different informants. Parcellated brain volume features were extracted from sMRI scans and used as input variables in supervised machine learning models, implemented using the XGBoost algorithm. The internalizing symptom domain was selected as the predictive target because it is assessed using parallel items across youth and parent reports, thus enabling a direct comparison of informant-specific prediction performance. Beyond evaluating overall model accuracy, the analysis specifically tested whether predictive performance varied systematically across informants. This focus on informant-related variation is motivated by prior findings indicating that internalizing symptoms are more likely to be reported inconsistently and subjectively across sources.

The primary objective of the present study was to investigate whether structural neuroimaging data could reliably predict the severity of OCD symptoms in youth, as reported by the youth themselves and their parents. In addition to this core inquiry, two supplementary questions were explored. First, the study examined whether structural brain features exhibit greater predictive accuracy for internalizing symptoms when derived from youth self-reports compared to parent-reports. Second, the analysis evaluated how these prediction models compare to the classification of OCD based on a clinically evaluated diagnosis. By integrating these components, the study seeks to elucidate the relationship between brain structure and subjective symptom expression across informants, thereby contributing to the development of individualized, informant-sensitive predictive models for OCD assessment in pediatric populations.

Methods

Data Source and Acquisition

The Adolescent Brain and Cognitive Development (ABCD) Study is a comprehensive decade-long research initiative in the United States (US) designed to enhance our understanding of physical and mental health and risk factors during adolescence. The study tracks children from ages 9-10 through late adolescence and into early adulthood. This age range includes a crucial developmental stage, when exposure to substance use, and the start of several mental health conditions take place. The repository includes around 12,000 children at baseline, recruited from 21 research sites (Karcher & Barch, 2021). To ensure the cohort is diverse and representative, the ABCD study employs a multi-stage probability sampling technique to minimize selection bias, thus increasing the generalizability of the findings (Garavan et al., 2018). The ABCD comprises a wide range of behavioral (Barch et al., 2018), multimodal brain imaging (Casey et al., 2018), and other evaluations (Zucker et al., 2018). It conducts annual lab-based evaluations and biannual imaging scans to assess mental and physical health metrics (Saragosa-Harris et al., 2022). The data utilized in this thesis is sourced from the ABCD Data Release 5.1 (Haist & Jernigan, 2023). To ensure consistency in data analysis, this thesis utilizes only the year two follow-up time point, as it provided the most comprehensive available data across all key variables of interest, including, neuroimaging data and questionnaire responses.

REK stuff hereeeee and ethical considerations

Structural MRI

The ABCD study collects MRI data from three different scanner platforms located at 21 collection sites across the US: Siemens Prisma, General Electric 750, and Philips scanners (Casey et al., 2018). T1-weighted (T1w) images are acquired using a 3D T1w inversion-prepared RF-spoiled gradient echo sequence with 1 mm isotropic resolution (Casey et al., 2018). Prospective motion correction is applied when available (currently only on Siemens and GE scanners; (Tisdall et al., 2012; White et al., 2010)). The Siemens scanner, acquisition parameters are TR = 2500 ms, TE = 2.88 ms, TI = 1060 ms, flip angle = 8°, with a 256 × 256 matrix, 176 slices, and a 256 mm FOV (acquisition time ~6:08). The Philips scanner used TR = 6.6 ms, TE = 3.1 ms, TI = 950 ms, flip angle = 9°, matrix size 256 × 256, 225 slices, and a FOV of 256 × 240 mm (acquisition time ~5:38). And, the GE scanner, parameters included TR = 2500 ms, TE = 2.0 ms, TI = 1060 ms, flip angle = 8°, with 208 slices and the same resolution and matrix size (acquisition time ~6:09).

The ABCD MRI acquisition protocol utilizes high-density phased array head coils, which can lead to significant variations in image intensity. Additionally, head motion presents a significant challenge—especially in pediatric populations, where increased movement is more common—since it can degrade image quality and distort derived metrics (Reuter et al., 2015; Satterthwaite et al., 2012). Therefore, although prospective motion correction techniques are implemented to mitigate the effects of motion in structural MRI scans, excessive head movement can still introduce substantial artefacts, hindering accurate brain segmentation (Tisdall et al., 2016). If these discrepancies are not adequately addressed, they may lead to inaccuracies in cortical surface reconstruction and brain segmentation (ref).

Due to the potential artifacts in MRI images, T1w quality control during the MRI acquisition includes three checks (ABCD Study, 2025; Hagler et al., 2019). Firstly, (1) automated checks for protocol compliance assess the completeness of the imaging series and ensure that they meet the specified parameters; these criteria include verifying whether key imaging characteristics such as voxel size and repetition time align with the expected values for each scanner. (2) Automated quality control procedures involve calculating signal-to-noise ratio and head motion statistics. Lastly, this is complemented by (3) a manual quality control process where trained technicians visually assess image quality, identifying and flagging significant artefacts. Series that fail to meet quality standards are excluded from further processing and analysis, and reviewers are required to document observable artefacts using standardized notations.

Preprocessing sMRI

All sMRI images used in this study were preprocessed and subjected to quality control by the ABCD Study prior to public data release, using their standardized in-house image processing pipeline (ABCD Study, 2025). While the ABCD team provides a set of recommended inclusion criteria—accounting for factors such as image quality, protocol compliance—it is the responsibility of the data user to select criteria that are appropriate for their specific analyses. In this thesis, a subset of these criteria was applied based on their relevance to the research objectives. Preprocessing T1w images involves steps to ensure the accuracy and reliability of the data, including (1) Correction for Gradient Nonlinearity Distortions, which addresses distortions in the MRI images introduced by the scanner's gradient system (Jovicich et al., 2006; Wald et al., 2001). These corrections are specific to each scanner model and are provided by MRI manufacturers to enhance image fidelity. (2) Bias Field Correction involves correcting brightness variations across the brain images, a phenomenon known as intensity non-uniformity. This distortion is often caused by the proximity of brain tissue to the MRI coils, leading to areas with extremely high-intensity values that may be erroneously identified as non-brain tissue (i.e., skull). To address this issue, T2-w images are registered to T1-w images using a technique called mutual information, which facilitates accurate alignment and overlay of the different scan types (Wells et al., 1996). Following this registration, the procedure includes tissue segmentation and the application of smoothly varying estimated B1-bias fields to adjust brightness levels, ensuring that each tissue type is represented consistently across the images (Sled et al., 1998). Lastly, (3) Resample to Isotropic: The final image preprocessing step standardizes the viewing and analysis of brain structures. The images are resized and aligned with an internally generated reference brain that features isotropic voxels of 1.0 mm and is approximately aligned with the anterior commissure/posterior commissure (AC/PC) axis (Friston et al., 1995).

Brain Segmentation

Cortical surface reconstruction and subcortical segmentation are conducted by the ABCD Study team using FreeSurfer version 7.1.1 (https://surfer.nmr.mgh.harvard.edu). FreeSurfer has been validated in adolescent samples (Biffen et al., 2020). The process begins with skull-stripping, which removes the skull and non-brain tissues from the MRI images (Ségonne et al., 2004). Simultaneously, white matter segmentation is conducted to identify white matter regions, while initial mesh creation produces a preliminary three-dimensional representation of the brain's surface (Dale et al., 1999). Following this, the correction of topological defects on the surface model is applied to address errors or irregularities (Fischl et al., 2001; Segonne et al., 2007). The surface model is optimized and refined (Dale et al., 1999; Dale & Sereno, 1993; Fischl & Dale, 2000). Lastly, the reconstructed brain surface undergoes nonlinear registration to a spherical atlas, aligning it with a standardized spherical model to facilitate consistent comparisons across different subjects (Fischl et al., 1999).

Regions of Interest

After completing cortical reconstruction, specific brain regions are labelled by two atlases. Cortical areas are labeled using the Destrieux atlas-based classification (Destrieux et al., 2010). It is widely used in structural MRI studies to analyze cortical volume in neurodevelopmental research ref. This atlas uses a sulco-gyral classification, distinguishing between exposed gyri and buried sulci based on mean curvature and convexity, thus providing 74 bilateral regions (148 total). Subcortical structures were analyzed using the Automated Segmentation of the Subcortical Structures (ASEG) provided by FreeSurfer (Fischl et al., 2002). This atlas allows the segmentation and volume measurement of subcortical areas and other intracranial structures, providing 46 regions in total. Combining these atlases facilitates a comprehensive analysis of cortical and subcortical regions. Once both cortical and subcortical structures are labeled, a total of 194 parcellated brain volumes are generated per individual. These data are provided by the ABCD Study in tabulated format and made accessible for downstream analyses by independent researchers.

*A screenshot of a computer screen

AI-generated content may be incorrect.Figure A. Parcellation of cortical and subcortical brain structures using the Destrieux and ASEG atlases.*

*Note: Visualization of tabulated cortical and subcortical regions was produced using the ggseg packages in R (Mowinckel & Vidal-Piñeiro, 2020). Cortical surfaces (top two rows) are segmented using the Destrieux atlas, which identifies 148 sulcal and gyral regions based on curvature and convexity features. The first row illustrates the right hemisphere in lateral and medial views, while the second row presents the corresponding views of the left hemisphere. Subcortical structures (bottom row) are segmented using FreeSurfer’s Automated Segmentation of Subcortical Structures (ASEG) atlas, displayed in sagittal (left) and coronal (right) views.*

Demographics and categorical diagnosis

To assess psychiatric symptoms and derive DSM-5-consistent diagnoses, this study utilized the Kiddie Schedule for Affective Disorders and Schizophrenia - Computerized Version (KSADS-COMP), as implemented in the ABCD Study. The KSADS-COMP is a self-administered, computerized diagnostic tool adapted from the gold-standard KSADS-PL (Present and Lifetime Version) and developed to align with DSM-5 diagnostic criteria (J. Kaufman et al., 1997; *KSADS-COMP*, n.d.). In ABCD, both caregivers and youth complete the KSADS-COMP independently on a tablet or computer. For the present analysis, we used KSADS-COMP diagnostic data from the Year 2 follow-up visit, which was selected due to its large sample size and concurrent sMRI data collection.  The instrument applies automated branching logic and standardized scoring algorithms to assess symptoms across a range of psychiatric disorders, including depression, anxiety, ADHD, conduct disorder, and substance use (Barch et al., 2018). This approach uses established criteria and diagnoses may be classified as "certain," "possible," "in remission," or "not present." While not administered by a clinician, the KSADS-COMP is clinically designed and validated for use in large-scale studies, offering reliable and scalable mental health phenotyping (Barch et al., 2019). Demographic questions, including age, sex, race, and ethnicity, are also retrieved from the KSADS-5.

Self and Parental Reports of dimensional psychopathology

The Achenbach System of Empirically Based Assessment (ASEBA) is a comprehensive evaluation tool developed after years of research and practical use (*ASEBA*, 2019). It is designed to assess behavioral, emotional, and social aspects, along with strengths, competencies, and adaptive functioning in individuals ranging from 1½ years old to over 90. ASEBA is widely applied in diverse areas such as mental health services, education, healthcare, research, and more. The Child Behavior Checklist (CBCL) and Brief Problem Monitor (BPM), two components of the ASEBA, provides a dimensional diagnostic/assessment approach that places behaviors along a continuum of frequency and/or severity. Raw scores from these instruments are converted into standardized scores using ASEBA-defined algorithms, which account for informant type, age, sex, and ethnicity (Achenbach et al., n.d.). These normed scores are expressed as T-scores, with a mean of 50 and a standard deviation of 10.

Parent-Reported Child Behavior Checklist

The CBCL is a component of the ASEBA first published in 2001, and is used to assess children's behavioral, emotional, or social problems (Achenbach, 2001). It is a 112-item parent-reported survey, which uses a 3-point Likert scale for responses: "Very True," "Somewhat True," or "Not True," where parents are asked to rate each item based on their child's behavior "now or within the past six months." As depicted in Figure 5, the CBCL consists of several dimensions categorized into Syndrome Scales and DSM-Oriented Scales (American Psychiatric Association, 2013; Nelson et al., 2001). The eight syndrome scales are established through factor analysis. They encompass clusters of common behaviors or symptoms. Furthermore, these scales are grouped into three high-level domains Internalizing, Externalizing, and Total Problems scales. These dimensions offer a detailed assessment of a child's emotional, social, and behavioral functioning, aiding in identifying areas that may benefit from therapeutic or educational interventions. The Internalizing Problems scale specifically captures emotional difficulties such as anxiety, depression, withdrawal, and somatic complaints—making it a key indicator of inwardly directed psychological distress in children. The CBCL/6–18 shows high reliability (α = 0.83–0.94; r = 0.88–0.92) and strong clinical validity (Achenbach, 2018).

Self-Reported Brief Problem Monitor

The BPM, another component of the ASEBA, was first published in 2011 (Achenbach et al., 2011). Developed to complement parental assessments, adolescents provide self-reports on higher-level domains and attention. It is a 19-item self-reported survey used to assess children's behavioral and emotional functioning and their responses to interventions (RTIs). It also uses a 3-point Likert scale for responses: "Very True", "Somewhat True," or "Not True." Children are instructed to rate each item based on their behavior "currently or within the past six months." As illustrated in Figure 5, the BPM assesses four domains—Internalizing, Attention Problems, Externalizing, and Total Problems—mirroring the structure of the CBCL/6–18 (Achenbach et al., 2017). Of particular relevance, the Internalizing scale captures symptoms of anxiety, depression, and withdrawal. Recent validation in a Norwegian sample confirmed good internal consistency for this domain (α = .76–.88) and supported the original three-factor structure, affirming its utility as a brief and valid tool for identifying internalizing problems in at-risk children (Pedersen et al., 2021).



***Figure 5.*** *The structure of the ASEBA specifically focusing on the parallel between CBCL and the BPM. The CBCL consists of Syndrome Scales including clusters of symptoms, which are further grouped into three high-level domains known as (1) Internalizing Problems, (2) Externalizing, and (3) Total Problems score that sums all problem items. The BPM is a shorter version that provides a rapid assessment parallel to dimensions in CBCL for monitoring behavioral and emotional functioning over time.*

Sample

Several criteria were applied to determine the final sample for inclusion in the study. Figure 4 illustrates the participant selection process. Beginning with the full ABCD Year 2 dataset (N = 10,730), participants were excluded if they lacked an MRI scan (n = 2,757) or if their imaging data did not meet quality control standards (n = 193). Of those with usable MRI data, additional exclusions were made for missing self- and/or parent-reported data on internalizing symptoms (n = 314). Moreover, to ensure that medication effects did not confound brain structure findings, participants who were currently prescribed common psychotropic medications (e.g., SSRIs, SNRIs, antipsychotics, stimulants) were excluded (n = 169). Furthermore, participants who met criteria for a current diagnosis of a psychiatric disorder other than obsessive-compulsive disorder (OCD) were excluded from the sample (n = 837). Consequently, healthy control participants were defined as individuals who did not meet diagnostic criteria for any current or past psychiatric disorder, as assessed by the KSADS-COMP. Note that participants with OCD were retained in the sample even if they presented with additional current psychiatric diagnoses, reflecting the high comorbidity typically observed in pediatric OCD (Geller & March, 2012).

Following all exclusions, the final study sample included 6,460 participants, comprising 6,054 healthy controls and 406 participants with obsessive-compulsive disorder (OCD), of whom 227 were classified as having only OCD and 179 as having comorbid OCD. Demographic characteristics for each group are presented in Table A. The mean age was comparable across groups, with the healthy control group averaging 9.47 years (SD = 0.51) and the OCD group averaging 9.46 years (SD = 0.50). In both groups, the sex distribution was identical: 47.5% female and 52.5% male. One participant in the healthy control group identified as intersex-male; no intersex participants were present in the OCD group. With respect to race and ethnicity, the healthy control group included 2.2% Asian, 13.3% Black, 19.2% Hispanic, 9.8% Other, and 55.6% White participants. In comparison, the OCD group consisted of 1.0% Asian, 15.3% Black, 19.5% Hispanic, 14.3% Other, and 50.0% White participants.

|  |  |  |
| --- | --- | --- |
| Demographic variable | Healthy Control (n=6,054) | OCD  (n=406) |
| Mean Age (SD) | 9.47(0.51) | 9.46(0.50) |
| Sex |  |  |
| Female | 47,5% (n=2,877) | 47.5% (n=193) |
| Intersex-Male | 0% (n=1) | - |
| Male | 52.5% (n=3,176) | 52.5% (n=213) |
| Race/Ethnicity |  |  |
| Asian | 2.2% (n=131) | 1.0% (n=4) |
| Black | 13.3% (n=806) | 15.3% (n=62) |
| Hispanic | 19.2% (n=1,160) | 19.5% (n=79) |
| Other | 9.8% (n=593) | 14.3% (n=58) |
| White | 55.6% (n=3,364) | 50.0% (n=203) |

**KSADS Lifetime Diagnosis**

|  |  |
| --- | --- |
| Comorbidities | OCD (n =406) |
| Any depressive disorder | 5 |
| Any anxiety disorder | 46 |
| ADHD | 71 |
| ODD/CD | 62 |
| Bipolar | 49 |
| Drug use disorder | 2 |
| Any suicidality | 39 |
| Any eating disorder | 14 |

**ASEBA T-Score**

|  |  |  |
| --- | --- | --- |
|  | Healthy Control | OCD (n =406) |
| Attention Problems |  |  |
| DSM-5 Anxiety |  |  |
| DSM-5 Depression |  |  |
| DSM-5 Somatic |  |  |
| DSM-5 ADHD |  |  |
| DSM-5 Opposite |  |  |
| DSM-5 Conduct |  |  |
| OCD |  |  |
| Total problem score |  |  |

***Figure 4.*** *Flowchart of participant selection and subgroup classification at 2-year follow-up.*



Modelling approach

Modeling Framework

All analyses were conducted in R version 4.3.3. and all models were implemented using the XGBoost 3.0 package. Under the XGBoost framework, model performance is optimized by minimizing a specified loss function which defines the objective of the model. The objective function serves as a core component of model specification, as it directly reflects the type of prediction task being addressed (Brownlee, 2021). To prevent data leakage, the dataset was stratified and split into three subsets prior to any model building or feature selection. Specifically, 70% (n=4522) of the data was allocated for training and tuning the hyperparameters of the model, 15% (n=969) for threshold calibration, and the remaining 15% (n=969) for model evaluation. This separation ensured that information from the evaluation set did not inadvertently influence model training or parameter tuning (S. Kaufman et al., 2012). Also did stratified sampling of target varaibles.

Initial Regression Approach

The prediction of the internalizing symptoms was initially formulated as a regression task, using the continuous T-score derived from the Achenbach scale ref. The model was trained to minimize the squared error loss function. However, the distribution of symptom severity was found to be highly skewed, with relatively few individuals classified within the clinical range. This imbalance in the outcome variable resulted in suboptimal model performance and limited interpretability, as the squared error loss function assumes equal importance across all prediction errors, irrespective of class prevalence (Hastie et al., 2009). In contrast, classification models are better equipped to handle imbalanced outcome distributions, particularly when combined with class weighting strategies that adjust for disparities in group sizes (He & Garcia, 2009).

Redefining the Task as Multiclass Classification

To improve clinical interpretability and model performance, the task was reframed as a multiclass classification problem by mapping T-scores to established clinical categories (Achenbach, 2009): Normal (<65), Borderline (65–69), and Clinical (≥70).

The model was trained using the multiclass log-loss (cross-entropy) objective, which provides a probabilistic output across all categories (XGBoost, 2022). This approach, mathematically equivalent to multinomial logistic regression, enables a symmetric and interpretable estimation of class membership probabilities (James et al., 2021). Multiclass log-loss is particularly suited for imbalanced classification settings, as it penalizes confident but incorrect predictions more heavily than uncertain ones. This results in a more informative loss signal than traditional accuracy metrics, which may obscure poor performance on minority classes (Niculescu-Mizil & Caruana, 2005).

Addressing Class Imbalance with Cost-Sensitive Learning

To enhance classification performance in the context of unequal class distributions, both data-level and algorithm-level strategies were employed. A commonly used approach for addressing class imbalance involves assigning greater weights to minority class samples during model training, thereby mitigating the tendency of the model to prioritize the majority class (Kuhn & Johnson, 2013; Ting, 2002). In this study, class weights were applied to increase the influence of participants categorized in the borderline and clinical symptom groups, relative to those in the normal range. This strategy was particularly important given the disproportionately smaller size of the clinically significant group and the high dimensionality of the neuroimaging feature space, both of which increase the risk of model bias, variance inflation, and overfitting. Weighting the minority classes ensured that the model remained sensitive to clinically meaningful patterns, even when those patterns were underrepresented in the training data.

Hyperparameter Tuning and Validation

Hyperparameter optimization was conducted using a grid search across a predefined set of key tuning parameters, including maximum tree depth, learning rate, regularization strength, and row and column sampling ratios (*Xgboost Grid Search - R*, n.d.). To evaluate model performance, a five-fold cross-validation procedure was implemented on the training dataset, with early stopping employed to prevent overfitting and reduce computational burden. Early stopping is a regularization technique that terminates model training when performance on a validation set ceases to improve after 10 iterations, thereby preventing overfitting and reducing training time (Prechelt, 1998). The grid search explored 864 unique parameter combinations, each evaluated across five validation folds, resulting in a total of 4,320 model fits. The configuration that achieved the lowest average multiclass log-loss across validation folds was selected as the final model. This tailored tuning procedure enabled the application of stronger regularization, controlled model complexity, and optimized sampling strategies, collectively contributing to a more robust, stable, and generalizable classifier suitable for high-dimensional neuroimaging data (James et al., 2021; Kuhn & Johnson, 2013).

Threshold Calibration

To improve classification performance and ensure that predicted labels aligned more closely with clinically meaningful groupings, a post-hoc threshold calibration was performed using the class probabilities generated by the XGBoost model (*3.3. Tuning the Decision Threshold for Class Prediction*, n.d.). In standard multiclass classification, labels are assigned using the argmax rule, which selects the class with the highest predicted probability. However, this default strategy can introduce bias toward the majority class, particularly in imbalanced datasets, which are common in clinical research contexts (Van Calster et al., 2019). To address this issue, a class-specific threshold calibration approach was implemented. Using a one-vs-rest (OvR) framework on a held-out validation set, the model evaluated the decision threshold for each class independently (Rifkin & Klautau, 2004). Accordingly, each category—Normal, Borderline, and Clinical—was assessed independently during threshold calibration. For each binary classification (e.g., Class A vs. not Class A), the optimal decision threshold was determined by maximizing balanced accuracy, which accounts for both sensitivity (true positive rate) and specificity (true negative rate), providing a more equitable measure of performance in imbalanced datasets (Brodersen et al., 2010). This process yielded individualized decision thresholds for each class, providing a more flexible alternative to the standard argmax-based decision rule. It addresses the model's inherent tendency to underrepresent minority classes—an effect rooted in the objective function of most machine learning classifiers, including XGBoost, which typically aim to minimize overall loss (e.g., log-loss) across all samples. In imbalanced datasets, this global loss is disproportionately influenced by the majority class, causing the model to prioritize accuracy on the more prevalent classes while neglecting performance on underrepresented ones (He & Garcia, 2009). As a result, the default argmax rule systematically favors majority class predictions, even when the minority class probabilities are clinically meaningful.

Evaluation Metrics

The final tuned and calibrated models’ performance was assessed on a held-out test set. The performance was assessed using a comprehensive set of evaluation metrics selected to address the imbalanced nature of the outcome classes, particularly the underrepresentation of individuals with clinically significant OCD symptoms. Overall accuracy was not used as the sole performance indicator, as it can obscure poor detection of minority classes in imbalanced datasets (Saito & Rehmsmeier, 2015).

Instead, a multidimensional evaluation strategy was implemented. The caret package was used to compute confusion matrices and extract key metrics, including balanced accuracy, sensitivity, specificity, and positive predictive value for each class (Kuhn, 2008). Balanced accuracy was used to account for class imbalance by averaging sensitivity and specificity (true negative rate) for each class, providing a more equitable evaluation of performance (Brodersen et al., 2010). Precision, recall, and F1 scores were calculated for each class using the MLmetrics package (Yan, 2016). Precision reflects the proportion of true positives among all positive predictions, while recall captures the proportion of actual positives correctly identified. The F1 score, defined as the harmonic mean of precision and recall, offers a balanced measure of classification performance that is particularly valuable when the costs of false positives and false negatives are asymmetric. To assess the model's ability to discriminate between classes, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed using a OvR approach with the pROC package (Robin et al., 2011). These metrics visualize and quantify the trade-off between sensitivity and specificity across thresholds and offer insight into the model’s separability across all classes.

*Figure. A Flowchart illustrating the machine learning pipeline*

*Note: The full sample was split into three stratified subsets. In the training phase, hyperparameter optimization was performed on the training set using grid search and 5-fold cross-validation. In the threshold calibration phase, here a threshold grid search for class-specific probability cutoffs was used. Finally, in the evaluation phase, the calibrated model was applied to the held-out test set to assess final performance metrics.*

Permutation Based Significance Testing

To evaluate whether the model's performance exceeded what could be expected by chance, a permutation test was conducted following established methods for assessing statistical significance in predictive modeling (Good, 2000; Ojala & Garriga, 2009). This non-parametric approach involves disrupting the relationship between input features and class labels while preserving the underlying feature distributions. For computational efficiency, hyperparameter tuning was performed once using the original labels and fixed for all permutation iterations. This approach may slightly underestimate the variance in the null distribution (Ojala & Garriga, 2009).

To simulate the null hypothesis of no association between features and target labels, class labels in the held-out test set were randomly permuted 1,000 times. For each permutation, the trained XGBoost model was used to generate predicted class probabilities, which were thresholded using the previously calibrated class-specific thresholds. These thresholded outputs were then converted into class predictions. The classification accuracy was computed for each of the 1,000 permutations, producing a null distribution of accuracies expected under chance. The observed accuracy, obtained using the true (non-permuted) labels, was then compared against this null distribution. A permutation-based p-value was calculated as the proportion of permuted accuracies that were greater than or equal to the observed accuracy, providing a robust estimate of statistical significance without assuming distributional properties of the data. A density plot was generated to visualize the null distribution, with the observed accuracy overlaid as a vertical reference line, highlighting its deviation from the permutation-based baseline.

Results

Distribution of Outcome Variables

The distribution of internalizing symptom scores was examined for both the CBCL and BPM scales (Figure A). Both measures exhibited strong positive skew, with the majority of participants falling within the normative range and relatively few classified in the borderline or clinical ranges. This skew was more pronounced in the BPM distribution, which demonstrated a sharp mode near the normative threshold, highlighting the low prevalence of clinical internalizing symptoms in the sample.

*Figure A. Distribution of Outcome Variable: internalizing symptom T-scores for parent- and child-reported measures.*

1. *b)*



*Note: Density plots showing the distribution of internalizing T-scores derived from the (a) parent-reported CBCL (green) and (b) child-reported BPM (purple).*

Model Parameter Optimization

As shown in Table A (in appendix), both models reached optimal performance at 50 boosting rounds. While core parameters such as learning rate and max depth were consistent, the parent reported model used a lower minimum child weight (1 vs. 5), higher gamma (2 vs. 0), and differed in subsampling strategies (subsample: 0.6 vs. 0.9; colsample by tree: 0.9 vs. 0.6). Hyperparameter tuning revealed variation in model performance across the parameter space (Figure A). For both the BPM (Figure Aa) and CBCL (Figure Ab) models, log loss values ranged from approximately 1.090 to 1.108, indicating sensitivity to tuning configurations. Performance was most strongly influenced by tree depth and learning rate, with lower log loss achieved under shallower trees and moderate learning rates. Despite the broad search space, both models converged on similar optimal configurations, yielding final multiclass log-loss values of 1.090 (CBCL) and 1.093 (BPM). Class weighting contributed to improved classification of underrepresented borderline and clinical categories.

*Figure A. Hyperparameter Optimization Results for internalizing symptom prediction models*

a) A diagram of a graph

AI-generated content may be incorrect.

b) A graph with lines and numbers

AI-generated content may be incorrect.

*Note:* *Parallel coordinate plots illustrating how performance varies across the hyperparameter space. Each line represents a unique combination of tuning parameters across the grid search, with axes corresponding to individual hyperparameters and the resulting multiclass log loss on the y axis. (a) . Parent-reported symptoms (CBCL); line color indicating yellow as lower loss and blue as higher loss. (b) Child-reported symptoms (BPM); line color indicates log loss, with red representing lower loss and purple representing higher loss*

The optimization history across 864 hyperparameter configurations is shown in Figure A. Each point represents a unique parameter set, with color indicating the resulting multiclass log loss. For the child-reported model (BPM; Figure Aa), performance gradually improved across the search, with the best configuration emerging near the end of the tuning process. In contrast, the parent-reported model (CBCL; Figure Ab) reached its lowest log loss earlier in the search, followed by diminishing returns in later trials. The red step lines indicate the running minimum log loss, highlighting the incremental identification of superior parameter combinations throughout both optimization processes.

*Figure. A Tuning Parameter Optimization History Across Iterations*

A graph with colorful dots

AI-generated content may be incorrect.*a)*

*A graph with dots and lines

AI-generated content may be incorrect.b)*

*Note: Optimization history of multiclass models across hyperparameter configurations. Each point represents a distinct hyperparameter set, with color indicating the corresponding multiclass log loss (lower values shown in red/yellow; higher in purple). The red step line tracks the running minimum log loss, illustrating how increasingly better-performing configurations were identified throughout the tuning process. (a) Optimization trajectory for the model predicting child-reported internalizing symptom. (b) Optimization trajectory for the model predicting parent-reported internalizing symptom.*

Threshold Calibration and Class-Specific Performance

For both CBCL and BPM models, the initial argmax classification approach yielded high overall accuracy (CBCL: 95.1%; BPM: 93.4%) but demonstrated extreme bias toward the majority class (Healthy), with zero sensitivity for the Borderline and Clinical groups. Despite the apparent accuracy, Cohen’s Kappa values near zero (CBCL: –0.0016; BPM: 0.000; see table A) and balanced accuracy around 0.50 indicated that performance on minority classes was at chance level. Threshold calibration using a one-vs-rest strategy improved the distribution of predictions but did not substantially enhance minority class detection. For the CBCL model, overall accuracy declined to 68.0%, and Kappa further dropped to –0.0289. Sensitivity for the Healthy class fell (0.725), while small improvements were observed for the Borderline (0.048) and Clinical (0.083) categories. However, these gains were minimal: the confusion matrix (Figure A) showed that only two cases from each minority class were correctly classified, with most predictions still concentrated in the Healthy category. For the BPM model, threshold calibration produced negligible change. Overall accuracy remained high (92.5%), but the model continued to misclassify all minority cases, resulting in a Kappa of –0.0168 and balanced accuracy values close to 0.50. While the confusion matrix showed more even distribution across predicted labels post-calibration, true positive rates for Borderline and Clinical remained zero.

These trends are further supported by ROC curves and class probability distributions (Figure A). The CBCL model exhibited near-diagonal ROC curves for both Borderline and Clinical classes and similarly compressed probability distributions, indicating near-random classification and low confidence for minority outcomes. The BPM model showed some upward curvature for the Clinical group ROC curve, suggesting limited discriminatory ability, whereas the Borderline group remained poorly differentiated. Predicted probabilities for Healthy cases were more confidently distributed near 1.0, while minority class probabilities clustered near zero, reflecting poor model certainty. Overall, these findings highlight the persistent difficulty of achieving clinically meaningful classification performance in imbalanced symptom categories, particularly for child-reported outcomes.

*Figure. A Confusion matrix for model predictions on the test set*

1. A graph of a patient's health

   AI-generated content may be incorrect.A graph of a health care patient

   AI-generated content may be incorrect.A graph of a patient's health

   AI-generated content may be incorrect.A graph of a health care patient

   AI-generated content may be incorrect. *b)*

*Note: Confusion matrices display true versus predicted class labels for models after threshold calibration. Color intensity indicates the number of cases per cell. (a) parent-reported model (CBCL) correctly identified a small number of Clinical and Borderline cases, but most predictions remained in the Healthy category. (b) child-reported model (BPM) showed more accurate identification of Borderline and Clinical cases but never predicted any cases as Clinical. Despite a more balanced distribution, both models remain biased toward predicting the majority (Healthy) class.*

*Figure A. ROC curves and class probability distributions for the test set*

*a)* *A graph of different types of graphs

AI-generated content may be incorrect.*

*b)* A collage of graphs

AI-generated content may be incorrect.

*Note: Model performance for each class is shown using OvR ROC curves (bottom row) and corresponding predicted probability distributions (top row). (a) Parent-reported model (CBCL): The Healthy class shows modest separability, with predicted probabilities skewed toward 1, but the ROC curve remains close to chance. Both the Borderline and Clinical classes exhibit poor separability, with predicted probabilities near zero and ROC curves that closely follow the diagonal, indicating near-random performance. (b) Child-reported model (BPM): The Healthy class is more clearly distinguishable, with higher predicted probabilities and a ROC curve above chance. The Borderline and Clinical classes also show modest discriminative ability, with ROC curves slightly above the diagonal, indicating better performance than CBCL for minority class classification — though separability remains limited.*

To evaluate whether model performance exceeded chance levels, permutation testing was conducted with 1,000 random label shuffles. As shown in Figure A, the null distribution of accuracy scores is represented by a density curve, with the observed accuracy overlaid as a vertical reference line. For the parent-reported model (CBCL; Figure Aa), the model achieved a high raw classification accuracy of 0.68. However, the observed performance closely overlapped with the null distribution, yielding a permutation-based p-value of 0.953. This suggests that the elevated accuracy was likely driven by class imbalance rather than a meaningful predictive signal. These findings align with previous ROC and confusion matrix analyses, which indicated that classification success was largely restricted to the majority (Healthy) class. A similar pattern was found for the child-reported model (BPM; Figure Ab). The model trained on true labels achieved a classification accuracy of 0.93. The null distribution, generated via label permutation, had a mean of 0.55 and a standard deviation of 0.03. The observed accuracy was near the center of this distribution, yielding a permutation-based p-value of 1.000, indicating that the model’s performance was not significantly above chance. As shown in Figure A, the observed accuracy falls well within the range expected under the null hypothesis.

*Figure A. Permutation test comparing model accuracy to null distribution.*

1. A graph of a graph

   AI-generated content may be incorrect.*A diagram of a normal distribution

   AI-generated content may be incorrect. b)*

*Note: Density plots show the null distribution of classification accuracies obtained from 1,000 random permutations of test set labels for the parent-reported model (CBCL; a) and child-reported model (BPM; b). The observed accuracy using the true (non-permuted) labels is marked by a vertical line (red for CBCL, green for BPM). In both models, the observed accuracy falls within the null distribution, indicating that model performance was not significantly better than chance.*

*Table A.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CBCL | | | | BPM | | | |
| Metric | Default | With Weights | Tuned + Argmax | Tuned + Threshold-Cal. | Default | With Weights | Tuned + Argmax | Tuned + Threshold-Cal. |
| Overall Performance | | | | | | | | |
| Accuracy | 0.3333 | 0.3323 | 0.9505 | 0.6801 | 0.3137 | 0.3158 | 0.934 | 0.9247 |
| 95% CI | 0.3037–0.364 | 0.3027–0.363 | 0.9349–0.9633 | 0.6497-0.7094 | 0.2846–0.344 | 0.2866–0.3461 | 0.9164–0.9488 | 0.9062–0.9405 |
| Kappa | -0.0074 | 0.0224 | -0.0016 | -0.0289 | -0.0136 | -0.009 | 0.0 | -0.0168 |
| Class Balance | | | | | | | | |
| Balanced Accuracy (Avg) | 0.4942 | 0.5687 | 0.4997 | 0.465 | 0.4745 | 0.4795 | 0.5 | 0.4966 |
| Sensitivity (Borderline) | 0.28571 | 0.45238 | 0.0 | 0.048 | 0.22727 | 0.29545 | 0.0 | 0.0 |
| Sensitivity (Clinical) | 0.41667 | 0.54167 | 0.0 | 0.083 | 0.388889 | 0.33333 | 0.0 | 0.0 |
| Specificity (Borderline) | 0.65049 | 0.64401 | 0.99894 | 0.865 | 0.66054 | 0.66811 | 1.0 | 0.98811 |
| Specificity (Clinical) | 0.68783 | 0.68042 | 1.0 | 0.859 | 0.657203 | 0.65089 | 1.0 | 1.0 |
| Per-class Discrimination | | | | | | | | |
| Precision (Borderline) | 0.03571 | 0.05444 | 0.0 | 0.016 | 0.03086 | 0.04062 | NaN | 0.0 |
| Precision (Clinical) | 0.03279 | 0.04127 | NaN | 0.015 | 0.021021 | 0.01775 | NaN | NaN |
| F1 Score (Borderline) | 0.0634 | 0.09669 | 0.0 | 0.0236 | 0.0538 | 0.07225 | 0.0 | 0.0 |
| F1 Score (Clinical) | 0.061 | 0.07634 | 0.0 | 0.0248 | 0.0382 | 0.03344 | 0.0 | 0.0 |

Accuracy – Overall proportion of correct predictions

95% Confidence Interval – The range in which the true accuracy likely falls

Kappa – Adjusts accuracy to account for chance agreement

Balanced Accuracy – Average performance across all classes

Sensitivity – How well the model detects true cases of a class (also called recall)

Specificity – How well the model avoids false positives

Precision – The proportion of predicted cases that are actually correct

F1 Score – A balance of precision and sensitivity; reflects overall class-specific performance

Discussion

The primary aim of this study was to evaluate whether structural brain data could be used to predict the severity of OCD internalizing symptoms as reported by both youths and their parents. In addition to answering the primary research question, we wanted to examine whether structural brain features were more predictive of internalizing symptoms when reported by youths as compared to their parents. To support this comparative analysis, all models were trained and evaluated using the same procedures, including class weighting, threshold calibration, and hyperparameter tuning, to ensure fairness and consistency across informants and symptom domains.

Externalizing just add to appendix

Why not find?

Who else find?

What other data can use ?

Sample ? 50 50 class imbalance

References

Appendix

*Table A. Best Tuning Parameters*

|  |  |  |
| --- | --- | --- |
| Parameter | CBCL | BPM |
| Booster | Gbtree | Gbtree |
| Objective | Multi:softprob | Multi:softprob |
| Evaluation Metric | mlogloss | mlogloss |
| Max Depth | 3 | 3 |
| Min Child Weight | 1 | 5 |
| Eta | 0.1 | 0.1 |
| Gamma | 2 | 0 |
| Subsample | 0.6 | 0.9 |
| Colsample by Tree | 0.9 | 0.6 |
| Number of Classes | 3 | 3 |
| Best number of rounds | 50 | 50 |
| Multiclass Log Loss | 1.090 | 1.093 |