An Integrative Approach for Subtyping Mental Disorders Using Multi-modality Data

Yinjun Zhao[1,*], Yuanjia Wang[2,], and Ying Liu[3,***]**

[1,2,3]Department of Biotatistics, Columbia University, New York, New York, U.S.A.

[3]Statistical Innovation Group, AbbVie Inc. Lake Country, IL 6006, U.S.A.

*email: yz3503@cumc.columbia.edu

**email: yw2016@cumc.columbia.edu

***email: summeryingl@gmail.com

Summary: Mental disorders exhibit significant heterogeneity, posing challenges for accurate subtyping and diagnosis. Traditional clustering methods do not integrate multi-modal data, limiting their clinical applicability. This study introduces the Mixed INtegrative Data Subtyping (MINDS) method, a Bayesian hierarchical joint model designed to identify subtypes of Attention-Deficit/Hyperactivity Disorder (ADHD) and Obsessive-Compulsive Disorder (OCD) in adolescents using multi-modality data from the Adolescent Brain Cognitive Development (ABCD) Study. MINDS integrates clinical assessments and neuro-cognitive measures while simultaneously performing clustering and dimension reduction. By leveraging Polya-Gamma augmentation, we propose an efficient Gibbs sampler to improve computational efficiency and provide subtype identification. Simulation studies demonstrate superior robustness of MINDS compared to traditional clustering techniques. Application to the ABCD study reveals more reliable and clinically meaningful subtypes of ADHD and OCD with distinct cognitive and behavioral profiles. These findings show the potential of multi-modal model-based clustering for advancing precision psychiatry in mental health.

Key words: Mental disorder subtyping; Multi-modality data; Bayesian hierarchical model; Polya-Gamma augmentation; ABCD study

1. Introduction

Over the past four decades, considerable efforts have been devoted to applying unsupervised machine learning algorithms to dissect the heterogeneity of mental disorders and discover subtypes (Schnack, 2019; Marquand et al., 2016). Unfortunately, the lack of reproducibility in clustering algorithms has been a major obstacle for these methods to provide clinically useful subtyping for any mental disorder (Marquand et al., 2019). Traditional methods for mental disorder subtyping often rely on a single data type, such as clinical symptoms or neuroimaging measures, which may fail to accurately capture the complex, overlapping functional domains of mental health conditions or their underlying biological and psychological mechanisms (Kozak and Cuthbert, 2016).

The Research Domain Criteria (RDoC) initiative, launched by the National Institute of Mental Health, proposes to integrate biological and behavioral measures from various sources of analysis and different domains of functioning to create dimensional latent constructs of mental disorders and improve the precision of diagnosis and treatment (Insel et al., 2010). The Adolescent Brain Cognitive Development (ABCD) study is the largest and most comprehensive nationwide initiative in the United States dedicated to research on adolescent brain and behavioral development. By collecting multi-modal data, including brain imaging, behavioral tasks, clinical evaluations, and lifestyle factors, the ABCD study captures a broad view of the biological and environmental influences during the critical developmental years of adolescents. This rich study presents an unprecedented opportunity to address the heterogeneity of mental health conditions through unsupervised learning techniques, such as clustering. In this paper, we fill a gap in mental health research by introducing a novel method that integrates multi-modality data of different distribution types while simultaneously performing clustering and dimension reduction. Our goal is to develop integrative subtyping techniques to identify Attention-Deficit/Hyperactivity Disorder (ADHD) and Obsessive-

Compulsive Disorder (OCD) subtypes in adolescents in the ABCD study. To achieve this, we follow the RDoC paradigm to integrate clinical assessments from the Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS) modality and the neuro-cognitive measures from the NIH Cognitive Toolbox, along with cognitive task data and cortical thickness in selected brain regions.

Traditional clustering and subtyping methods cannot effectively integrate multi-modal data and identify meaningful subtypes. For example, distance-based approaches such as K-means and hierarchical clustering are widely applied in mental health research. However, they rely on predefined distance measures that are sensitive to noise and outliers to form clusters, which reduces their statistical robustness (Chen et al., 2022). Additionally, they lack the capability to estimate latent constructs, which is essential in RDoC to reveal patterns of measures of mental disorders and capture core mental functions (Lanza et al., 2007). Model-based subtyping methods, such as latent class analysis (Magidson and Vermunt, 2002), normative models (Marquand et al., 2016), and finite mixture models (McLachlan, 2000), are generally considered more robust than distance-based clustering because they use probabilistic frameworks to account for uncertainty and the distribution of data (Little, 2013), rather than arbitrary distance measures. By modeling the data under probabilistic frameworks, one can enhance reliability and reproducibility in identifying clinically meaningful subtypes. However, these models are typically designed to handle a single type of data distribution and cannot integrate multiple modalities or data types (e.g., neuroimaging, clinical symptoms, and behavioral data), limiting their applicability in studies that require a comprehensive, multi-modal analysis (Sinha et al., 2021).

To accommodate mixed-type data, distance-based subtyping methods can be adapted to using different similarity measures for each type (e.g., Euclidean for continuous variables, Dice or Jaccard for binary variables). For instance, Gower's distance computes similarity

across mixed data types and is widely used for this purpose (Gower, 1971). Another common approach for clustering mixed-type data is a two-step feature-based method. The first step involves feature transformation or dimension reduction, such as binary-to-continuous transformation (e.g., using logistic regression to convert binary variables into probability to estimate liability scores (McLachlan, 2000)), factor analysis with polychoric or polyserial correlations, or principal component analysis (PCA) with Gower's distance. This creates a compact, continuous representation of the mixed-type data. The second step applies conventional clustering to these transformed features. While these methods leverage the strengths of each data type, using only intermediate features instead of the original data may lead to the loss of shared information, reducing the efficiency of clustering. Latent class analysis (LCA) can jointly model both continuous and categorical/binary modalities; however, fitting LCA models may pose computational challenges because the joint likelihood (e.g., for combining multinomial and Gaussian distributions) often lacks a closed-form solution. Numeric integration is required in the estimation process, which becomes infeasible when the dimension of the latent constructs is high (Hagenaars and McCutcheon, 2002).

In this paper, we develop a Mixed INtegrative Data Subtyping (MINDS) method for subtyping mental disorders and clustering individuals by integrating measures across mixed-type modalities. We propose Bayesian hierarchical joint models with latent variables based on item response theory (IRT) and finite mixture modeling. For models integrating categorical data such as KSADS, we utilize Pólya-Gamma augmentation (Polson et al., 2013), which converts the logistic likelihood into a conditionally conjugate form. This transformation enables efficient Gibbs sampling and makes it feasible and efficient to handle the integration of diverse data types within a Bayesian framework. Our method tackles several challenges in clustering multi-modal data. First, MINDS can handle diverse data types, such as categorical and continuous, into a unified clustering model. Second, beyond assigning cluster membership,

MINDS estimates latent constructs for each individual to assist in explaining the unique characteristics of each cluster. This is essential for disease subtyping, identifying subject-specific disease risks, and developing targeted interventions. Third, our model extends the IRT to perform clustering and dimension reduction at the same time, which improves both interpretability and computational efficiency. Fourth, we propose to use Pólya-Gamma augmentation to facilitate Gibbs sampling, which is computationally more efficient than many other Bayesian methods. Lastly, by directly integrating original data instead of pre-processed features, MINDS captures subtle patterns and minimizes information loss, thereby enhancing clustering efficiency and robustness compared to the feature-based methods.

The rest of this paper is organized as follows. Section 2 introduces the proposed models and Pólya-Gamma augmentation for integrating categorical/binary modality, as well as the details of the algorithm. Section 3 presents simulation studies conducted to assess the consistency of the estimator and the efficiency of our method. Here, we also compare the performance of our approach to several traditional methods. In Section 4, we apply our methods to the ABCD study, subtyping adolescent participants based on symptoms and neuro-cognitive measures related to ADHD and OCD. We conclude with limitations and extensions in Section 5.

2. Model

2.1 *The IRT model*

The Item Response Theory (IRT; De Ayala, 2013) is a framework used to model the relationship between latent traits, i.e., unobserved characteristics or attributes, such as attention or cognitive control, and their manifestations in observed variables, such as responses to items in questionnaires or clinical instruments. The IRT is widely used in educational testing, psychology, psychometrics, and other fields where assessments and questionnaires are utilized

to measure underlying traits, including ability, attitude, or personality (Van Der Linden and Hambleton, 1997). It is a suitable model to capture the relationship between binary responses in the KSADS to identify latent traits of attention and cognitive control. MINDS addresses multi-modal integration by extending the IRT with Bayesian hierarchical joint modeling for clustering and integrating different data types. Below, we briefly introduce commonly used IRT models and how they are adapted in MINDS.

In the Rasch model, for a binary item, the probability that a person i with a latent trait θ_i (e.g., attention) endorsing item j with difficulty parameter δ_j is given by a logistic regression with random effects,

$$P(Y_{ij} = 1|\theta_i, \delta_j) = \frac{e^{\theta_i - \delta_j}}{1 + e^{\theta_i - \delta_j}}. \tag{1}$$

The relationship specified in Equation (1) is also called the item response function (IRF). To accommodate the non-unity slope of IRF, an additional parameter α is introduced to generalize the Rasch model to a one-parameter logistic (1PL) model such that

$$P(Y_{ij} = 1|\theta_i, \alpha, \delta_j) = \frac{e^{\alpha(\theta_i - \delta_j)}}{1 + e^{\alpha(\theta_i - \delta_j)}}.$$

When α varies across items, the 1PL model becomes a two-parameter logistic (2PL) model as

$$P(Y_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta_i - \delta_j)}}{1 + e^{\alpha_j(\theta_i - \delta_j)}}.$$

An item's α_j characterizes how well the item can differentiate among individuals located at different points on the continuum. Thus, α_j is also referred to as the item j's discrimination parameter.

The multidimensional IRT (MIRT) model extends the traditional uni-dimensional IRT to account for multiple latent traits, such as attention and cognitive control, as

$$P(Y_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \gamma_j) = \frac{e^{\sum_{l=1}^{L} \alpha_{jl}\theta_{il} + \gamma_j}}{1 + e^{\sum_{l=1}^{L} \alpha_{jl}\theta_{il} + \gamma_j}} = \frac{e^{\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i + \gamma_j}}{1 + e^{\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i + \gamma_j}}, \tag{2}$$

where $\boldsymbol{\theta}_i = \{\theta_{il}\}_{l=1,\cdots,L}$ is the L-dimensional vector of latent traits for person i, $\boldsymbol{\alpha}_j =$

$\{\alpha_{jl}\}_{l=1,\cdots,L}$ is the L-dimensional vector of discrimination parameters for item j, and $\gamma_j = -\sum_{l=1}^{L} \alpha_{jl}\delta_{jl}$ is the difficulty parameter for item j, also referred to as the item j's threshold. This is useful when test items are conceptualized to assess multiple underlying abilities or attributes.

2.2 *Latent mixture model with single modality*

The traditional MIRT models can only conduct dimension reduction for categorical data and do not accommodate mixture distributions. Here, we extend them to incorporate continuous variables and conduct clustering simultaneously. We first describe the modeling for binary, ordinal, or continuous measures, respectively, and then introduce how to integrate different data types.

Let Y_{ij} denotes the binary modality's jth item in a clinical instrument from the ith subject, where $i = 1, \cdots, N_b$, and $j = 1, \cdots, N_d$. We propose a latent mixture model inspired by MIRT for binary modality as

$$Y_{ij}|X, a_j, Z_i, b_i, V_{.j} \sim \text{Bernoulli}\left(\frac{e^{(Z_i^T X + b_i^T)V_{.j} - a_j}}{1 + e^{(Z_i^T X + b_i^T)V_{.j} - a_j}}\right), \tag{3}$$

where X is a $N_c \times N_t$ matrix representing the N_c cluster centers, each of which is a N_t-dimensional vector. Here, Z_i is an indicator of the ith subject's unobserved cluster membership assumed to follow a categorical distribution, $p(Z_i = k) = \theta_k$, and b_i is the ith subject's latent construct, e.g., cognitive control ability that deviates from his or her cluster mean ability. Furthermore, $V_{.j}$ is an N_t-dimensional loading vector on latent construct for the jth item representing how well the items can differentiate among individuals at different points of latent constructs, and a_j represents the jth item's threshold. Note that the above model reduces to the MIRT model when there is a single cluster. A similar method under a cumulative model can be developed for ordinal measures as

$$P(Y_{ij} \leqslant \gamma_t | X, a_{jt}, Z_i, b_i, V_{.j}) = \frac{e^{(Z_i^T X + b_i^T)V_{.j} - a_{jt}}}{1 + e^{(Z_i^T X + b_i^T)V_{.j} - a_{jt}}},$$

where $\gamma_t = \sum_{s=1}^{t} \pi_s, \quad t = 1, \cdots, T_j$, and π_t represents the probability that Y_{ij} belongs to tth category, and a_{jt} is the threshold parameter for jth item belonging to tth category. For modeling N_d-dimensional continuous measures Y_{ij}, we propose a multivariate Gaussian model as

$$Y_{ij}|X, a_j, Z_i, b_i, U_{.j} \sim \mathrm{MVN}((Z_i^T X + b_i^T)U_{.j} - a_j, \sigma_j^2).$$

The interpretations of $a_j, Z_i, b_i, U_{.j}$ are similar to $X, a_j, Z_i, b_i, V_{.j}$ in model (3). In these models, individuals in different subtypes have distinct mean latent traits (i.e., cluster center X) while sharing other parameters such as discriminant and threshold values.

Since the binary modality's distribution does not belong to the Gaussian family, the posterior distribution is not conjugate to the prior, making Gibbs sampling inapplicable. Pólya-Gamma augmentation has been proposed to yield a Gibbs sampler for the Bayesian logistic model (Polson et al., 2013). A random variable Y has a Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$, denoted as $Y \sim \mathrm{PG}(b, c)$, if

$$Y \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)},$$

where $g_k \sim \mathrm{Ga}(b, 1)$ are independent gamma random variables, and $\overset{D}{=}$ indicates equality in distribution. Let $p(\omega)$ denote the density of the random variable $\omega \sim \mathrm{PG}(b, 0)$, $b > 0$. Then the following integral identity holds for all $a \in \mathbb{R}$:

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2}\, p(\omega)\, d\omega, \tag{4}$$

where $\kappa = a - b/2$. Moreover, the conditional distribution $p(\omega|\psi)$ is also in the Pólya-Gamma class, i.e.,

$$(\omega \,|\, \psi) \sim \mathrm{PG}(b, \psi). \tag{5}$$

The properties of (4) (5) suggest a simple strategy for Gibbs sampling across a wide class of binomial models, that is, Gaussian draws for the main parameters and Pólya-Gamma draws for a single layer of latent variables. Under model (3), the conditional distribution of Y given

all the rest of parameters is

$$P(Y|X, Z_i, b_i, a_j, V_{.j}) = \prod_{i=1,\cdots,N_b} \prod_{j=1,\cdots,N_d} \frac{e^{\psi_{ij} y_{ij}}}{1 + e^{\psi_{ij}}},$$

by Equation (4), the conditional distribution can be expressed as

$$P(Y|X, Z_i, b_i, a_j, V_{.j}) = \frac{1}{2} \prod_i \prod_j \exp\left\{(y_{ij} - 1/2)\psi_{ij}\right\} E_{\omega_{ij} \sim PG(1,0)} \exp\left\{-\psi_{ij}^2 \omega_{ij}/2\right\}, \tag{6}$$

where $\psi_{ij} = (Z_i^T X + b_i^T)V_{.j} - a_j$.

Given ω_{ij}, the expectation in Equation (6) is $\exp\left\{-\psi_{ij}^2 \omega_{ij}/2\right\}$, which leads the conditional probability of Y_{ij} in model (3) follows the Gaussian distribution.

$$P(Y|\omega_{ij}, X, Z_i, b_i, a_j, V_{.j}) = \frac{1}{2} \prod_i \prod_j \exp\left\{(y_{ij} - 1/2)\psi_{ij}\right\} \exp\left\{-\psi_{ij}^2 \omega_{ij}/2\right\}. \tag{7}$$

By the property of Pólya-gamma distribution in Equation (5), we have

$$\omega_{ij}|X, Z_i, b_i, a_j, V_{.j} \sim PG(1, \psi_{ij}).$$

In this sense, all parameters' posterior distributions are conjugate to their prior distributions, so the Gibbs sampling can be adopted for parameter estimation. Details of the posterior distributions are in section 2.4.

2.3 *Latent mixture model with multi-modalities*

Model (3) only handles a single modality. Here, we extend the single-modality model to handle more complex multi-modality data using discrete clinical and continuous behavioral measures. We propose a joint model that integrates binary and continuous modalities as follows

$$Y_{i.}^{(1)}|X, b_i, V, Z_i, a^{(1)} \sim \text{Bernoulli}\left(\frac{e^{(Z_i^T X + b_i^T)V - a^{(1)}}}{1 + e^{(Z_i^T X + b_i^T)V - a^{(1)}}}\right), \tag{8}$$

$$Y_{i.}^{(2)}|X, b_i, U, Z_i, a^{(2)}, \Sigma \sim \text{Normal}\left((Z_i^T X + b_i^T)U - a^{(2)}, \Sigma\right), \tag{9}$$

where X is a matrix of cluster centers shared by these two modalities, Z_i is subject ith's cluster membership, b_i is N_t-dimensional subject's latent construct, V and U are $N_t \times N_{d_1}$-dimensional and $N_t \times N_{d_2}$-dimensional loading matrix on cluster center X for binary and

continuous modality, respectively. N_{d_1} represents the number of items in binary modality, and N_{d_2} represents the number of measures in continuous modality. $a^{(1)}$ and $a^{(2)}$ represent item difficulty for two different modalities. For continuous modality, Σ is a diagonal covariance matrix with diagonal entries σ_j^2, $j = 1, \cdots, N_{d_2}$. In this model, we assume the two data-type modalities share the cluster center X, and each subject belongs to the same cluster under different modalities.

We use a Bayesian hierarchical algorithm to estimate the parameters. We assume the following prior distributions for the parameters in (8) and (9) as that Z_i follows categorical distribution with hyperparameter θ; b_i follows $N(0, \mathcal{I}(\sigma_b^2))$; X, V, U are from multivariate normal distributions, $N(\mu_X, \mathcal{I}(\sigma_x^2))$, $N(\mu_V, \mathcal{I}(\sigma_v^2))$, and $N(\mu_U, \mathcal{I}(\sigma_u^2))$, respectively; $a^{(1)}, a^{(2)}$ are from multivariate normal, $N(0, \mathcal{I}(\sigma_a^2))$, respectively. We assume θ follows Dirichlet distribution with parameters of equal weight $1/N_c$, and σ_b^2 and σ_j^2 follow Inverse Gamma distribution $IG(\alpha_b, \beta_b)$ and $IG(\alpha_j, \beta_j)$, respectively.

The conditional distribution of Y given all the rest parameters in the joint model (8)-(9) is

$$P(Y|X, Z_i, b_i, a^{(1)}, a^{(2)}, V, U) = \prod_{i=1,\cdots,N_b} \prod_{j=1,\cdots,N_{d_1}} p\left(y_{ij}^{(1)}|rest\right) \cdot \prod_{i=1,\cdots,N_b} \prod_{j=1,\cdots,N_{d_2}} p\left(y_{ij}^{(2)}|rest\right).$$

We use a similar method as for model (3), which introduces a latent variable ω following the Pólya-Gamma distribution such that $\omega_{ij}|X, Z_i, b_i, a_j^{(1)}, V_{.j} \sim PG(1, \psi_{ij})$, where $\psi_{ij} = (Z_i^T X + b_i^T)V_{.j} - a_j^{(1)}$. Similar to equation (7), the conditional distribution of $Y_{ij}^{(1)}$ given ω and all rest parameters in model (8) is from a Gaussian family. In this sense, the posterior distributions of all the parameters $X, b_i, a_j^{(1)}, a_j^{(2)}, V_{.j}, U_{.j}, \sigma_j^2$ in the joint model are conjugate to their prior distributions. The conditional probability that the ith subject belongs to the kth cluster, given the rest of the parameters, denoted as π_{ik}, is expressed as:

$$\pi_{ik} = \prod_{j=1,\cdots,N_{d_1}} \exp\left(-\frac{\omega_{ij}}{2}\left[(X_k + b_i^T)V_{.j} - \frac{\kappa_{ij}}{\omega_{ij}} - a_j^{(1)}\right]^2\right) \cdot$$

$$\prod_{j=1,\cdots,N_{d_2}} \frac{1}{\sigma_j} \exp\left(-\frac{1}{2\sigma_j^2}\left[(X_k + b_i^T)U_{.j} - y_{ij}^{(2)} - a_j^{(2)}\right]^2\right) \cdot P(Z_i = k|\theta) \cdot P(\theta),$$

which indicates that the posterior distribution of Z_i is from Multi$(1, \pi_i)$, $\quad \pi_i = (\pi_{i1}, \cdots, \pi_{iN_c})$. All the parameters' posterior distributions can be expressed in closed forms, so the Gibbs sampling can be adopted for parameter estimation. Details of the posterior distributions are in section 2.4.

This model has metric indeterminacy in the sense that we do not have a metric for V and U with intrinsic origin or unit. In addition, the model has a rotational indeterminacy in the sense that the axes of X that represent latent variables are free to rotate about their origin because there is no external reference to fix their orientation. To address these issues and ensure identifiability, we rescale b_i to have a mean of zero and a variance of 0.01. Also, we set X_{11} to be one and estimate X, U and V by solving equation system that $X * V = X_{sample} * V_{sample}$, $X * U = X_{sample} * U_{sample}$, where $X_{sample}, V_{sample}, U_{sample}$ are sampled values in each iteration. The equation system has unique solution when the number of parameters, $N_c * N_t + N_t * (N_{d_1} + N_{d_2})$ is not greater than the number of equations, $N_c * (N_{d_1} + N_{d_2})$, i.e. $N_c * N_t \leqslant (N_c - N_t) * (N_{d_1} + N_{d_2})$.

2.4 *Algorithm*

We derive Gibbs sampling steps for the joint model (8)-(9) in Algorithm 1. This algorithm can be easily modified to accommodate the single modality model (3), a simpler version of the joint model. Information Criterion (IC), a selection information criterion for Baysesian model (Ando, 2011), which avoids the problems of the over-fitting of the DIC (Spiegelhalter et al., 2014) is used to determine the optimal number of clusters N_c.

$$IC = -2E_{\theta|y}[logf(y|\theta)] + 2P_D,$$

where P_D is defined as difference between the posterior mean of deviance and the deviance estimated at the posterior mean of the parameters, $2[logf(y|\hat{\theta}_n))] - 2E_{\theta|y}[logf(y|\theta)]$, and $\hat{\theta}_n$ is the posterior mean. The stationary trace plot is used to examine convergence.

Algorithm 1: Gibbs sampling steps for joint model

Data: Binary items $y^{(1)}$, continuous measures $y^{(2)}$

Result: Estimators $\boldsymbol{\lambda} = (X, Z, b, a^{(1)}, a^{(2)}, V, U, \theta, \Sigma)$

Initialize $\boldsymbol{\lambda}$;

repeat

$\omega_{ij}|X, Z_i, b_i, a_j^{(1)}, V_{.j} \sim PG(1, (Z_i^T X + b_i^T)V_{.j} - a_j^{(1)})$;

$X_t|\text{rest} \sim N\left(V_{X_t}\left(\sum_{i=1}^{N_b} Z_i U_t \Sigma \Phi_i^{(t)} + \sum_{i=1}^{N_b} Z_i V_t \Omega_i \phi_i^{(t)} + B_t^{-1}\mu_{X_t}\right), V_{X_t}\right)$;

$Z_i|\text{rest} \sim \text{categorical}(\pi_i)$;

$V_{tj}|\text{rest} \sim N(B_{tj}Q_{tj}, B_{tj})$;

$U_{tj}|\text{rest} \sim N(D_{tj}R_{tj}, D_{tj})$;

$b_t|\text{rest} \sim N(\mu_{b_t}, V_{b_t})$;

$a_j^{(1)}|\text{rest} \sim N\left(V_{a_j^{(1)}} \sum_{i=1}^{N_b} \sigma_j^{-2} r_{ij}^{(1)}, V_{a_j^{(1)}}\right)$;

$a_j^{(2)}|\text{rest} \sim N\left(V_{a_j^{(2)}} \sum_{i=1}^{N_b} \omega_{ij} r_{ij}^{(2)}, V_{a_j^{(2)}}\right)$;

$\theta|\text{rest} \sim Dir(\sum_{i=1}^{N_b} \mathbb{1}_{\{Z_i=k\}} + \frac{1}{Nc}), k = 1, \cdots, Nc$;

$\sigma_b^2|\text{rest} \sim IG(\frac{N_b}{2} + \alpha_b, \frac{1}{2}\sum_{i=1}^{N_b} b_i^2 + \beta_b)$;

$\sigma_j^2|\text{rest} \sim IG(\frac{N_b}{2} + \alpha_j, \frac{1}{2}\sum_{i=1}^{N_b} M_{ij}^2 + \beta_j)$, where $M_{ij} = y_{ij}^{(2)} - (Z_i^T X + b_i^T)U_{.j} - a_j^{(2)}$;

until $\boldsymbol{\lambda}$ *converges*;

In Algorithm 1, for step 4,

$$
\begin{aligned}
\Omega_i &= \mathcal{I}(\omega_{i1}, \cdots, \omega_{iN_{d_1}}),\ \Sigma = \mathcal{I}(\sigma_1^{-2}, \cdots, \sigma_{N_{d_2}}^{-2}), \\
U_t &= (U_{t1}, \cdots, U_{tN_{d_2}})^T,\ V_t = (V_{t1}, \cdots, V_{tN_{d_1}})^T, \\
X_t &= (X_{1t}, \cdots, X_{N_c t})^T, \\
\phi_{ij}^{(t)} &= \frac{\kappa_{ij}}{\omega_{ij}} + a_j^{(1)} - (Z_i^T X \cdot \mathbb{1}_{\{s\neq t\}}^T + b_i^T) V_{.j}, s = 1, \cdots, N_t, \\
\Phi_{ij}^{(t)} &= y_{ij}^{(2)} + a_j^{(2)} - (Z_i^T X \cdot \mathbb{1}_{\{s\neq t\}}^T + b_i^T) U_{.j}, s = 1, \cdots, N_t, \\
V_{X_t}^{-1} &= \sum_{i=1}^{N_b} Z_i V_t \Omega_i V_t^T Z_i^T + B_t^{-1},\ B_t = \mathcal{I}_{N_c}(\sigma_x^{-2}).
\end{aligned}
$$

For step 5, the conditional probability of ith subject belongs to the kth cluster center, i.e., $Z_i^T = \mathbb{1}_{\{s=k\}}, s = 1, \cdots, N_c$, given the rest parameters is

$$
\begin{aligned}
\pi_{ik} = P(Z_i = k|\theta) = \prod_{j=1}^{N_{d_1}} \exp\left(-\frac{\omega_{ij}}{2}\left[(X_k + b_i^T)V_{.j} - \frac{\kappa_{ij}}{\omega_{ij}} - a_j^{(1)}\right]^2\right) \cdot \\
\prod_{j=1}^{N_{d_2}} \frac{1}{\sigma_j} \exp\left(-\frac{1}{2\sigma_j^2}\left[(X_k + b_i^T)U_{.j} - y_{ij}^{(2)} - a_j^{(2)}\right]^2\right).
\end{aligned}
$$

For steps 6 and 7,

$$
\begin{aligned}
B_{tj}^{-1} &= \sigma_v^{-2} + \sum_{i=1}^{N_b} \omega_{ij} A_{it}^2,\ Q_{tj} = \frac{\mu_v}{\sigma_v^2} + \sum_{i=1}^{N_b} \omega_{ij} A_{it} \psi_{ij} \\
D_{tj}^{-1} &= \sigma_u^{-2} + \sum_{i=1}^{N_b} \sigma_j^{-2} A_{it}^2,\ R_{tj} = \frac{\mu_u}{\sigma_u^2} + \sum_{i=1}^{N_b} \sigma_j^{-2} A_{it} \Psi_{ij},
\end{aligned}
$$

where

$$
\begin{aligned}
A_{it} &= Z_i^T X_t + b_{it}, \\
\psi_{ij}^{(t)} &= \frac{\kappa_{ij}}{\omega_{ij}} + a_j^{(1)} - (Z_i^T X + b_i^T) \cdot \mathbb{1}_{\{s\neq t\}}^T V_{.j}, s = 1, \cdots, N_t, \\
\Psi_{ij}^{(t)} &= y_{ij}^{(2)} + a_j^{(2)} - (Z_i^T X + b_i^T) \cdot \mathbb{1}_{\{s\neq t\}}^T U_{.j}, s = 1, \cdots, N_t.
\end{aligned}
$$

For step 8,

$$\begin{aligned}
\mu_{b_t} &= \left\{ V_{b_{it}} \left(\sum_{j=1}^{N_{d_1}} V_{tj} C_{ij}^{(t)} \omega_{ij} + \sum_{j=1}^{N_{d_2}} U_{tj} D_{ij}^{(t)} \sigma_j^{-2} \right) \right\}_{i=1,\cdots,N_b}, \\
V_{b_t}^{-1} &= \mathcal{I}\{\sigma_{b_t}^{-2} + \sum_{j=1}^{N_{d_1}} V_{tj}^2 \omega_{ij} + \sum_{j=1}^{N_{d_2}} U_{tj}^2 \sigma_j^{-2}\}, \\
C_{ij}^{(t)} &= \frac{\kappa_{ij}}{\omega_{ij}} + a_j^{(1)} - (Z_i^T X + b_i^T \cdot \mathbb{1}_{\{s \neq t\}}^T) V_{.j}, \\
D_{ij}^{(t)} &= y_{ij}^{(2)} + a_j^{(2)} - (Z_i^T X + b_i^T \cdot \mathbb{1}_{\{s \neq t\}}^T) U_{.j}, s = 1, \cdots, N_t.
\end{aligned}$$

For steps 9 and 10,

$$\begin{aligned}
r_{ij}^{(1)} &= (Z_i^T X + b_i^T) V_{.j} - \frac{\kappa_{ij}}{\omega_{ij}}, \ V_{a_j^{(1)}}^{-1} = \sum_{i=1}^{N_b} \omega_{ij} + \sigma_a^{-2}, \\
r_{ij}^{(2)} &= (Z_i^T X + b_i^T) U_{.j} - y_{ij}^{(2)}, \ V_{a_j^{(2)}}^{-1} = \sigma_a^{-2} + N_b \sigma_j^{-2}.
\end{aligned}$$

3. Simulation studies

We conducted extensive simulation studies to evaluate the performance of our proposed algorithm for the joint model. This involved examining the consistency of latent variable estimates, calculating test errors on testing data when the cluster membership is known, and comparisons with alternative methods.

We included three groups of settings with a combination of different numbers of binary items and continuous measures, (10, 10), (20, 10), (30, 10), where the first element represents the number of binary items and the second element represents the number of continuous measures. For example, (10, 20) represents $N_{d_1} = 10$, $N_{d_2} = 20$. Three latent constructs were considered for the cluster center X and the ith subject's latent construct b_i, i.e., $N_t = 3$. The elements of true cluster center X and subject-specific ability b_i were randomly drawn from a uniform distribution $U(0, 2)$ and a Gaussian distribution $N(0, 0.01)$, respectively. Subjects were generated from four clusters, i.e., $N_c = 4$, with cluster membership weights θ to be (0.3, 0.3, 0.2, 0.2), based on which we generated true membership Z_i from Categorical(θ)

for subject i. The item difficulty parameters $a_j^{(1)}$ for binary items and $a_j^{(2)}$ for continuous measures were randomly drawn from U(-0.5, 0.5) and U(-5, 5), respectively. All elements of the loading matrix V for binary items and U for continuous measures were independently randomly drawn from uniform(0, 2). All the prior distributions were non-informative, and all the estimators' initial values were random. We conducted 200 replications for each group setting with 1,000 and 4,000 subjects, respectively, which are in the range of the sample sizes for various groups in the motivating ABCD study.

As shown in Web Figures 1-4, the medians of the estimated values for cluster center X, loading matrix U and V, and item difficulty $a^{(1)}$ and $a^{(2)}$ are close to their true values except for $U_{14}, V_{15}, V_{16}, V_{18}$ and V_{110}. As the sample size increases from 1000 to 4000, the variance of the estimators decreases. We compared the root mean squared errors (RMSE) and bias of all parameters for 1,000 sample size and 4,000 sample size. As shown in Table 1, all parameters' RMSEs decrease as the sample size increases from 1,000 to 4,000. The other two settings have similar performance, and results are omitted.

We simulated a testing dataset of size 10,000 using the same true values of parameters X, U, V, $a^{(1)}$, and $a^{(2)}$ as used in generating the training data. For the subject-specific parameters Z_i and b_i, the same distributions with the same parameters θ and σ_b as in the training data were used in generating the testing data, respectively. To get the predicted cluster membership, we plugged the estimated values of X, U, V, $a^{(1)}$, and $a^{(2)}$ from training data, and iterated the remaining unknown parameters ω, Z and b_i until converge. The estimated Z returns the predicted cluster membership. Test errors were calculated using three metrics to evaluate the clustering performance. The three metrics are (a) Bayes error rate, which is defined as the expectation of the event that the predicted category (the one having maximum membership weight) is not equal to the true category; (b) Classification error, the proportion that the predicted category is not equal to the true category; (c) Jaccard distance, defined

as the one minus the size of the intersection divided by the size of the union of two label sets. For all three metrics, smaller scores imply better performance.

We compared the results of the MINDS method with four other alternative approaches: K-means clustering, Hierarchical clustering (shorted for Hclust), two-step K-means clustering, and two-step Hclust. In the two-step K-means, we first conducted factor analysis based on the polychoric correlation of the testing data, and in the second step, we used K-means to cluster the factor scores into N_c clusters. The two-step Hclust differs from the two-step K-means in that the second step uses Hclust to cluster the factor scores. As shown in Table 2 and Web Table 1, the test errors of MINDS are lower than all the other four methods in all three metrics, which shows that the performance of MINDS is satisfactory.

Computing time for one iteration of a joint model with 2,000 samples when the number of binary items varies from 100, 200, and 400 items were 0.7, 1.2, and 2.2 seconds, respectively.

4. Disease Subtyping in ABCD Study

We applied our proposed method, MINDS, to subtype individuals by integrating their ADHD- and OCD-related symptoms and neuro-cognitive measures in the ABCD study. The binary modalities include 18 KSADS ADHD-related symptoms and 11 KSADS OCD-related symptoms. Three neuro-cognitive scores, the Nback, an SST, and the Flanker Effect, and seven NIH Cognition Toolbox scores were included in the continuous modality. Nback Score and SST Score are defined as the standard deviation of reaction time divided by the mean reaction time in the EN-back Task and the Stop Signal Task, respectively. The Flanker effect is the mean reaction time for incongruent trials minus that for congruent trials in the Flanker task. These three scores show the children's ability in attention, also related to response inhibition and working memory (Casey et al., 2018; Siemann et al., 2018). The seven NIH Cognition Toolbox scores include Picture Vocabulary, Flanker, List Sort, Card Sort, Picture Sequence, Reading, and Pattern. These measures examine subjects' ability in

response inhibition, working memory, attention, language, reading, and so on (Thompson et al., 2019). See details in Table 3.

We considered subjects at baseline and excluded 6,038 individuals without any ADHD/OCD-related symptoms and 1,652 individuals with missing measures, which yielded 2,427 subjects. In the model fitting, a negative sign is applied to each of the three variability scores, with a larger measure indicating a higher ability in the child, consistent with the seven NIH scores. N_c=7 is suggested by the Elbow plot of IC, as shown in Web Figure 8; however, when N_c=7, two clusters are similar. Thus, for parsimony and interpretability, we choose N_c=6.

The six identified groups are ADHD&OCD I, ADHD&OCD II, ADHD I, ADHD II, OCD, and MILD. The six clusters show distinct differences in terms of the distributions of ADHD/OCD symptoms and neuro-cognitive measures. As shown in Figure 1, both the ADHD&OCD I and ADHD&OCD II groups exhibit a high concentration of symptoms related to both ADHD and OCD. However, the ADHD&OCD I group shows a slightly higher density of symptoms. The ADHD I and ADHD II groups both show a high density of ADHD symptoms. However, in the ADHD I group, symptoms related to inattention are more prominent, while in the ADHD II group, symptoms related to hyperactivity and impulsivity are more pronounced. The OCD group has dense OCD symptoms; the symptoms in the MILD group are less severe.

Although the ADHD&OCD I group and the ADHD&OCD II group have similar patterns in the distribution of ADHD and OCD symptoms to other groups, the characteristics of their neuro-cognitive measures are quite different, as shown in Figure 1. ADHD&OCD II group has the best performance in List Sort, Pattern, Picture, SST Score, and Card Sort compared to other groups. In contrast, the ADHD&OCD I group has the best performance in Picture Vocabulary, Reading, Flanker, and FlankerEffect but has the worst performance in SST Score, Card Sort, and Nback Score. The distributions of the neuro-cognitive measures

in the ADHD I group and the ADHD II group are different. Compared to other groups, the ADHD II group performs worse on Nback Score and Reading and has the worst performance on Flanker, Flanker Effect, SST Score, and Picture and List Sort, while measures in the ADHD I group perform neutral except for List Sort, which is similar to the MILD group.

To compare MINDS with DSM diagnosis, which is the standard criteria for diagnosing ADHD and OCD, we calculated the Calinski-Harabasz Index (CH) as

$$CH = \frac{B}{W} * \frac{N-K}{K-1},$$

where B is the sum of squares between clusters, W is the sum of squares within clusters, N is the total number of data points, K is the number of clusters. A higher score of CH indicates larger between-cluster variation and smaller within-cluster variation. The MINDS method is better in capturing homogeneity and distinguishing heterogeneity between clusters than clustering by ADHD/OCD DSM diagnosis, which groups the participants into ADHD&OCD, ADHD, OCD, and None. The CHs of neuro-cognitive measures for the MINDS method vs the ADHD/OCD DSM diagnosis are 2.2 vs 1.5, where MINDS is 50% better. The CHs of symptoms for the MINDS method vs ADHD/OCD diagnosis are 322 vs 194, where MINDS is 70% better.

Next, we compare MINDS with several other traditional clustering approaches, including K-means, hierarchical clustering (Hclust), two-step K-means, and two-step Hclust. In the two-step K-means, we first conducted factor analysis based on the polychoric correlation of the testing data, and in the second step, we used K-means to cluster the factor scores into N_c clusters. The two-step Hclust differs from the two-step K-means in that the second step uses Hclust to cluster the factor scores. We compared the methods using JCC similarity values (1 - Jaccard distance), where higher values represent greater robustness. The mean (sd) values of the JCC similarity are MINDS: 0.4 (0.05), K-means: 0.19 (0.014), Hclust: 0.3 (0), Two-step K-Means:0.29 (0.016), Two-step Hclust:0.24 (0), based on 100 non-parametric

bootstrap samples. Our proposed methods have much higher JCC similarity values, which suggest a greater robustness.

To examine the external validity of the subtypes from MINDS, Figure 3 shows the distribution of self-reported grades by parents among these subtypes in the third follow-up year. Our analysis indicates that the MILD and OCD groups exhibit the highest academic achievement. In contrast, the ADHD&OCD II group, ADHD I, and ADHD II groups demonstrate the lowest academic performance. These results were further confirmed by results (as shown in Figure 2) of the proportional odds logistic regression of the third-year self-reported grades on participants' subtypes adjusted by age, gender, race, education, family income, etc. The ADHD&OCD II group (OR: 1.92, 95% CI: 1.1-3.35), the ADHD I group (OR: 1.86, 95% CI: 1.36-2.54), and the ADHD II group (OR: 1.76, 95% CI:1.23-2.52) are all significantly associated with poor academic performance compared to the MILD group, whereas the OCD group (OR: 0.85, 95% CI: 0.62-1.15) and the ADHD&OCD I group (OR: 1.62, 95% CI: 0.84-3.08) have no significant difference from the MILD group in the academic performance.

These findings shed new light on previous research suggesting that OCD symptoms are associated with academic underperformance (Fischer-Terworth, 2013; Pérez-Vigil et al., 2018). Interestingly, our results show that the OCD group demonstrates good academic performance. On the other hand, the ADHD I and ADHD II groups, characterized by significant ADHD symptoms, exhibit the poorest academic outcomes. This suggests that ADHD symptoms, rather than OCD symptoms, may directly contribute to lower grades. Additionally, neuro-cognitive factors could also play a role in academic achievement. Though the ADHD&OCD I and ADHD&OCD II groups both have dense ADHD-related symptoms, the ADHD&OCD I group has the highest Reading score among all groups, which may serve as a compensatory factor of the ADHD-related symptoms in enhancing their overall academic performance. The poor academic performance of the ADHD&OCD II, ADHD I, and ADHD

II groups highlights the need for increased social awareness. The DSM diagnosis subtyping also finds one significant group, its ADHD group (OR:1.62, 95%CI:1.14-2.31), associated with academic performance as shown in Figure 2. However, the ADHD I and ADHD II groups by the MINDS method have higher ORs, and this method integrating both symptom and neuro-cognitive measures shows that these two subgroups are most at risk for poor performance and require psychological and educational interventions.

The MINDS subtyping results additionally reveal that the ADHD II group has borderline significantly higher odds of alcohol use compared to the MILD group (OR:1.55, 95%CI:0.98-2.45), whereas the DSM subtyping did not find any potentially risky groups. The ADHD II group exhibits a greater concentration of hyperactivity and impulsivity-related symptoms, which may explain this association. See Web Figure 5 for the proportional odds regression results.

5. Discussions

Disease subtyping is essential for achieving personalized medicine, particularly in mental health, due to the high variability of disorders. The ABCD study offers a comprehensive dataset for investigating multi-modal integration techniques to understand the heterogeneity of adolescent development. We introduce MINDS, a Bayesian hierarchical joint model with latent variables, to integrate clinical symptoms and neuro-cognitive measures, utilizing Pólya-Gamma augmentation for posterior approximation, which facilitates Gibbs sampling. Extensive simulations are performed to assess the consistency of the estimators and the clustering efficiency of our method. We applied MINDS to the ABCD study, where our method demonstrates greater robustness compared to traditional clustering approaches, providing a more reliable classification of clinical subtypes. By identifying meaningful distinctions among ADHD and OCD subgroups, our findings contribute to the development of

future psychological and educational interventions tailored to specific clinical presentations in mental health.

There are several challenges of MINDS. The estimation of some elements of the loading matrix can have large variances when the initial values are non-informative and the dimensions of binary items and continuous outcomes are high. Using informative initial values, for example, by fitting single-modality models, may reduce variability. As the number of items within each modality increases, computational demands increase, posing challenges in handling high-dimensional data. This can significantly slow down clustering and inference, particularly when there are many variables across modalities. Exploring variational inference methods may mitigate these computational challenges in a Bayesian approach. Determining the number of clusters may require both data-driven methods (e.g., DIC or IC) and domain knowledge (e.g., clinical interpretability).

Lastly, several extensions may be of interest. Introducing a penalty to the loading matrix could help manage high-dimensional data more effectively. Additionally, our method could be adapted to integrate other data types, such as zero-inflated or complex-omics data, making it even more applicable to diverse datasets in precision medicine. Validating our identified subtypes in an independent study such as All*of*Us study (All of Us Research Program Investigators, 2019) or other clinical samples would be important.

Acknowledgments

Data used in the preparation of this manuscript were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (DOI: 10.15154/z563-zd24), held in the National Institute of Mental Health (NIMH) Data Archive (NDA). NDA is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in mental health. This manuscript reflects the views of the authors

and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDA.

This work is funded in part by MH123487, NS073671 and TL1TR001875. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This publication was neither originated nor managed by AbbVie, and it does not communicate the results of AbbVie-sponsored Scientific Research. Thus, it is not in the scope of the AbbVie Publication Procedure (PUB-100).

References

All of Us Research Program Investigators (2019). The "all of us" research program. *New England Journal of Medicine* **381,** 668–676.

Ando, T. (2011). Predictive bayesian model selection. *American Journal of Mathematical and Management Sciences* **31,** 13–38.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. (2018). The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience* **32,** 43–54.

Chen, Z. S., Galatzer-Levy, I. R., Bigio, B., Nasca, C., Zhang, Y., et al. (2022). Modern views of machine learning for precision psychiatry. *Patterns* **3,**.

De Ayala, R. J. (2013). *The theory and practice of item response theory.* Guilford Publications.

Fischer-Terworth, C. (2013). Obsessive-compulsive disorder in children and adolescents: Impact on academic and psychosocial functioning in the school setting. *Life Span and Disability* **16,** 127–55.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* pages 857–871.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Kozak, M. J. and Cuthbert, B. N. (2016). The nimh research domain criteria initiative: background, issues, and pragmatics. *Psychophysiology* **53,** 286–297.

Lanza, S. T., Collins, L. M., Lemmon, D. R., and Schafer, J. L. (2007). Proc lca: A sas procedure for latent class analysis. *Structural equation modeling: a multidisciplinary journal* **14,** 671–694.

Little, T. D. (2013). *The Oxford handbook of quantitative methods in psychology: Vol. 2: statistical analysis*, volume 2. OUP USA.

Magidson, J. and Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian journal of marketing research* **20,** 36–43.

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological psychiatry* **80,** 552–561.

Marquand, A. F., Wolfers, T., and Dinga, R. (2019). Phenomapping: methods and measures for deconstructing diagnosis in psychiatry. *Personalized Psychiatry: Big Data Analytics in Mental Health* pages 119–134.

McLachlan, G. (2000). Finite mixture models. *A wiley-interscience publication* .

Pérez-Vigil, A., de la Cruz, L. F., Brander, G., Isomura, K., Jangmo, A., Feldman, I., Hesselmark, E., Serlachius, E., Lázaro, L., Rück, C., et al. (2018). Association of

obsessive-compulsive disorder with objective indicators of educational attainment: a nationwide register-based sibling control study. *JAMA psychiatry* **75,** 47–55.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association* **108,** 1339–1349.

Schnack, H. G. (2019). Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia research* **214,** 34–42.

Siemann, J., Herrmann, M., and Galashan, D. (2018). The effect of feature-based attention on flanker interference processing: An fmri-constrained source analysis. *Scientific reports* **8,** 1580.

Sinha, P., Calfee, C. S., and Delucchi, K. L. (2021). Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. *Critical care medicine* **49,** e63–e79.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76,** 485–493.

Thompson, W. K., Barch, D. M., Bjork, J. M., Gonzalez, R., Nagel, B. J., Nixon, S. J., and Luciana, M. (2019). The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: Findings from the abcd study's baseline neurocognitive battery. *Developmental cognitive neuroscience* **36,** 100606.

Van Der Linden, W. J. and Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory*, pages 1–28. Springer.

Supporting Information

Appendix

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

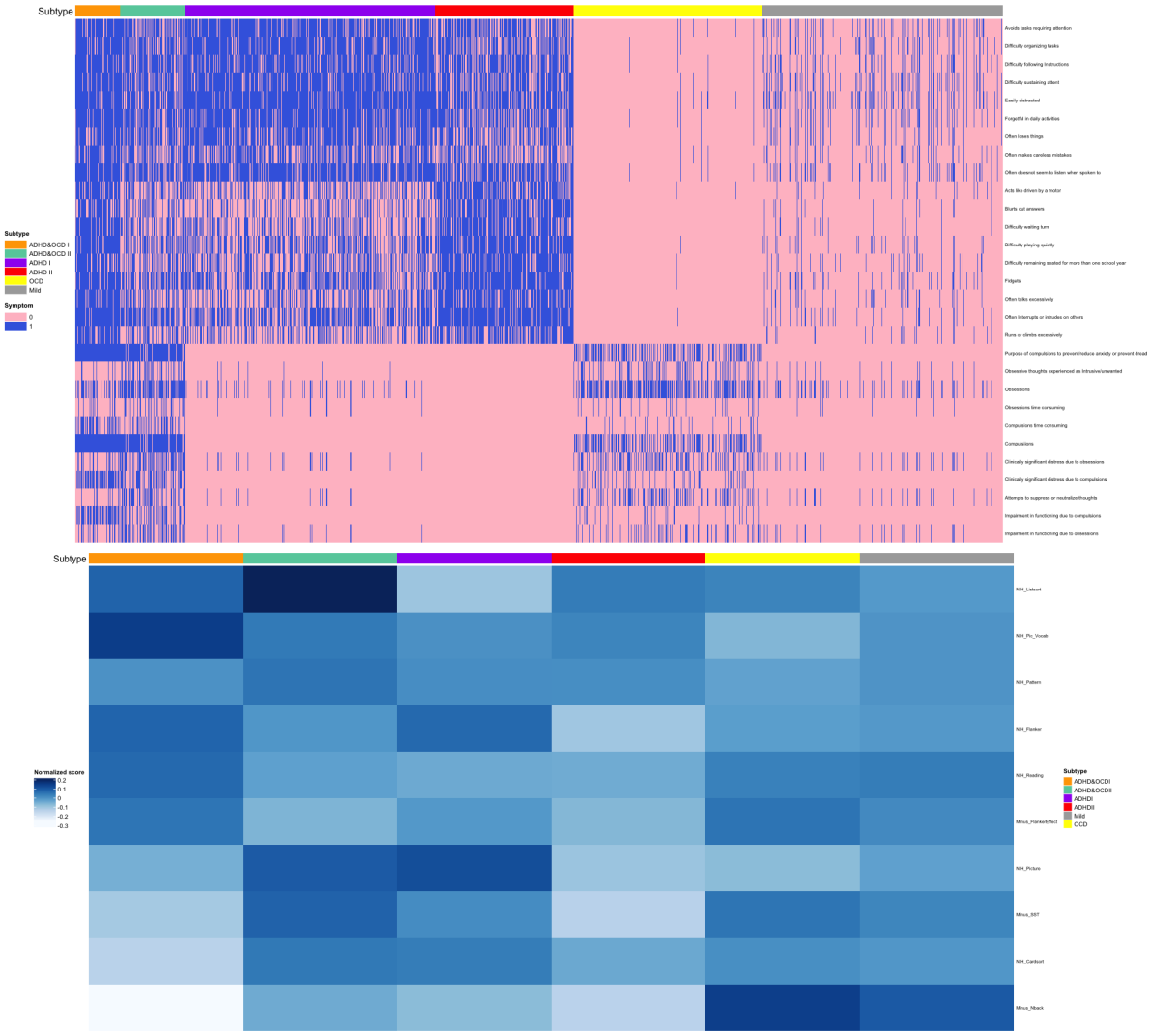

Figure 1: Heatmap of binary items and continuous measures across subtypes by the MINDS method; integrating ADHD&OCD related symptoms and selected neuro-cognitive measures; colors of orange to grey represent participants in ADHD&OCD I, ADHD&OCD II, ADHD I, ADHDII, OCD, MILD groups, respectively, scores are normalized by row.

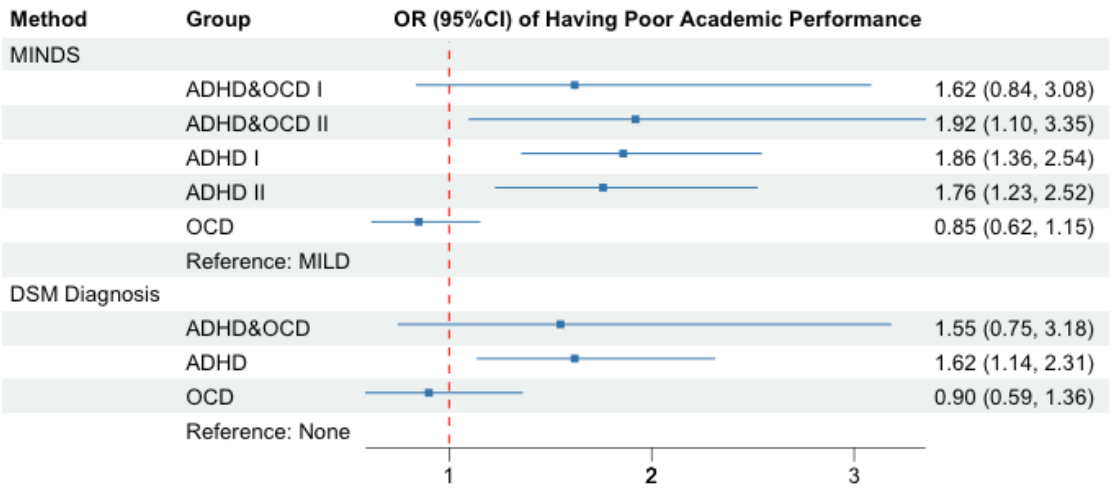

Figure 2: Odds ratio and 95% CI of each subtype vs healthy group in terms of having poor academic performance; subtypes were made by the MINDS method and by ADHD&OCD DSM diagnosis, respectively.

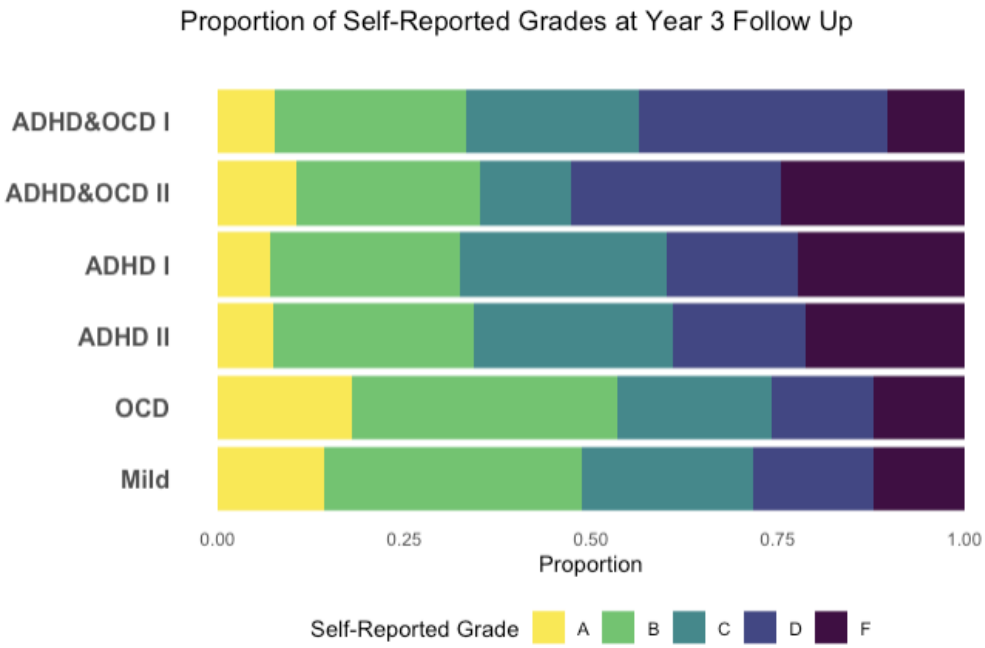

Figure 3: Proportion of self-reported grades by parents in different subtypes at the third follow-up year.

Table 1: Simulation results from joint model for integrating one binary modality and one continuous modality based on 200 replicates; RMSE and bias are calculated for 1,000 vs. 4,000 sample size; (10, 10) represents the scenario that the model has 10 binary items and 10 continuous measures.

	Sample Size	Number of Items	X	V	U	a	θ
RMSE	1000	(10,10)	0.133	0.14	0.12	0.046	0.016
	4000	(10,10)	0.128	0.112	0.096	0.028	0.005
	1000	(20,10)	0.292	0.312	0.313	0.042	0.014
	4000	(20,10)	0.237	0.308	0.296	0.03	0.004
	1000	(30,10)	0.316	0.241	0.228	0.089	0.012
	4000	(30,10)	0.295	0.224	0.19	0.07	0.004
Bias	1000	(10,10)	0.368	0.388	0.318	0.123	0.016
	4000	(10,10)	0.337	0.325	0.295	0.081	0.005
	1000	(20,10)	0.422	0.371	0.328	0.15	0.014
	4000	(20,10)	0.373	0.355	0.319	0.109	0.004
	1000	(30,10)	0.526	0.348	0.269	0.308	0.012
	4000	(30,10)	0.486	0.309	0.237	0.22	0.004

Table 2: Bayes test error results from joint model for integrating one binary modality and one continuous modality based on 200 replicates; training set varies from 1,000 to 4,000 subjects; (20, 10) represents the scenario that the model has 20 binary items and 10 continuous measures.

Number of items	**Training set size**	**Bayes error**				
		MINDS	K-means	Hclust	Two-step K-means	Two-step Hclust
(10,10)	1000	0.01(0.03)	0.08(0.09)	0.02(0.03)	0.04(0.06)	0.04(0.06)
(10,10)	2000	0.01(0.03)	0.08(0.09)	0.02(0.04)	0.04(0.07)	0.04(0.06)
(10,10)	3000	0.01(0.02)	0.08(0.09)	0.02(0.04)	0.04(0.07)	0.04(0.06)
(10,10)	4000	0.01(0.03)	0.08(0.09)	0.03(0.06)	0.05(0.07)	0.04(0.06)
(20,10)	1000	0.01(0.02)	0.08(0.09)	0.03(0.01)	0.02(0.05)	0.07(0.06)
(20,10)	2000	0.01(0.02)	0.09(0.09)	0.03(0.01)	0.02(0.04)	0.07(0.06)
(20,10)	3000	0.01(0.02)	0.08(0.09)	0.03(0.01)	0.02(0.04)	0.06(0.06)
(20,10)	4000	0.01(0.01)	0.06(0.08)	0.03(0.01)	0.02(0.05)	0.07(0.06)
(30,10)	1000	0.03(0.06)	0.12(0.09)	0.11(0.04)	0.04(0.06)	0.13(0.13)
(30,10)	2000	0.02(0.04)	0.1(0.09)	0.11(0.04)	0.05(0.07)	0.14(0.11)
(30,10)	3000	0.02(0.03)	0.11(0.09)	0.11(0.04)	0.04(0.07)	0.14(0.12)
(30,10)	4000	0.02(0.05)	0.11(0.09)	0.11(0.04)	0.04(0.07)	0.15(0.13)

Table 3: Neuro-cognitive measures used in the ABCD study participants subtyping

Variable	**Task/Test**	**What it measures**
SST Score	The Stop Signal Task	Response inhibition
Nback Score	The EN-back Task	Working memory
Flanker Effect	Toolbox Flanker Task	Attention, inhibition
Pic Vocab	Toolbox Picture Vocabulary Task	Language
Flanker	Toolbox Flanker Task	Attention, inhibition of automatic response
List Sort	Toolbox List Sorting Working Memory Test	Working memory, information processing
Card Sort	Toolbox Dimensional Change Card Sort Task	Executive function
Pattern	Toolbox Pattern Comparison Processing Speed Test	Information processing
Picture	Toolbox Picture Sequence Memory Test	Episodic memory, sequencing
Reading	Toolbox Oral Reading Recognition Task	Langurage, reading/decoding skills