ABCD Multiview Step 1. Proposal for Data Analysis and Simulation Study

Aidan Neher

2024-05-17

## Motivating Questions

Data Analysis: Cross-sectionally, at baseline, which early life adversity (ELA) and neuroimaging features are important for the prediction of continuous clinical outcomes “internalizing problems” and “externalizing problems”? From these features, how well can we predict these clinical outcomes?

Simulation Study: Does extension of the BIP framework to a mixed outcome model predict well? Specifically, does the extension provide predictive value above and beyond the original BIP framework and the 2 step (i) dimension reduction and (ii) prediction method commonly employed in an ABCD Study-like setting and other realistic settings?

## Data Analysis Plan

Cross-sectionally, at baseline, we will conduct an integrative analysis of the following data types (“views”) in relation to the listed covariates and clinical outcomes.

### About the Data

The Adolescent Brain and Cognitive Development Study (ABCD Study®) is the largest long-term brain development and child health study in the United States. Starting with subjects aged 9-10 years old and repeated until 19-20 years, brains scans are taken (bi-annually), biosamples, paper and pen tests, and iPad tasks (annually), and interviews (every 3-6 months). We use baseline data from the 5.1 release. We are most interested in the following views:

* **Early life adversity (ELA)**: We consider all ELA variables that Brieant et al. 2023 and Orendain et al. 2023 investigate. Similar to Brieant et al. 2023, we exclude ELA variables with >50% missingness and <0.05% endorsment. The endorsement filtering step led to 1 variable from Orendain et al. 2023 getting dropped from our dataset. Additionally, we removed ELA variables represented in the covariates. This results in a view with size 88 variables.
* **Structural MRI (sMRI)**: We use data parcellated with the Destrieux Atlas, which has 74 regions of interest (ROIs) defined. With 74 ROIs across 2 hemispheres, we have 148 measures for each of 2 sMRI metrics, Cortical Thickness (CT) and Surface Area (SA). These data are vectorized and concatenated to form an sMRI view. This view consequently has size 296 variables.
* **Functional MRI (fMRI)**: We plan to use pre-tabulated Functional Connectivity (FC), which measures pairwise correlation in functional activity across ROIs. This data is not readily available at present so will need to be procured or generated. The total number of variables for this view is not identified here since whichever Atlas is available will define how large this view ultimately is.

**Covariates:** We include: Sex (At Birth), Age (Months), White (Yes/No), Black (Yes/No), American Indian or Native American (Yes/No), Native Hawaiian or Pacific Islander (Yes/No), Asian (Yes/No), Other Race (Yes/No), Missing Race (Yes/No), Indigenous Race (Yes/No), Total Family Income (Past 12 Months), Highest Parent Education Completed, Parent Marital Status. Total n covariates = 13. **Outcomes:** Continuous random variables Internalizing Problems (T-Score), Externalizing Problems (T-Score) were studied by both Brieant et al. 2023 and Orendain et al. 2023. Total n outcomes = 2.

Table 1. Number of Variables per Data Type

| data\_type | n\_variables |
| --- | --- |
| outcomes | 2 |
| covariates | 13 |
| ela\_view | 88 |
| sMRI\_view | 296 |

#### Missing Data

We describe the number of observations missing per data type (Table 2), and notice ELA and covariates account for the most missingness. On exploring the top 5 variables by number of missing observations (Table S2), we realize household combined income has the most missingness (n\_missing=1017), and ELA variables represent the other top 4 missing variables.

Table 2. Observations with Missingness per Data Type

| data\_type | n\_missing |
| --- | --- |
| outcomes | 8 |
| covariates | 1096 |
| ela\_view | 3376 |
| sMRI\_view | 13 |

We plan for a complete case analysis, excluding observations with missingness in any of the data types mentioned though we are ignoring FC since we lack access at present. This would result in an overall sample size N = 7706.

#### Observational Clustering

From the complete case data, we explore the clustering by family and family within site that we will account for by mixed modeling. In Equation 1, we propose use of the an outcome model to account for this observational clustering. For the th view (), so the model can be expressed after suppressing the subscript and introducing random effects in the intercept of the th unit. Indices subject, family, and site.

where

Furthermore, we describe the ABCD Study’s nested clustering () empirically in Figure S1. This distribution will be referenced in simulation study scenario development. Note, in complete case analysis, presently, one site has 26 families available, which might challenge random effect estimation.

### Model Training & Testing

We will split our data 80:20 train:test. This will result in roughly 6164.8 training samples and length(sample\_key$src\_subject\_id) \* .2 test samples. We perform hyperparameter tuning using 5-fold cross-validation in the training set. The test set is reserved for final model performance evaluation. Our primary performance metric is mean square prediction error (MSPE) in the test set. Secondarily, we will explore overlap between the ELA features we find with substantial loadings (threshold TBD) by BIP modeling and those found/ used by Brieant et al. 2023 and Orendain et al. 2023.

**Table 1:** Schematic of data analysis prediction performance results. Confidence intervals will be estimated by bootstrapping.

| Outcome | MSPE |
| --- | --- |
| Internalizing Problems (T-Score) |  |
| Externalizing Problems (T-Score) |  |

## Simulation Study

We will test our mixed modeling framework. TBD if random effect nesting will be achieved/ not. Regardless, the simulation study proposed below is relevant.

### Competing Methods

* BIPmixed: Proposed BIP framework with a mixed outcome model.
* BIP: BIP method originally proposed by Chekouo and Safo 2021.
* 2-Step Regression: Brieant et al. 2023 and Orendain et al. 2023 (i) perform factor analysis for dimension reduction and then (ii) fit linear regression models in 2 steps. In particular, we consider Brieant et al. 2023’s (i) factor analysis and then (ii) fit a Bayesian linear regression model. *Note, the number of factors pass from (i) to (ii) is a hyperparameter that will need to be tuned. Also, for (i), we will either need to concatenate views/ consider another multiview factor analysis method.*
* Other Methods from BIP Paper: *Are there methods from BIP paper that we should carry forward to this simulation? Are there methods more usable/ more suitable for comparison to the BIPmixed approach?*

### Simulation Settings

We fix our outcome model, include 10 covariates, which is roughly the data analysis number, and 2 views where the ELA-analogous view is smaller (p=100) and the more traditionally considered -omics views are larger (p=300). We fix the ABCD Study site count = 21.

* **Sample Size:** Small (We can use traditional neuroimaging studies as a benchmark for this), Medium, and Large (e.g. , like the ABCD Study).
* **Number of Variables:** Small (Can be based on the number we are using, roughly 400 in total, 100/ 300 per view) to large number of variables in the -omics view (The size to vary to can be based on if we were to include gene expression data/ SNPs) **Study Design:** Balanced (Number of families/ site is fixed to median number observed in ABCD Study) vs. Unbalanced (Distribution of can be empirically derived from ABCD Study. See Figure S1 for details).

3 Sample Size x 2 Number of Variables x 2 Study Design conditions = 12 total simulation scenarios.

We will generate datasets and estimate performance of each competing method on all sets.

### Performance Metrics

**Table 2: Primary performance metric** mean square prediction error in simulated test sets where confidence intervals are based on performance across datasets schematic.

| Scenario | BIPmixed MSPE | BIP MSPE | … |
| --- | --- | --- | --- |
| Scenario 1 |  | … | … |
| Scenario 2 |  | … | … |
| … |  | … | … |
| Scenario 12 |  | … | … |

**Table 3: Secondary performance metric** AUC for variable selection against the truth in simulated test sets where confidence intervals are based on performance across datasets schematic.

| Scenario | BIPmixed AUC | BIP AUC | … |
| --- | --- | --- | --- |
| Scenario 1 |  | … | … |
| Scenario 2 |  | … | … |
| … |  | … | … |
| Scenario 12 |  | … | … |

## Relevant References & Why

1. Chekouo, T. & Safo, S. E. Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. Biostatistics 24, 124–139 (2021).
2. Orendain, N., Anderson, A., Galván, A., Bookheimer, S. & Chung, P. J. A data-driven approach to categorizing early life adversity exposure in the ABCD Study. BMC Med Res Methodol 23, 164 (2023).
3. Brieant, A. et al. Characterizing the dimensional structure of early-life adversity in the Adolescent Brain Cognitive Development (ABCD) Study. Developmental Cognitive Neuroscience 61, 101256 (2023).
4. Saragosa-Harris, N. M. et al. A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. Developmental Cognitive Neuroscience 55, 101115 (2022).
5. Wiglesworth, A. et al. Sex and age variations in the impact of puberty on cortical thickness and associations with internalizing symptoms and suicidal ideation in early adolescence. Developmental Cognitive Neuroscience 59, 101195 (2023).
6. Gelman, A. et al. Bayesian Data Analysis Third edition (with errors ﬁxed as of 15 February 2021).
7. Gelman, A. Prior distributions for variance parameters in hierarchical models.
8. Banerjee, S. Bayesian Linear Model: Gory Details.

Why?

1. This is the framework we are extending, which can be referenced, especially BIP’s performance relative to competing methods.
2. Authors define ABCD Study ELA variables we use in the data analysis.
3. Authors define ABCD Study ELA variables we use in the data analysis. Also, we follow similar data processing decisions, except for excluding/ consolidating highly correlated ELA variables.
4. Useful description of the ABCD Study that can be cited in our brief introduction to the study.
5. This paper mirror our covariate coding approach and which covariates we include in relation to outcomes internalizing and externalizing problems.
6. This textbook will be referenced in hierarchical model development.
7. This paper will be referenced in hierarchical model development, especially prior choice for variance parameters. Authors recommend our initial uniform prior choice for the variance parameters as a starting point.
8. While not an academic article that will be cited/ citable in our paper, this has been referenced for understanding Bayesian linear model development and result derivations.

## Appendix 1. Supplementary Materials

**Software:** All analyses will be conducted using the R statistical software, version 4.2.0, and relevant R packages.

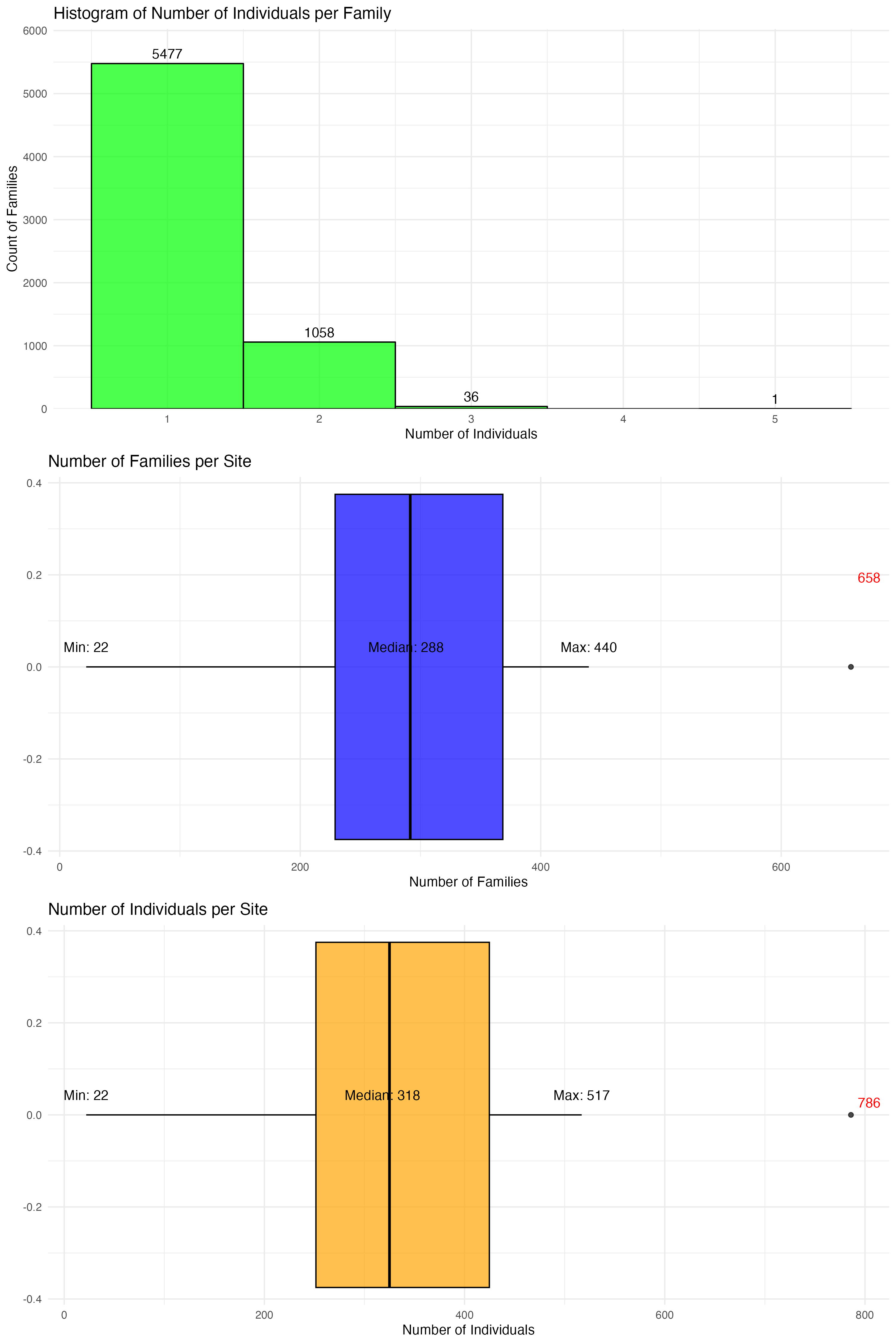
**Table S1:** Summary statistics of outcomes Internalizing Problems (T-Score) and Externalizing Problems (T-Score) and covariates for complete case sample.

|  | level | Overall |
| --- | --- | --- |
| n |  | 7706 |
| Internalizing Problems (T-Score) (mean (SD)) |  | 47.99 (10.35) |
| Externalizing Problems (T-Score) (mean (SD)) |  | 45.09 (9.96) |
| Sex (At Birth) (%) | Male | 4054 (52.6) |
|  | Female | 3651 (47.4) |
|  | Intersex-Male | 1 ( 0.0) |
|  | Intersex-Female | 0 ( 0.0) |
| Age (Months) (mean (SD)) |  | 119.00 (7.49) |
| White (Yes/No) (%) | No | 1609 (20.9) |
|  | Yes | 6097 (79.1) |
| Black (Yes/No) (%) | No | 6375 (82.7) |
|  | Yes | 1331 (17.3) |
| American Indian or Native American (Yes/No) (%) | No | 7467 (96.9) |
|  | Yes | 239 ( 3.1) |
| Native Hawaiian or Pacific Islander (Yes/No) (%) | No | 7657 (99.4) |
|  | Yes | 49 ( 0.6) |
| Asian (Yes/No) (%) | No | 7210 (93.6) |
|  | Yes | 496 ( 6.4) |
| Other Race (Yes/No) (%) | No | 7239 (93.9) |
|  | Yes | 467 ( 6.1) |
| Missing Race (Yes/No) (%) | No | 7689 (99.8) |
|  | Yes | 17 ( 0.2) |
| Indigenous Race (Yes/No) (%) | No | 7423 (96.3) |
|  | Yes | 283 ( 3.7) |
| Total Family Income (Past 12 Months) (%) | Less than $5,000 | 247 ( 3.2) |
|  | $5,000 - $11,999 | 265 ( 3.4) |
|  | $12,000 - $15,999 | 178 ( 2.3) |
|  | $16,000 - $24,999 | 308 ( 4.0) |
|  | $25,000 - $34,999 | 402 ( 5.2) |
|  | $35,000 - $49,999 | 617 ( 8.0) |
|  | $50,000 - $74,999 | 1064 (13.8) |
|  | $75,000 - $99,999 | 1146 (14.9) |
|  | $100,000 - $199,999 | 2498 (32.4) |
|  | $200,000 and greater | 981 (12.7) |
| Highest Parent Education Completed (%) | 1st grade | 0 ( 0.0) |
|  | 2nd grade | 0 ( 0.0) |
|  | 3rd grade | 3 ( 0.0) |
|  | 4th grade | 4 ( 0.1) |
|  | 5th grade | 0 ( 0.0) |
|  | 6th grade | 19 ( 0.2) |
|  | 7th grade | 4 ( 0.1) |
|  | 8th grade | 15 ( 0.2) |
|  | 9th grade | 47 ( 0.6) |
|  | 10th grade | 40 ( 0.5) |
|  | 11th grade | 68 ( 0.9) |
|  | 12th grade, no diploma | 81 ( 1.1) |
|  | High school graduate | 418 ( 5.4) |
|  | GED or equivalent | 192 ( 2.5) |
|  | Less than 1 year of college credit/post-secondary education | 890 (11.5) |
|  | One year or more of college credit, no degree | 546 ( 7.1) |
|  | Associate degree: Occupational, Technical, or Vocational | 409 ( 5.3) |
|  | Associate degree: Academic Program | 2090 (27.1) |
|  | Bachelor's degree (e.g., BA, AB, BS, BBA) | 1976 (25.6) |
|  | Master's degree (e.g., MA, MS, MEng, MEd, MBA) | 439 ( 5.7) |
|  | Professional School degree (e.g., MD, DDS, DVM, JD) | 465 ( 6.0) |
|  | Doctoral degree (e.g., PhD, EdD) | 0 ( 0.0) |
| Parent Marital Status (%) | Married | 5667 (73.5) |
|  | Widowed | 47 ( 0.6) |
|  | Divorced | 606 ( 7.9) |
|  | Separated | 231 ( 3.0) |
|  | Never married | 748 ( 9.7) |
|  | Living with partner | 407 ( 5.3) |

Table S2. Top 5 Variables by Missing Observations

| data\_type | variable | n\_missing |
| --- | --- | --- |
| covariates | demo\_comb\_income\_v2\_bl | 1017 |
| ela\_view | crpbi\_caregiver14\_y | 917 |
| ela\_view | crpbi\_caregiver16\_y | 916 |
| ela\_view | crpbi\_caregiver12\_y | 915 |
| ela\_view | crpbi\_caregiver15\_y | 915 |

**Figure S1.** Summary of ABCD Study nested clustering of observations (study subject/ individual within family & family within study site).



# Appendix 2: Code for Document

library(tidyverse)  
library(kableExtra)  
# Define the directory, date, processor, and file suffixes  
data\_dir <- "data"  
figures\_dir <- "figures"  
data\_processing\_date <- "2024-05-16"  
data\_processor <- "AN"  
# Construct file paths  
missing\_per\_datatype\_path <- file.path(data\_dir,  
 paste0(data\_processing\_date,  
 "\_", data\_processor,  
 "\_n\_missing\_per\_datatype.csv"))  
missing\_per\_variable\_path <- file.path(data\_dir, paste0(data\_processing\_date,   
 "\_", data\_processor,  
 "\_n\_missing\_per\_variable.csv"))  
variables\_per\_datatype\_path <- file.path(data\_dir, paste0(data\_processing\_date,   
 "\_", data\_processor,  
 "\_n\_variables\_per\_datatype.csv"))  
# Read in the CSV files  
missing\_per\_datatype\_df <- read.csv(missing\_per\_datatype\_path)   
missing\_per\_variable\_df <- read.csv(missing\_per\_variable\_path)   
variables\_per\_datatype\_df <- read.csv(variables\_per\_datatype\_path)   
  
# Show only the top 5 variables with the most missing data  
top\_5\_missing\_per\_variable\_df <- missing\_per\_variable\_df %>%  
 arrange(desc(n\_missing)) %>%  
 slice\_head(n = 5)  
  
# Table 1: Variables per Data Type  
variables\_per\_datatype\_table <- kable(variables\_per\_datatype\_df,   
 booktabs = TRUE,   
 align = "c",   
 caption = "Table 1. Number of Variables per Data Type") %>%  
 kable\_styling(latex\_options = c("striped", "repeat\_header"),  
 stripe\_color = "blue!15")  
  
# Table 2: Missing per Data Type  
missing\_per\_datatype\_table <- kable(missing\_per\_datatype\_df,   
 booktabs = TRUE,   
 align = "c",   
 caption = "Table 2. Observations with Missingness per Data Type") %>%  
 kable\_styling(latex\_options = c("striped", "repeat\_header"),  
 stripe\_color = "blue!15")  
  
# Table S2: Top 5 Missing per Variable  
top\_5\_missing\_per\_variable\_table <- kable(top\_5\_missing\_per\_variable\_df,   
 booktabs = TRUE,   
 align = "c",   
 caption = "Table S2. Top 5 Variables by Missing Observations") %>%  
 kable\_styling(latex\_options = c("striped", "repeat\_header"),  
 stripe\_color = "blue!15")  
  
covar\_labels <- c(  
 "Sex (At Birth)",  
 "Age (Months)",  
 "White (Yes/No)",  
 "Black (Yes/No)",  
 "American Indian or Native American (Yes/No)",  
 "Native Hawaiian or Pacific Islander (Yes/No)",  
 "Asian (Yes/No)",  
 "Other Race (Yes/No)",  
 "Missing Race (Yes/No)",  
 "Indigenous Race (Yes/No)",  
 "Total Family Income (Past 12 Months)",  
 "Highest Parent Education Completed",  
 "Parent Marital Status"  
)  
  
outcome\_labels <- c(  
 "Internalizing Problems (T-Score)",  
 "Externalizing Problems (T-Score)"  
)  
  
# Show the number of variables included from each data type   
variables\_per\_datatype\_table  
missing\_per\_datatype\_table  
sample\_key <- read.csv("data/2024-05-16\_AN\_sample\_key.csv")  
library(tableone)  
library(labelled) # for set\_variable\_levels  
library(kableExtra)  
  
covariates <- file.path(data\_dir, paste0(data\_processing\_date, "\_", data\_processor, "\_covariates.csv")) %>%  
 read.csv()  
outcomes <- file.path(data\_dir, paste0(data\_processing\_date, "\_", data\_processor, "\_outcomes.csv")) %>%  
 read.csv()  
  
# Merge outcomes and covariates data frames and filter to common subjects  
data <- inner\_join(outcomes, covariates, by = "src\_subject\_id") %>%  
 filter(src\_subject\_id %in% sample\_key$src\_subject\_id)  
  
# Convert all columns except 'src\_subject\_id' to numeric  
data <- data %>%  
 mutate\_at(vars(-src\_subject\_id), as.numeric)  
  
# Relevel race variables to be factors with levels "No" and "Yes"  
race\_vars <- colnames(data)[grepl("\_race$", colnames(data))]  
data <- data %>%  
 mutate\_at(vars(one\_of(race\_vars)), ~ factor(., levels = c(0, 1), labels = c("No", "Yes")))  
  
# Relevel demo\_sex\_v2  
data <- data %>%  
 mutate(demo\_sex\_v2 = factor(demo\_sex\_v2, levels = c(1, 2, 3, 4), labels = c(  
 "Male", "Female", "Intersex-Male", "Intersex-Female"  
 )))  
  
# Define labels  
table1\_labels <- c(  
 "Internalizing Problems (T-Score)",  
 "Externalizing Problems (T-Score)",  
 "Sex (At Birth)",  
 "Age (Months)",  
 "White (Yes/No)",  
 "Black (Yes/No)",  
 "American Indian or Native American (Yes/No)",  
 "Native Hawaiian or Pacific Islander (Yes/No)",  
 "Asian (Yes/No)",  
 "Other Race (Yes/No)",  
 "Missing Race (Yes/No)",  
 "Indigenous Race (Yes/No)",  
 "Total Family Income (Past 12 Months)",  
 "Highest Parent Education Completed",  
 "Parent Marital Status"  
)  
  
# Define factor levels and labels  
data <- data %>%  
 mutate(  
 demo\_comb\_income\_v2\_bl = factor(demo\_comb\_income\_v2\_bl, levels = 1:10, labels = c(  
 "Less than $5,000", "$5,000 - $11,999", "$12,000 - $15,999",   
 "$16,000 - $24,999", "$25,000 - $34,999", "$35,000 - $49,999",  
 "$50,000 - $74,999", "$75,000 - $99,999", "$100,000 - $199,999",  
 "$200,000 and greater")),  
 highest\_demo\_ed\_bl = factor(highest\_demo\_ed\_bl, levels = c(1:22), labels = c(  
 "1st grade", "2nd grade", "3rd grade", "4th grade", "5th grade",   
 "6th grade", "7th grade", "8th grade", "9th grade", "10th grade", "11th grade",   
 "12th grade, no diploma", "High school graduate", "GED or equivalent",   
 "Less than 1 year of college credit/post-secondary education", "One year or more of college credit, no degree",   
 "Associate degree: Occupational, Technical, or Vocational", "Associate degree: Academic Program",   
 "Bachelor's degree (e.g., BA, AB, BS, BBA)", "Master's degree (e.g., MA, MS, MEng, MEd, MBA)",   
 "Professional School degree (e.g., MD, DDS, DVM, JD)", "Doctoral degree (e.g., PhD, EdD)")),  
 demo\_prnt\_marital\_v2\_bl = factor(demo\_prnt\_marital\_v2\_bl, levels = c(1:6), labels = c(  
 "Married", "Widowed", "Divorced", "Separated", "Never married", "Living with partner"))  
 )  
  
# Apply labels to variables  
colnames(data)[-1] <- table1\_labels  
  
# Create Table 1 using the table1 package  
tbl\_1 <- CreateTableOne(data = data, vars = colnames(data)[-1],   
 test = FALSE)  
tbl\_1\_printed <- print(tbl\_1, varLabels = TRUE, showAllLevels = TRUE)  
kable(tbl\_1\_printed, booktabs = TRUE,   
 align = "c", longtable = TRUE) %>%  
 kable\_styling(latex\_options = c("striped", "repeat\_header"),  
 stripe\_color = "blue!15")  
top\_5\_missing\_per\_variable\_table