Exploratory Analysis of Direct and Indirect Prenatal Smoking Exposure Effects on Children Externalizing Behavior

1. Introduction

The primary research goal of this report is to delve into the effects of smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) on various aspects of adolescent development, specifically focusing on self-regulation, substance use, and externalizing behaviors. The data for this analysis were sourced from Dr. Lauren Micalizzi of the Brown University Department of Behavioral and Social Sciences. The study initially recruited a cohort of low-income pregnant women, totaling 738 individuals, as part of a smoke avoidance intervention program to reduce maternal smoking and ETS exposure during pregnancy. Furthermore, the study also assessed children's exposure to ETS in the immediate postpartum period. For this research, a subset of 100 adolescents and their mothers was randomly selected for recruitment, forming the core dataset.

The data can be broadly categorized into two sections: child and parent data to gain a comprehensive understanding of the adolescent and parent dynamics. The child section has essential demographic information, including race, age, and sex. Additionally, there are scores related to attention, internalizing, externalizing problems, and emotion regulation attributes such as cognitive reappraisal and expressive suppression. Moreover, the dataset contains information on child substance use and the parent-child relationship. The parent section mirrors many of the child's data categories. It also includes additional information such as the child's ADHD status, parental demographic details like income, employment, and education, and data related to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS).

Despite the richness of the dataset, it's important to acknowledge certain limitations that may influence the scope and interpretation of the analysis findings. First, cotinine levels, which serve as a biomarker of nicotine exposure, were only measured at two-time points—34 weeks gestation and 6 months after birth. This limitation necessitates a reliance on self-reported data for nicotine exposure in other instances. Second, the data analyzed in this report are limited to baseline measurements for both children and parents. Therefore, this study does

not cover longitudinal analysis, limiting the ability to draw conclusions about changes over time. Finally, this research's inclusion criteria required data for both the parent and child, resulting in a reduced dataset of 49 parent-child pairs. This reduced sample size may impact the generalizability of the findings to a broader population.

2. Preliminary Analysis

In the initial stages of our research, a crucial process involved getting the raw research data ready for analysis. This meant performing tasks like aggregating answers from related questionnaires to create summarized scores (e.g., parental knowledge score). Furthermore, we combined various pieces of data to develop meaningful indicators (e.g., smoke exposure one year after a child's birth). We also removed irrelevant variables or overly complex data to compact the dataset. We determined the parent-child pairs using the `parent_id`, and any non-paired individuals were removed.

Following the initial data preprocessing, we were left with a dataset consisting of 49 records and 78 variables. As we delved into the data, we noticed certain anomalies in some variables that raised questions. For instance, within the variable indicating the biological sex of the parents, there was one record indicating a male. However, given that all respondents were mothers, it is likely that this occurrence resulted from a data entry error or a misunderstanding by the respondent. Additionally, we encountered an unusual entry in the income variable, with a record stating '250,000,' which appeared to be a potential data entry error. Moreover, some irregularities surfaced in the daily number of cigarettes smoked by mothers, with entries like '2 black and miles a day,' '44989,' and '20-25.' These data issues are noted for further fixing in our subsequent preprocessing steps, where we will also address the creation of new variables that can be derived from the existing dataset.

3. Missing Data

Furthermore, during our data examination, we observed that 54 variables contained some missing data, and 765 observations had missing values, roughly 20% of the data being incomplete. However, the extent of the missing data varied among the variables. Notably, five variables displayed a missing data proportion exceeding 50%. Four of these variables pertained to the number of cigarettes, marijuana, e-cigarettes, and alcohol consumed by the child in the last 30 days, with 90% or more of their data missing. However, we determined that this missing data could be explained by the fact that only a few children responded to the substance consumption questionnaire. On the other hand, one variable, the autism spectrum disorder indicator, showed a missing data proportion exceeding 50% without a clear explanation, leading us to exclude this particular data from our analysis. We considered the extent of missing data manageable for the remaining variables with varying degrees of missing data, ranging from 0% to 32%. However, we found that the missing data actually contributed by 8 IDs who had more than 40 missing variables (exceeding 50% of the total variables), including critical information

like externalizing behavior scores and self-reported substance usage. We decided to dropped these IDs.

4. Enhancing Dataset

Considering our initial exploration of the data, we have decided to introduce several new variables to enhance our dataset's utility. These new variables consolidate related variables, simplify the values of certain existing variables, and rectify anomalies identified during our preliminary analysis.

One key addition is the variable `num_substance_used`, which summarizes the number of different substances consumed by the children from the four types present in the dataset (cigarettes/e-cigarettes, marijuana, alcohol). For instance, if a child has consumed cigarettes and alcohol, their `num_substance_used` would be recorded as 2. A parallel variable, `pnum_substance_used`, was created to capture a similar summary for the parents.

In relation to smoking during pregnancy (SDP), we introduced variables such as `mom_prenatal_smoke` to indicate if the mother self-reported smoking during any of the follow-ups in gestational weeks. A similar set of variables, `mom_postnatal_smoke`, represents smoking during the postnatal period. To gauge the intensity of mom SDP, we created `mom_prenatal_smoke_consistency` and `mom_postnatal_smoke_consistency`, which count how many times the mother reported smoking during prenatal and postnatal follow-ups, respectively. Furthermore, we recorded the pattern of maternal smoking during both prenatal and postnatal periods in the variable `mom_smoke_pattern`. For example, if a respondent reported smoking during the prenatal period but not during the postnatal period, they would have a value of '1 0' in this column. A parallel set of variables was created to represent environmental tobacco smoke (ETS) exposure. For the ETS we use the self-reported smoke exposure from year 1 through year 5.

5. Univariate Analysis

We conduct univariate analysis for most of the variables used in this study we start with a sanity check using summary statistics (minimum, maximum, median, 1st quartile, 3rd quartile) to get an initial idea of their distribution. If anything unusual is observed, we proceed to plot their distribution. The study sample primarily comprises individuals from Hispanic/Latino and White ethnic backgrounds (children and parents alike), reflecting a diverse population. Most children are male and in the middle school age range, indicating a focus on this particular demographic for adolescent development research. Regarding the parents of the children under study, they usually had the children in the study in their young adult phase (20-25). Most of the parents in the dataset have education levels exceeding high school. While the parents are predominantly employed, their income levels tend to fall below the threshold of $38,133.

Both children and parents exhibit similar distribution patterns in attention problem scores, internalizing problem scores, and externalizing scores, which are right-skewed. This implies that most children and parents scored relatively low on these metrics. Additionally, they share a similar distribution in cognitive reappraisal scores, which is left-skewed. However, there is a difference between children and parents when it comes to expressive suppression. Children's scores show a left-skewed distribution, while parents' scores are right-skewed. It means that children are more likely to have expressive suppression compared to the parents.

Regarding ADHD, the inattentive type exhibits a left-skewed distribution, while the hyperactive type is right-skewed. Furthermore, both children and parents display a right-skewed distribution in the number of substance types used, indicating that most participants consumed relatively few types of substances.

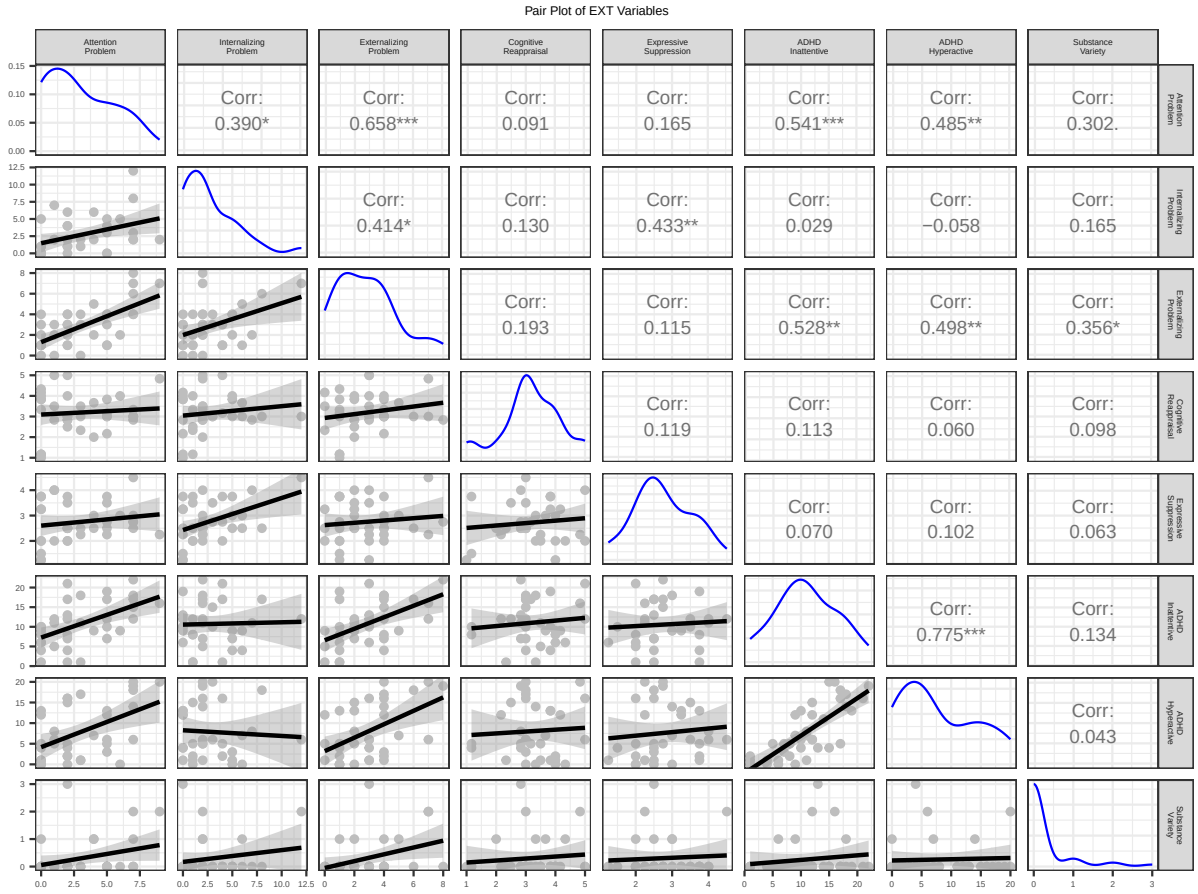
However, it's worth noting that there were minimal reports of substance usage among children, with alcohol being the most reported, totaling 5 records. In contrast, a larger proportion of parents reported consuming both tobacco and alcohol.

The parental monitoring attributes within the dataset exhibit a left-skewed distribution, whether viewed from the children's or the parent's perspective. This finding suggests a harmonious condition in the families, with both displaying a similar level of agreement regarding monitoring behaviors.

Regarding smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS), the data trends show a slight inclination towards "no smoking" and "no smoking exposure" during both prenatal and postnatal periods. However, it is worth noting an interesting observation: there are more reports of "no exposure" compared to "mom smoking" claims. This phenomenon raises questions about the accuracy of the exposure data. Some errors in data entry or misclassification within the exposure data may contribute to this discrepancy between reported maternal smoking and reported exposure levels. Further investigation may be needed to reconcile these disparities and ensure the reliability of the exposure data.

6. Interrelatedness between Children Externalizing Behavior

In our analysis, we delved into the correlation between various Externalizing Behaviors variables, including Attention Problem Score, Internalizing Problem Score, Externalizing Problem Score, Cognitive Reappraisal, Expressive Suppression, ADHD Inattentive SWAN Score, ADHD Hyperactive SWAN Score, and Number of Substance Types Used. Given the nature of these variables, it was anticipated that some of them would be interrelated, such as the potential correlation between ADHD Inattentive and Attention Problem Score or between Internalizing Problems and Expressive Suppression. To explore these relationships, we employed pair plots as seen in Figure 1.


Pair Plot of EXT Variables
Attention Problem
Internalizing Problem
Externalizing Problem
Cognitive Reappraisal
Expressive Suppression
ADHD Inattentive
ADHD Hyperactive
Substance Variety
Corr: 0.390*
Corr: 0.658***
Corr: 0.091
Corr: 0.165
Corr: 0.541***
Corr: 0.485**
Corr: 0.302.
Corr: 0.414*
Corr: 0.130
Corr: 0.433**
Corr: 0.029
Corr: −0.058
Corr: 0.165
Corr: 0.193
Corr: 0.115
Corr: 0.528**
Corr: 0.498**
Corr: 0.356*
Corr: 0.119
Corr: 0.113
Corr: 0.060
Corr: 0.098
Corr: 0.070
Corr: 0.102
Corr: 0.063
Corr: 0.775***
Corr: 0.134
Corr: 0.043

Our analysis revealed notable correlations among these variables: 1) A strong correlation was observed between ADHD Inattentive and ADHD Hyperactive, with both also showing correlations between Attention and Externalizing Problems. This implies that children with high scores in one problem area will likely have elevated scores in related problems. Also, it means that most of the case, children have both ADHD Inattentive and ADHD Hyperactive.
2) We found that both the Attention Problem and Externalizing Problem correlated with the Internalizing Problem, indicating a potential link between these variables.
3) Our analysis confirmed our initial assumption that the Internalizing Problem has a substantial correlation with Expressive Suppression.
4) The variable significantly correlated with the Number of Substance Types Used was the Externalizing Problem.
5) However, it's important to note that Cognitive Reappraisal did not exhibit any significant correlation with the other variables, indicating its distinct nature in the context of this study. Based on these findings, it becomes apparent that externalizing behaviors are interconnected, suggesting a common underlying issue contributing to their correlation.

7. Relationship between Children Externalizing Behavior, Smoking During Pregnancy, and Other Variables.

In our pursuit to address the primary goal of this analysis, we conducted an examination of the relationship between variables related to smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) with the variables linked to externalizing behaviors (EXT). To illustrate these relationships, we initiated our investigation by comparing the characteristics of EXT variables across distinct groups (e.g., children exposed to smoke during prenatal stages versus those not exposed). The main variables used in this analysis are Mom Smoking (Prenatal) Indicator and Prenatal Smoke Exposure (Only prenatal are chosen since smoking during pregnancy refers to prenatal period). To facilitate this comparison, we constructed summary tables that provided a comprehensive view of key statistics, including medians, interquartile ranges, and p-values for each group. Wilcoxon rank sum test for 2 levels group are used to obtain the p-values. It is crucial to note that the p-values obtained from this analysis are based on a relatively small sample size. Therefore, it is advisable to consider larger sample sizes for more precise p-value estimations.

The results of our investigation unveiled several noteworthy relationships:
1) ADHD Hyperactive scores connected with Mom Smoking (Prenatal) suggest maternal smoking during pregnancy may be associated with children's hyperactivity. Children with smoking moms (prenatal) have 13 ADHD Hyperactive scores in the median compared to 5 ADHD Hyperactive scores in the median.
2) Internalizing problem scores were found to be related to Prenatal Smoke Exposure, with the group that had prenatal smoke exposure scoring 3 in the median and the non-exposed group scoring 1 in the median.
3) Externalizing problem scores exhibited relationships with Prenatal Smoke Exposure, with

Variable	N	Overall, N = 38	Maternal Prenatal Smoking		p-value
			0, N = 24	1, N = 14	
Attention Problem	29	3.00 (1.00, 5.00)	2.50 (0.25, 5.00)	5.00 (2.00, 7.00)	0.24
Internalizing Problem	27	2.00 (0.00, 4.00)	2.00 (0.75, 4.25)	2.00 (0.00, 3.50)	0.92
Externalizing Problem	29	3.00 (1.00, 4.00)	2.50 (1.00, 4.00)	3.00 (1.50, 5.00)	0.49
Cognitive Reappraisal	28	3.00 (2.83, 3.88)	3.00 (2.33, 3.88)	3.00 (3.00, 3.83)	0.41
Expressive Suppression	28	2.63 (2.25, 3.31)	2.50 (2.00, 3.00)	3.25 (2.50, 3.63)	0.084
ADHD Inattentive	30	11.5 (9.0, 15.8)	11.0 (8.0, 14.5)	13.0 (10.0, 16.5)	0.39
ADHD Hyperactive	30	6.0 (3.3, 13.0)	5.0 (1.5, 9.5)	13.0 (6.0, 17.0)	0.025
Substance Variety	38	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.36

[1] Median (IQR)

[2] Wilcoxon rank sum test

Variable	N	Overall, N = 42	Prenatal Environmental Smoking Exposure		p-value
			0, N = 27	1, N = 15	
Attention Problem	35	2.00 (1.00, 5.00)	2.00 (0.00, 4.50)	4.00 (2.00, 7.00)	0.072
Internalizing Problem	33	2.00 (1.00, 4.00)	1.00 (0.00, 3.50)	3.00 (2.00, 5.00)	0.012
Externalizing Problem	35	3.00 (1.00, 4.00)	2.00 (1.00, 3.00)	4.00 (3.00, 4.00)	0.009
Cognitive Reappraisal	34	3.08 (2.88, 3.83)	3.33 (2.83, 4.00)	3.00 (3.00, 3.42)	0.97
Expressive Suppression	34	2.63 (2.25, 3.44)	2.50 (2.00, 3.13)	3.00 (2.50, 3.75)	0.058
ADHD Inattentive	39	11.0 (7.5, 15.0)	10.0 (7.5, 13.5)	12.0 (9.3, 17.0)	0.36
ADHD Hyperactive	39	6.0 (2.5, 13.0)	5.0 (3.0, 12.0)	6.0 (2.8, 16.3)	0.36
Substance Variety	42	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.90

[1] Median (IQR)

[2] Wilcoxon rank sum test

Variable	N	Overall, N = 43	Postnatal Environmental Smoking Exposure		p-value
			0, N = 26	1, N = 17	
Attention Problem	35	2.00 (1.00, 5.00)	2.00 (0.00, 4.75)	3.00 (2.00, 5.00)	0.16
Internalizing Problem	33	2.00 (1.00, 4.00)	2.00 (0.00, 4.00)	2.50 (2.00, 4.25)	0.14
Externalizing Problem	35	3.00 (1.00, 4.00)	2.00 (1.00, 4.00)	3.00 (2.00, 4.00)	0.21
Cognitive Reappraisal	34	3.00 (2.83, 3.83)	3.33 (2.83, 4.00)	3.00 (3.00, 3.58)	0.76
Expressive Suppression	34	2.50 (2.25, 3.44)	2.50 (2.00, 3.00)	3.50 (2.50, 3.75)	0.011
ADHD Inattentive	38	10.5 (7.3, 14.8)	9.0 (7.0, 12.8)	12.0 (10.8, 17.0)	0.12
ADHD Hyperactive	38	6.0 (2.3, 13.0)	5.0 (2.3, 11.8)	9.5 (5.0, 16.3)	0.14
Substance Variety	43	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.54

[1] Median (IQR)

[2] Wilcoxon rank sum test

the group with prenatal smoke exposure scoring 4 in the median and the non-exposed group scoring 2 in the median.
5) Expressive Suppression scores displayed connections with Smoke Exposure (Postnatal). The group with postnatal smoke exposure scored 3.5 in the median, and the non-exposed group scored 2.5 in the median.

It seems that both prenatal smoke exposure and mom smoking during prenatal has an impact on Externalizing Behavior such as ADHD Hyperactive, Internalizing Problem, Externalizing Problem, and Expressive Suppression. Indirectly, it could impact substance use, too, since Externalizing Problems and Substance Use are related. We can conclude that the group exposed to smoke (from mom or secondary exposure) was associated with higher externalizing behavior scores and, therefore, more prone to worse externalizing behavior.

We also compared the Externalizing Behavior of different races (children & parents), sexes (children), higher education indicators, income levels, and employment status using the same method. Groups associated with higher attention problem scores and internalizing problems compared to other races children are white children. Surprisingly, children with part-timing moms have lower internalizing problem scores than other groups. It could be because that group only has a small sample size (7 children), and we need more samples to get more accurate conclusions. These relationship can be seen in the Figure 2.

Table 1: Correlation Between EXT and Non EXT Variables

Other Variables	Externalizing Behavior Variables	Correlation
Parent Attention Problem	Substance Variety	0.44
Maternal Postnatal Smoking Consistency	Substance Variety	0.38
Parental Knowledge	Substance Variety	-0.56
Child Disclosure	Substance Variety	-0.40
Parental Control	Substance Variety	-0.37
Prenatal Environmental Smoking Exposure Consistency	Internalizing Problem	0.36
Parent Attention Problem	Externalizing Problem	0.56
Parent Internalizing Problem	Externalizing Problem	0.39
Prenatal Environmental Smoking Exposure Consistency	Expressive Suppression	0.39
Postnatal Environmental Smoking Exposure Consistency	Expressive Suppression	0.54
Parent Attention Problem	Attention Problem	0.60
Parent Internalizing Problem	Attention Problem	0.37
Parent Attention Problem	ADHD Inattentive	0.36
Parent Internalizing Problem	ADHD Inattentive	0.37
Child Disclosure	ADHD Inattentive	-0.40
Parent Attention Problem	ADHD Hyperactive	0.37
Parent Internalizing Problem	ADHD Hyperactive	0.53
Maternal Prenatal Smoking Consistency	ADHD Hyperactive	0.41

Furthermore, we also want to explore the other continuous variables influencing externalizing behavior. We want to know if factors such as the age gap between parents and children could impact these scores or if there might be an inheritance of externalizing behavior from parents to children. Additionally, we aimed to investigate whether the consistency of smoking behavior in the prenatal and postnatal periods affected externalizing behavior. To accomplish

Child Internalizing Problem Score vs SDP & Other Potential Confounders

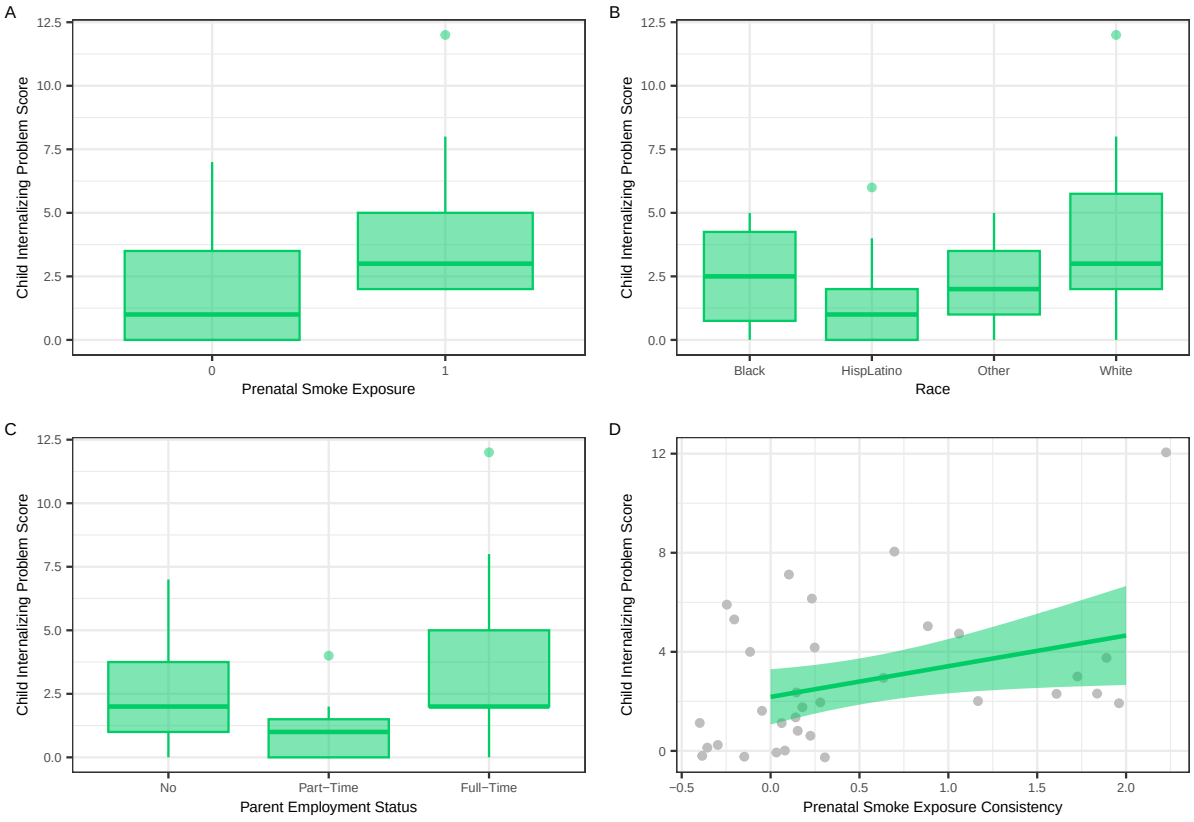


Figure 2. Child Internalizing Problem Score vs SDP & Other Potential Confounders

these objectives, we conducted correlation analyses between externalizing behavior variables and the other continuous variables in our dataset. The result can be seen in Table 3.

Our analysis revealed intriguing findings. We observed multiple correlations between parental attention and internalizing problem scores with children's externalizing behavior variables. This suggests that parental attributes related to attention and internalizing problems influence the externalizing behavior of their children. Similarly, we identified a correlation between smoking behavior consistency and externalizing behavior, implying that the consistency of parental smoking habits in prenatal and postnatal may play a role in shaping their children's externalizing behavior. Also, we could see that the higher the parental knowledge, child disclosure, and parental control, the children are less likely to try any substances. We can see this relationships in the Figure 3.

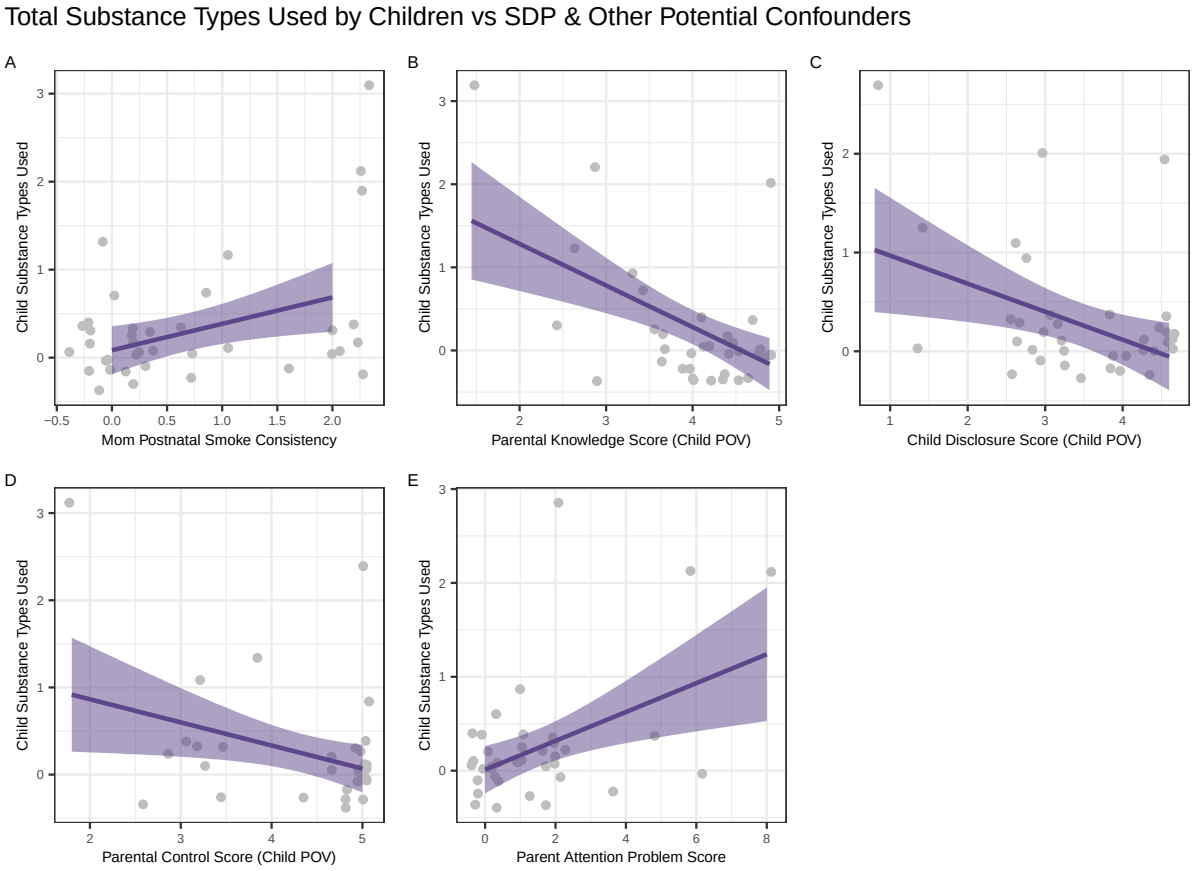


Figure 3. Total Substance Types Used by Children vs SDP & Other Potential Confounders

8. Conclusion

Based on our exploratory analysis, it becomes evident that smoking during pregnancy influences certain aspects of externalizing behavior. In general, exposure to smoking during pregnancy, whether from mother smoking or secondary exposure within the environment, tends to correlate with elevated externalizing behavior scores among children. Additionally, we observed a positive correlation between parental externalizing behavior and that of their children. However, it's important to acknowledge that the sample size utilized in this analysis is relatively small. To gain a more comprehensive understanding of the relationship between externalizing behavior scores and potential confounding factors, acquiring larger sample sizes would facilitate more extensive analyses, including regression analyses. Such analyses could provide deeper insights into the intricate relationships between externalizing behavior and the various factors in this context.

Supplemental Material

Supplemental material can be seen in this github page

Code Appendix

```
##############
### SETUP ###
##############

library(formatR)

knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(fig.align="center")
knitr::opts_chunk$set(fig.width=8, fig.height=6)

###############
### LIBRARY ###
###############

library(tidyverse)
library(ggplot2)
library(naniar)
library(gt)
library(gtsummary)
library(kableExtra)
library(GGally)
library(corrplot)
library(patchwork)
############################
### PREPROCESS RAW DATA ###
############################

# first process child data
setwd("/Users/amirahff/Documents/Brown Biostatistics/PHP 2550/project1")
child_df <- read.csv("K01BB.csv")

child_df <- child_df %>%
  select(c(participant_id:su_interview_complete))  %>%
  filter(redcap_event_name == "child_baseline_arm_1")

# select demographic variables
child_df <- child_df %>%
```

```
  select(-c(participant_id, part, lastgrade, redcap_event_name, famid,
            visit_date, time, redcap_survey_identifier, enroll_timestamp,
            handednesst, tgender, sexorient, whichlang, nativelang, traceoth,
            usborn, relation, guardian, livewith___0:livewith___7,
            attendance, demographics_complete, langpref, pacemaker,
            longlive)) %>%
  rename(taian = trace___0, tasian = trace___1, tnhpi = trace___2,
         tblack = trace___3, twhite = trace___4, trace_other = trace___5)

# drop brief because scoring difficult
child_df <- child_df %>%
  select(-c(brief_ysr_timestamp:brief_ysr_complete))

# cigarette usage summarize
child_df <- child_df %>%
  mutate(cig_ever = suc1, num_cigs_30 = suc11) %>%
  select(-c(suc1:honc10))

# e-cig usage summarize
child_df <- child_df %>%
  mutate(e_cig_ever = ecig1, num_e_cigs_30 = ecig4) %>%
  select(-c(ecig1:ehonc10))

# marijuana usage summarize
child_df <- child_df %>%
  mutate(mj_ever = mj1, num_mj_30 = mj8) %>%
  select(-c(mj1:mpi29))

# alchohol usage summarize
child_df <- child_df %>%
  mutate(alc_ever = alc2, num_alc_30 = alc7) %>%
  select(-c(alc1:alcsus3))

# other drugs and norms - dropping
child_df <- child_df %>%
  select(-c(odrg1:othdrglist,
            perceived_norms_peers_timestamp:perceived_norms_peers_complete,
            substance_use_cigarettes_timesta:substance_use_other_drug_use_com))

# brief problem monitor scoring
child_df <- child_df %>%
```

```
  mutate(bpm_att = rowSums(dplyr::select(., c(bpm1,bpm3,bpm4,bpm5,bpm10))),
         bpm_ext = rowSums(dplyr::select(., c(bpm2,bpm6,bpm7,bpm8,bpm15,
                                              bpm16,bpm17))),
         bpm_int = rowSums(dplyr::select(., c(bpm9,bpm11,bpm12,bpm13,bpm18,
                                              bpm19)))) %>%
  select(-c(brief_problem_monitor_timestamp:brief_problem_monitor_complete))

# emotional regulation
child_df <- child_df %>%
  mutate(erq_cog = rowMeans(dplyr::select(., c(erq1,erq3,erq5,erq7,
                                               erq8,erq10))),
         erq_exp = rowMeans(dplyr::select(., c(erq2,erq4,erq6,
                                               erq9)))) %>%
  select(-c(emotion_regulation_questionnaire:emotion_regulation_questionnair1))

# physical - dropping for the purpose of this research
child_df <- child_df %>%
  select(-c(physical_development_scale_ysr_t:physical_development_scale_ysr_c,
            height1:body_measurements_complete))

# life stress - dropping for the purpose of this research
child_df <- child_df %>%
  select(-c(life_stress_ysr_timestamp:life_stress_ysr_complete))

# parental monitoring scoring
child_df <- child_df %>%
  mutate(pmq_parental_knowledge = (pmq1+pmq2+pmq3+pmq4+pmq5+pmq6+
                                     pmq7+pmq8+(5-pmq9))/9,
         pmq_child_disclosure = (pmqcd1+pmqcd2+(5-pmqcd3)+(5-pmqcd4)+pmqcd5)/5,
         pmq_parental_solicitation = rowMeans(dplyr::select(., pmqps1:pmqps5)),
         pmq_parental_control = rowMeans(dplyr::select(., pmqpc1:pmqpc5))) %>%
  select(-c(parental_monitoring_questionnair:parental_monitoring_questionnai1))

# dysregulation - drop to simplify analysis
child_df <- child_df %>%
  select(-c(dysregulation_inventory_ysr_time:dysregulation_inventory_ysr_comp))

# early adolescent temperament - drop to simplify analysis
child_df <- child_df %>%
  select(-c(early_adolescent_temperament_que:early_adolescent_temperament_qu1))
```

```r
# alcohol and substance abuse - too few observed so remove
child_df <- child_df %>%
  select(-c(miniaud1:minikid_sud_2_complete))

# remove remaining diet questions for purposes of this research
child_df <- child_df %>%
  select(-c(intuitive_eating_scale_timestamp:su_interview_complete))

# parent df
parent_df <- read.csv("K01BB.csv") %>%
  filter(redcap_event_name == "parent_baseline_arm_2") %>%
  select(c(parent_id, page:chart23))

# demographics
parent_df <- parent_df %>%
  select(-c(pgender, marstat, handednessp, plang1:plang3,
            praceoth, ppacemaker, pusa, pedu1:pedu3,
            prelation:parent_demographics_complete, govtasst___0:govtasst___5,
            parent_demographics_asd_timestam,
            parent_demographics_asd_complete)) %>%
  rename(paian = prace___0, pasian = prace___1, pnhpi = prace___2,
         pblack = prace___3, pwhite = prace___4, prace_other = prace___5)

# brief - dropping for difficulty scoring
parent_df <- parent_df %>%
  select(-c(brief_p_on_c_timestamp:brief_p_on_c_complete))

# swan - p on c
parent_df <- parent_df %>%
  mutate(swan_inattentive = rowSums(dplyr::select(., swan1:swan9),
                                    na.rm=TRUE),
         swan_hyperactive = rowSums(dplyr::select(., swan10:swan18),
                                    na.rm=TRUE)) %>%
  select(-c(swan_p_on_c_timestamp:swan_p_on_c_complete))

# connors - drop because swan will be similar
parent_df <- parent_df %>%
  select(-c(connors_p_on_c_timestamp:connors_p_on_c_complete))

# pbpm - parent answering about child
parent_df <- parent_df %>%
```

```
  mutate(bpm_att_p = rowSums(dplyr::select(., c(pbpm1,pbpm3,pbpm4,pbpm5,pbpm10))),
         bpm_ext_p = rowSums(dplyr::select(., c(pbpm2,pbpm6,pbpm7,pbpm8,pbpm15,
                                                pbpm16,pbpm17))),
         bpm_int_p = rowSums(dplyr::select(., c(pbpm9,pbpm11,pbpm12,pbpm13,pbpm18,
                                                pbpm19)))) %>%
  select(-c(bpm_p_on_c_timestamp:bpm_p_on_c_complete))

# alc and drug use
parent_df <- parent_df %>%
  mutate(magic2 = ifelse(magic1 == 0, 0, magic2),
         magic5 = ifelse(magic4 == 0, 0, magic5),
         smoke_exposure_6mo = pmax(magic2, magic5),
         magic8 = ifelse(magic7 == 0, 0, magic8),
         magic11 = ifelse(magic10 == 0, 0, magic11),
         smoke_exposure_12mo = pmax(magic8, magic11),
         magic14 = ifelse(magic13 == 0, 0, magic14),
         magic17 = ifelse(magic16 == 0, 0, magic17),
         smoke_exposure_2yr = pmax(magic14, magic17),
         magic20 = ifelse(magic19 == 0, 0, magic20),
         magic23 = ifelse(magic22 == 0, 0, magic23),
         smoke_exposure_3yr = pmax(magic20, magic23),
         magic26 = ifelse(magic25 == 0, 0, magic26),
         magic29 = ifelse(magic28 == 0, 0, magic29),
         smoke_exposure_4yr = pmax(magic26, magic29),
         magic32 = ifelse(magic31 == 0, 0, magic32),
         magic35 = ifelse(magic34 == 0, 0, magic35),
         smoke_exposure_5yr = pmax(magic32, magic35)
          ) %>%
  select(-c(nidaliftetime___1:inject,penncig2:penn_state_ecigarette_dependenc1,
            penn_state_cigarette_dependence_,
            nida_quick_screen_timestamp,
            nida_quick_screen_complete, magic_timestamp:magic_complete)) %>%
  rename(mom_numcig = penncig1)



# brief - dropping because difficulty scoring
parent_df <- parent_df %>%
  select(-c(briefa_timestamp:briefa_complete))

# parental monitoring - parent answering on child
```

```
parent_df <- parent_df %>%
  mutate(ppmq_parental_knowledge = (ppmq1+ppmq2+ppmq3+ppmq4+ppmq5+ppmq6+
                                      ppmq7+ppmq8+(5-ppmq9))/9,
         ppmq_child_disclosure = (ppmqcd1+ppmqcd2+(5-ppmqcd3)+(5-ppmqcd4)
                                  +ppmqcd5)/5,
         ppmq_parental_solicitation = rowMeans(dplyr::select(., ppmqps1:ppmqps5)),
         ppmq_parental_control = rowMeans(dplyr::select(., ppmqpc1:ppmqpc5))) %>%
  select(-c(ppmq1:ppmqps5,parental_monitoring_questionnai2,
            parental_monitoring_questionnai3))

# chaos - dropping for purposes of this research
parent_df <- parent_df %>%
  select(-c(chaos_timestamp:chaos_complete))

# bpm adult
parent_df <- parent_df %>%
  mutate(bpm_att_a = rowSums(dplyr::select(., c(abpm1,abpm6,abpm7,abpm8,abpm9,
                                                abpm12))),
         bpm_ext_a = rowSums(dplyr::select(., c(abpm3,abpm13,abpm14,abpm17,
                                                abpm18))),
         bpm_int_a = rowSums(dplyr::select(., c(abpm2,abpm4,abpm5,abpm10,abpm15,
                                                abpm16)))) %>%
  select(-c(brief_problem_monitoradult_times:brief_problem_monitoradult_compl))

# parent emotional regulation
parent_df <- parent_df %>%
  mutate(erq_cog_a = rowMeans(dplyr::select(., c(perq1,perq3,perq5,perq7,
                                                 perq8,perq10))),
         erq_exp_a = rowMeans(dplyr::select(., c(perq2,perq4,perq6,
                                                 perq9)))) %>%
  select(-c(emotion_regulation_questionnair2:emotion_regulation_questionnair3))

# adult temperament - drop to simplify analysis
parent_df <- parent_df %>%
  select(-c(adult_temperament_questionnaire_:adult_temperament_questionnaire1))

# etq - drop to simplify analysis
parent_df <- parent_df %>%
  select(-c(eatq_p_on_c_timestamp:eatq_p_on_c_complete))

# stress - dropping for purposes of this research
```

```r
parent_df <- parent_df %>%
  select(-c(nih_toolbox_stress_timestamp:teen_birthday_complete))

# reported smoking during pregnancy and postpartum
parent_df <- parent_df %>%
  select(-c(BBID:ethn2, bl_6:bl_280, s2_10:s2_280, s3_6:s3_280,
            s4_6:s4_280,  s5_6:s5_280, s6_6:s6_280, s7_6:s7_280,
            chart21A:chart23) )   %>%
  rename(mom_smoke_16wk = bl_5,
         mom_smoke_22wk = s2_5,
         mom_smoke_32wk = s3_5,
         mom_smoke_pp1 = s4_5,
         mom_smoke_pp2 = s5_5,
         mom_smoke_pp12wk = s6_5,
         mom_smoke_pp6mo = s7_5,
         cotimean_34wk = wk34cot_cotimean,
         cotimean_pp6mo = mo6momcot_cotimean,
         cotimean_pp6mo_baby = mo6babcot_cotimean)


new_df <- inner_join(parent_df, child_df, by = "parent_id")
# write.csv(new_df, "project1.csv", row.names=FALSE)

##############################
### GENERATE NEW VARIABLES ###
##############################

# Generate new variables, modify and rename some variables
new_df2 = new_df %>%
         # Simplify mom smoke indicator
  mutate(mom_smoke_16wk = factor(mom_smoke_16wk, levels=c('2=No','1=Yes')
                                 , labels = c(0,1))
         , mom_smoke_22wk = factor(mom_smoke_22wk, levels=c('2=No','1=Yes')
                                   , labels = c(0,1))
         , mom_smoke_32wk = factor(mom_smoke_32wk, levels=c('2=No','1=Yes')
                                   , labels = c(0,1))
         , mom_smoke_pp1 = factor(mom_smoke_pp1, levels=c('2=No','1=Yes')
                                  , labels = c(0,1))
         , mom_smoke_pp2 = factor(mom_smoke_pp2, levels=c('2=No','1=Yes')
                                  , labels = c(0,1))
         , mom_smoke_pp12wk = factor(mom_smoke_pp12wk, levels=c('2=No','1=Yes')
```

```
                              , labels = c(0,1))
, mom_smoke_pp6mo = factor(mom_smoke_pp6mo, levels=c('2=No','1=Yes')
                             , labels = c(0,1))
# Make new indicator combining cig & ecig
, cig = pmax(cig_ever,e_cig_ever)
, drug = mj_ever
, alc = alc_ever
# Make tobacco consumption indicator
, pcig = ifelse(nidatob > 0, 1, 0)
# Make drug consumption indicator
, pdrug = ifelse(pmax(nidapres,nidaill) > 0, 1, 0)
# Make alcohol consumption indicator
, palc = ifelse(nidaalc > 0, 1, 0)
# Simplify child race
, race = case_when(tethnic == 1 ~ 'HispLatino'
                   , twhite == 1 ~ 'White'
                   , taian == 1 ~ 'Other'
                   , tasian == 1 ~ 'Other'
                   , tnhpi == 1 ~ 'Other'
                   , tblack == 1 ~ 'Black'
                   , trace_other == 1 ~ 'Other'
                   , TRUE ~ NA)
# Simplify parent race
, prace = case_when(pethnic == 1 ~ 'HispLatino'
                   , pwhite == 1 ~ 'White'
                   , paian == 1 ~ 'Other'
                   , pasian == 1 ~ 'Other'
                   , pnhpi == 1 ~ 'Other'
                   , pblack == 1 ~ 'Black'
                   , prace_other == 1 ~ 'Other'
                   , TRUE ~ NA)
# Rectify swan record for ids mentioned
, swan_inattentive = ifelse(parent_id %in% c(50502,51202,51602,52302
                                             ,53002,53502,53902,54402
                                             ,54602,54702)
                            ,NA
                            ,swan_inattentive)
, swan_hyperactive = ifelse(parent_id %in% c(50502,51202,51602,52302
                                             ,53002,53502,53902,54402
                                             ,54602,54702)
                            ,NA
```

```
                                    ,swan_hyperactive)
        # Rectify income record
        , income = case_when(income == '' ~ NA
                             , income == '250, 000' ~ 250000
                             , TRUE ~ as.numeric(income))
        # Make higher edu (education after high school) indicator
        , phigheredu = case_when(pedu %in% c(0,1,2) ~ 0
                             , pedu %in% c(3,4,5,6) ~ 1
                             , TRUE ~ NA)
        ) %>%
         # Transform mom smoke indicators into numeric
mutate(mom_smoke_16wk = as.numeric(as.character(mom_smoke_16wk))
        , mom_smoke_22wk = as.numeric(as.character(mom_smoke_22wk))
        , mom_smoke_32wk = as.numeric(as.character(mom_smoke_32wk))
        , mom_smoke_pp1 = as.numeric(as.character(mom_smoke_pp1))
        , mom_smoke_pp2 = as.numeric(as.character(mom_smoke_pp2))
        , mom_smoke_pp12wk = as.numeric(as.character(mom_smoke_pp12wk))
        , mom_smoke_pp6mo = as.numeric(as.character(mom_smoke_pp6mo))
        # Make indicator whether child use any substance at all (cig/drug/alc)
        , substance_at_all = case_when(cig == 1 | alc == 1 | drug == 1 ~ 1
                                       , cig == 0 & alc == 0 & drug == 0 ~ 0
                                       , TRUE ~ NA
                                       )
        # Make indicator whether parent use any substance at all (cig/drug/alc)
        , psubstance_at_all = case_when(pcig == 1 | palc == 1 | pdrug == 1 ~ 1
                                       , pcig == 0 & palc == 0 & pdrug == 0 ~ 0
                                       , TRUE ~ NA
                                       )
        # Track how many types of substance used by child (0-3)
        , num_substance_used = rowSums(dplyr::select(., cig:alc), na.rm=TRUE)
        # Track how many types of substance used by parent (0-3)
        , pnum_substance_used = rowSums(dplyr::select(., pcig:palc), na.rm=TRUE)
        # Total child's parental monitoring score
        , pmq_total = pmq_parental_knowledge + pmq_child_disclosure
                      + pmq_parental_solicitation + pmq_parental_control
        # Total parent's parental monitoring score
        , ppmq_total = ppmq_parental_knowledge + ppmq_child_disclosure
                      + ppmq_parental_solicitation + ppmq_parental_control
        # Average of brief problem monitor score (child)
        , bpm_summary = ifelse(is.na(bpm_int)
                               , (bpm_att+bpm_ext)/2
```

```
                            , (bpm_att+bpm_ext+bpm_int)/3)
        # Average of brief problem monitor score (parent)
        , pbpm_summary = ifelse(is.na(bpm_att_a)
                            , (bpm_int_a+bpm_ext_a)/2
                            , ifelse(is.na(bpm_ext_a)
                                    ,(bpm_int_a+bpm_att_a)/2
                                    ,(bpm_att_a+bpm_ext_a+bpm_int_a)/3))
        # Average of emotion regulation score (child)
        , erq_summary = ifelse(is.na(erq_cog)
                            , erq_exp
                            , ifelse(is.na(erq_exp)
                                    , erq_cog
                                    , (erq_exp+erq_cog)/2))
        # Average of emotion regulation score (child)
        , perq_summary = ifelse(is.na(erq_cog_a)
                            , erq_exp_a
                            , ifelse(is.na(erq_exp_a)
                                    , erq_cog_a
                                    , (erq_exp_a+erq_cog_a)/2))
        # Average of ADHD score (child)
        , swan_summary = (swan_hyperactive + swan_inattentive)/2
        # Age gap between parent and child
        , age_gap = abs(tage - page)
        # Simplify income based on this range (https://money.usnews.com/money/
        #    +personal-finance/family-finance/articles/
        #    +where-do-i-fall-in-the-american-economic-class-system)
        , income = case_when(income < 38133 ~ 1
                            , (income >= 38133 & income <57200) ~ 2
                            , (income >= 57200 & income <114000) ~ 3
                            , income >= 114000 ~ 4
                            , TRUE ~ NA)
        ) %>%
        # Indicator whether mom smoked at all during pregnancy or not
mutate(mom_prenatal_smoke = case_when((mom_smoke_16wk == 0 & mom_smoke_22wk == 0
                                    & mom_smoke_32wk == 0) ~ 0
                                    , (mom_smoke_16wk == 1 | mom_smoke_22wk == 1
                                    | mom_smoke_32wk == 1)  ~ 1
                                    , TRUE ~ NA)
        # Indicator whether mom smoked at all post pregnancy or not
        , mom_postnatal_smoke = case_when((mom_smoke_pp12wk == 0
                                        & mom_smoke_pp6mo == 0) ~ 0
```

```
                                    , (mom_smoke_pp12wk == 1
                                       | mom_smoke_pp6mo == 1) ~ 1
                                    , TRUE ~ NA)
        # Indicator whether mom consistently smoked during pregnancy or not
        #    number get bigger if mom consistently reported smoking every followup
        , mom_prenatal_smoke_consistency = mom_smoke_16wk + mom_smoke_22wk + mom_smoke_32
        # Indicator whether mom consistently smoked post pregnancy or not
        #    number get bigger if mom consistently reported smoking every followup
        , mom_postnatal_smoke_consistency = mom_smoke_pp12wk + mom_smoke_pp6mo
        # Indicator whether child exposed by smoked at all during pregnancy or not
        , prenatal_exposure = case_when(smoke_exposure_6mo == 0
                                      & smoke_exposure_12mo == 0 ~ 0
                                      , smoke_exposure_6mo == 1
                                      | smoke_exposure_12mo == 1 ~ 1
                                      , TRUE ~ NA
                                      )
        # Indicator whether child exposed by smoked at all post pregnancy or not
        , postnatal_exposure = case_when(smoke_exposure_2yr == 0
                                       & smoke_exposure_3yr == 0
                                       & smoke_exposure_4yr == 0
                                       & smoke_exposure_5yr == 0 ~ 0
                                       , smoke_exposure_2yr == 1
                                       | smoke_exposure_3yr == 1
                                       | smoke_exposure_4yr == 1
                                       | smoke_exposure_5yr == 1 ~ 1
                                       , TRUE ~ NA)
        # Indicator whether child consistently exposed by smoked during pregnancy or not
        #    number get bigger if child consistently exposed every followup
        , prenatal_exposure_consistency =  smoke_exposure_6mo + smoke_exposure_12mo
        # Indicator whether child consistently exposed by smoked post pregnancy or not
        #    number get bigger if child consistently exposed every followup
        , postnatal_exposure_consistency = smoke_exposure_2yr + smoke_exposure_3yr
                                         + smoke_exposure_4yr + smoke_exposure_5yr
        # Difference between child & parent parental monitoring score
        , pc_pmq_disharmony = abs(pmq_total - ppmq_total)
        ) %>%
        # impute some missing value in prenatal_exposure if mom_prenatal_smoke == 1
    mutate(prenatal_exposure = ifelse(is.na(prenatal_exposure) & mom_prenatal_smoke == 1
                                    ,1
                                    ,prenatal_exposure)
        # impute some missing value in postnatal_exposure if mom_postnatal_smoke == 1
```

```
        , postnatal_exposure = ifelse(is.na(postnatal_exposure)
                                      & mom_postnatal_smoke == 1
                                      ,1
                                      ,postnatal_exposure)
        ) %>%
        # Get prenatal & postnatal smoke exposure combination
        #   (e.g., 01 meaning no exposure during pregnancy but exposure after pregnancy)
  mutate(exposure_pattern = case_when(is.na(prenatal_exposure) ~ NA
                                      , is.na(postnatal_exposure) ~ NA
                                      , TRUE ~ paste(prenatal_exposure,postnatal_exp
                                      )
        # Get prenatal & postnatal mom smoke combination
        #   (e.g., 01 meaning mom didn't smoke during pregnancy but smoke after pregnancy
        , mom_smoke_pattern = case_when(is.na(mom_prenatal_smoke) ~ NA
                                      , is.na(mom_postnatal_smoke) ~ NA
                                      , TRUE ~ paste(mom_prenatal_smoke,mom_postnata
                                      )
  )

####################
### FILTER DATA ###
####################

# Only take necessary variables for analysis
new_df3 = new_df2 %>%
  dplyr::select(-c(plang:prace_other,childasd:cotimean_34wk,bpm_att_p:smoke_exposure_5yr
                   ,language:trace_other,num_cigs_30,num_e_cigs_30,num_mj_30,num_alc_30))
categorical_var = c('psex','employ','pedu','income','tsex','cig_ever'
                    ,'e_cig_ever','mj_ever','alc_ever','cig','drug','alc'
                    ,'pcig','pdrug','palc','race','prace','phigheredu'
                    ,'substance_at_all','psubstance_at_all','mom_prenatal_smoke'
                    ,'mom_postnatal_smoke','prenatal_exposure','postnatal_exposure'
                    ,'exposure_pattern','mom_smoke_pattern')

# Make tableone to get the population characteristic
pop_characteristic = CreateTableOne(data = new_df3, factorVars = categorical_var)
pop_characteristic_tb = print(pop_characteristic)
save(pop_characteristic_tb,file='pop_characteristic_tb.Rda')

# Statistic summary for continuous variables
summary(new_df3 %>% dplyr::select(-categorical_var))
```

```
####################################
### INTERRELATEDNESS BETWEEN EXT ###
####################################

# Make pair plot for all EXT variables
externalizing_behaviors = c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp'
                            ,'swan_inattentive','swan_hyperactive'
                            ,'num_substance_used')

ext_behavior = new_df3[,externalizing_behaviors] %>%
  rename('Attention\nProblem' = 'bpm_att'
         ,'Internalizing\nProblem' = 'bpm_int'
         ,'Externalizing\nProblem' = 'bpm_ext'
         ,'Cognitive\nReappraisal' = 'erq_cog'
         ,'Expressive\nSuppression' = 'erq_exp'
         ,'ADHD\nInattentive' = 'swan_inattentive'
         ,'ADHD\nHyperactive' = 'swan_hyperactive'
         ,'Substance\nVariety' = 'num_substance_used')

# Make pair plot for all EXT variables
ggpairs(ext_behavior
        ,lower = list(continuous=wrap("smooth", colour="grey"))
        ,diag = list(continuous = wrap("densityDiag", colour = "blue"))
        ,upper = list(continuous = wrap("cor", size = 3))
        ,progress = NULL) +
  theme_bw() +
  labs(title = 'Pair Plot of EXT Variables') +
  theme(legend.position = 'bottom'
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5)
        , text=element_text(size=5))

#####################################################################
### COMPARE EXT CHARACTERISTICS BETWEEN DIFFERENT GROUPS (SDP) ###
#####################################################################

# Mom Prenatal Smoke
# Filter dataset
included = c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inattentive'
             ,'swan_hyperactive','num_substance_used','mom_prenatal_smoke')
prenatalComparison = new_df3[,included]
```

```r
# Stratify summary by group
prenatalComparison %>%
  tbl_summary(by = mom_prenatal_smoke
              , type = list(c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inat
                              ,'swan_hyperactive','num_substance_used') ~ 'continuous')
              , statistic = list(all_continuous() ~ "{median} ({p25}, {p75})")
              , missing = 'no'
              , label = list(
                'bpm_att' = 'Attention Problem'
                ,'bpm_int' = 'Internalizing Problem'
                ,'bpm_ext' = 'Externalizing Problem'
                ,'erq_cog' = 'Cognitive Reappraisal'
                ,'erq_exp' = 'Expressive Suppression'
                ,'swan_inattentive' = 'ADHD Inattentive'
                ,'swan_hyperactive' = 'ADHD Hyperactive'
                ,'num_substance_used' = 'Substance Variety'
              )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Maternal Prenatal Smoking**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , latex_options = 'scale_down')

# Prenatal Exposure
# Filter dataset
included = c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inattentive'
             ,'swan_hyperactive','num_substance_used','prenatal_exposure')
prenatalComparison = new_df3[,included]

# Stratify summary by group
prenatalComparison %>%
  tbl_summary(by = prenatal_exposure
              , type = list(c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inat
                              ,'swan_hyperactive','num_substance_used') ~ 'continuous')
              , statistic = list(all_continuous() ~ "{median} ({p25}, {p75})")
              , missing = 'no'
```

```
              , label = list(
                'bpm_att' = 'Attention Problem'
                ,'bpm_int' = 'Internalizing Problem'
                ,'bpm_ext' = 'Externalizing Problem'
                ,'erq_cog' = 'Cognitive Reappraisal'
                ,'erq_exp' = 'Expressive Suppression'
                ,'swan_inattentive' = 'ADHD Inattentive'
                ,'swan_hyperactive' = 'ADHD Hyperactive'
                ,'num_substance_used' = 'Substance Variety'
              )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Prenatal Environmental Smoking Exposur
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , latex_options = 'scale_down')

# Postnatal Exposure
# Filter dataset
included = c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inattentive'
             ,'swan_hyperactive','num_substance_used','postnatal_exposure')
# prenatalComparison = new_df3[,c(9:10,26:30,46,56)]
prenatalComparison = new_df3[,included]

# Stratify summary by group
prenatalComparison %>%
  tbl_summary(by = postnatal_exposure
              , type = list(c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inat
                              ,'swan_hyperactive','num_substance_used') ~ 'continuous')
              , statistic = list(all_continuous() ~ "{median} ({p25}, {p75})")
              , missing = 'no'
              , label = list(
                'bpm_att' = 'Attention Problem'
                ,'bpm_int' = 'Internalizing Problem'
                ,'bpm_ext' = 'Externalizing Problem'
                ,'erq_cog' = 'Cognitive Reappraisal'
                ,'erq_exp' = 'Expressive Suppression'
```

```
                ,'swan_inattentive' = 'ADHD Inattentive'
                ,'swan_hyperactive' = 'ADHD Hyperactive'
                ,'num_substance_used' = 'Substance Variety'
              )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Postnatal Environmental Smoking Exposu
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , latex_options = 'scale_down')

### CHILD INTERNALIZING PROBLEM SCORE ###
bpm_int_p1 = ggplot(subset(new_df3, !is.na(prenatal_exposure))) +
  geom_boxplot(aes(x=as.factor(prenatal_exposure), y=bpm_int)
               , alpha = 0.5, fill = 'springgreen3', color = 'springgreen3') +
  theme_bw() +
  labs(x = 'Prenatal Smoke Exposure'
       ,y = 'Child Internalizing Problem Score'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# bpm_int_p1

bpm_int_p2 = ggplot(subset(new_df3, !is.na(race))) +
  geom_boxplot(aes(x=as.factor(race), y=bpm_int)
               , alpha = 0.5, fill = 'springgreen3', color = 'springgreen3') +
  theme_bw() +
  labs(x = 'Race'
       ,y = 'Child Internalizing Problem Score'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# bpm_int_p2
```

```r
bpm_int_p3 = ggplot(subset(new_df3, !is.na(employ))) +
  geom_boxplot(aes(x=as.factor(employ), y=bpm_int)
               , alpha = 0.5, fill = 'springgreen3', color = 'springgreen3') +
  scale_x_discrete(labels=c('No', 'Part-Time', 'Full-Time')) +
  theme_bw() +
  labs(x = 'Parent Employment Status'
       ,y = 'Child Internalizing Problem Score'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# bpm_int_p3

bpm_int_p4 = ggplot(new_df3) +
  geom_jitter(aes(x=prenatal_exposure_consistency, y=bpm_int), color = 'grey') +
  geom_smooth(aes(x=prenatal_exposure_consistency, y=bpm_int), method = 'lm'
               , alpha = 0.5, fill = 'springgreen3', color = 'springgreen3') +
  theme_bw() +
  labs(x = 'Prenatal Smoke Exposure Consistency'
       ,y = 'Child Internalizing Problem Score'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# bpm_int_p4

# Combine EXT variable plots
patchwork4 = (bpm_int_p1 + bpm_int_p2 + bpm_int_p3 + bpm_int_p4)

options(repr.plot.width=6, repr.plot.height=4)
patchwork4 +
  plot_annotation(tag_levels = 'A'
                  ,title = 'Child Internalizing Problem Score vs SDP & Other Potential Con
                  ,caption = 'Figure 2. Child Internalizing Problem Score vs SDP & Other P
                  ,theme = theme(plot.title = element_text(size = 12)
                                 ,plot.tag = element_text(size = 10)))
############################################################
### CORRELATION OF EXT WITH OTHER CONTINUOUS VARIABLE ###
############################################################
```

```r
# Variables to be compared
EXT = c('bpm_att','bpm_int','bpm_ext','erq_cog','erq_exp','swan_inattentive'
        ,'swan_hyperactive','num_substance_used')
non_EXT = c('bpm_att_a','bpm_int_a','bpm_ext_a','erq_cog_a','erq_exp_a'
            ,'pnum_substance_used','mom_prenatal_smoke_consistency'
            ,'mom_postnatal_smoke_consistency','prenatal_exposure_consistency'
            ,'postnatal_exposure_consistency','pmq_parental_knowledge'
            ,'pmq_child_disclosure','pmq_parental_solicitation'
            ,'pmq_parental_control','tage','age_gap')

# Make correlation matrix
cor_mat = cor(new_df3[,c(EXT,non_EXT)]
              , use="pairwise.complete.obs")

# Remove diagonal, redundant values and rename the values
cor_mat[!lower.tri(cor_mat)] = NA
cor_df = data.frame(cor_mat) %>%
  rownames_to_column() %>%
  gather(key="variable", value="correlation", -rowname) %>%
  filter(abs(correlation) > 0.35)
# Only take non EXT and EXT pair with abs(correlation) > 0.35
cor_df %>%
  rename('variable1' = 'rowname', 'variable2' = 'variable') %>%
  filter(!(variable1 %in% EXT & variable2 %in% EXT)
         , !(variable1 %in% non_EXT & variable2 %in% non_EXT)) %>%
  mutate(variable2 = case_when(variable2=='bpm_att' ~ 'Attention Problem'
                               ,variable2=='bpm_int' ~ 'Internalizing Problem'
                               ,variable2=='bpm_ext' ~ 'Externalizing Problem'
                               ,variable2=='erq_cog' ~ 'Cognitive Reappraisal'
                               ,variable2=='erq_exp' ~ 'Expressive Suppression'
                               ,variable2=='swan_inattentive' ~ 'ADHD Inattentive'
                               ,variable2=='swan_hyperactive' ~ 'ADHD Hyperactive'
                               ,variable2=='num_substance_used' ~ 'Substance Variety')
         ,variable1 = case_when(variable1=='bpm_att_a' ~ 'Parent Attention Problem'
                                ,variable1=='bpm_int_a' ~ 'Parent Internalizing Problem'
                                ,variable1=='bpm_ext_a' ~ 'Parent Externalizing Problem'
                                ,variable1=='erq_cog_a' ~ 'Parent Cognitive Reappraisal'
                                ,variable1=='mom_prenatal_smoke_consistency'
                                ~ 'Maternal Prenatal Smoking Consistency'
                                ,variable1=='mom_postnatal_smoke_consistency'
                                ~ 'Maternal Postnatal Smoking Consistency'
```

```
                                ,variable1=='prenatal_exposure_consistency'
                                ~ 'Prenatal Environmental Smoking Exposure Consistency'
                                ,variable1=='postnatal_exposure_consistency'
                                ~ 'Postnatal Environmental Smoking Exposure Consistency'
                                ,variable1=='pnum_substance_used' ~ 'Parental Substance Va
                                ,variable1=='pmq_parental_knowledge' ~ 'Parental Knowledge
                                ,variable1=='pmq_child_disclosure' ~ 'Child Disclosure'
                                ,variable1=='pmq_parental_solicitation' ~ 'Parental Solici
                                ,variable1=='pmq_parental_control' ~ 'Parental Control'
                                ,variable1=='cotimean_34wk' ~ 'Prenatal Maternal Cotinine'
                                ,variable1=='cotimean_pp6mo_baby' ~ 'Baby Cotinine 6mo'
                                ,variable1=='cotimean_pp6mo' ~ 'Postnatal Maternal Cotinin
                                ,variable1=='age_gap' ~ 'Parent Child Age Gap')) %>%
  arrange(desc(variable2)) %>%
  rename('Other Variables' = 'variable1'
         , 'Externalizing Behavior Variables' = 'variable2'
         , 'Correlation' = 'correlation') %>%
  mutate(Correlation = round(Correlation, 2)) %>%
  kableExtra::kbl(caption = 'Correlation Between EXT and Non EXT Variables'
                  , booktabs = T
                  , escape = F
                  , align = 'c') %>%
  kableExtra::kable_classic(full_width = F
                            ,font_size = 7
                            , html_font = 'Cambria'
                            , latex_options = 'HOLD_position')


### CHILD SUBSTANCE TYPES USED ###
subs_used_p1 = ggplot(subset(new_df3, !is.na(mom_postnatal_smoke_consistency))) +
  geom_jitter(aes(x=mom_postnatal_smoke_consistency, y=num_substance_used), color = 'grey'
  geom_smooth(aes(x=mom_postnatal_smoke_consistency, y=num_substance_used), method = 'lm'
              , alpha = 0.5, fill = 'mediumpurple4', color = 'mediumpurple4') +
  theme_bw() +
  labs(x = 'Mom Postnatal Smoke Consistency'
       ,y = 'Child Substance Types Used'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
```

```
# subs_used_p1

subs_used_p2 = ggplot(subset(new_df3, !is.na(pmq_parental_knowledge))) +
  geom_jitter(aes(x=pmq_parental_knowledge, y=num_substance_used), color = 'grey') +
  geom_smooth(aes(x=pmq_parental_knowledge, y=num_substance_used), method = 'lm'
              , alpha = 0.5, fill = 'mediumpurple4', color = 'mediumpurple4') +
  theme_bw() +
  labs(x = 'Parental Knowledge Score (Child POV)'
       ,y = 'Child Substance Types Used'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# subs_used_p2

subs_used_p3 = ggplot(subset(new_df3, !is.na(pmq_child_disclosure))) +
  geom_jitter(aes(x=pmq_child_disclosure, y=num_substance_used), color = 'grey') +
  geom_smooth(aes(x=pmq_child_disclosure, y=num_substance_used), method = 'lm'
              , alpha = 0.5, fill = 'mediumpurple4', color = 'mediumpurple4') +
  theme_bw() +
  labs(x = 'Child Disclosure Score (Child POV)'
       ,y = 'Child Substance Types Used'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# subs_used_p3

subs_used_p4 = ggplot(subset(new_df3, !is.na(pmq_parental_control))) +
  geom_jitter(aes(x=pmq_parental_control, y=num_substance_used), color = 'grey') +
  geom_smooth(aes(x=pmq_parental_control, y=num_substance_used), method = 'lm'
              , alpha = 0.5, fill = 'mediumpurple4', color = 'mediumpurple4') +
  theme_bw() +
  labs(x = 'Parental Control Score (Child POV)'
       ,y = 'Child Substance Types Used'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
```

```
        , plot.caption = element_text(hjust = 0.5))
# subs_used_p4

subs_used_p5 = ggplot(subset(new_df3, !is.na(bpm_att_a))) +
  geom_jitter(aes(x=bpm_att_a, y=num_substance_used), color = 'grey') +
  geom_smooth(aes(x=bpm_att_a, y=num_substance_used), method = 'lm'
              , alpha = 0.5, fill = 'mediumpurple4', color = 'mediumpurple4') +
  theme_bw() +
  labs(x = 'Parent Attention Problem Score'
       ,y = 'Child Substance Types Used'
       ,) +
  theme(legend.position = 'bottom'
        ,text = element_text(size=7)
        , plot.title = element_text(hjust = 0.5)
        , plot.caption = element_text(hjust = 0.5))
# subs_used_p5

# Combine EXT variable plots
patchwork7 = (subs_used_p1 + subs_used_p2 + subs_used_p3 + subs_used_p4
              +subs_used_p5)

options(repr.plot.width=6, repr.plot.height=4)
patchwork7 +
  plot_annotation(tag_levels = 'A'
                  ,title = 'Total Substance Types Used by Children vs SDP & Other Potentia
                  ,caption = 'Figure 3. Total Substance Types Used by Children vs SDP & Ot
                  ,theme = theme(plot.title = element_text(size = 12)
                                 ,plot.tag = element_text(size = 10)))
```