Examining the Effects of Smoking During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) on Adolescent Self-Regulation, Externalizing Behavior, and Substance Use

Project 1: Exploratory Data Analysis

Yiwen Liang

12/15/2023

Abstract

Background: Smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure are major public health concerns associated with multiple negative outcomes for babies. This project delves into the exploration and analysis of data from Dr. Lauren Micalizzi's research, focusing on examining the association between SDP and ETS and adolescent self-regulation, externalizing behavior, and substance use.

Methods: An initial step involves creating summary variables for prenatal and postnatal smoke exposure indicators (SDP and ETS) and their intensities. Then we pre-process the data (cleaning and formatting) before assessing the missing data pattern. The main body of the exploratory analysis will be performed through univariate analysis, bivariate analysis, and investigating interrelations between targeted variables.

Results: The inherent self-report bias arising from the use of self-reported surveys, is acknowledged. There're substantial amount of missing values in the dataset. Significant differences in substance use status and dosage, and assessments scores between children with varying smoke exposure during and post-pregnancy are observed.

Conclusions: The results supports the idea that there's an association between SDP and ETS and adolescent self-regulation, externalizing behavior, and substance use. The report identifies variables requiring further attention in subsequent research, and highlights unique patterns crucial to addressing Dr. Lauren Micalizzi's research questions. The limitations of the study include bias due to data collection mechanisms and the smalle size of the data, which should be addressed in future anlaysis.

---

1. Introduction

According to the Centers for Disease Control and Prevention (CDC) (CDC 2020), smoking during pregnancy (SDP) is a major public health concern associated with multiple negative outcomes for babies. SDP increases the risks of miscarriage, preterm delivery, and tissue damage in the lung and brain of the unborn baby. Babies whose mothers smoked during pregnancy are approximately three times more likely to die from sudden infant death syndrome (SIDS) than those not exposed to smoke during pregnancy. Furthermore, SDP has more long-term adverse consequences, with children exposed to SDP at higher risks of attention-deficit/hyperactivity disorder (ADHD) (Nomura Y 2010), conduct disorder (Cornelius MD 2009), substance use issues, and self-regulatory problems as they grow up. In addition to maternal smoking during pregnancy, scientists have identified a significant inverse association between environmental tobacco smoke (ETS) exposure and children's self-regulation and educational achievement (Rashidi M 2020).

This project involves an exploratory analysis of data from Dr. Lauren Micalizzi's research, aiming to examine the association between smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) and adolescent self-regulation, externalizing behavior, and substance use. The mothers in this study were randomly selected from participants in a previous study on video intervention to reduce smoking and ETS during and post-pregnancy (Risica PM 2017). The children included in this study are those carried by the mothers during their participation in the previous study. The dataset comprises 49 observations and 78 variables, with each pair of mother and child sharing a unique identification number (`parent_id`). The dataset contains information collected from mothers during the previous study, as well as recently collected data from both mothers and children in the current study.

The data used in this project is not available to the public, hence only the code used for analysis and codebook for the dataset are available at Github, https://github.com/yiwen-liang/PHP_2550_Final_Portfolio.

2. Methods

2.1 Data

The data for this project is sourced from Dr. Lauren Micalizzi, and the mothers participating in this study were randomly selected from the participants in a prior study titled "*A Tailored Video Intervention to Reduce Smoking and Environmental Tobacco Exposure During and After Pregnancy: Rationale, Design and Methods of Baby's Breath*" (Risica PM 2017). The original study's inclusion criteria required mothers to be: 1) smoke-exposed (current smokers, those who quit smoking themselves, or those exposed to smoke from others), 2) pregnant with one baby, and 3) having access to a telephone or video player to receive the intervention (for both control and treatment groups). Participants received newsletters containing content aimed at smoking cessation and avoidance (5 letters during and 3 letters after pregnancy), along with individually tailored videos based on behavioral theory-based survey questions (3 videos during and 2 videos after pregnancy). The study's outcomes included salivary cotinine levels for both mother and baby at 32 weeks gestation and 6 months postpartum, along with self-reported environmental tobacco smoke (ETS) exposure and avoidance behaviors.

In Dr. Micalizzi's current study, a subset of 100 mothers and their children, who were carried by the mothers during their participation in the previous study, are randomly selected for recruitment. Both parents and children undergo laboratory examinations at baseline, 6-month, and 12-month intervals, and complete self-regulatory assessments, self-reported substance use, self-regulation, and electrocardiogram (EKG) on several behaviors. The primary objective of this study is to examine the effects of smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) on multiple child outcomes. These outcomes fall into three main categories: self-regulation, substance use, and externalizing behavior.

- *Emotion Regulation Questionnaire* is provided to measure children's tendency to regulate their emotions, and their average responses related to cognitive reappraisal and expressive suppression are collected separately.

- The study will focus on four common substances: alcohol, cigarettes, e-cigarettes, and marijuana. Children are asked whether or not to ever use a substance, and the number of days using the substance in the past 30 days.

- Variables related to externalizing problems are hard to clearly classified. Parents and children's answers on the *Brief Problem Monitor* on items related to externalizing problems are easy to identify. Also, since we know that externalizing behaviors included Attention-Deficit/Hyperactivity Disorder, responses on *SWAN Rating Scale* indicating whether the child is of hyperactive/impulsive or inattentive type of ADHD should also be considered relevant. *Brief Problem Monitor* also included summary of responses on items related to attention problems, so we may also use it to evaluate externalizing issues. Individuals with Autism Spectrum Disorder are also likely to develop externalizing problems, so we should consider its impact as well.

Table 1: Example of Self-Report Bias

	Self Reported Smoke Exposure from 0 to 6 Months	
	No	Yes
Baby's Urine Cotinine < 1	12	1
Baby's Urine Cotinine > 1	12	6

As described earlier, a significant portion of the data relies on self-reported information, covering aspects such as substance use in children, the mother's smoking status during pregnancy and postpartum, and children's smoke exposure in their early years. Self-report bias, indicative of the disparity between self-reporting and actual events, is a pertinent consideration. In our initial investigation, we specifically address this form of bias. Table 1 illustrates a potential bias example, where columns represent self-reported smoke exposure of children from the mother or partner from 0 to 6 months, and rows show dichotomous urine cotinine levels at 6 months postpartum in infants. According to the Health Encyclopedia of the University of Rochester Medical Center (Rochester Medical Center 2023), cotinine levels in a non-smoker are generally less than 10 nanograms per milliliter (ng/mL), though this criterion might not

be directly applicable to infants. Since there is no universally recognized cotinine level for infants, we chose 1 ng/mL for illustrative purposes.

Cotinine, a chemical that persists in the body after exposure to smoke (nicotine), theoretically increases with higher smoke exposure. However, among children whose parents reported no smoke exposure, half of them tested positive for cotinine levels greater than 1 unit. Even when adjusting the threshold to 3 units, a quarter of the children reported as having no exposure still exhibited cotinine levels exceeding 3 units. This prompts concerns about the accuracy of self-reported variables, highlighting a tendency for parents to under-report their children's smoke exposure in an effort to present themselves as responsible parents.

2.2 Creating Summary Variables of Prenatal and Postnatal Smoke Exposure

Before conducting the actual analysis, we need to create summary variables for both prenatal and postnatal smoke exposure. For the prenatal exposure indicator `SDP`, we set it to 1 if the mother reported smoking at least once at 16, 22, and 32 weeks during pregnancy. This indicates positive smoke exposure during pregnancy. The prenatal exposure intensity `SDP_intensity` is calculated by summing up the smoking status of the mother at 16, 22, and 32 weeks. This variable will be used to assess the intensity of the mother's smoking during pregnancy.

Similarly, for the postnatal exposure indicator `ETS`, we set it to 1 if any of the following conditions are met:

1. The child has been exposed to smoke from 0 to 6 months (`smoke_exposure_6mo`=1),

2. The child has been exposed to smoke from 7 to 12 months (`smoke_exposure_12mo`=1).

3. The mother reported herself as current smoker at 6 months postpartum (`smoke_exposure_12mo`=1).

4. The cotimean level of mother at 6 months postpartum if greater than 10 units.

5. The cotimean level of baby at 6 months postpartum if greater than 5 units.

The postnatal exposure intensity `ETS_intensity` is calculated by summing up the self-reported exposure from the mother or partner from 0 to 6 months, from 7 to 12 months, and in the 2nd to 5th years. This variable will be used to evaluate the intensity of children's smoke exposure after birth.

2.3 Analysis Plan

We briefly talk about the data collection and design of the current study in **2.1**, as well as the potential self-report bias in the data. In **3.1**, we first need to pre-process the data, and then explore the missing data pattern in the given dataset in **3.2**. Further, we'll conduct univariate and bivariate analysis of variables in **3.3** and **3.5**. Additionally, we also investigate the interrelations between prenatal and postnatal exposure and among self-regulation variables as suggested by Dr. Micalizzi in **3.4**. All analyses are performed in R v4.2.2.

3. Results

3.1 Pre-processing

There are several noteworthy variables that warrant attention and modification. Firstly, there are 16 parents and 21 adolescents who identified themselves with more than one ethnicity. All 16 parents and 18 adolescents are biracial, and 3 adolescents indicated affiliation with three ethnicity groups. It's common for individuals to select more than one racial category if their parents are from different racial groups. However, this complicates the process of combining all race-related variables.

Another issue concerns the variable `mom_numcig`, which represents the number of cigarettes the mom usually smokes per day. While attempting to summarize this variable, we identify 10 missing values and encounter 4 non-numeric or unreasonable answers: "`2 black and miles a day`", "`20-25`", "`44989`" and "`None`". From these responses, it appears that answers are provided in a free-text format rather than choosing from closed-ended multiple-choice options. We can interpret "`None`" as simply recording 0, handle "`20-25`" by using some statistics of this range, but "`2 black and miles a day`" is challenging to understand and "`44989`" is clearly too large for a one-day smoking consumption count.

Before exploring the pattern of missing values, there are 4 variables that need manual imputation because of the design of the questionnaire: `num_cigs_30`, `num_e_cigs_30`, `num_mj_30`, and `num_alc_30`. If the subject answered

"No" to the question of whether or not to ever use a substance, then the corresponding dosage for this substance should be 0, instead of missing (`NA`).

Another set of variables need assessing simultaneously are `mom_smoke_pp6mo` and `smoke_exposure_6mo`. `mom_smoke_pp6mo` is collected at 6 months postpartum, while `smoke_exposure_6mo` is reported retrospectively in the later study. The results from the two variables might differ for some mothers. We will combine the two variables, and if at least one of the answers is "yes", then we will record it as smoking (exposed to smoke) at 6 months postpartum.

3.2 Missing Data

One characteristic of longitudinal studies is the potential for missing information at different time points for various reasons. Despite providing gift cards as rewards for follow-up visit, there remains a substantial proportion of missing values in the dataset. There are 63 variables with missing values, and the number and proportion of missingness of each variable are presented in Table 2. `mom_smoke_pp1` has the highest number of missingness, and `childasd` also has more than 50% of missing value.

Table 2: Summary of Missing Values in Each Variable

Variables	N	Proportion (%)	Variables	N	Proportion (%)	Variables	N	Proportion (%)
page	8	16.33	cotimean_34wk	11	22.45	tage	12	24.49
psex	8	16.33	cotimean_pp6mo_baby	11	22.45	tsex	13	26.53
plang	8	16.33	cotimean_pp6mo	11	22.45	language	12	24.49
pethnic	8	16.33	bpm_att_p	13	26.53	tethnic	12	24.49
employ	8	16.33	bpm_ext_p	12	24.49	cig_ever	12	24.49
pedu	8	16.33	bpm_int_p	10	20.41	num_cigs_30	12	24.49
income	12	24.49	smoke_exposure_6mo	10	20.41	e_cig_ever	12	24.49
childasd	28	57.14	smoke_exposure_12mo	10	20.41	num_e_cigs_30	13	26.53
nidaalc	10	20.41	smoke_exposure_2yr	10	20.41	mj_ever	12	24.49
nidatob	10	20.41	smoke_exposure_3yr	11	22.45	num_mj_30	12	24.49
nidapres	11	22.45	smoke_exposure_4yr	11	22.45	alc_ever	13	26.53
nidaill	10	20.41	smoke_exposure_5yr	10	20.41	num_alc_30	14	28.57
momcig	10	20.41	ppmq_parental_knowledge	12	24.49	bpm_att	12	24.49
mom_numcig	11	22.45	ppmq_child_disclosure	12	24.49	bpm_ext	12	24.49
mom_smoke_16wk	1	2.04	ppmq_parental_solicitation	15	30.61	bpm_int	14	28.57
mom_smoke_22wk	7	14.29	ppmq_parental_control	12	24.49	erq_cog	13	26.53
mom_smoke_32wk	9	18.37	bpm_att_a	11	22.45	erq_exp	13	26.53
mom_smoke_pp1	39	79.59	bpm_ext_a	11	22.45	pmq_parental_knowledge	14	28.57
mom_smoke_pp2	20	40.82	bpm_int_a	10	20.41	pmq_child_disclosure	13	26.53
mom_smoke_pp12wk	7	14.29	erq_cog_a	10	20.41	pmq_parental_solicitation	14	28.57
mom_smoke_pp6mo	9	18.37	erq_exp_a	10	20.41	pmq_parental_control	16	32.65

Through further analyses, we observe that parent's age, gender, language spoke, ethnicity, employment and level of education are missing together. Similarly, there's also a substantial overlap in the children's missingness of these variables. Variables starting with "`nida-`" which inquire about the frequency of substance use, and `mom_numcig` are almost missing simultaneously. Additionally, self-reported smoking exposure variables at postpartum are largely missing together. Cotinine level for parent and baby at 6 months postpartum are also missing together, indicating that the mother didn't bring the child in for testing at this time point. Most responses from the same questionnaire or rating scales also appeared to be missing together, suggesting that participants likely skip these questionnaires.

Table 3: Summary of Patients with More Than 50 Missing Values

	50502	51202	51602	52302	53902	54402	54602	54702
N	56	54	60	54	57	58	58	55

The "missing together" observation mentioned earlier is approximate, as we observe that the number of missing values in variables that are missing together are not exactly the same. For such surveys, it's crucial to simultaneously consider the impact of individual differences. Therefore, we also assess the missingness by mother-child pair (`parent_id`). We notice that all pairs of mother and child have at least 1 missing value, and Table 3 lists 8 mothers and their kids who have more than 50 missing values out of 77 variables. These high proportions of missing data

may be attributed to individual reasons, suggesting that some families may be uninterested in participating this study or highly sensitive to their privacy.

3.3 Univariate Analysis

For basic demographic information, I would like to highlight the gender of the parent. Among all parents that answered, one checked male for the gender. Since this variable captures the biological sex assigned at birth, and assuming that it is filled out by the mother, it's surprising to observe. More information is needed to determine if this is due to a filling error or if there is a special circumstance.

Furthermore, we have information on employment status, the highest level of education, and estimated annual household income to provide insights into participants' socioeconomic status. Less than half of the parents are fully-employed, one-fourth of the parents didn't attend college, and the average annual household income is $44,424. These findings align with the design of the original study, where low-income women are intentionally recruited.

In terms of responses on the questionnaires, parents reported higher average scores than their children on the *Emotion Regulation Questionnaire*, both related to cognitive reappraisal and expressive suppression. Average parent-reported scores on their children on the *Brief Problem Monitor* related to attention, externalizing and internalizing are lower than average self-report responses from the children themselves. Furthermore, parent's average self-report scores on self are even lower. On the *Parental Knowledge Questionnaire*, parents provide higher average responses for all four groups compared to the average responses from children. It's evident that there are discrepancies between parents' and children's answers to the same questions. Therefore, if we intend to include the results of one of the questionnaires mentioned, it would be advisable to consider both parent and child responses to eliminate possible bias.

Table 4: Summary of Variables Related to Self-Regulation and Externalizing Issues

Variables	N	Min	Median	Mean	Max	SD
erq_cog	36	1	3.00	3.19	5	0.97
erq_exp	36	1	2.50	2.75	4	0.80
erq_cog_a	39	1	5.67	5.38	7	1.30
erq_exp_a	39	1	3.25	3.46	7	1.58
bpm_att	37	0	2.00	3.00	9	2.62
bpm_ext	37	0	3.00	2.81	8	2.01
bpm_att_p	36	0	1.00	2.06	8	2.22
bpm_ext_p	37	0	1.00	1.68	11	2.50
bpm_att_a	38	0	1.00	1.47	8	1.96
bpm_ext_a	38	0	1.00	1.24	6	1.57
swan_hyperactive	49	0	5.00	6.20	20	6.51
swan_inattentive	49	0	9.00	8.69	22	6.57

The summary statistics of variables related to adolescents' self-regulation (first 4 rows) and externalizing problems (the rest 8 rows) are presented in Table 4. Adolescents' self-regulation is assessed through the average responses of children and parents on the *Emotion Regulation Questionnaire*, focusing on cognitive reappraisal and expressive suppression. Externalizing issues are summarized with additional variables. Variables start with `bpm_` indicate the sum of responses on the *Brief Problem Monitor* related to attention and externalizing problems for both children and parents. The last two rows, where variables start with `swan_`, capture the sum of responses on *SWAN Rating Scale*, where a score of greater than or equal to 6 indicates that the child is likely of ADHD-hyperactive/impulsive or inattentive type.

The other essential component of the dataset is about SDP and ETS. Firstly, we would like to report the statistics of urine cotimean levels in mothers and children at different time points. The average urine cotimean of mothers at week 34 gestation is 49.75, with standard deviation 97.69. The average urine cotimean of mothers at 6 months postpartum is 100.49, with standard deviation 179.34. The average urine cotimean of children at 6 months postpartum is 4.04,

with standard deviation 7.62. However, for the primary exposure of interest, cotimean levels at two time points are far from enough. We create `SDP`

3.4 Interrelations

One of the analytical perspectives suggested by Dr. Micalizzi is to examine the interrelations between prenatal and postnatal exposure and among self-regulation variables. The correlation plot for SDP and ETS intensities is presented in Figure 1. We can see that the correlation between SDP and ETS intensities are 0.45, with p-value = 0.014.
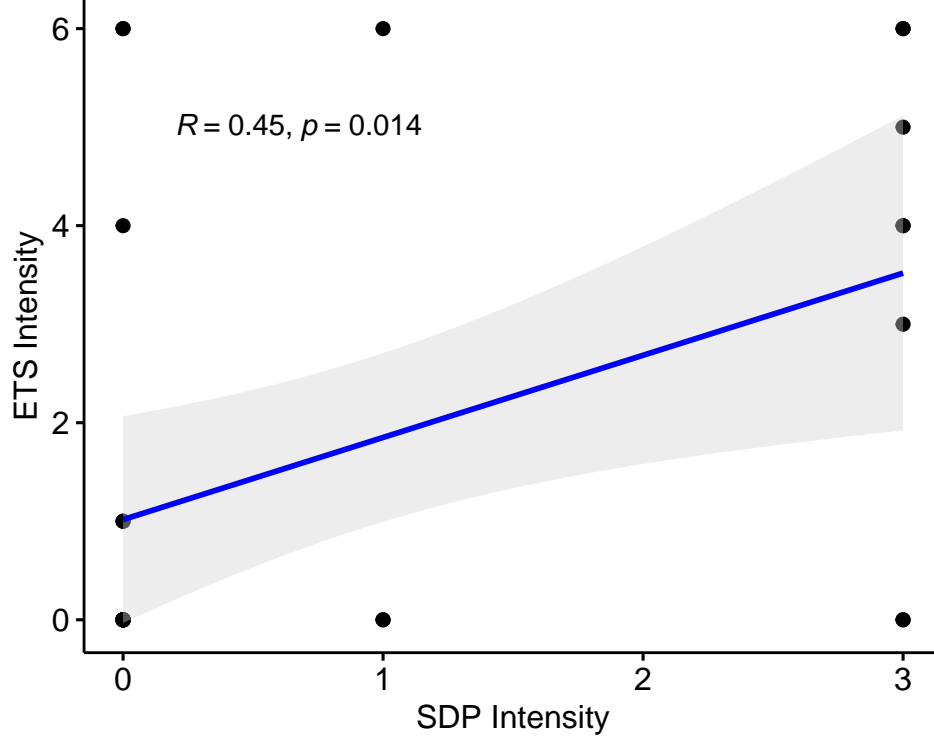

Figure 1: Interrelations Between SDP and ETS Intensity and Among Self-Regulation Variables

Regarding interrelations among self-regulation variables, we notice that some responses from parents and children on the same question are highly correlated. Additionally, `bpm_att` (Brief Problem Monitor responses on attention problems) and `bpm_ext` (Brief Problem Monitor responses on externalizing problems) are correlated, with $\rho = 0.63$, and responses on SWAN Rating Scales on child's ADHD types (hyperactive/impulsive or inattentive) are also correlated, with $\rho = 0.76$.

3.5 Bivariate Analysis

3.5.1 Smoking During Pregnancy (SDP)

Exposure to smoking during pregnancy (SDP) is the first exposure we would like to evaluate that may have an impact on children having attention deficit, conduct disorder, substance using and self-regulatory problems. For SDP, we have 3 variables recording self-report smoking status at 16, 22, 32 weeks gestation and the results from cotinine urine test at 34 weeks gestation. 14 of the mother reported as a smoker at least once during pregnancy. For better presentation, we opt for the summary variable `SDP` instead of a great number of comparisons across different time points and substances.

Figure 2 exhibits the distribution of children's substances use status with respect to smoking status of mothers during pregnancy. Since the proportion of missing values is not negligible, the missingness is also presented in the barplots. It's obvious that more mothers didn't smoke during their pregnancy. Overall, the proportion of kids using substance is higher among those who are exposed to smoke during pregnancy. More studies need to be conducted, controlling for more factors, before we're able to draw the conclusion.
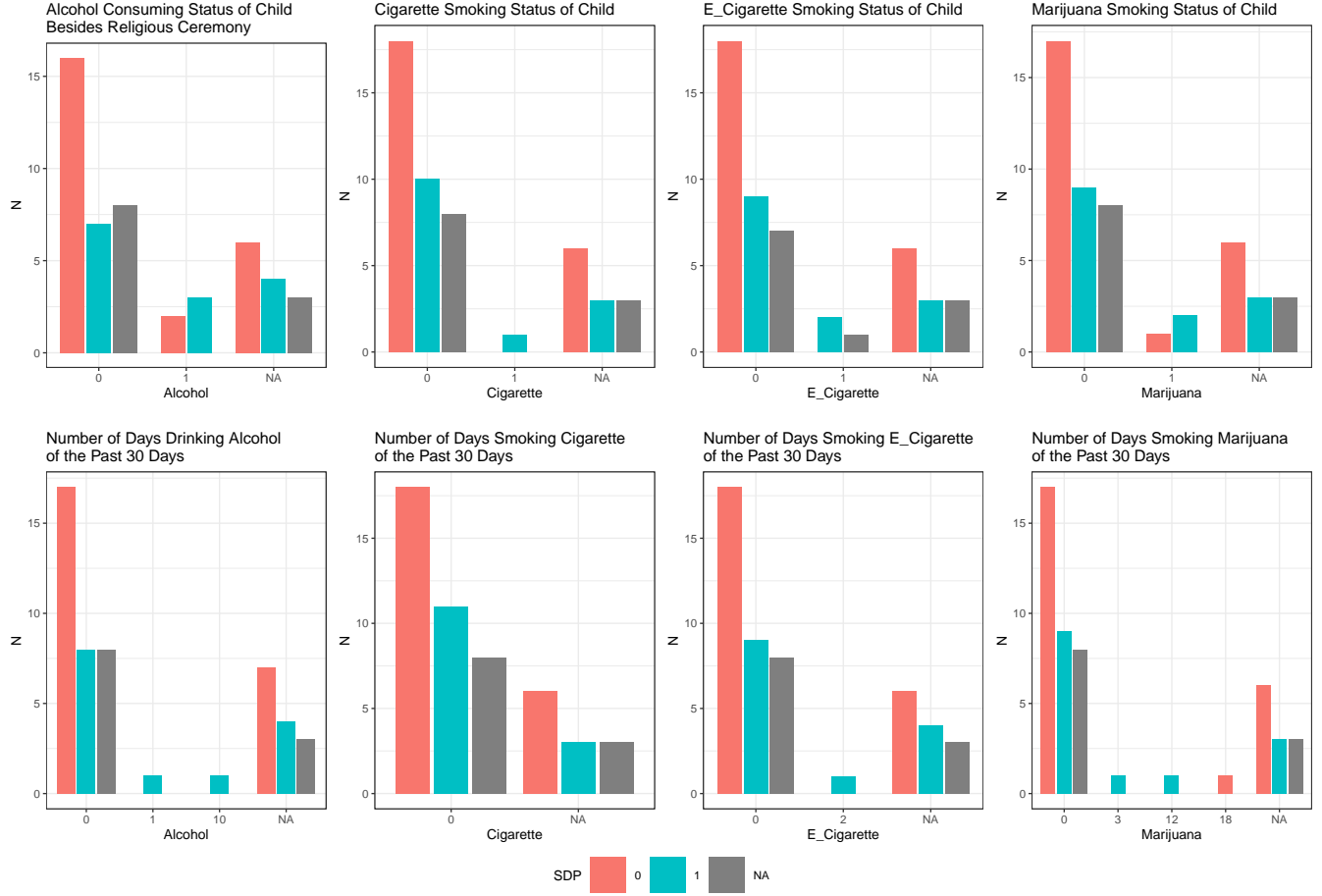
Figure 2: Child's Substance Use Status and Dosage With Respect to SDP

The distributions of dosage of substance use in children with respect to smoking status of mothers with respect to smoke exposure during pregnancy are also shown in Figure 2. Though the dosage (i.e. on how many of the past 30 days did you use a substance) is recorded as continuous variable, but given the limited observed values and concentrated distribution of values (0 or `NA`), I still chose barplots for visualization. Alcohol and marijuana have higher dosages among children, compared to cigarette or e_cigarette. It's surprising that none of the 49 kids had smoked in the 30 days prior to taking the survey. Instead of closed-ended question, dosage had a wider range and it's hard for us to compare the difference in dosage based solely on figures. Again, more analyses need to be performed before we're able to have a clearer understanding on the association between SDP and dosage of substance use.

Table 5: Summary of Responses Related to Child Outcome by SDP at 16, 22, and 32 Weeks

	16 Weeks			22 Weeks			32 Weeks		
Variables	No Smoke	Smoked	p-value	No Smoke	Smoked	p-value	No Smoke	Smoked	p-value
erq_cog	3.12 (2.75, 3.83)	3.46 (3.00, 3.92)	0.5	3.17 (2.83, 4.00)	3.37 (3.00, 3.83)	>0.9	3.10 (2.83, 3.88)	3.36 (3.00, 3.42)	>0.9
erq_exp	2.68 (2.00, 3.13)	2.97 (2.50, 3.50)	0.3	2.55 (2.00, 3.00)	3.13 (2.50, 3.69)	0.085	2.72 (2.13, 3.25)	2.94 (2.50, 3.56)	0.5
bpm_att	3 (0, 4)	4 (2, 7)	0.11	3 (0, 5)	5 (2, 7)	0.070	3 (0, 5)	5 (2, 7)	0.037
bpm_att_p	2 (0, 2)	3 (1, 6)	0.12	2 (0, 2)	3 (1, 6)	0.039	2 (0, 2)	4 (1, 6)	0.030
bpm_ext	3 (1, 4)	3 (1, 4)	0.6	3 (1, 4)	4 (2, 6)	0.2	3 (1, 4)	4 (3, 5)	0.3
bpm_ext_p	1 (0, 2)	2 (0, 4)	0.4	2 (0, 2)	2 (0, 4)	0.6	2 (0, 2)	3 (0, 4)	0.2

The outcome variables indicating self-regulation and externalizing problems of children are mainly responses and score of questionnaires and rating scale which can be summarized and presented in tables. Table 5 are the summaries

of these responses by the smoking status of mother at three time points during pregnancy. All 6 scores in the table are higher for those whose mothers smoked during pregnancy. Sum of responses on the *Brief Problem Monitor* from parents on items related to attention problems on child are significantly different between groups at 22 weeks and 32 weeks, and sum of responses on the *Brief Problem Monitor* from kids on themselves on items related to attention problems is significantly different between groups at 32 weeks.

3.5.2 Environmental Tobacco Smoke (ETS)

Environmental tobacco smoke (ETS) is the other exposure we would like to evaluate that may have an impact on child outcomes. For ETS, we had self-report smoking status at first, second, 12 weeks and 6 months postpartum, and the results from cotinine urine test at 6 months postpartum for both parents and children from the original study. From this study, the retrospective answers to smoke exposure to children in the first 5 years are also collected. Similarly, we'll use the summary variable ETS instead of a great number of comparisons across different time points and substances for better presentation.

Figure 3 exhibits the distribution of children's substances use status with respect to environmental smoke exposure. Similar to what we observe in Figure 2, the proportion of kids using substance is higher among those who are exposed to environmental smoke after being born.
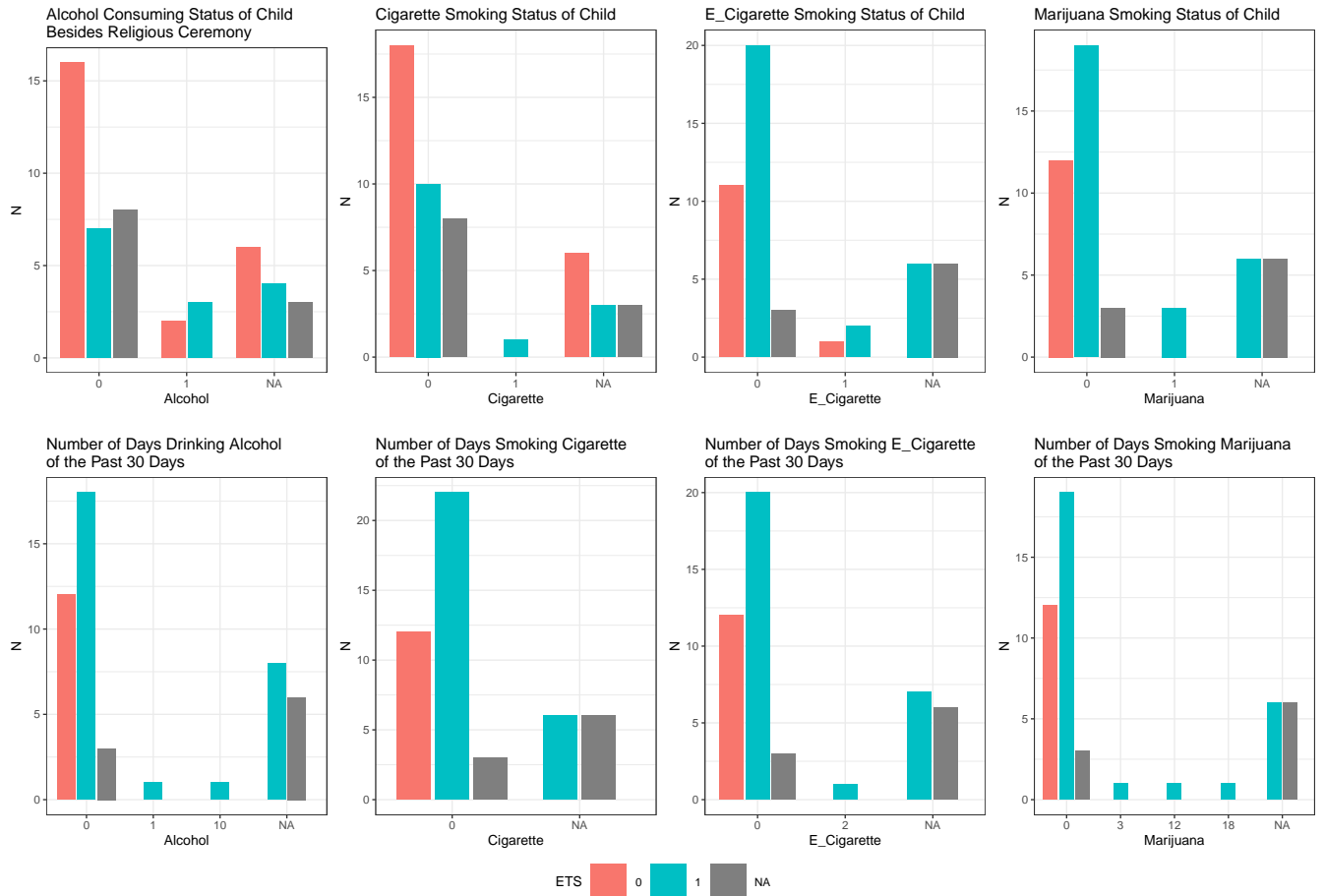

Figure 3: Child's Substance Use Status and Dosage With Respect to ETS

The distributions of dosage of substance use in children with respect to environmental smoke exposure postpartum are shown in Figure 4. Apparently, all children with relatively higher doses of substance in the past 30 days have been exposed to smoke postpartum.

Table 6: Summary of Responses Related to Child Outcome by ETS at 1st and 2nd Postparpum Visit

	1st Postpartum			2nd Postpartum		
Variables	Didn't Smoke	Smoked	p-value	Didn't Smoke	Smoked	p-value
erq_cog	3.14 (2.88, 3.63)	3.92 (3.46, 4.38)	0.4	3.21 (2.83, 4.00)	2.96 (2.96, 3.00)	0.3
erq_exp	3.33 (2.69, 3.81)	2.50 (2.38, 2.63)	0.2	2.56 (2.00, 3.00)	2.95 (2.50, 3.50)	0.4
bpm_att	3 (0, 5)	8 (8, 9)	0.092	2 (0, 5)	4 (2, 5)	0.3
bpm_att_p	1 (0, 1)	6 (5, 6)	0.076	2 (0, 2)	3 (1, 2)	0.5
bpm_ext	3 (2, 4)	6 (5, 6)	0.2	3 (1, 4)	3 (3, 4)	0.4
bpm_ext_p	0 (0, 1)	3 (2, 3)	0.068	2 (0, 3)	2 (0, 4)	0.8

Table 7: Summary of Responses Related to Child Outcome by ETS at 12 Weeks and 6 Months Postpartum

	12 Weeks Postpartum			6 Months Postpartum		
Variables	Didn't Smoke	Smoked	p-value	Didn't Smoke	Smoked	p-value
erq_cog	3.17 (2.83, 3.96)	3.31 (3.00, 3.21)	>0.9	2.95 (2.71, 3.71)	3.62 (3.00, 4.58)	0.2
erq_exp	2.61 (2.00, 3.00)	3.11 (2.50, 3.75)	0.2	2.58 (2.13, 3.00)	3.00 (2.44, 3.75)	0.3
bpm_att	2 (0, 4)	5 (2, 7)	0.009	3 (0, 5)	4 (2, 6)	0.2
bpm_att_p	2 (0, 2)	4 (1, 6)	0.024	2 (0, 2)	3 (1, 4)	0.050
bpm_ext	2 (1, 4)	4 (3, 6)	0.080	2 (1, 4)	3 (2, 4)	0.3
bpm_ext_p	1 (0, 2)	3 (0, 5)	0.14	1 (0, 1)	2 (0, 4)	0.15

Table 6 and Table 7 are the summaries of the responses by the smoke exposure at four time points postpartum: 1st and 2nd postpartum visit, 12 weeks and 6 months postpartum. Differences in 6 scores between with and without tobacco exposure are more complicated for postpartum. Sum of responses on the *Brief Problem Monitor* from parents on items related to attention problems on child are significantly different between groups at 12 weeks and 6 months postpartum, and sum of responses on the *Brief Problem Monitor* from kids on themselves on items related to attention problems is significantly different between groups at 12 weeks postpartum.

4. Discussion and Conclusions

This analysis begins with a thorough examination of the dataset, encompassing a quality check to assess potential bias arising from data collection methods, particularly the use of self-reported surveys. The inherent self-report bias, especially regarding smoking during pregnancy, is acknowledged. Additionally, the study addresses issues related to open-ended questions, necessitating data cleaning and formatting.

I believed that the dataset is missing not at random (MNAR). To mitigate missing values, imputation and modification of variables are performed. The extent of missingness, notably in variables like `mom_smoke_pp1` and `childasd`, poses challenges for analysis, especially with binary variables and multiple groups. Table 2 presents the missing values for each variable, revealing the highest proportions in `mom_smoke_pp1` and `childasd`. Notably, eight mother-child pairs have more than 50 missing values out of 77 variables (excluding `parent_id`). The MNAR nature of missing data complicates analysis, particularly with binary variables and multiple groups.

Subsequent univariate and bivariate analyses further complete the exploratory data analysis. We also investigate the interrelations among prenatal and postnatal exposures and self-regulation variables. Significant differences in substance use and assessments scores between groups with varying smoke exposure during and post-pregnancy are observed. However, these exploratory analyses lack control for confounding factors. While conclusions cannot be drawn, the report identifies variables requiring further attention in subsequent research, highlighting unique patterns crucial to addressing Dr. Lauren Micalizzi's research questions.

The primary limitation of this study is the relatively small sample size, consisting of only 49 mother-child pairs at baseline, coupled with a considerable amount of missing values (66 out of 77 variables). The limited sample size may introduce bias and, in some cases, could alter the direction and strength of the true associations between variables. Another noteworthy limitation is the presence of self-report bias. Given the well-established negative impact of smoking during pregnancy and exposing infants to smoke, it is reasonable to discover that parents might

under-report their smoking behaviors during pregnancy and their children's smoke exposure during early years. This bias could potentially affect the accuracy of the study's findings.

Reference

CDC. 2020. "Smoking During Pregnancy." https://www.cdc.gov/tobacco/basic_information/health_effects/pregnancy/index.htm#:~:text=Smoking%20during%20pregnancy%20can%20cause,maternal%20smoking%20and%20cleft%20lip.&text=Studies%20also%20suggest%20a%20relationship%20between%20tobacco%20and%20miscarriage.

Cornelius MD, Day NL. 2009. "Developmental Consequences of Prenatal Tobacco Exposure." *Curr Opin Neurol.* 22 (2): 121–25. https://doi.org/10.1097/WCO.0b013e328326f6dc.

Nomura Y, Halperin JM, Marks DJ. 2010. "Prenatal Exposure to Maternal and Paternal Smoking on Attention Deficit Hyperactivity Disorders Symptoms and Diagnosis in Offspring." *J Nerv Ment Dis.* 198 (9): 672–78. https://doi.org/10.1097/NMD.0b013e3181ef3489.

Rashidi M, Sahebihagh MH, Mohammadpoorasl A. 2020. "Environmental Tobacco Smoke and Educational Self-Regulation and Achievement in First Grade High School Students." *J Med Life* 13 (2): 229–34. https://doi.org/10.25122/jml-2020-0020.

Risica PM, Parker DR, Gavarkovs A. 2017. "A Tailored Video Intervention to Reduce Smoking and Environmental Tobacco Exposure During and After Pregnancy: Rationale, Design and Methods of Baby's Breath." *Contemp Clin Trials* 52: 1–9. https://doi.org/10.1016/j.cct.2016.10.010.

Rochester Medical Center, University of. 2023. "Nicotine Cotinine (Urine)." urmc.rochester.edu/encyclopedia/content.aspx?content

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# Load the packages
library(cowplot)
library(tidyverse)
library(knitr)
library(kableExtra)
library(corrplot)
library(gt)
library(gtsummary)
library(ggplot2)
library(ggpubr)
library(janitor)
library(mice)

# Load the data
df <- read.csv("project1.csv", na.strings = c("", "NA"))

tbl1 <- df %>%
  mutate(pp6mo_bb = ifelse(cotimean_pp6mo_baby<1, 0, 1)) %>%
  dplyr::select(pp6mo_bb, smoke_exposure_6mo) %>%
  group_by_all() %>%
  tally() %>%
  spread(key = smoke_exposure_6mo, value = n) %>%
  as.data.frame()

tbl1 <- tbl1[1:2, 2:3]
rownames(tbl1) <- c("Baby's Urine Cotinine < 1",
                    "Baby's Urine Cotinine > 1")
colnames(tbl1) <- c("No", "Yes")

tbl1 %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Example of Self-Report Bias",
      row.names = TRUE, booktabs=T, escape=T, align = "c") %>%
  add_header_above(c("", "Self Reported Smoke Exposure from 0 to 6 Months"=2)) %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'))

# Race variables
df %>%
  group_by(parent_id) %>%
  summarise(sum=sum(pethnic, paian, pasian, pnhpi, pblack, pwhite, prace_other)) %>%
  filter(sum!=1)

df %>%
  group_by(parent_id) %>%
  summarise(sum=sum(tethnic, taian, tasian, tnhpi, tblack, twhite, trace_other)) %>%
  filter(sum!=1)

# mom_numcig
df %>%
  filter(mom_numcig %in% c("2 black and miles a day", "20-25", "44989", "None")) %>%
  group_by(mom_numcig) %>%
```

```r
  summarise(N=n()) %>%
  mutate_all(linebreak) %>%
  kbl(col.names = linebreak(c("Contents", "N")),
      caption = "Contents of mom\\_numcig in the Wrong Format",
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('HOLD_position'),
                font_size = 7)

sum(is.na(df$mom_numcig))

df$mom_numcig[df$mom_numcig=="2 black and miles a day"] = NA
df$mom_numcig[df$mom_numcig=="20-25"] = 25
df$mom_numcig[df$mom_numcig=="244989"] = NA
df$mom_numcig[df$mom_numcig=="None"] = 0

# Impute num_cigs_30=0 if cig_ever=0
df$num_cigs_30[ df$cig_ever==0 & is.na(df$num_cigs_30) ] <- 0

# Impute num_e_cigs_30=0 if e_cig_ever=0
df$num_e_cigs_30[ df$e_cig_ever==0 & is.na(df$num_e_cigs_30) ] <- 0

# Impute num_mj_30=0 if mj_ever=0
df$num_mj_30[ df$mj_ever==0 & is.na(df$num_mj_30) ] <- 0

# Impute num_alc_30=0 if alc_ever=0
df$num_alc_30[ df$alc_ever==0 & is.na(df$num_alc_30) ] <- 0

df$smoke_pp6mo[df$mom_smoke_pp6mo == "2=No"] <- 0
df$smoke_pp6mo[df$mom_smoke_pp6mo == "1=Yes"] <- 1
df$smoke_pp6mo <- as.numeric(df$smoke_pp6mo)

for (i in 1:nrow(df)) {
  if (is.na(df$smoke_pp6mo[i])&is.na(df$smoke_exposure_6mo[i])) {
    df$mom_pp6mo[i] = NA
  } else {
    df$mom_pp6mo[i] = max(df$smoke_pp6mo[i], df$smoke_exposure_6mo[i], na.rm=TRUE)
  }
}

# 63 out of 78 variables have missing value
descript1 <- df %>%
  summarise(
    N = colSums(is.na(df)),
    prop = round(colMeans(is.na(df))*100, 2)) %>%
  mutate(Variables = colnames(df)) %>%
  filter(N != 0) %>%
  as.data.frame()

descript1 <- descript1[,c(3,1,2)]

# Display missing data summary table using kable
knitr::kable(
  list(descript1[1:21,], descript1[22:42,], descript1[43:63,]),
  caption = "Summary of Missing Values in Each Variable",
  col.names = linebreak(c("Variables", "N", "Proportion (%)")),
```

```r
    row.names = FALSE,
    booktabs = TRUE,
    escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'),
                font_size = 5.5)

# There're sets of variables with similar number of missingness
# Is there any pattern of personal missingness?
descript2 <- data.frame(parent_id = df$parent_id,
                        N = rowSums(is.na(df))) %>%
  filter(N > 50) %>%
  t() %>%
  as.data.frame() %>%
  row_to_names(row_number = 1)

# Display using kable
descript2 %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Summary of Patients with More Than 50 Missing Values",
      row.names = TRUE,
      booktabs = TRUE, escape = TRUE, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('HOLD_position')) %>%
  column_spec(1, bold = TRUE)

# age, gender, race of both parents and child
df$income <- as.numeric(gsub(",", "", df$income))

tb_demo <- df %>%
  dplyr::select(page, psex, plang, pethnic, paian, pnhpi, pblack, pwhite,
                prace_other, tage, tsex, language, tethnic, taian, tnhpi,
                tblack, twhite, trace_other) %>%
  tbl_summary(
    missing = "no",
    type = list(page ~ 'continuous',
                psex ~ 'categorical',
                plang ~ 'categorical',
                pethnic ~ 'categorical',
                tage ~ 'continuous',
                tsex ~ 'categorical',
                language ~ 'categorical',
                tethnic ~ 'categorical'),
    statistic = all_continuous() ~ c("{mean} ({p25}, {p75})")) %>%
  modify_header(label = "**Variable**",
                stat_1 = "**Statistics**, N={n}") %>%
  modify_caption(caption = "Summary of Basic Demographics") %>%
  as_tibble()

tb_demo[is.na(tb_demo)] <- ""

# Summary of responses and scores
tb_response <- df %>%
  dplyr::select(childasd, erq_cog, erq_exp, erq_cog_a, erq_exp_a,
                bpm_att, bpm_ext, bpm_int,
                bpm_att_p, bpm_ext_p, bpm_int_p,
```

```r
                bpm_att_a, bpm_ext_a, bpm_int_a,
                ppmq_parental_knowledge, ppmq_child_disclosure,
                ppmq_parental_solicitation, ppmq_parental_control,
                pmq_parental_knowledge, pmq_child_disclosure,
                pmq_parental_solicitation, pmq_parental_control,
                swan_hyperactive, swan_inattentive,) %>%
  tbl_summary(
    missing = "no",
    type = list(bpm_att_p ~ 'continuous',
                bpm_ext_p ~ 'continuous',
                ppmq_parental_control ~ 'continuous',
                bpm_att_a ~ 'continuous',
                bpm_ext_a ~ 'continuous',
                bpm_int_a ~ 'continuous',
                bpm_att ~ 'continuous',
                bpm_ext ~ 'continuous'),
    statistic = all_continuous() ~ c("{mean} ({p25}, {p75})")) %>%
  modify_header(label = "**Variable**",
                stat_1 = "**Statistics**, N={n}") %>%
  modify_caption(caption = "Summary of Responses") %>%
  as_tibble()

tb_response[is.na(tb_response)] <- ""

# employ, pedu, income
tb_ses <- df %>%
  dplyr::select(employ, pedu, income) %>%
   tbl_summary(
     missing = "no",
     type = list(income ~ 'continuous')) %>%
  modify_header(label = "**Variable**",
                stat_1 = "**Statistics**, N={n}") %>%
  modify_caption(caption = "Summary of Socioeconomic Status") %>%
  as_tibble()

tb_ses[is.na(tb_ses)] <- ""

# Display using kable
knitr::kable(
  list(tb_demo, tb_response, tb_ses),
  caption = "Summary of Basic Demographics, Responses Related to Child Outcome,
  and Socioeconomic Status",
  col.names = linebreak(c("Variables", "Statistics")),
  row.names = FALSE,
  booktabs = TRUE,
  escape = TRUE,
  align = "c") %>%
  footnote(general = "n(%), Mean (IQR)") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'),
                font_size = 7)

summary(df$cotimean_34wk)
sd(df$cotimean_34wk, na.rm = TRUE)
summary(df$cotimean_pp6mo)
sd(df$cotimean_pp6mo, na.rm = TRUE)
```

```r
summary(df$cotimean_pp6mo_baby)
sd(df$cotimean_pp6mo_baby, na.rm = TRUE)


temp <- df %>%
  dplyr::select(erq_cog, erq_exp, erq_cog_a, erq_exp_a,
                bpm_att, bpm_ext,
                bpm_att_p, bpm_ext_p,
                bpm_att_a, bpm_ext_a,
                swan_hyperactive, swan_inattentive)

stat_temp <- temp %>%
  map_df(.f = ~ broom::tidy(summary(.x)), .id = "variable") %>%
  dplyr::select(-c(q1, q3, na)) %>%
  mutate(sd = c(apply(temp, 2, sd, na.rm=TRUE)),
         N = c(colSums(!is.na(temp))))

stat_temp <- stat_temp[,c(1,7,2:6)]
stat_temp[,c(2,3,6)] <- round(stat_temp[,c(2,3,6)])
stat_temp[,c(4,5,7)] <- round(stat_temp[,c(4,5,7)], 2)


stat_temp %>%
  kbl(caption = "Summary of Variables Related to Self-Regulation and Externalizing Issues",
      col.names=linebreak(c("Variables", "N", "Min", "Median",
                            "Mean", "Max", "SD")),
      booktabs=T, escape=T, align="c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position')) %>%
  row_spec(4, hline_after = TRUE)
df <- df %>%
  mutate(mom_smoke_16wk = case_when(mom_smoke_16wk=="1=Yes" ~ 1,
                                     mom_smoke_16wk=="2=No" ~ 0),
         mom_smoke_22wk = case_when(mom_smoke_22wk=="1=Yes" ~ 1,
                                     mom_smoke_22wk=="2=No" ~ 0),
         mom_smoke_32wk = case_when(mom_smoke_32wk=="1=Yes" ~ 1,
                                     mom_smoke_32wk=="2=No" ~ 0),
         mom_smoke_pp1 = case_when(mom_smoke_pp1=="1=Yes" ~ 1,
                                     mom_smoke_pp1=="2=No" ~ 0),
         mom_smoke_pp2 = case_when(mom_smoke_pp2=="1=Yes" ~ 1,
                                     mom_smoke_pp2=="2=No" ~ 0),
         mom_smoke_pp12wk = case_when(mom_smoke_pp12wk=="1=Yes" ~ 1,
                                     mom_smoke_pp12wk=="2=No" ~ 0),
         mom_smoke_pp6mo = case_when(mom_smoke_pp6mo=="1=Yes" ~ 1,
                                     mom_smoke_pp6mo=="2=No" ~ 0),
         num_cigs_30 = case_when(cig_ever=="0" ~ 0,
                                     cig_ever=="1" ~ num_cigs_30),
         num_e_cigs_30 = case_when(e_cig_ever=="0" ~ 0,
                                     e_cig_ever=="1" ~ num_e_cigs_30),
         num_mj_30 = case_when(mj_ever=="0" ~ 0,
                                     mj_ever=="1" ~ num_mj_30),
         num_alc_30 = case_when(alc_ever=="0" ~ 0,
                                     alc_ever=="1" ~ num_alc_30))

# Create SDP indicator and corresponding intensity variables
df$SDP <- as.factor(ifelse(df$mom_smoke_16wk==1 | df$mom_smoke_22wk==1 |
                           df$mom_smoke_32wk==1, 1, 0))
```

```r
df <- df %>%
  mutate(SDP_intensity = mom_smoke_16wk + mom_smoke_22wk + mom_smoke_32wk)

# Create ETS indicator and corresponding intensity variables
df$ETS <- as.factor(ifelse(df$smoke_exposure_6mo==1 | df$mom_smoke_pp12wk==1 |
                             df$mom_smoke_pp6mo==1 | df$cotimean_pp6mo>=10 |
                             df$cotimean_pp6mo_baby>5, 1, 0))

df <- df %>%
  mutate(ETS_intensity = smoke_exposure_6mo + smoke_exposure_12mo +
           smoke_exposure_2yr + smoke_exposure_3yr +
           smoke_exposure_4yr + smoke_exposure_5yr)

ggscatter(df, x = "SDP_intensity", y = "ETS_intensity",
          add = "reg.line", add.params = list(color = "blue", fill = "lightgray"),
          conf.int = TRUE, cor.coef = TRUE,
          cor.coeff.args = list(method = "pearson", label.x = 0.2, label.y = 5),
          xlab = "SDP Intensity", ylab = "ETS Intensity")

# Correlation plot: interaction
p_corr <- cor(select_if(df, is.numeric) %>% dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bp
               use = "complete.obs")

corrplot(p_corr, method = "shade", shade.col = NA,
         tl.col = "black", addCoef.col = "black", tl.srt = 45,
         cl.pos = "r", number.cex = 0.3, tl.cex = 0.3, cl.cex = 0.4)

temp <- dplyr::select( df, c(cig_ever, e_cig_ever, mj_ever, alc_ever,
                              SDP, SDP_intensity, ETS, ETS_intensity) ) %>%
  pivot_longer(
    cols = c( cig_ever, e_cig_ever, mj_ever, alc_ever ),
    names_to = "Substance",
    values_to = "child_substance")

p_alc <- ggplot( temp[temp$Substance=="alc_ever",],
                 aes(factor(child_substance), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Alcohol Consuming Status of Child\nBesides Religious Ceremony",
        x = "Alcohol",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_cig <- ggplot( temp[temp$Substance=="cig_ever",],
                 aes(factor(child_substance), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Cigarette Smoking Status of Child",
        x = "Cigarette",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_e_cig <- ggplot( temp[temp$Substance=="e_cig_ever",],
                   aes(factor(child_substance), fill = factor(SDP)) ) +
```

```r
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "E_Cigarette Smoking Status of Child",
        x = "E_Cigarette",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_mj <- ggplot( temp[temp$Substance=="mj_ever",],
                 aes(factor(child_substance), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Marijuana Smoking Status of Child",
        x = "Marijuana",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

temp <- dplyr::select( df, c(num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30, SDP, SDP_intensity, ETS, E'
  pivot_longer(
    cols = c( num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30 ),
    names_to = "Substance",
    values_to = "dosage")

p_alc_dosage <- ggplot( temp[temp$Substance=="num_alc_30",],
                 aes(factor(dosage), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Drinking Alcohol\nof the Past 30 Days",
        x = "Alcohol",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_cig_dosage <- ggplot( temp[temp$Substance=="num_cigs_30",],
                 aes(factor(dosage), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking Cigarette\nof the Past 30 Days",
        x = "Cigarette",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_e_cig_dosage <- ggplot( temp[temp$Substance=="num_e_cigs_30",],
                 aes(factor(dosage), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking E_Cigarette\nof the Past 30 Days",
        x = "E_Cigarette",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

p_mj_dosage <- ggplot( temp[temp$Substance=="num_mj_30",],
                 aes(factor(dosage), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking Marijuana\nof the Past 30 Days",
```

```
        x = "Marijuana",
        y = "N", fill = "SDP" ) +
  theme_bw( base_size = 6 )

ggarrange(p_alc, p_cig, p_e_cig, p_mj, ncol = 4)

ggarrange(p_alc_dosage, p_cig_dosage, p_e_cig_dosage, p_mj_dosage,
          ncol = 4, common.legend = TRUE, legend = "bottom")

df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk) %>%
  pivot_longer(cols = c(mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk),
               names_to = "date",
               values_to = "date_value") %>%
  tbl_strata(
    strata = date,
    ~.x %>%
      tbl_summary(
        missing = "no",
        by = date_value,
        type = list(erq_cog ~ 'continuous',
                erq_exp ~ 'continuous',
                bpm_att_p ~ 'continuous',
                bpm_ext_p ~ 'continuous',
                bpm_att ~ 'continuous',
                bpm_ext ~ 'continuous'),
        statistic = all_continuous() ~ c("{mean} ({p25}, {p75})")
      ) %>%
      add_p()
    ) %>%
  kbl(caption = "Summary of Responses Related to Child Outcome by SDP at 16, 22, and 32 Weeks",
      col.names=linebreak(c("Variables", "No Smoke", "Smoked", "p-value",
                            "No Smoke", "Smoked", "p-value",
                            "No Smoke", "Smoked", "p-value")),
      booktabs=T, escape=T, align="c") %>%
  add_header_above(c("", "16 Weeks"=3, "22 Weeks"=3, "32 Weeks"=3 )) %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'),
                font_size = 7) %>%
  column_spec(c(2:3, 5:6, 8:9), width = "6em")
temp <- dplyr::select( df, c(cig_ever, e_cig_ever, mj_ever, alc_ever,
                             SDP, SDP_intensity, ETS, ETS_intensity) ) %>%
  pivot_longer(
    cols = c( cig_ever, e_cig_ever, mj_ever, alc_ever ),
    names_to = "Substance",
    values_to = "child_substance")

p_alc <- ggplot( temp[temp$Substance=="alc_ever",],
                 aes(factor(child_substance), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Alcohol Consuming Status of Child\nBesides Religious Ceremony",
        x = "Alcohol",
        y = "N", fill = "ETS" ) +
```

```r
    theme_bw( base_size = 6 )

p_cig <- ggplot( temp[temp$Substance=="cig_ever",],
                 aes(factor(child_substance), fill = factor(SDP)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Cigarette Smoking Status of Child",
        x = "Cigarette",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

p_e_cig <- ggplot( temp[temp$Substance=="e_cig_ever",],
                   aes(factor(child_substance), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "E_Cigarette Smoking Status of Child",
        x = "E_Cigarette",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

p_mj <- ggplot( temp[temp$Substance=="mj_ever",],
                aes(factor(child_substance), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  labs( title = "Marijuana Smoking Status of Child",
        x = "Marijuana",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

temp <- dplyr::select( df, c(num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30, SDP, SDP_intensity, ETS, ET
  pivot_longer(
    cols = c( num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30 ),
    names_to = "Substance",
    values_to = "dosage")

p_alc_dosage <- ggplot( temp[temp$Substance=="num_alc_30",],
                        aes(factor(dosage), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Drinking Alcohol\nof the Past 30 Days",
        x = "Alcohol",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

p_cig_dosage <- ggplot( temp[temp$Substance=="num_cigs_30",],
                        aes(factor(dosage), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking Cigarette\nof the Past 30 Days",
        x = "Cigarette",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )
```

```r
p_e_cig_dosage <- ggplot( temp[temp$Substance=="num_e_cigs_30",],
                  aes(factor(dosage), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking E_Cigarette\nof the Past 30 Days",
        x = "E_Cigarette",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

p_mj_dosage <- ggplot( temp[temp$Substance=="num_mj_30",],
                  aes(factor(dosage), fill = factor(ETS)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  theme(legend.position = "none") +
  labs( title = "Number of Days Smoking Marijuana\nof the Past 30 Days",
        x = "Marijuana",
        y = "N", fill = "ETS" ) +
  theme_bw( base_size = 6 )

ggarrange(p_alc, p_cig, p_e_cig, p_mj, ncol = 4)

ggarrange(p_alc_dosage, p_cig_dosage, p_e_cig_dosage, p_mj_dosage,
          ncol = 4, common.legend = TRUE, legend = "bottom")

df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                mom_smoke_pp1, mom_smoke_pp2) %>%
  pivot_longer(cols = c(mom_smoke_pp1, mom_smoke_pp2),
               names_to = "date",
               values_to = "date_value") %>%
  tbl_strata(
    strata = date,
    ~.x %>%
      tbl_summary(
        missing = "no",
        by = date_value,
        type = list(erq_cog ~ 'continuous',
                erq_exp ~ 'continuous',
                bpm_att_p ~ 'continuous',
                bpm_ext_p ~ 'continuous',
                bpm_att ~ 'continuous',
                bpm_ext ~ 'continuous'),
        statistic = all_continuous() ~ c("{mean} ({p25}, {p75})")
      ) %>%
      add_p()
    ) %>%
  kbl(caption = "Summary of Responses Related to Child Outcome by ETS at 1st and 2nd Postparpum Visit",
      col.names=linebreak(c("Variables", "Didn't Smoke", "Smoked", "p-value",
                            "Didn't Smoke", "Smoked", "p-value")),
      booktabs=T, escape=T, align="c") %>%
  add_header_above(c("", "1st Postpartum"=3, "2nd Postpartum"=3)) %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'HOLD_position'),
                font_size = 8)

df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
```

```r
                    mom_smoke_pp12wk, smoke_pp6mo) %>%
pivot_longer(cols = c(mom_smoke_pp12wk, smoke_pp6mo),
             names_to = "date",
             values_to = "date_value") %>%
tbl_strata(
  strata = date,
  ~.x %>%
    tbl_summary(
      missing = "no",
      by = date_value,
      type = list(erq_cog ~ 'continuous',
                  erq_exp ~ 'continuous',
                  bpm_att_p ~ 'continuous',
                  bpm_ext_p ~ 'continuous',
                  bpm_att ~ 'continuous',
                  bpm_ext ~ 'continuous'),
      statistic = all_continuous() ~ c("{mean} ({p25}, {p75})")
    ) %>%
    add_p()
  ) %>%
kbl(caption = "Summary of Responses Related to Child Outcome by ETS at 12 Weeks and 6 Months Postpartum"
    col.names=linebreak(c("Variables", "Didn't Smoke", "Smoked", "p-value",
                          "Didn't Smoke", "Smoked", "p-value")),
    booktabs=T, escape=T, align="c") %>%
add_header_above(c("", "12 Weeks Postpartum"=3, "6 Months Postpartum"=3)) %>%
kable_styling(full_width = FALSE,
              latex_options = c('striped', 'HOLD_position'),
              font_size = 8)
```