PHP2550 Project 1: Exploratory Data Analysis

Yiwen Liang

10/08/2023

In this project, we mainly conduct an exploratory analysis of the data from Dr. Lauren Micalizzi's research. The mothers recruited in this study were randomly selected from participants of a previous study on video intervention to reduce smoking and ETS during and post pregnancy. The children included in this study were those carried by the mother when they were participating the previous study. There are 49 observations and 78 variables in the given dataset. Each pair of mother and child shared one identification number (`parent_id`). The dataset includes data from mothers collected in previous studies, and data collected from mothers and children in this study.

This project will be focusing on the aim of examining the effects of smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) on adolescent self-regulation, externalizing behavior and substance use. We'll first check the quality of the data set and then based on the main research objectives, we attempt to answer these questions by performing an exploratory data analysis on the data.

The complete code used to analyze this dataset and codebook for the pre-processed dataset are available at Github, https://github.com/yiwen-liang/PHP_2550_Project_1.

Self-Report Bias

This data contains a large portion of the self-report data, including substance use of children, the mother's smoking status during pregnancy and postpartum, and smoke exposure of children in their early years. Self-report bias is the deviation between the self-report and reality. Based on the nature of the data, we first checked this type of bias.

Table 1 shows an example of the potential bias. Columns represent the self-reported smoke exposure of children from mom or partner from 0 to 6 months, and rows are dichotomous urine cotinine levels at 6 months postpartum from baby. According to Health Encyclopedia of University of Rochester Medical Center, the cotinine levels in a non-smoker are generally less than 10 nanograms per milliliter (ng/mL), but this criteria does not apply to infants. Since a commonly recognized level of cotinine for infants has not been found, I chose 1 ng/mL here for a rough illustration.

Cotinine is the chemical remained in one's body once exposed to smoke (nicotine). Theoretically, the greater the exposure to smoke, the higher the level of cotinine in the urine. However, among children whose parents reported no smoke exposure, half of the children were tested to have cotinine level greater than 1 unit. Even if we change the threshold to 3 unit, there are still 1/4 of the reported-no-exposure children with cotinine level greater than 3. We therefore suspect the accuracy of self-reported variables. Not surprisingly, parents tend to under-report their children's smoke exposure to try to look like responsible parents.

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

Variable Imputation and Modification

Before exploring the pattern of missing values, there are 4 variables that need 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.

Another set of variables need assessing simultaneously are `mom_smoke_pp6mo` and `smoke_exposure_6mo`. `mom_smoke_pp6mo` was collected at 6 months postpartum, while `smoke_exposure_6mo` was reported retrospectively in the later study. The results from 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.

Missing Value Pattern

Another feature of longitudinal study is that there can be missing information at different time point for a variety of reasons. Though gift cards are given as rewards for follow-up visit, there are still a great proportion of missing values in the dataset. There are 63 variables had missing values, and the number and proportion of missingness of each variable were presented in Table 2. `mom_smoke_pp1` had the highest number of missingness, and `childasd` also had 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	10	20.41	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 noticed that parent's age, gender, language spoke, ethnicity, employment and level of education were missing together. Similarly, there was also a great proportion of overlap in the children's missingness of these variables. Variables start with "`nida-`" which asked parents about the frequency of using substance, and `mom_numcig` were almost missing simultaneously. Besides, self-reported smoking exposure variables at postpartum were almost missing together. Cotinine level for parent and baby at 6 months postpartum were also missing together, which indicated that the mother didn't bring the child in for testing at this time point. Most of the responses from the same questionnaire or rating scales also appeared to be missing together. That is, it's likely that the participant skipped this questionnaire.

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" discussed above was rough, where we saw that the number of missingness in variables that were missing together were not exactly the same. For such survey, we must at the same time take into account the impact of individual differences. So, we also evaluated the missingness by mother-child pair (`parent_id`). We

noticed that all pairs of mother and child had at least 1 missing value, and Table 3 listed 8 mothers and their kids that had more than 50 missingness out of 77 variables. These high proportions of missing data may be due to individual reasons. That is, some families may be uninterested in participating this study, or highly sensitive to their privacy.

"Outliers"

There are also multiple "outliers" in some variables that worth noting. First, there are 16 parents and 21 adolescents identified themselves with more than one ethnicity. All 16 parents and 18 adolescents were biracial, and 3 adolescents checked three ethnicity groups. It's common for people to select more than one racial categories if their parents were from different racial groups. Meanwhile, this made it more difficult for us to combine all of our race-related variables.

Another issue is about `mom_numcig`, the number of cigarettes the mom usually smokes per day. While trying to summarizing this variable, we noticed 10 missing values and 4 non-numeric or unreasonable answers: "`2 black and miles a day`", "`20-25`", "`44989`" and "`None`". Based on these responses, we assume that answers were filled in, instead of closed-ended multiple choice. "`None`" can be simply recorded as 0, "`20-25`" can be recorded using some statistics of this range, while "`2 black and miles a day`" is hard to understand and "`44989`" is definitely too large for one-day smoke.

Demographic of Parents and Children

Table 4: Summary of Basic Demographics, Responses Related to Child Outcome, and Socioeconomic Status

Variables	Statistics
page	38 (35, 39)
psex	
0	1 (2.4%)
1	40 (98%)
plang	
0	26 (63%)
1	15 (37%)
pethnic	
0	28 (68%)
1	13 (32%)
paian	4 (8.2%)
pnhpi	8 (16%)
pblack	0 (0%)
pwhite	26 (53%)
prace_other	6 (12%)
tage	14 (13, 15)
tsex	
0	23 (64%)
1	13 (36%)
language	
0	26 (70%)
1	11 (30%)
tethnic	
0	21 (57%)
1	15 (41%)
2	1 (2.7%)
taian	5 (10%)
tnhpi	0 (0%)
tblack	15 (31%)
twhite	19 (39%)
trace_other	5 (10%)

Variables	Statistics
childasd	
0	19 (90%)
1	1 (4.8%)
2	1 (4.8%)
erq_cog	3.19 (2.83, 3.83)
erq_exp	2.75 (2.25, 3.31)
erq_cog_a	5.38 (4.67, 6.50)
erq_exp_a	3.46 (2.50, 4.25)
bpm_att	3 (1, 5)
bpm_ext	3 (1, 4)
bpm_int	3 (1, 4)
bpm_att_p	2 (1, 2)
bpm_ext_p	2 (0, 2)
bpm_int_p	2 (0, 4)
bpm_att_a	1 (0, 2)
bpm_ext_a	1 (0, 2)
bpm_int_a	2 (0, 3)
ppmq_parental_knowledge	4.26 (3.89, 4.67)
ppmq_child_disclosure	3.68 (3.20, 4.20)
ppmq_parental_solicitation	4.18 (3.80, 4.75)
ppmq_parental_control	4.58 (4.60, 5.00)
pmq_parental_knowledge	3.99 (3.67, 4.56)
pmq_child_disclosure	3.43 (2.80, 4.40)
pmq_parental_solicitation	2.98 (2.00, 4.10)
pmq_parental_control	4.35 (3.40, 5.00)
swan_hyperactive	6 (0, 12)
swan_inattentive	9 (1, 13)

Variables	Statistics
employ	
0	12 (29%)
1	7 (17%)
2	22 (54%)
pedu	
0	3 (7.3%)
1	3 (7.3%)
2	5 (12%)
3	15 (37%)
4	3 (7.3%)
5	10 (24%)
6	2 (4.9%)
income	44,424 (20,000, 66,250)

Note:
n(%), Mean (IQR)

The information in the dataset can be categorized into **basic demographic information** including age, gender, language, ethnicity, parents and children's **responses** related to the dependent variables of our interests, **socio-economic status (SES)**, and variables about **smoking** and **substance use** status. Table 4 below presented basic demographic information, responses and SES in three columns respectively. Smoking and substance use conditions

would be discussed later along with figures. The missingness of the relevant variables has been shown in Table 2, so the following tables only characterized the non-missing values.

For the first column, I would like to focus on the gender of parent. Among all parents that answered, one checked male for the gender. Given this variable captures the biological sex assigned at birth, and I assumed that this was filled out by the mother, this was surprising to notice. More information is needed for us to tell if this is due to a filling error or if there is a special circumstance.

In the second column, responses from the same questionnaires or rating scales were presented in clusters. Parents had higher average response than children on the *Emotion Regulation Questionnaire*, both related to cognitive reappraisal and expressive suppression. Average parent-report scores on their children on the *Brief Problem Monitor* related to attention, externalizing and internalizing were lower than average self-report responses from children themselves. Parent's average self-report scores on self were even lower. Parent's average responses on the *Parental Knowledge Questionnaire* for all four groups were higher than the average responses from children. We can see that parents and children's answers on the same questions were different. Therefore, if we would like to include the results of one of the questionnaires above, we should include both parent and child responses to eliminate possible bias.

We have employment status, highest level of education and estimated annual household income to help us understand participants' SES. Less than half of the parents were fully-employed, one-fourth of the parents didn't go to college, and the average annual household income was $44,424. All these findings were consistent with the design of the original study, in which the low-income women were recruited on purpose.

Hazardous Environmental Exposures and Children's Outcome

Our main purpose in this study is to examine the effects of **smoking during pregnancy (SDP)** and **environmental tobacco smoke (ETS)** on multiple child's outcome. Since a number of these longitudinal smoke exposure variables were binary, I chose to present the relationship between child's outcomes and two exposures in plots respectively, instead of in tables. Outcomes consisted of three main category: self-regulation, substance use, and externalizing.

- *Emotion Regulation Questionnaire* was provided to measure children's tendency to regulate their emotions, and their average responses related to cognitive reappraisal and expressive suppression were collected separately.
- The study will focus on four common substances: cigarettes, e-cigarettes, marijuana, and alcohol. Children were 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 were hard to clearly classified. Parents and children's answers on the *Brief Problem Monitor* on items related to externalizing problems were 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. From Table 3, however, we can see that only 1 child was reported diagnosed and 1 was suspected of having ASD. We'll not include plot related to ASD in the following section, but it should be considered carefully in future analyses.

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 had three 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. Figure 1 exhibited the distribution of children's substances use status with respect to smoking status of mothers during pregnancy with barplots. Since the proportion of missing values was not negligible, the missingness was also presented.

It's obvious that more mothers didn't smoke during their pregnancy, and hence the number of children with certain substance use who were exposed to smoke during pregnancy was lower than those whose mother didn't smoke. But when we considered proportion, we noticed that the proportion of kids using substance whose mothers smoked or didn't at different trimester of pregnancy was different. Overall the proportion of kids using substance was higher

among those who were exposed to smoke during pregnancy. More studies need to be conducted, controlling for more factors, before we're able to draw the conclusion.

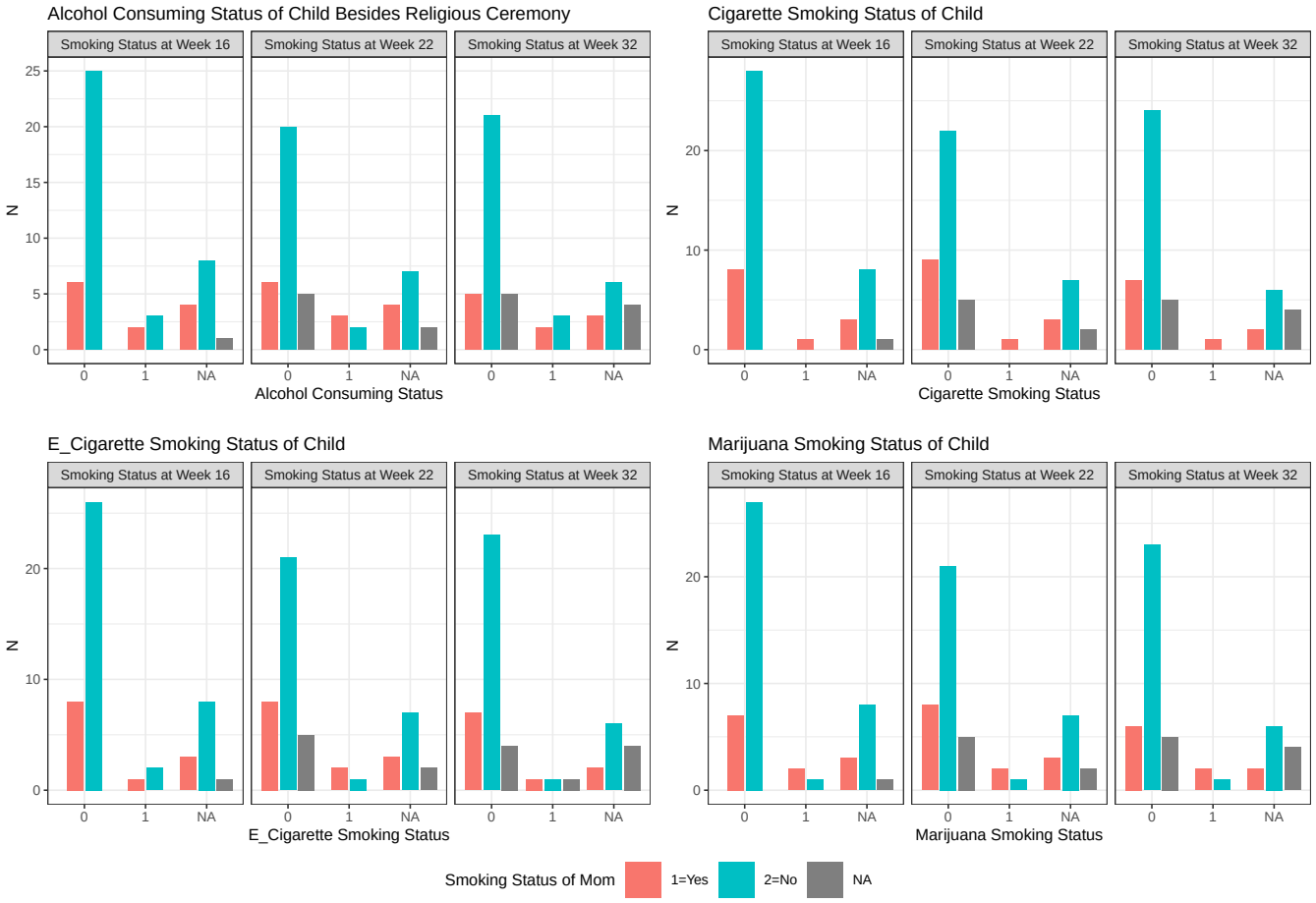


Figure 1: Child's Substance Use Status With Respect to Mom's Smoking Status During Trimesters of Pregnancy

The distribution of dosage of substance use in children with respect to smoking status of mothers at three time point during pregnancy was shown in Figure 2. Though the dosage (i.e. on how many of the past 30 days did you use a substance) was recorded as continuous variable, but given the limited observed values and concentrated distribution of values (0 or `NA`), I still chose barplots for visualization. The dosage was higher for alcohol and marijuana, than 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.

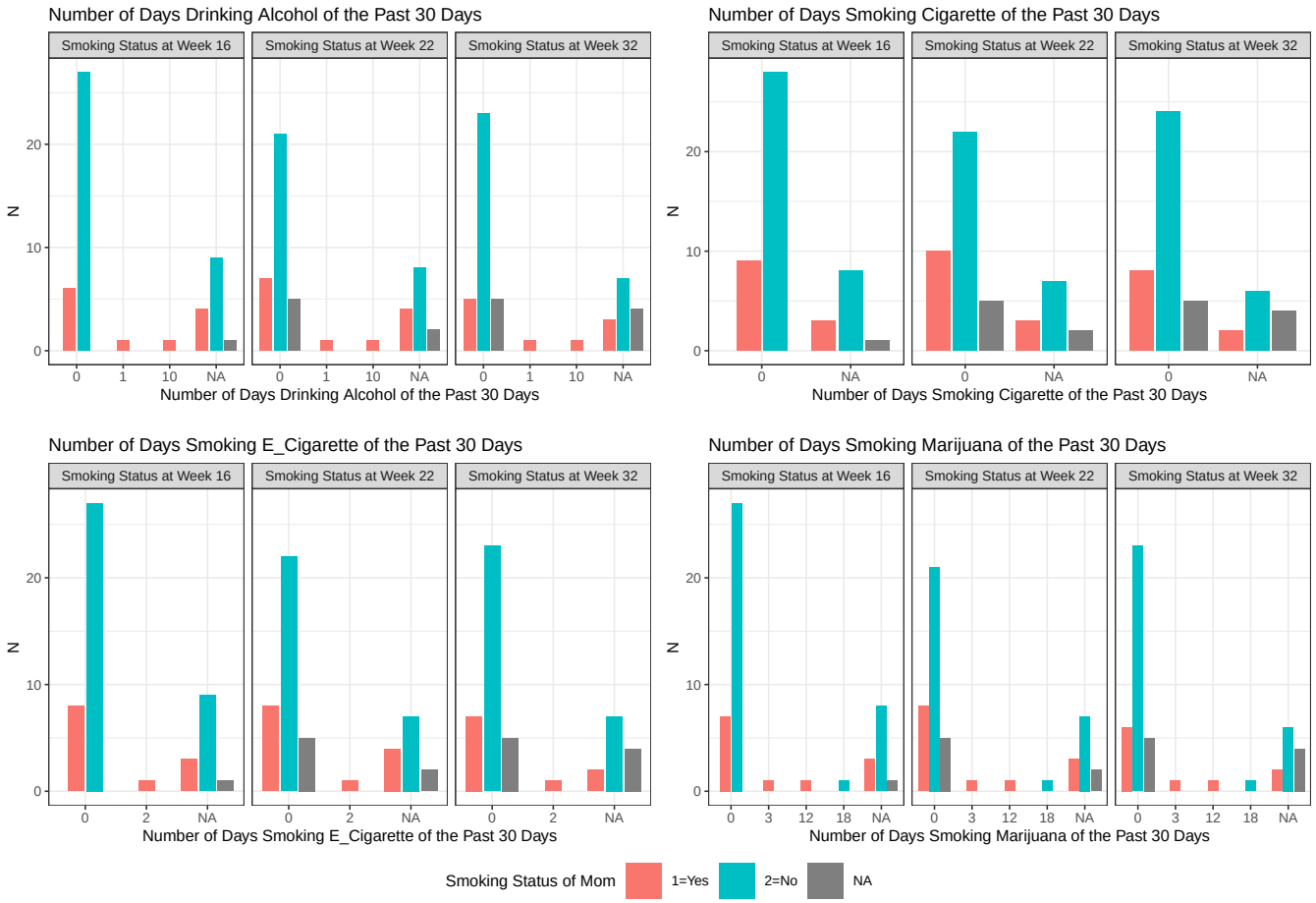


Figure 2: Dosage of Substance Use in Children With Respect to Mom's Smoking Status During Pregnancy

The outcome variables indicating self-regulation and externalizing problems of children were mainly responses and score of questionnaires and rating scale which can be summarized and presented in tables. Table 5-7 were the summaries of these responses by the smoking status of mother at three time points during pregnancy. All 7 scores in three tables were higher for those whose mother smoked during pregnancy, but only the cotinine level was significantly different between two groups at all three time points. Sum of responses on the *Brief Problem Monitor* from parents on items related to attention problems on child were 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 was significant ly different between groups at 32 weeks.

Table 5: Summary of Responses Related to Child Outcome by SDP at 16 Weeks

Variables	Smoked at 16 Weeks	Didn't Smoke at 16 Weeks	p-value
erq_cog	3.46 (3.00, 3.92)	3.12 (2.75, 3.83)	0.5
erq_exp	2.97 (2.50, 3.50)	2.68 (2.00, 3.13)	0.3
bpm_att	4 (2, 7)	3 (0, 4)	0.11
bpm_att_p	3 (1, 6)	2 (0, 2)	0.12
bpm_ext	3 (1, 4)	3 (1, 4)	0.6
bpm_ext_p	2 (0, 4)	1 (0, 2)	0.4
cotimean_34wk	166 (98, 227)	2 (0, 1)	<0.001

Table 6: Summary of Responses Related to Child Outcome by SDP at 22 Weeks

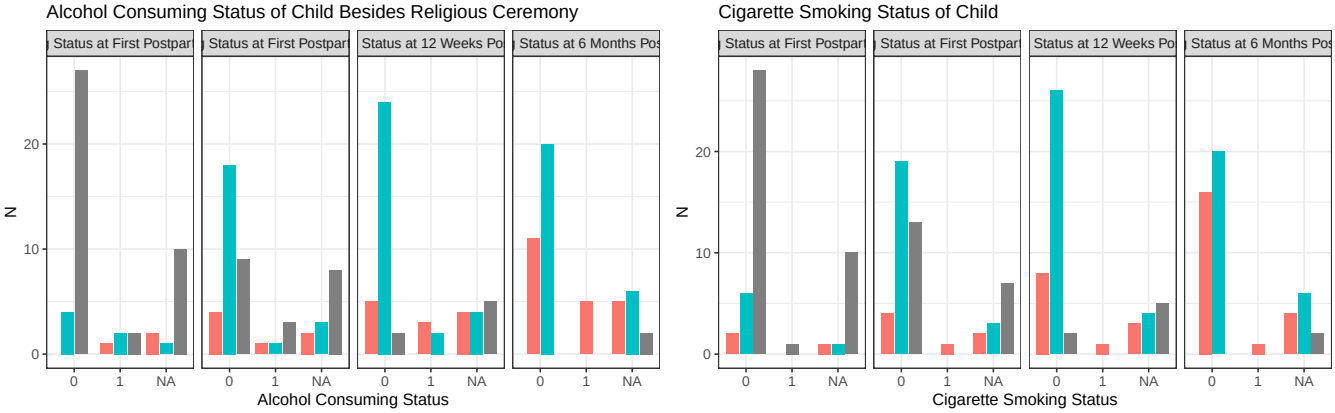
Variables	Smoked at 22 Weeks	Didn't Smoke at 22 Weeks	p-value
erq_cog	3.37 (3.00, 3.83)	3.17 (2.83, 4.00)	>0.9
erq_exp	3.13 (2.50, 3.69)	2.55 (2.00, 3.00)	0.085
bpm_att	5 (2, 7)	3 (0, 5)	0.070
bpm_att_p	3 (1, 6)	2 (0, 2)	0.039
bpm_ext	4 (2, 6)	3 (1, 4)	0.2
bpm_ext_p	2 (0, 4)	2 (0, 2)	0.6
cotimean_34wk	154 (71, 199)	2 (0, 2)	<0.001

Table 7: Summary of Responses Related to Child Outcome by SDP at 32 Weeks

Variables	Smoked at 32 Weeks	Didn't Smoke at 32 Weeks	p-value
erq_cog	3.36 (3.00, 3.42)	3.10 (2.83, 3.88)	>0.9
erq_exp	2.94 (2.50, 3.56)	2.72 (2.13, 3.25)	0.5
bpm_att	5 (2, 7)	3 (0, 5)	0.037
bpm_att_p	4 (1, 6)	2 (0, 2)	0.030
bpm_ext	4 (3, 5)	3 (1, 4)	0.3
bpm_ext_p	3 (0, 4)	2 (0, 2)	0.2
cotimean_34wk	183 (118, 255)	2 (0, 2)	<0.001

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 were also collected. Figure 3 exhibited the distribution of children's substances use status with respect to smoke exposure in the first five years after they were born with barplots.



We did notice a great number of missingness in the response at first postpartum visit. The first to the third months may be the hardest time for new parents, so it's very likely that they were simply too busy to make that visit. Another interesting pattern was that the number of postpartum mothers who smoked increased over time. This may be due to postpartum depression or other complex reasons that deserves attention in subsequent studies.

The distribution of dosage of substance use in children with respect to smoking exposure years after been born was shown in Figure 4. The dosage was still presented with barplots. Apparently, the distribution of dosage of each substance was roughly the constant with no big change. Combined with Figure 3, these figures suggested that the impact of environmental tobacco smoke exposure on substance use was greater for the first year. Its effect became less significant after the infant was greater than 1 year old. Hence the timing of the exposure was very important will performing regression analysis.

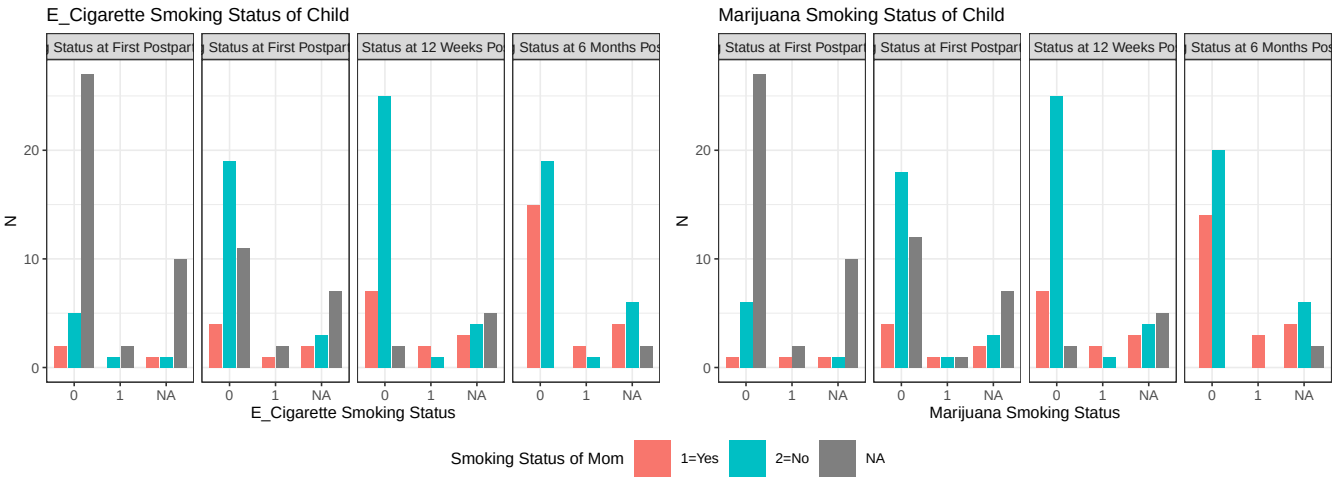


Figure 3: Child's Substance Use Status With Respect to Mom's Postpartum Smoking Status

According to what we observed from Table 5-7, only time points within one year postpartum were evaluated. Table 8-11 were the summaries of these responses by the smoke exposure at four time points postpartum: 1st and 2nd postpartum visit, 12 weeks postpartum, and 6 months postpartum. Differences in 7 scores between with and without tobacco exposure at four time points were more complicated postpartum. The cotinine level was still significantly different between groups at all four time points. Sum of responses on the *Brief Problem Monitor* from parents on items related to attention problems on child were 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 was significantly different between groups at 12 weeks postpartum.

Table 8: Summary of Responses Related to Child Outcome by ETS at 1st Postpartum Visit

Variables	Smoked at 1st Postpartum Visit	Didn't Smoke at 1st Postpartum Visit	p-value
erq_cog	3.92 (3.46, 4.38)	3.14 (2.88, 3.63)	0.4
erq_exp	2.50 (2.38, 2.63)	3.33 (2.69, 3.81)	0.2
bpm_att	8 (8, 9)	3 (0, 5)	0.092
bpm_att_p	6 (5, 6)	1 (0, 1)	0.076
bpm_ext	6 (5, 6)	3 (2, 4)	0.2
bpm_ext_p	3 (2, 3)	0 (0, 1)	0.068
cotimean_34wk	215 (157, 250)	3 (0, 2)	0.017

Table 9: Summary of Responses Related to Child Outcome by ETS at 2nd Postpartum Visit

Variables	Smoked at 2nd Postpartum Visit	Didn't Smoke at 2nd Postpartum Visit	p-value
erq_cog	2.96 (2.96, 3.00)	3.21 (2.83, 4.00)	0.3
erq_exp	2.95 (2.50, 3.50)	2.56 (2.00, 3.00)	0.4
bpm_att	4 (2, 5)	2 (0, 5)	0.3
bpm_att_p	3 (1, 2)	2 (0, 2)	0.5
bpm_ext	3 (3, 4)	3 (1, 4)	0.4
bpm_ext_p	2 (0, 4)	2 (0, 3)	0.8
cotimean_34wk	178 (92, 251)	7 (0, 1)	<0.001

Figure 4: Child's Substance Use Status With Respect to Mom's Postpartum Smoking Status

Table 10: Summary of Responses Related to Child Outcome by ETS at 12 Weeks Postpartum

Variables	Smoked at 12 Weeks Postpartum	Didn't Smoke at 12 Weeks Postpartum	p-value
erq_cog	3.31 (3.00, 3.21)	3.17 (2.83, 3.96)	>0.9
erq_exp	3.11 (2.50, 3.75)	2.61 (2.00, 3.00)	0.2
bpm_att	5 (2, 7)	2 (0, 4)	0.009
bpm_att_p	4 (1, 6)	2 (0, 2)	0.024
bpm_ext	4 (3, 6)	2 (1, 4)	0.080
bpm_ext_p	3 (0, 5)	1 (0, 2)	0.14
cotimean_34wk	133 (61, 163)	2 (0, 1)	<0.001

Table 11: Summary of Responses Related to Child Outcome by ETS at 6 Months Postpartum

Variables	Smoke Exposure at 6 Months Postpartum	No Smoke Exposure at 6 Months Postpartum	p-value
erq_cog	2.95 (2.71, 3.71)	3.62 (3.00, 4.58)	0.2
erq_exp	2.58 (2.13, 3.00)	3.00 (2.44, 3.75)	0.3
bpm_att	3 (0, 5)	4 (2, 6)	0.2
bpm_att_p	2 (0, 2)	3 (1, 4)	0.050
bpm_ext	2 (1, 4)	3 (2, 4)	0.3
bpm_ext_p	1 (0, 1)	2 (0, 4)	0.15
cotimean_34wk	2 (0, 2)	126 (6, 167)	0.004

Conclusion

I first check the quality of the dataset, including assessing the potential bias due to data collection methods (i.e. self-reported survey), imputing and modifying variables and then describing the missing data pattern. We're all aware that smoking during pregnancy can have a negative impact on child, and we're not surprising to find out that parents under-report the smoke exposure. This leads to the self-report bias. Another issue is with open-ended questions that the answers received may not be of the format we want, hence data cleaning and formatting are required before performing further analyses. The variables including answers related to the answers of previous question also needs imputation to reduce missingness due to the design of the survey.

The missing values of each variables were presented in Table 2. We noticed that `mom_smoke_pp1` and `childasd` had the most proportion of missing, and 8 pairs of mother and child had more than 50 missing values out of 77 variables (exclude `parent_id`). According to the interpretation next to Table 1, I believed that the dataset is missing not at random (MNAR). Missing values in this dataset indeed made it difficult for us to perform analysis, especially because we have some binary variables and multiple groups to compare. The amount of missingness in both baseline and follow-up information also impede us from performing imputations on all missing values.

Then I created tables and figures to summarize the variables we have and their associations between each other, especially for those of our research interests. We did observe significant difference in some scores between those with different smoke exposure during and post pregnancy. But the exploratory analyses performed above didn't control for other factors while assessing the relationship between two variables. Though we cannot draw conclusions, this report identified a number of variables that need attention in further research and some unique patterns that may worth testing and may be essential in answering Dr. Lauren Micalizzi's research questions.

Code Appendix

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

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

# Load the data
setwd("/Users/liangxiaohu/Desktop/PHP 2550/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'),
                font_size = 7)

# 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
```

```r
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 (%)")),
  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'),
                font_size = 7) %>%
  column_spec(1, bold = TRUE)

# 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) %>%
  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

# 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,
                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)

# SDP on substance use
temp <- dplyr::select( df, c(cig_ever, e_cig_ever, mj_ever, alc_ever, mom_smoke_16wk,
```

```
                              mom_smoke_22wk, mom_smoke_32wk) ) %>%
  pivot_longer(
    cols = c( cig_ever, e_cig_ever, mj_ever, alc_ever ),
    names_to = "Substance",
    values_to = "child_substance") %>%
  pivot_longer(
    cols = c( mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk ),
    names_to = "Trimester",
    values_to = "mom_smoke")

tri.labs <- c("Smoking Status at Week 16", "Smoking Status at Week 22",
              "Smoking Status at Week 32")
names(tri.labs) <- c("mom_smoke_16wk", "mom_smoke_22wk", "mom_smoke_32wk")

p_alc <- ggplot( temp[temp$Substance=="alc_ever",],
                 aes(factor(child_substance), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Alcohol Consuming Status of Child Besides Religious Ceremony",
        fill = "Smoking Status of Mom", x = "Alcohol Consuming Status", y = "N" ) +
  theme_bw( base_size = 7 )

p_cig <- ggplot( temp[temp$Substance=="cig_ever",],
                 aes(factor(child_substance), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single"),
            show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Cigarette Smoking Status of Child", fill = "Smoking Status of Mom",
        x = "Cigarette Smoking Status", y = "N" ) +
  theme_bw( base_size = 7 )

p_e_cig <- ggplot( temp[temp$Substance=="e_cig_ever",],
                   aes(factor(child_substance), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "E_Cigarette Smoking Status of Child", fill = "Smoking Status of Mom",
        x = "E_Cigarette Smoking Status", y = "N" ) +
  theme_bw( base_size = 7 )

p_mj <- ggplot( temp[temp$Substance=="mj_ever",],
                aes(factor(child_substance), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Marijuana Smoking Status of Child", fill = "Smoking Status of Mom",
        x = "Marijuana Smoking Status", y = "N" ) +
  theme_bw( base_size = 7 )

ggarrange(p_alc, p_cig, ncol = 2)

ggarrange(p_e_cig, p_mj, ncol = 2, common.legend = TRUE, legend = "bottom")

# SDP on substance dosage
temp <- dplyr::select( df, c(num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30, mom_smoke_16wk,
```

```
                              mom_smoke_22wk, mom_smoke_32wk) ) %>%
  pivot_longer(
    cols = c( num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30 ),
    names_to = "Substance",
    values_to = "dosage") %>%
  pivot_longer(
    cols = c( mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk ),
    names_to = "Trimester",
    values_to = "mom_smoke")

p_alc_dosage <- ggplot( temp[temp$Substance=="num_alc_30",],
                  aes(factor(dosage), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single"), show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Number of Days Drinking Alcohol of the Past 30 Days",
        x = "Number of Days Drinking Alcohol of the Past 30 Days",
        y = "N", fill = "Smoking Status of Mom" ) +
  theme_bw( base_size = 7 )

p_cig_dosage <- ggplot( temp[temp$Substance=="num_cigs_30",],
                  aes(factor(dosage), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single"), show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Number of Days Smoking Cigarette of the Past 30 Days",
        x = "Number of Days Smoking Cigarette of the Past 30 Days",
        y = "N", fill = "Smoking Status of Mom" ) +
  theme_bw( base_size = 7 )

p_e_cig_dosage <- ggplot( temp[temp$Substance=="num_e_cigs_30",],
                    aes(factor(dosage), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Number of Days Smoking E_Cigarette of the Past 30 Days",
        x = "Number of Days Smoking E_Cigarette of the Past 30 Days",
        y = "N", fill = "Smoking Status of Mom" ) +
  theme_bw( base_size = 7 )

p_mj_dosage <- ggplot( temp[temp$Substance=="num_mj_30",],
                 aes(factor(dosage), fill = factor(mom_smoke)) ) +
  geom_bar( position = position_dodge2(preserve = "single") ) +
  facet_wrap( vars(Trimester), ncol = 3, labeller = labeller(Trimester=tri.labs) ) +
  labs( title = "Number of Days Smoking Marijuana of the Past 30 Days",
        x = "Number of Days Smoking Marijuana of the Past 30 Days",
        y = "N", fill = "Smoking Status of Mom" ) +
  theme_bw( base_size = 7 )

ggarrange(p_alc_dosage, p_cig_dosage, ncol = 2)

ggarrange(p_e_cig_dosage, p_mj_dosage, ncol = 2, common.legend = TRUE, legend = "bottom")

# Tables for all scores at 16wk
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_16wk) %>%
```

```
  tbl_summary(
    missing = "no",
    by = mom_smoke_16wk,
    type = list(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 Weeks",
      col.names=linebreak(c("Variables",
                            "Smoked at 16 Weeks",
                            "Didn't Smoke at 16 Weeks", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'hold_position'),
                font_size = 8) %>%
  row_spec(7, color = "blue")

# Tables for all scores at 22wk
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_22wk) %>%
  tbl_summary(
    missing = "no",
    by = mom_smoke_22wk,
    type = list(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 22 Weeks",
      col.names=linebreak(c("Variables",
                            "Smoked at 22 Weeks",
                            "Didn't Smoke at 22 Weeks", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'hold_position'),
                font_size = 8) %>%
  row_spec(c(4, 7), color = "blue")

# Tables for all scores at 32wk
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_32wk) %>%
  tbl_summary(
    missing = "no",
    by = mom_smoke_32wk,
    type = list(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 32 Weeks",
```

```
      col.names=linebreak(c("Variables",
                            "Smoked at 32 Weeks",
                            "Didn't Smoke at 32 Weeks", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c('striped', 'hold_position'),
                font_size = 8) %>%
  row_spec(c(3, 4, 7), color = "blue")

# SDP on substance use
df$mom_pp6mo[df$mom_pp6mo==0] <- "2=No"
df$mom_pp6mo[df$mom_pp6mo==1] <- "1=Yes"

temp <- dplyr::select(df, c(cig_ever, e_cig_ever, mj_ever, alc_ever, mom_smoke_pp1,
                            mom_smoke_pp2, mom_smoke_pp12wk, mom_pp6mo)) %>%
  pivot_longer(
    cols = c(cig_ever, e_cig_ever, mj_ever, alc_ever),
    names_to = "Substance",
    values_to = "child_substance") %>%
  pivot_longer(
    cols = c(mom_smoke_pp1, mom_smoke_pp2, mom_smoke_pp12wk, mom_pp6mo),
    names_to = "time",
    values_to = "mom_smoke")

temp$time <- factor(temp$time, levels = c("mom_smoke_pp1", "mom_smoke_pp2",
                                          "mom_smoke_pp12wk", "mom_pp6mo"))
time.labs <- c("Smoking Status at First Postpartum Visit",
               "Smoking Status at First Postpartum Visit",
               "Smoking Status at 12 Weeks Postpartum",
               "Smoking Status at 6 Months Postpartum")
names(time.labs) <- c("mom_smoke_pp1", "mom_smoke_pp2", "mom_smoke_pp12wk", "mom_pp6mo")

p_alc <- ggplot(temp[temp$Substance=="alc_ever",],
                aes(factor(child_substance), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single"),
           show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap( vars(time), ncol = 4, labeller = labeller(time=time.labs) ) +
  labs(title = "Alcohol Consuming Status of Child Besides Religious Ceremony",
       fill = "Smoking Status of Mom",
       x = "Alcohol Consuming Status", y = "N" ) +
  theme_bw( base_size = 7 )

p_cig <- ggplot(temp[temp$Substance=="cig_ever",],
                aes(factor(child_substance), fill = factor(mom_smoke)) ) +
  geom_bar(position = position_dodge2(preserve = "single"),
           show.legend = FALSE ) +
  theme(legend.position = "none") +
  facet_wrap(vars(time), ncol = 4, labeller = labeller(time=time.labs)) +
  labs(title = "Cigarette Smoking Status of Child", fill = "Smoking Status of Mom",
       x = "Cigarette Smoking Status", y = "N") +
  theme_bw(base_size = 7)

p_e_cig <- ggplot(temp[temp$Substance=="e_cig_ever",],
                  aes(factor(child_substance), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single")) +
```

```
  facet_wrap(vars(time), ncol = 4, labeller = labeller(time=time.labs)) +
  labs(title = "E_Cigarette Smoking Status of Child", fill = "Smoking Status of Mom",
       x = "E_Cigarette Smoking Status", y = "N") +
  theme_bw( base_size = 7 )

p_mj <- ggplot(temp[temp$Substance=="mj_ever",],
               aes(factor(child_substance), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single")) +
  facet_wrap(vars(time), ncol = 4, labeller = labeller(time=time.labs)) +
  labs(title = "Marijuana Smoking Status of Child", fill = "Smoking Status of Mom",
       x = "Marijuana Smoking Status", y = "N") +
  theme_bw(base_size = 7)

ggarrange(p_alc, p_cig, ncol = 2)

ggarrange(p_e_cig, p_mj, ncol = 2, common.legend = TRUE, legend = "bottom")

# SDP on substance use
df$mom_pp6mo[df$mom_pp6mo==0] <- "2=No"
df$mom_pp6mo[df$mom_pp6mo==1] <- "1=Yes"

temp <- dplyr::select(df, c(num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30,
                            smoke_exposure_12mo, smoke_exposure_2yr, smoke_exposure_3yr,
                            smoke_exposure_4yr, smoke_exposure_5yr)) %>%
  pivot_longer(
    cols = c(num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30),
    names_to = "Substance",
    values_to = "dosage") %>%
  pivot_longer(
    cols = c(smoke_exposure_12mo, smoke_exposure_2yr, smoke_exposure_3yr,
             smoke_exposure_4yr, smoke_exposure_5yr),
    names_to = "time",
    values_to = "mom_smoke")

temp$time <- factor(temp$time, levels = c("smoke_exposure_12mo", "smoke_exposure_2yr",
                                          "smoke_exposure_3yr", "smoke_exposure_4yr",
                                          "smoke_exposure_5yr"))
time.labs <- c("Exposure from 7 to 12 months", "Exposure in the 2nd year",
               "Exposure in the 3rd year", "Exposure in the 4th year",
               "Exposure in the 5th year")
names(time.labs) <- c("smoke_exposure_12mo", "smoke_exposure_2yr", "smoke_exposure_3yr",
                      "smoke_exposure_4yr", "smoke_exposure_5yr")

p_alc_d <- ggplot(temp[temp$Substance=="num_alc_30",],
                  aes(factor(dosage), fill = factor(mom_smoke)) ) +
  geom_bar(position = position_dodge2(preserve = "single"),
           show.legend = FALSE) +
  theme(legend.position = "none") +
  facet_wrap( vars(time), ncol = 5, labeller = labeller(time=time.labs)) +
  labs(title = "Number of Days Drinking Alcohol of the Past 30 Days", fill = "Smoke Exposure",
       x = "Number of Days Drinking Alcohol of the Past 30 Days", y = "N" ) +
  theme_bw(base_size = 7)

p_cig_d <- ggplot(temp[temp$Substance=="num_cigs_30",],
                  aes(factor(dosage), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single"),
```

```
          show.legend = FALSE) +
  theme(legend.position = "none") +
  facet_wrap( vars(time), ncol = 5, labeller = labeller(time=time.labs)) +
  labs(title = "Number of Days Smoking Cigarette of the Past 30 Days",
       fill = "Smoke Exposure",
       x = "Number of Days Smoking Cigarette of the Past 30 Days", y = "N") +
  theme_bw(base_size = 7)

p_e_cig_d <- ggplot(temp[temp$Substance=="num_e_cigs_30",],
                    aes(factor(dosage), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single")) +
  facet_wrap(vars(time), ncol = 5, labeller = labeller(time=time.labs)) +
  labs(title = "Number of Days Smoking E_Cigarette of the Past 30 Days",
       fill = "Smoke Exposure",
       x = "Number of Days Smoking E_Cigarette of the Past 30 Days",
       y = "N") +
  theme_bw( base_size = 7 )

p_mj_d <- ggplot(temp[temp$Substance=="num_mj_30",],
                 aes(factor(dosage), fill = factor(mom_smoke))) +
  geom_bar(position = position_dodge2(preserve = "single")) +
  facet_wrap(vars(time), ncol = 5, labeller = labeller(time=time.labs))  +
  labs(title = "Number of Days Smoking Marijuana of the Past 30 Days",
       fill = "Smoke Exposure",
       x = "Number of Days Smoking Marijuana of the Past 30 Days",
       y = "N") +
  theme_bw(base_size = 7)

ggarrange(p_alc_d, p_cig_d, p_e_cig_d, p_mj_d, ncol = 1, common.legend = TRUE, legend = "bottom")

# Tables for all scores at 1st postpartum visit
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_pp1) %>%
  tbl_summary(
    missing = "no",
    by = mom_smoke_pp1,
    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 Postpartum Visit",
      col.names = linebreak(c("Variables", "Smoked at 1st Postpartum Visit",
                              "Didn't Smoke at 1st Postpartum Visit", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options =c('striped','hold_position'),
                font_size = 8) %>%
  row_spec(7, color = "blue")

# Tables for all scores at 2nd postpartum visit
```

```r
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_pp2) %>%
  tbl_summary(
    missing = "no",
    by = mom_smoke_pp2,
    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 2nd Postpartum Visit",
      col.names = linebreak(c("Variables", "Smoked at 2nd Postpartum Visit",
                              "Didn't Smoke at 2nd Postpartum Visit", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options =c('striped','hold_position'),
                font_size = 8) %>%
  row_spec(7, color = "blue")

# Tables for all scores at 12 weeks postpartum
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, mom_smoke_pp12wk) %>%
  tbl_summary(
    missing = "no",
    by = mom_smoke_pp12wk,
    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 Postpartum",
      col.names = linebreak(c("Variables", "Smoked at 12 Weeks Postpartum",
                              "Didn't Smoke at 12 Weeks Postpartum", "p-value")),
      booktabs=T, escape=T, align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options =c('striped','hold_position'),
                font_size = 8) %>%
  row_spec(c(3, 4, 7), color = "blue")

# Tables for all scores at 6 months postpartum
df %>%
  dplyr::select(erq_cog, erq_exp, bpm_att, bpm_att_p, bpm_ext, bpm_ext_p,
                cotimean_34wk, smoke_pp6mo) %>%
  tbl_summary(
    missing = "no",
    by = smoke_pp6mo,
```

```
  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 6 Months Postpartum",
    col.names = linebreak(c("Variables", "Smoke Exposure at 6 Months Postpartum",
                            "No Smoke Exposure at 6 Months Postpartum", "p-value")),
    booktabs=T, escape=T, align = "c") %>%
kable_styling(full_width = FALSE,
              latex_options =c('striped','hold_position'),
              font_size = 8) %>%
row_spec(c(4, 7), color = "blue")
```