Exploring Adolescent Outcomes: A Comprehensive Explanatory Data Analysis of Smoking During Pregnancy, Environmental Tobacco Smoke Exposure, and Self-Regulation

2023-10-08

Abstract

This study employs explanatory data analysis to scrutinize the interplay between smoking during pregnancy (SDP), environmental tobacco smoke (ETS) exposure, and various outcomes related to self-regulation, externalizing behavior, and substance use among adolescents. Explanatory data analysis serves as the methodological foundation, unraveling intricate relationships, patterns, and contributing factors to elucidate the obtained results. By employing a comprehensive approach, this study aims to achieve three primary objectives:

Aim 1: Investigate the impact of SDP/ETS on adolescent self-regulation, substance use, and externalizing behavior.

Aim 2: Explore the connections between baseline self-regulation and subsequent substance use and externalizing behavior at 6- and 12-month follow-ups.

Aim 3: Identify self-regulation challenges acting as mediators in the relationship between SDP/ETS exposure and the severity, as well as the trajectory, of substance use and externalizing behavior over time.

The study cohort comprises women recruited from a previous intervention study focused on smoke avoidance among low-income women (N=738) during pregnancy and immediate postpartum periods, aimed at reducing both smoking and ETS exposure. From this cohort, a subset of adolescents (N=100) and their mothers are randomly selected for inclusion in this investigation. Baseline data are available, and two longitudinal follow-up assessments occur at 6- and 12-month intervals post-baseline. This research provides a comprehensive exploration of the multifaceted relationships between prenatal tobacco exposure, adolescent outcomes, and self-regulation, contributing valuable insights to inform public health strategies and interventions. The prevalence of (SDP) and exposure to (ETS) is widespread and represents significant environmental factors impacting children. Such exposure during the early stages of development has been linked to an increased likelihood of externalizing behaviors in children, including conditions such as Attention-Deficit/Hyperactivity Disorder (ADHD), along with elevated rates of substance abuse issues.

Introduction

Smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) stand out as prevalent and harmful environmental exposures for children. Early exposure to smoke is associated with heightened occurrences of externalizing behaviors, including Attention-Deficit/Hyperactivity Disorder (ADHD), and substance use problems, carrying significant public health implications. Moreover, early smoke exposure is linked to self-regulation issues, encompassing challenges in emotional regulation.

The baseline data is filtered and partially pre-processed, resulting in 78 variables from 49 participants. This includes participant ID and demographic characteristics of the child and parent. The dataset encompasses self-reported smoking during pregnancy and postpartum, urine nicotine metabolite levels, emotional regulation results for parents and children, attention and externalizing problems, and current substance use. Smoking status is based on self-reports, with fewer lab results, representing Smoking During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) within a five year period after the child's birth.. The variables assessed

Table 1: Table 1: Patients Missing

	Patient ID	% missing
5	50502	74.35897
12	51202	71.79487
16	51602	76.92308
22	52302	71.79487
38	53902	73.07692
43	54402	76.92308
45	54602	76.92308
46	54702	73.07692

in the study pertain to self-regulation, externalizing problems, and substance use for both parents and children.

Pre-Processing

In the initial data reformatting process, a systematic approach was undertaken to address potential biases and skewed results. The baseline characteristics were analyzed by factoring race, sex, education, income, and time. Identification of a row with predominantly NA values led to its removal. Subsequently, the time variables for smoking status at all points were factored, with a focus on detecting and addressing outliers. An anomaly in the "mom_numcig" variable, specifically a reported value of 44,989 cigarettes in a single day, was deemed erroneous and replaced with an NA value to prevent skewing of results. String-to-integer conversions were applied to "2 black and miles a day" and "none." NA values were introduced to columns in the Smoking During Pregnancy (SDP) time variables, and NA values were set to 0 if the child reported never having consumed a substance. Additionally, "swan_inattentive" and "swan_hyperactive" values of 0 were replaced with NA values, aligning with the original dataset's representation.

Missing Data Processing

In the subsequent phase, an examination of missing data patterns was conducted across various variables, including baseline characteristics, Smoking During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) time variables, survey score responses, and overall patient data. Remarkably high percentages of missingness were identified, notably in variables such as "mom_smoke_pp1" (79.59%), "childasd" (57.14%), and "mom_smoke_pp2" (40.81%). Smoking time variables for SDP and ETS exhibited missing percentages ranging from 22.45% to 14.29%. Addressing the cumulative nature of cigarette smoking and nicotine addiction, assumptions were made to fill data gaps. Consistency in smoking behavior was assumed across specific time points, providing a basis for imputations.

Notably, postpartum visits 1 & 2 were excluded from the analysis due to substantial missing data and uncertainty about their precise timing in the timeline. Imputation or sensitivity analysis options were considered but deemed less appropriate than a complete case analysis.

The parental knowledge questions answered by the child exhibited high percentages of missingness (around 28.5% to 26.5%), potentially attributed to the perspective of not being a parent/adult. It is crucial to emphasize that assumptions were avoided for Environmental Tobacco Smoke (ETS) exposure variables, given their more fluctuating nature compared to parental smoking behavior.

The patient-wise analysis revealed nine cases with missing data exceeding 70% whcih can be seen in Table 1. While a larger dataset might warrant the removal of such cases, the decision was made to retain eight out of nine patients due to the dataset's limited size. The exclusion of one patient (ID 53902) lacking baseline measurements, particularly in SDP (mom_smoke_16wk), aimed to mitigate potential issues like bias, reduced statistical power, and inaccurate estimates. The approach chosen prioritizes a complete case analysis, ensuring robustness in the subsequent analyses.

Table 2: Table 1:Summary of SDP

SDP characteristics	#
mom_smoke_16wk	NA
1	36 (75%)
2	12 (25%)
mom_smoke_22wk	NA
1	29 (69%)
2	13 (31%)
Missing	6
mom_smoke_32wk	NA
1	30 (75%)
2	10 (25%)
Missing	8

Table 3: Table 2:Summary of Post-Partum

Smoking PP characteristics	#
mom_smoke_pp12wk	NA
1	30 (71%)
2	12 (29%)
Missing	6
mom_smoke_pp6mo	NA
1	24 (60%)
2	16 (40%)
Missing	8

Table 4: Table 3:Summary of ETS

Smoking Follow-Up characteristics	#
smoke_exposure_12mo	NA
0	30 (77%)
1	9 (23%)
Missing	9
smoke_exposure_2yr	NA
0	28 (72%)
1	11 (28%)
Missing	9
smoke_exposure_3yr	NA
0	27 (71%)
1	11 (29%)
Missing	10
smoke_exposure_4yr	NA
0	28 (74%)
1	10 (26%)
Missing	10
smoke_exposure_5yr	NA
0	29 (74%)
1	10 (26%)
Missing	9

Table 5: Table 4:Summary of Cotimean

Cotimean characteristics	#
cotimean_34wk	1 (0, 37)
Missing	10
cotimean_pp6mo_baby	1.5 (0.6, 4.0)
Missing	10
cotimean_pp6mo	15 (1, 119)
Missing	10

Table 6: Table 4:Summary of Self-Regulation

Self-Regulation characteristics	#
erq_cog_a	5.67 (4.67, 6.50)
Missing	9
erq_exp_a	3.25 (2.50, 4.25)
Missing	9
erq_cog	3.00 (2.83, 3.83)
Missing	12
erq_exp	2.50 (2.25, 3.31)
Missing	12

The absence of any complete entries in this dataset underscores a pervasive issue of data incompleteness, where each record is afflicted by missing values. This deficiency is particularly notable in the case of 8 records, each exhibiting the absence of at least 50 variables. The extensive nature of these missing data points poses a considerable challenge in conducting a thorough analysis and obtaining accurate estimates. Ultimately, the identified issue is not merely a technical inconvenience but a fundamental concern for data quality. Recognizing this as part of a broader data quality problem is imperative, prompting a reassessment of data collection methodologies, potential biases, and strategies for addressing missing data.

#Univariate Relationships We look at Tables 4,5 and 6 which gives us a peek into the summary statistics for the self-regulation and externalizing factors. I also have included a table (Table 7) of descriptive statistics reflecting the demographic of the parents offer insights into the central tendency and variability of SDP and ETS, aiding in summarizing and describing the characteristics of the variables.

Table 7: Table 6:Summary of Externalizing

Externalizing characteristics	#
bpm_att_p	NA
0	9 (25%)
1	11 (31%)
2	7 (19%)
3	1 (2.8%)
4	1 (2.8%)
5	3 (8.3%)
6	2 (5.6%)
7	1 (2.8%)
8	1 (2.8%)
Missing	12
bpm_ext_a	NA
0	17 (45%)
1	9 (24%)
2	5 (13%)
3	3 (7.9%)
4	2 (5.3%)
5	1 (2.6%)
6	1 (2.6%)
Missing	10
bpm_int_a	NA
0	16 (41%)
1	8 (21%)
2	5 (13%)
3	3 (7.7%)
4	5 (13%)
5	1 (2.6%)
8	1 (2.6%)
Missing	9
bpm_ext_p	NA
0	18 (49%)
1	5 (14%)
2	5 (14%)
3	3 (8.1%)
4	2 (5.4%)
5	1 (2.7%)
7	2 (5.4%)
11	1 (2.7%)
Missing	11
bpm_int_p	1 (0, 4)
Missing	9
swan_inattentive	11 (8, 15)
Missing	9
swan_hyperactive	6 (4, 14)
Missing	13

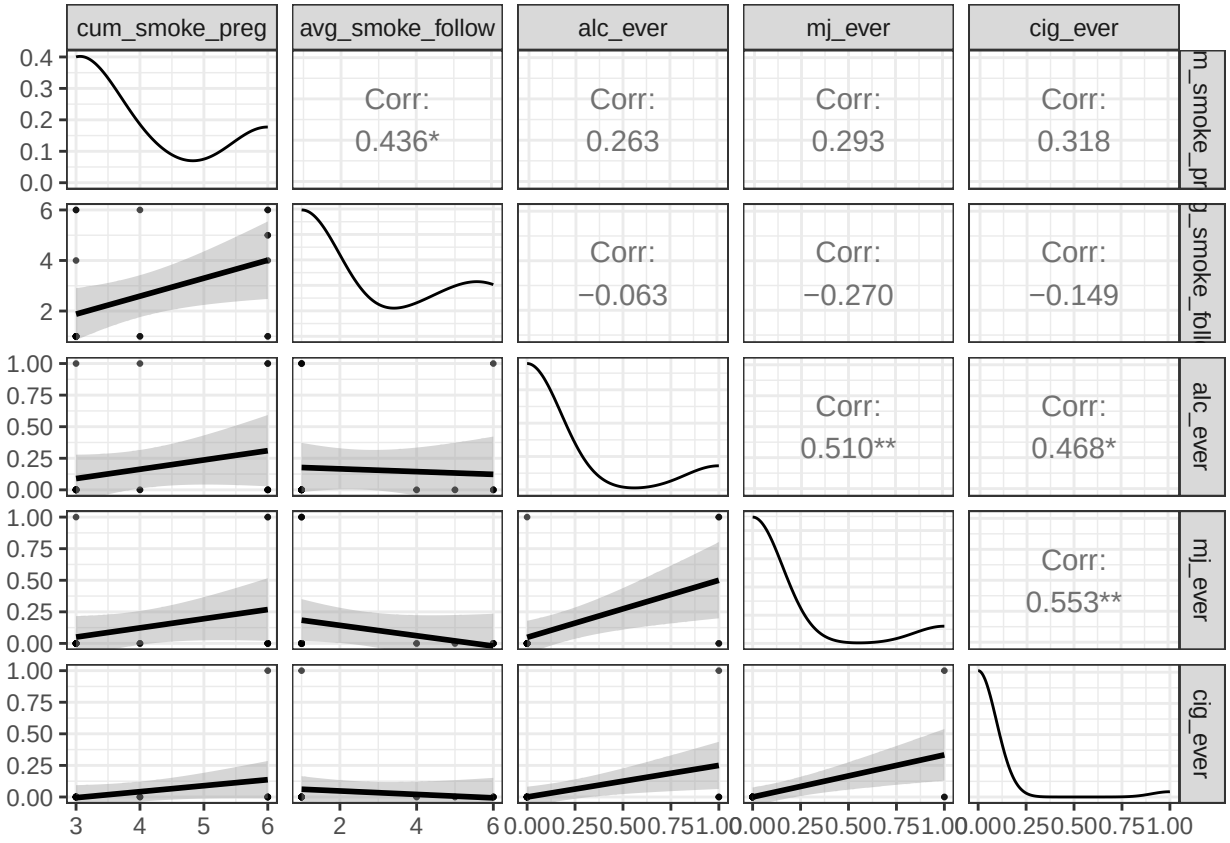
	Overall
n	48
page (mean (SD))	37.62 (3.57)
psex = 1 (%)	39 (97.5)
plang = 1 (%)	15 (37.5)
pethnic = 1 (%)	13 (32.5)
paian = 1 (%)	4 (8.3)
pasian = 0 (%)	48 (100.0)
pnhpi = 1 (%)	8 (16.7)
pblack = 0 (%)	48 (100.0)
pwhite = 1 (%)	25 (52.1)
prace_other = 1 (%)	6 (12.5)
employ (%)	
0	11 (27.5)
1	7 (17.5)
2	22 (55.0)
pedu (%)	
0	2 (5.0)
1	3 (7.5)
2	5 (12.5)
3	15 (37.5)
4	3 (7.5)
5	10 (25.0)
6	2 (5.0)
income (mean (SD))	59174.51 (51825.15)
tage (mean (SD))	13.62 (1.21)
tsex = 1 (%)	13 (36.1)
language = 1 (%)	11 (29.7)
tethnic (%)	
0	21 (56.8)
1	15 (40.5)
2	1 (2.7)
taian = 1 (%)	5 (10.4)
tasian = 0 (%)	48 (100.0)
tnhpi = 0 (%)	48 (100.0)
tblack = 1 (%)	15 (31.2)
twhite = 1 (%)	19 (39.6)
trace_other = 1 (%)	5 (10.4)

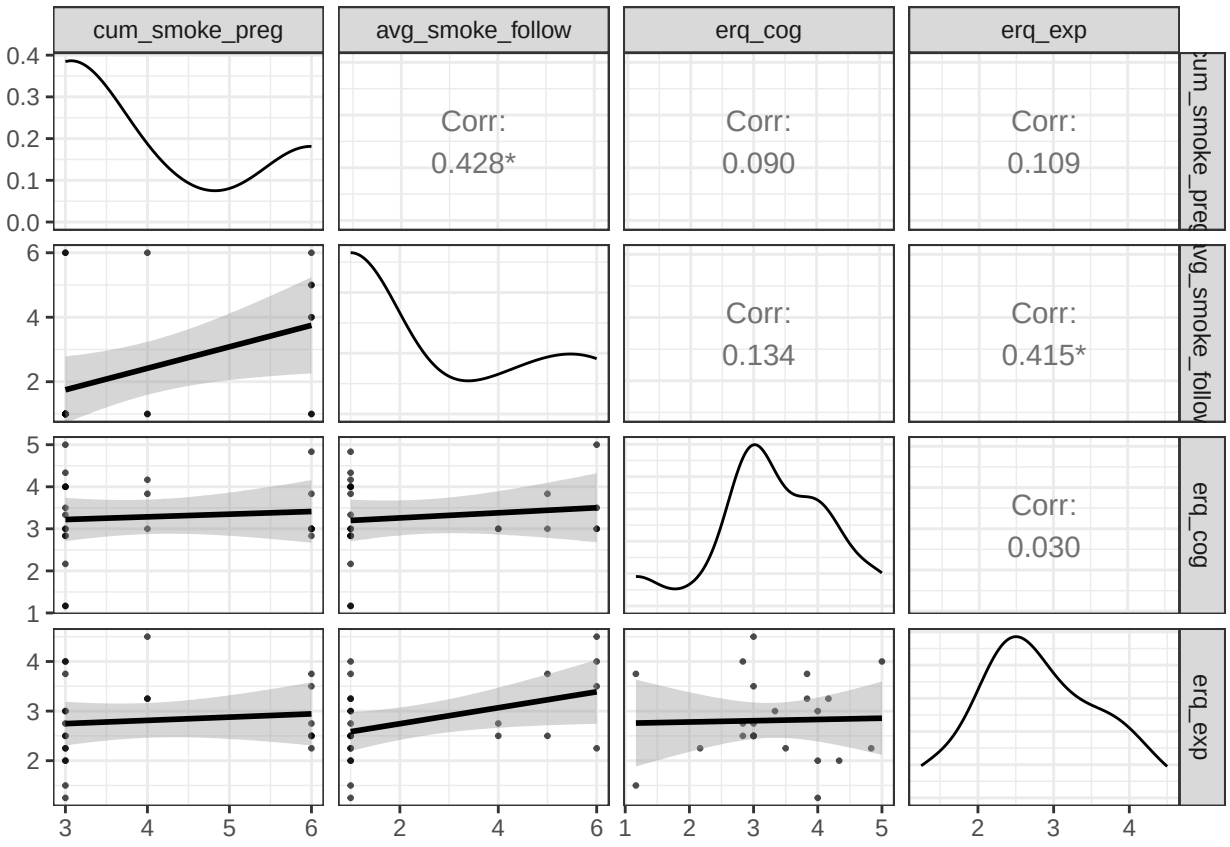
After the initial data transformation, the generation of composite scores aimed to streamline the representation of smoking variables. The "cum_smoke_preg" variable was introduced to amalgamate smoking occurrences during pregnancy, summing instances at 16, 22, and 32 weeks. Additionally, the "avg_smoke_follow" variable calculated the average of smoking exposures during follow-ups at 6 months, 12 months, 2 years, 3 years, 4 years, and 5 years. To enhance the smoke exposure variable, the code leveraged information from other relevant variables. The "smoke_exposure_immediate" variable, for instance, was revised based on information from "smoke_exposure_6mo," "mom_smoke_pp12wk," "mom_smoke_pp6mo," "cotimean_pp6mo," and

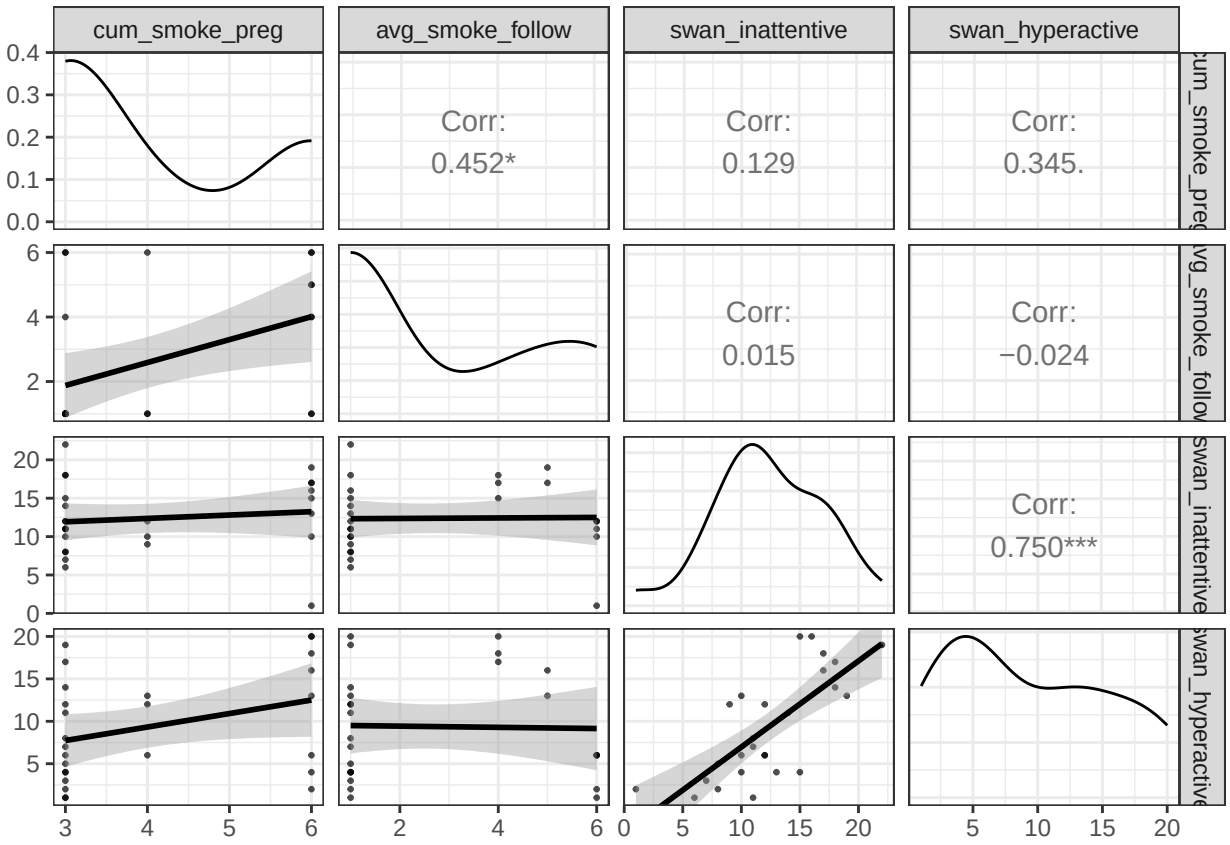
"cotimean_pp6mo_baby," resulting in a more refined representation of smoking exposure. These adjusted and newly created variables contribute to a more nuanced and informative analysis of smoking-related patterns across different time points.

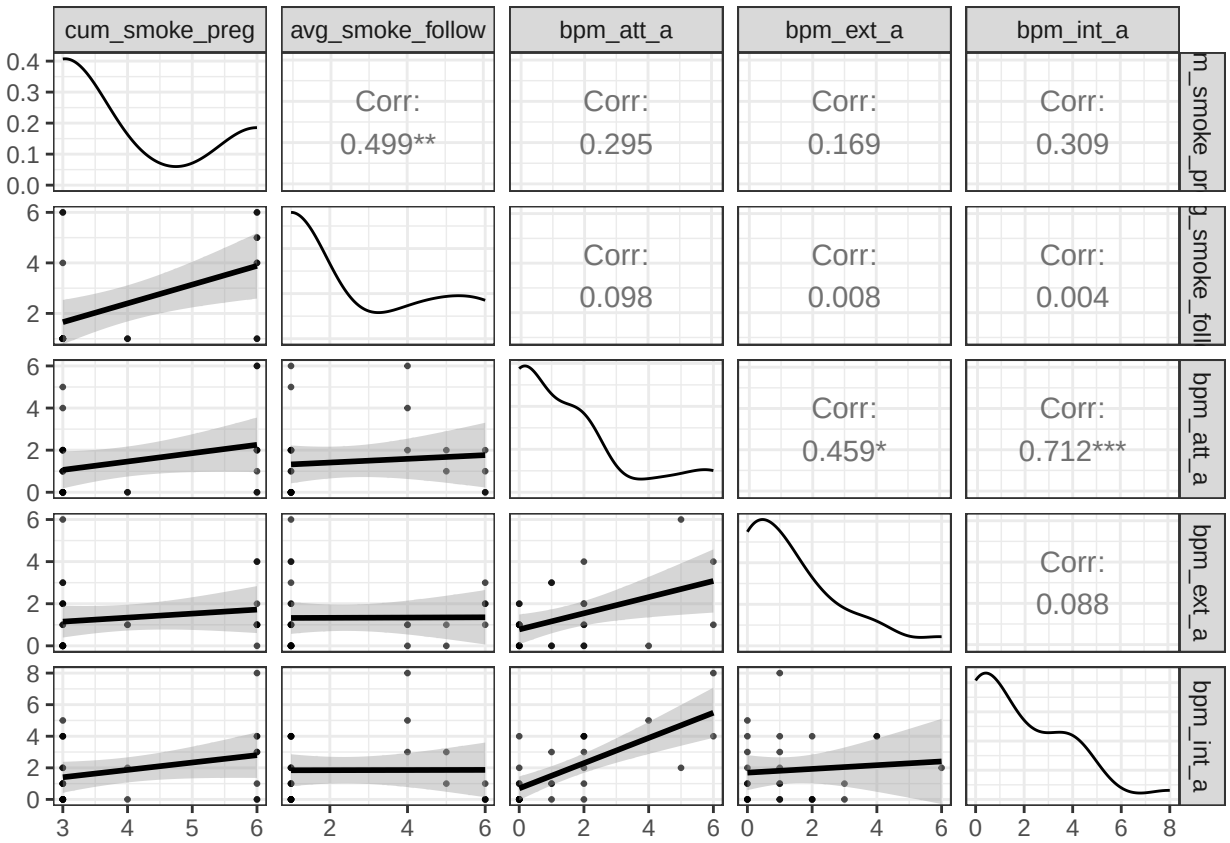
The correlation results indicate a moderately positive relationship between "cum_smoke_preg" (cumulative smoking during pregnancy) and "avg_smoke_follow" (average smoking exposure during follow-ups) with a correlation coefficient of approximately 0.48. This suggests that individuals who reported higher cumulative smoking during pregnancy also tended to exhibit higher average smoking exposure during subsequent follow-up periods. However, the correlation is not exceptionally strong, indicating that other factors may contribute to smoking behaviors during pregnancy and follow-up periods. It must also be noted that the follow up smoke doesn't strictly just focus on the smoking status of the mother but everyone in the household as well.

Bivariate Relationships






cum_smoke_preg
avg_smoke_follow
swan_inattentive
swan_hyperactive
Corr:
0.452*
Corr:
0.129
Corr:
0.345.
Corr:
0.015
Corr:
−0.024
Corr:
0.750***



The GGally package provides a powerful tool for analyzing bivariate relationships through its scatterplot matrix and correlation coefficients. When examining the intricate connections between Smoke During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) exposure with subsequent outcomes, such as substance use, externalizing factors, and self-regulation, several noteworthy patterns emerge. Notably, ETS demonstrates a negative relationship with substance use in children, indicating that those exposed to environmental tobacco smoke after birth are less likely to engage in substance use as teenagers. Conversely, SDP exhibits a positive relationship with a child's substance use, suggesting a potential association between prenatal smoke exposure and increased likelihood of substance use in adolescence.

In examining the relationship between Smoke During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) exposure and their impact on self-regulation, specifically in the cognitive and expressive domains, interesting patterns emerge. Surprisingly, the cognitive aspect of self-regulation does not display a substantial correlation with either SDP or ETS, suggesting that these maternal smoking factors might not significantly influence this particular facet of self-regulation.

On the other hand, the expressive dimension of self-regulation demonstrates a moderately strong correlation with ETS, registering at 0.415. This finding implies that children exposed to environmental tobacco smoke after birth may experience a slightly heightened suppression in expressing their emotions. However, it's essential to note that the correlation, while statistically significant, is not exceptionally robust. This suggests that while ETS may contribute to a child's reluctance to express emotions, other factors could play a role in shaping the complexity of these emotional dynamics within the household environment. Exploring these nuances is crucial for a more comprehensive understanding of the various influences on children's self-regulation processes.

In the examination of Environmental Tobacco Smoke (ETS) exposure, the correlation analysis reveals minimal associations with the SWAN (Strengths and Weaknesses of ADHD Symptoms and Normal Behavior) variables of inattentiveness and hyperactivity. The correlation estimates for ETS with SWAN inattentive and hyperactive variables are 0.015 and -0.024, respectively, indicating virtually no substantial correlation.

Similarly, when scrutinizing Smoke During Pregnancy (SDP), there is only limited correlation observed with the SWAN variables. The correlation estimates for SDP with SWAN inattentive and hyperactive variables are 0.129 and 0.345, respectively, signifying modest correlations. These findings suggest that neither ETS nor SDP is significantly linked to the inattentive and hyperactive behaviors assessed by the SWAN scale in the study population.

Lastly the examination of Environmental Tobacco Smoke (ETS) exposure, the correlation analysis reveals minimal associations with the Brief Problem Monitor variables. The correlation estimates for ETS with BPM variables are 0.098,0.008, and 0.004 respectively, indicating virtually no substantial correlation. Similarly, when scrutinizing Smoke During Pregnancy (SDP), there is only limited correlation observed with the BPM variables. The correlation estimates for SDP with BPM variables are 0.169,0.295, and 0.309, respectively, signifying some modest correlations. These findings suggest that neither ETS nor SDP is significantly linked to the BPM behaviors in the study population.

Next, a GLM model was also chosen due to the presence of outliers and non-normality assumptions in the data, as GLMs are suitable for capturing nonlinear relationships between predictors and responses. GLM models were applied to the Smoke During Pregnancy (SDP) variables, revealing that for mj_ever, alc_ever, and ecig_ever, the baseline risk was statistically significant ($p = 0.00124$, $p = 0.000472$, $p = 0.000595$), indicating significant differences from zero when considering no other predictors. However, the cumulative score of SDP does not appear to be significant in predicting substance use in children. The intercept for cig_ever was not statistically significant.

For the subsequent analysis, it is deemed more appropriate to explore effects through summary tables rather than GLMs. Utilizing summary statistics aids in identifying potential issues, such as skewness, outliers, or heteroscedasticity, and allows for a thorough examination of assumptions before applying complex models. This approach was adopted in the evaluation of the Environmental Tobacco Smoke (ETS) follow-up on substance use in adolescents.

Characteristic	**1**, N = 25	**4**, N = 3	**5**, N = 2	**6**, N = 6	**p-value**
alc_ever	0.14 (0.35)	0.00 (0.00)	0.00 (0.00)	0.17 (0.41)	0.88
Missing	3	1	0	0	
cig_ever	0.05 (0.21)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.92
Missing	3	0	0	0	
e_cig_ever	0.09 (0.29)	0.00 (0.00)	0.00 (0.00)	0.17 (0.41)	0.83
Missing	3	0	0	0	
mj_ever	0.14 (0.35)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.66
Missing	3	0	0	0	
num_cigs_30	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
Missing	3	0	0	0	
num_e_cigs_30	0.09 (0.43)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.93
Missing	3	0	0	1	
num_mj_30	1.50 (4.51)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.66
Missing	3	0	0	0	
num_alc_30	0.50 (2.13)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.82
Missing	3	1	0	0	

This summary table displays the effects of the average follow up ETS on the substance use variables. We see that none of the variables alc-ever,cig_ever, e_cig_ever, mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30, for the children showed any statistical significance with the environmental tobacco smoke in their early years of childhood. We can see that there is a lot of missingness once again and that most people have an average score of 0 for the follow-up. There was a lot of variation in the the data.

Table 8: Compare Externalizing Swan factor score with SDP

Characteristic	**3**, N = 24	**4**, N = 3	**6**, N = 10	**p-value**
swan_hyperactive	6 (4, 11)	12 (9, 13)	15 (6, 19)	0.2
Unknown	8	0	2	
swan_inattentive	11 (8, 15)	10 (10, 11)	16 (12, 17)	0.3
Unknown	5	0	2	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

Table 9: Compare Externalizing ERQ factors score with SDP

Characteristic	**3**, N = 24	**4**, N = 3	**6**, N = 10	**p-value**
erq_exp	2.50 (2.00, 3.00)	3.25 (3.25, 3.88)	2.63 (2.50, 3.56)	0.10
Unknown	7	0	2	
erq_cog	3.00 (2.33, 3.88)	3.83 (3.42, 4.00)	3.00 (3.00, 3.42)	0.5
Unknown	6	0	3	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

Characteristic	**3**, N = 24	**4**, N = 3	**6**, N = 10	**p-value**
bpm_att	2.83 (2.53)	2.33 (4.04)	4.88 (2.70)	0.17
Missing	6	0	2	
bpm_att_p	1.71 (2.08)	1.00 (1.00)	3.75 (2.82)	0.12
Missing	7	0	2	
bpm_ext	2.72 (1.81)	3.00 (3.61)	3.63 (2.13)	0.59
Missing	6	0	2	
bpm_ext_p	2.00 (2.92)	0.33 (0.58)	2.63 (2.62)	0.32
Missing	7	0	2	
bpm_int	2.44 (2.28)	4.00 (6.93)	2.75 (2.55)	0.84
Missing	8	0	2	
bpm_int_p	2.37 (2.75)	1.33 (1.15)	3.50 (2.88)	0.36
Missing	5	0	2	

First, an examination of statistics from the "swan_inattentive" and "swan_hyperactive" variables revealed median scores and their confidence intervals. For instance, when the cumulative Smoke During Pregnancy (SDP) is 1 for "swan_hyperactive," the median is 12 with a narrow confidence interval (CI) ranging from 9 to 13. Contrarily, a cumulative SDP of 0 for the hyperactive variable has a wider range (4 to 15) with a median of 5, and cumulative score 3 exhibits an even broader interval (6 to 19) with a median of 15. The non-significant p-value of 0.14 suggests no statistically significant difference in the "swan_hyperactive" variable between the three SDP groups. Similarly, for "swan_inattentive," the p-value of 0.2 and the different medians and confidence intervals are detailed in the summary table above.

Subsequently, an analysis of Emotion Regulation Questionnaire (ERQ) responses between cumulative SDP and "erq_exp" revealed a small p-value (0.081) that, nevertheless, proved to be non-significant. When cumulative SDP is 1 for "erq_exp," the median is 3.25 with a narrow CI ranging from the median value 3.25 to 3.88. Conversely, for the "erq_exp" variable, cumulative SDP of 0 displays a wider range (2 to 3) with a median of 2.5. The "erq_cog" variable had a higher p-value of 0.5, indicating non-significance.

In the context of Brief Problem Monitor (BPM) responses, none of the variables showed significance with cumulative SDP. However, noteworthy small p-values were observed for "bpm_att" and "bpm_att_p." It is important to highlight that "bpm_int" had the highest missingness among all variables, and the following table breaks down each confidence interval for SDP cumulative scores.

Table 10: Compare Externalizing Swan factor score with ETS

Characteristic	**1**, N = 25	**4**, N = 3	**5**, N = 2	**6**, N = 6	**p-value**
swan_hyperactive	7 (4, 12)	18 (18, 19)	15 (14, 15)	4 (1, 6)	0.011
Unknown	4	0	0	0	
swan_inattentive	9 (7, 13)	17 (16, 18)	18 (18, 19)	11 (6, 12)	0.038

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

Table 11: Compare Externalizing ERQ factors score with ETS

Characteristic	**1**, N = 25	**4**, N = 3	**5**, N = 2	**6**, N = 6	**p-value**
erq_exp	2.50 (2.00, 3.00)	2.63 (2.56, 2.69)	3.13 (2.81, 3.44)	3.75 (3.56, 3.94)	0.054
Unknown	3	1	0	0	
erq_cog	3.33 (2.83, 4.00)	3.00 (3.00, 3.00)	3.42 (3.21, 3.63)	3.50 (3.00, 3.83)	0.7
Unknown	3	0	0	1	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

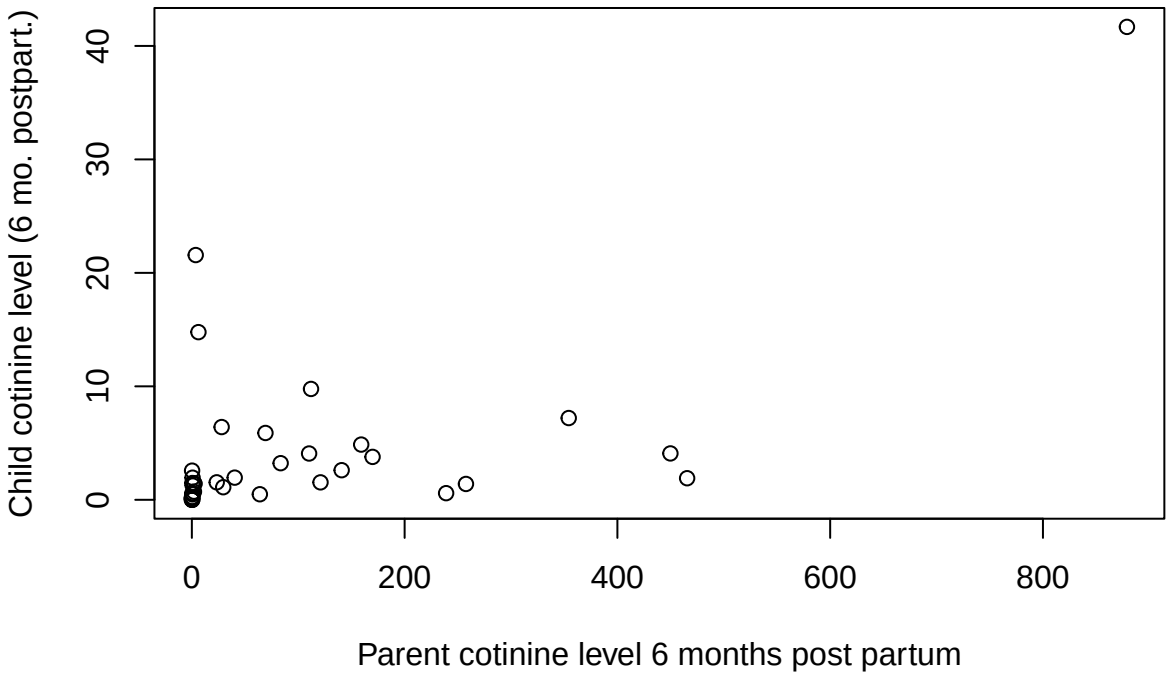
Characteristic	**1**, N = 25	**4**, N = 3	**5**, N = 2	**6**, N = 6	**p-value**
bpm_att	2.64 (2.75)	5.67 (2.31)	3.50 (2.12)	3.00 (2.61)	0.27
Missing	3	0	0	0	
bpm_att_p	1.52 (2.00)	6.50 (2.12)	4.00 (2.83)	1.33 (0.52)	0.054
Missing	2	1	0	0	
bpm_ext	2.64 (2.15)	4.67 (1.15)	2.50 (2.12)	3.50 (1.97)	0.26
Missing	3	0	0	0	
bpm_ext_p	1.04 (1.72)	6.33 (4.16)	5.50 (2.12)	0.67 (1.21)	0.007
Missing	2	0	0	0	
bpm_int	2.24 (2.30)	5.50 (3.54)	1.00 (1.41)	4.67 (3.78)	0.091
Missing	4	1	0	0	
bpm_int_p	1.56 (2.18)	5.67 (1.53)	5.00 (5.66)	2.17 (1.47)	0.039

The analysis of the Externalizing Swan factor scores in relation to ETS characteristics reveals interesting patterns. For the "swan_hyperactive" variable, the median scores (with 95% confidence intervals) for different ETS characteristics are as follows: 7 (4, 12) for characteristic 1 (N = 25), 18 (18, 19) for characteristic 4 (N = 3), 15 (14, 15) for characteristic 5 (N = 2), and 4 (1, 6) for characteristic 6 (N = 6). The p-value of 0.011 suggests a statistically significant difference in the "swan_hyperactive" variable among the different PP characteristics. Notably, characteristic 4 exhibits the highest median score, indicating potential concerns in that subgroup. The "swan_inattentive" variable also displays significance, with median scores and confidence intervals of 9 (7, 13), 17 (16, 18), 18 (18, 19), and 11 (6, 12) for characteristics 1, 4, 5, and 6, respectively. The p-value of 0.038 highlights a statistically significant difference in "swan_inattentive" scores across the ETS characteristics. It's crucial to acknowledge the presence of unknown values (Unknown) in characteristic 1, which might impact the interpretation of results.

The examination of Externalizing Emotion Regulation Questionnaire (ERQ) factor scores in relation to Environmental Tobacco Smoke (ETS) characteristics reveals noteworthy findings. For the "erq_exp" variable, the median scores (with 95% confidence intervals) for different ETS characteristics are as follows: 2.50 (2.00, 3.00) for characteristic 1 (N = 25), 2.63 (2.56, 2.69) for characteristic 4 (N = 3), 3.13 (2.81, 3.44) for characteristic 5 (N = 2), and 3.75 (3.56, 3.94) for characteristic 6 (N = 6). Although the p-value is 0.054, suggesting a borderline significance, it's crucial to note the presence of unknown values in characteristics 1 and 4, which may influence the results. For the "erq_cog" variable, the median scores and confidence

intervals are 3.33 (2.83, 4.00), 3.00 (3.00, 3.00), 3.42 (3.21, 3.63), and 3.50 (3.00, 3.83) for characteristics 1, 4, 5, and 6, respectively. The p-value of 0.7 indicates no statistically significant difference in "erq_cog" scores across the ETS characteristics. The presence of unknown values in characteristic 1 and 6 should be considered when interpreting these results.

For the Brief Problem Monitor responses only two variables showed statistical significance which was bpm_ext_p with a p-value of 0.007 and bpm_int_p with a p-value of 0.039. With the Cumulative ETS I do want to highlight other the small but non-significant variables which are bpm_att and bpm_att_p.



To look at the distribution of the contimean score between the mothers smoking status in her last trimester. We see that when answered Yes for their smoking status there is more of range and a higher average which would be predicted since continine does measure and test screens for nictoine which is present in the yes. This shows that the mothers answered truthfully about their smoking status in the last trimester. The same can be said for no with its small to none range and the couple of outliers in which can be accounted for fi they smoked recently prior to the testing in this last trimester of pregnancy.

The child continine levels at 6 months post-partum compared the mother's shows a high concentration at when the mother's measurement is 0 the child's is also 0. Even when the parents have high measurements of continine it is good to see that the tobacco screening is not reflected in child's urine meaning that it has not been passed onto the child.

Conclusion

The findings reveal a modest positive correlation between SDP and ETS, although the strength of this correlation is not particularly robust. This disparity may be attributed to the fact that prenatal smoke exposure is solely derived from the mother's smoking status, while postnatal smoke exposure integrates information from other individuals in the household. Furthermore, the prenatal data originates from the original study, whereas the postnatal data predominantly emanates from the current times. Notably, exposure

to smoke during the early stages of development has been linked to a heightened prevalence of externalizing behaviors in children, including Attention-Deficit/Hyperactivity Disorder (ADHD), along with elevated rates of substance abuse problems. Through bivariate analysis, distinct stages of early smoke exposure exhibit associations with self-regulation, externalizing problems, and substance use, respectively.

Limitations

With a limited sample size of 49 observations at baseline, the dataset exhibits a considerable number of missing values, particularly when scrutinizing specific variables that demand complete data. This small sample size poses a risk of biased and statistically insignificant results, given the potential impact of specific data records on analytical outcomes. Moreover, the dataset includes both self-reported and reflective data, contributing to data quality challenges. Additionally, our emphasis on analyzing univariate and bivariate relationships underscores the importance of acknowledging the limitations associated with unadjusted results, which can introduce bias and hinder comparability among diverse groups with varying levels of smoking exposure.

Code Appendix

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

# load the packages
library(tidyverse)
library(dplyr)
library(kableExtra)
library(gtsummary)
library(ggplot2)
library(naniar)
library(psych)
library(mice)
library(webshot2)
library(webshot)
library(GGally)
library(tinytex)
library(tableone)
library(gridExtra)

# load the data
new_df<- read.csv("C:/Users/VGSCHOOL/project1.csv")

# factorize the categorical data
new_df[,c(3:13)]<- new_df[,c(3:13)] %>%
mutate_if(is.numeric, factor)

new_df[,c(53:61)]<- new_df[,c(53:61)] %>%
mutate_if(is.numeric, factor)

new_df[,c(37:42)]<- new_df[,c(37:42)] %>%
mutate_if(is.numeric, factor)

new_df <- new_df %>%
mutate(mom_smoke_16wk = factor(case_when(mom_smoke_16wk=="1=Yes"~1,
mom_smoke_16wk=="2=No"~0)),
mom_smoke_22wk = factor(case_when(mom_smoke_22wk=="1=Yes"~1,
mom_smoke_22wk=="2=No"~0)),
mom_smoke_32wk = factor(case_when(mom_smoke_32wk=="1=Yes"~1,
mom_smoke_32wk=="2=No"~0)),
mom_smoke_pp1= factor(case_when(mom_smoke_pp1=="1=Yes"~1,
mom_smoke_pp1=="2=No"~0)),
mom_smoke_pp2= factor(case_when(mom_smoke_pp2=="1=Yes"~1,
mom_smoke_pp2=="2=No"~0)),
mom_smoke_pp12wk =factor(case_when(mom_smoke_pp12wk=="1=Yes"~1,
mom_smoke_pp12wk=="2=No"~0)),
mom_smoke_pp6mo = factor(case_when(mom_smoke_pp6mo=="1=Yes"~1,
mom_smoke_pp6mo=="2=No"~0)))


#Create NA values for missingness in smoking during pregnancy and post partum variables
new_df[,c(22:28)]<-lapply(new_df[,c (22:28)],function(x) ifelse(x=="",NA,x))
```

```r
new_df[,c(22:28)]<- new_df[,c(22:28)] %>%
mutate_if(is.integer, factor)

#Remove Outliers in Dataset in which we have viewed and change to standardize them with rest of data

new_df$mom_numcig <- as.character(new_df$mom_numcig)

new_df$mom_numcig <- ifelse(new_df$mom_numcig == "None", "0", new_df$mom_numcig )
new_df$mom_numcig <- ifelse(new_df$mom_numcig == "2 black and miles a day", "2",new_df$mom_numcig)

#average 20-25
new_df$mom_numcig <- ifelse(new_df$mom_numcig == "20-25","22.5",new_df$mom_numcig)
#change 44989 to NA
new_df$mom_numcig <- ifelse(new_df$mom_numcig == "44989",NA,new_df$mom_numcig)

#Remove comma to standardize to rest of values
new_df$income<- as.character(new_df$income)
new_df$income<- ifelse(new_df$income == "250,000", "250000", new_df$income )

#Change 0 values to NA for Swan scores in data
new_df$swan_inattentive<- ifelse(new_df$swan_inattentive == 0,NA,new_df$swan_inattentive)
new_df$swan_hyperactive<- ifelse(new_df$swan_hyperactive == 0,NA,new_df$swan_hyperactive)

#Change number of cigs  and income into numeric values
new_df$mom_numcig<-as.numeric(new_df$mom_numcig)
new_df$income<-as.numeric(new_df$income)



#If child answered no to consuming substances in the ever variables than put 0 as the value instead of
new_df$num_cigs_30 <- ifelse(new_df$cig_ever == 0, 0, new_df$num_cigs_30)
new_df$num_alc_30 <- ifelse(new_df$alc_ever == 0, 0, new_df$num_alc_30)
new_df$num_mj_30 <- ifelse(new_df$mj_ever == 0, 0, new_df$num_mj_30)
new_df$num_e_cigs_30 <- ifelse(new_df$e_cig_ever == 0, 0, new_df$num_e_cigs_30)


overall_missing<-miss_var_summary(new_df)


 #Look at missing data for each parent
pct_na_r <- rowSums(is.na(new_df)) / ncol(new_df) * 100
row_na <- data.frame(parent_id = new_df$parent_id, pct_na = pct_na_r)
row_na <- row_na[row_na$pct_na > 70,]
knitr::kable(row_na, caption = "Table 1: Patients Missing", col.names = c("Patient ID", "% missing"))



# table <- kable(row_na, format = "html", caption = "Parent ID with Percentage of Missing Values") %>%
#   kable_styling(bootstrap_options = "striped", full_width = FALSE)
# table
#Remove row with baseline values missing
row_to_remove <- 38
new_df <- new_df[-row_to_remove, ]
```

```r
#Create imputation for smoking status for post-partum and during pregnancy
new_df$mom_smoke_32wk<- ifelse(new_df$mom_smoke_16wk == "1=Yes" &
                                 new_df$mom_smoke_22wk == "1=Yes",
                  "1=Yes", as.character(new_df$mom_smoke_32wk) )

new_df$mom_smoke_32wk<-factor(new_df$mom_smoke_32wk)

new_df$mom_smoke_pp1<- ifelse(new_df$mom_smoke_22wk == "1=Yes" &
                                new_df$mom_smoke_32wk == "1=Yes" &
                                new_df$mom_smoke_pp2 == "1=Yes",
                  "1=Yes", as.character(new_df$mom_smoke_pp1))
new_df$mom_smoke_pp1<-factor(new_df$mom_smoke_pp1)

table1_SDP<-new_df %>% select(mom_smoke_16wk,mom_smoke_22wk, mom_smoke_32wk) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 1:Summary of SDP",

col.names = c("SDP characteristics","#"))

table1_SDP

table2_PP<-new_df %>% select(mom_smoke_pp12wk,mom_smoke_pp6mo) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 2:Summary of Post-Partum",

col.names = c("Smoking PP characteristics","#"))
table2_PP

table_ETS<-new_df %>% select(smoke_exposure_12mo,smoke_exposure_12mo,smoke_exposure_2yr,smoke_exposure_
                             smoke_exposure_4yr,smoke_exposure_5yr) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 3:Summary of ETS",

col.names = c("Smoking Follow-Up characteristics","#"))
table_ETS
table4_contimean<-new_df %>% select(cotimean_34wk,cotimean_pp6mo_baby,cotimean_pp6mo) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 4:Summary of Cotimean",

col.names = c("Cotimean characteristics","#"))
table4_contimean

table5_erq<-new_df %>% select(erq_cog_a,erq_exp_a,erq_cog,erq_exp) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 4:Summary of Self-Regulation",

col.names = c("Self-Regulation characteristics","#"))
table5_erq

table6_ext<-new_df %>% select(bpm_att_p,bpm_ext_a,bpm_int_a,bpm_att_p,bpm_ext_p,bpm_int_p,swan_inattenti
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 6:Summary of Externalizing",
```

```
col.names = c("Externalizing characteristics","#"))
table6_ext


kableone(CreateTableOne(data=new_df,vars=names(new_df)[c(2:14,52:61)])) %>%
kable_styling(full_width=T, font_size=12,latex_options = c('HOLD_position'))
new_df$SDP <- factor(ifelse(new_df$mom_smoke_16wk==1 |
new_df$mom_smoke_22wk==1 |
new_df$mom_smoke_32wk==1,1,0))
new_df <- new_df %>%
mutate(cum_smoke_preg=as.numeric(as.character(mom_smoke_16wk))+
as.numeric(as.character(mom_smoke_22wk))+
as.numeric(as.character(mom_smoke_32wk)))
# revise the smoke exposure variable by leverage the other variable's information
new_df$smoke_exposure_immediate<- factor(ifelse(new_df$smoke_exposure_6mo==1|
new_df$mom_smoke_pp12wk==1|new_df$mom_smoke_pp6mo==1|
new_df$cotimean_pp6mo >10 | new_df$cotimean_pp6mo_baby>10,1,0))
# smoke exposure intensity
new_df <- new_df %>%
mutate(avg_smoke_follow=as.numeric(as.character(new_df$smoke_exposure_immediate))+
as.numeric(as.character(new_df$smoke_exposure_12mo))+
as.numeric(as.character(new_df$smoke_exposure_2yr))+
as.numeric(as.character(new_df$smoke_exposure_3yr))+
as.numeric(as.character(new_df$smoke_exposure_4yr))+as.numeric(as.character(new_df$smoke_exposure_5yr))

# Assuming your data frame is named 'new_df'

library(corrplot)

# Select relevant variables
smoking_variables <- new_df[, c("cum_smoke_preg", "avg_smoke_follow")]

# Check for missing values and remove them if necessary
smoking_variables <- na.omit(smoking_variables)

# Calculate correlation
correlation <- cor(smoking_variables)

# Print the correlation matrix
#print(correlation)
plt1<-ggpairs(new_df %>% select(cum_smoke_preg,avg_smoke_follow,alc_ever,mj_ever,cig_ever)%>%na.omit(),
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
plt1

plt2<-ggpairs(new_df %>% select(cum_smoke_preg,avg_smoke_follow,erq_cog,erq_exp)%>%na.omit(),
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
plt2

plt3<-ggpairs(new_df %>% select(cum_smoke_preg,avg_smoke_follow,swan_inattentive,swan_hyperactive)%>%na
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
```

```
plt3

plt4<-ggpairs(new_df %>% select(cum_smoke_preg,avg_smoke_follow,bpm_att_a,bpm_ext_a,bpm_int_a)%>%na.omi
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
plt4
#fit new model using cumulative count with alcohol use
model_cum_smoke_alc<- glm(alc_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
#summary(model_cum_smoke_alc)

#create plot showing the probability predictions
new_cum_data <- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data$Predicted_Probabilities<- predict(model_cum_smoke_alc, newdata = new_cum_data, type = "resp


# Create a plot for baseline and alcohol use predictions
alc_plot<-ggplot(data = new_cum_data, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child Alcohol Use",
    title = "Graph 1: Logistic Regression Predicted Probabilities"
  )

model_follow_smoke_alc<- glm(alc_ever ~ avg_smoke_follow, data = new_df, family = binomial(link = "logi
#summary(model_follow_smoke_alc)

#create plot showing the probability predictions
new_cum_data_fol<- data.frame(avg_smoke_follow = new_df$avg_smoke_follow)

# Predict probabilities using the model
new_cum_data_fol$Predicted_Probabilities<- predict(model_follow_smoke_alc, newdata = new_cum_data_fol,


# Create a plot for baseline and alcohol use predictions
alc_plot1<-ggplot(data = new_cum_data_fol, aes(x = avg_smoke_follow, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child Alcohol Use",
    title = "Graph 1: Logistic Regression Predicted Probabilities"
  )
#fit new model using cumulative count with marijuana use
model_cum_smoke_mj<- glm(mj_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
#summary(model_cum_smoke_mj)

#create plot showing the probability predictions
new_cum_data_mj<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_mj$Predicted_Probabilities<- predict(model_cum_smoke_mj, newdata = new_cum_data_mj, type =
```

```r
# Create a plot for baseline and alcohol use predictions
mj_plot<-ggplot(data = new_cum_data_mj, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child Marijuana Use",
    title = "Graph 2: Logistic Regression Predicted Probabilities"
  )
#fit new model using cumulative count with ecig use
model_cum_smoke_ecig<- glm(e_cig_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
#summary(model_cum_smoke_ecig)

#create plot showing the probability predictions
new_cum_data_ecig<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_ecig$Predicted_Probabilities<- predict(model_cum_smoke_ecig, newdata = new_cum_data_ecig, t


# Create a plot for baseline and alcohol use predictions
ecig_plot<-ggplot(data = new_cum_data_ecig, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child E-cigarette Use",
    title = "Graph 3:Logistic Regression Predicted Probabilities"
  )
#fit new model using cumulative count with cigarette use

model_cum_smoke_cig<- glm(cig_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
#summary(model_cum_smoke_cig)

#create plot showing the probability predictions
new_cum_data_cig<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_cig$Predicted_Probabilities<- predict(model_cum_smoke_cig, newdata = new_cum_data_cig, type


# Create a plot for baseline and alcohol use predictions
cig_plot<-ggplot(data = new_cum_data_cig, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child cigarette Use",
    title = "Graph 4: Logistic Regression Predicted Probabilities"
)
new_df %>% select(avg_smoke_follow, alc_ever, cig_ever, e_cig_ever,

mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_30) %>%

tbl_summary( by = avg_smoke_follow, missing_text = "Missing",
```

```
type = list (c(alc_ever, cig_ever, e_cig_ever,mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_30
c(avg_smoke_follow) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(alc_ever, cig_ever, e_cig_ever,mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_3
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))

#stratified by cum SDP
tbl1_swan<- new_df[,c(32,33,80)] %>%
  tbl_summary(by = cum_smoke_preg) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing Swan factor score with SDP")
as_kable_extra(tbl1_swan)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))


tbl1_erq<- new_df[,c(73,74,80)] %>%
  tbl_summary(by = cum_smoke_preg) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing ERQ factors score with SDP")
as_kable_extra(tbl1_erq)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))


new_df %>% select(cum_smoke_preg, bpm_att, bpm_att_p, bpm_ext,

bpm_ext_p, bpm_int, bpm_int_p) %>%

tbl_summary( by = cum_smoke_preg, missing_text = "Missing",
type = list (c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ "continuous",
c(cum_smoke_preg) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ c(2, 2))) %>%
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
#stratified by cum PP
tbl2_swan<- new_df[,c(32,33,82)] %>%
  tbl_summary(by = avg_smoke_follow) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing Swan factor score with ETS")
as_kable_extra(tbl2_swan)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))


tbl2_erq<- new_df[,c(73,74,82)] %>%
  tbl_summary(by = avg_smoke_follow) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing ERQ factors score with ETS")
as_kable_extra(tbl2_erq)%>%
```

```
  kable_styling(full_width=T, latex_options = c('HOLD_position'))


new_df %>% select(avg_smoke_follow, bpm_att, bpm_att_p, bpm_ext,

bpm_ext_p, bpm_int, bpm_int_p) %>%

tbl_summary( by = avg_smoke_follow, missing_text = "Missing",
type = list (c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ "continuous",
c(avg_smoke_follow) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ c(2, 2))) %>%
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
contimean_post<-plot(new_df$cotimean_pp6mo, new_df$cotimean_pp6mo_baby,
xlab = "Parent cotinine level 6 months post partum",
ylab = "Child cotinine level (6 mo. postpart.)")
```