Prenatal Tobacco Exposure & Child Outcomes: Exploratorary Data Analysis

Background

Maternal smoking during pregnancy (SDP) is a significant public health concern, affecting 7-15% of U.S. infants annually and costing the economy billions due to healthcare expenses. SDP exposes children to externalizing behaviors, attention-deficit/hyperactivity disorder, conduct issues, substance use, and self-regulation problems.

Self-regulation, crucial for child development, encompasses executive function, emotion regulation, effortful control, and vagal tone. SDP's link to self-regulation deficits makes it an important area of study. This research focuses on approximately 800 pregnant individuals exposed to SDP or environmental tobacco smoke (ETS) and their offspring, aged 12-16 years. It aims to:

- AIM 1: Investigate SDP's impact on adolescent self-regulation, substance use, and externalizing behaviors. Hypothesis: Greater early smoke exposure leads to poorer self-regulation and increased SU and EXT.
- AIM 2: Assess timing and dosage effects of SDP on adolescent outcomes. Hypotheses: Cumulative SDP exposure and exposure during vulnerable trimesters affect self-regulation, SU, and EXT.
- AIM 3: Identify self-regulation deficits mediating SDP's link to SU and EXT. Hypothesis: Hot executive function and emotion regulation mediate this connection.

The goal of this paper is to conduct a comprehensive exploratory data analysis that serves as the cornerstone for unraveling the multifaceted relationships inherent in the dataset.

Data pre-processing

In the initial stages of preparing the dataset for analysis, a series of rigorous data preprocessing steps were meticulously carried out. These steps were aimed at ensuring the data's accuracy, consistency, and suitability for subsequent analytical procedures. This section elucidates the pivotal procedures employed during the data preprocessing phase.

First and foremost, an examination was conducted to validate the correctness of data import and to identify any potential formatting inconsistencies. This process involved scrutinizing the data for any anomalies and ensuring that data values adhered to a consistent format. Specific attention was paid to variables where values required standardization to enhance uniformity.

Furthermore, in an effort to align variable names with both the dataset and the accompanying codebook, certain variable names underwent renaming. This not only facilitated clarity but also ensured that the variables were consistent across all pertinent documents.

Addressing missing values was another crucial aspect of data preprocessing. Systematic procedures were implemented to handle missing data within relevant variables. Variables that encompassed binary responses ("1=Yes" and "2=No") underwent a transformation into numeric format, with "1=Yes" being assigned the value 1 and "2=No" denoting 2; empty entries were coded as NA to indicate missing values.

For a specific variable, "mom_numcig," which originally displayed diverse numeric representations, a comprehensive standardization process was applied. This involved harmonizing entries for consistency. For instance, "2 black and miles a day" was harmonized to 2, "44989" was transformed into "NA," "20-25" was adjusted to 22 to maintain consistency, "None" was converted to 0, and blank entries were standardized as "NA."

Furthermore, to ensure coherence and conformity, variable names that exhibited inconsistencies between the dataset and the codebook were meticulously harmonized. For instance, "paiain" was renamed as "paian," "taiain" was aligned with "taian," and "nidaalc" was adjusted to "nidalc."

Additionally, numeric transformations were applied to variables like "income," converting entries such as "250,000" to 250000. Blank entries in this variable were also replaced with NA to signify missing income information.

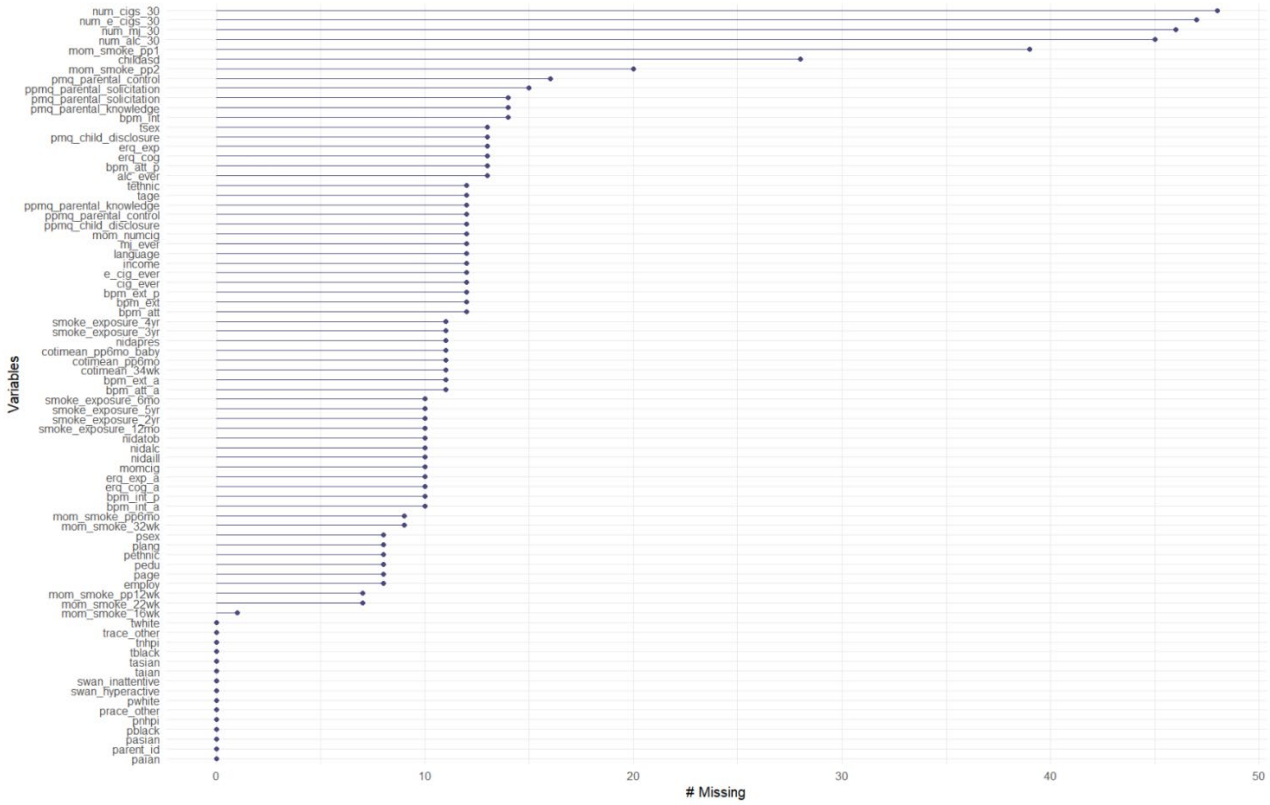

Figure 1. Summary of missing data

Turning to the issue of missing data, it was observed that approximately 23% of the dataset contained missing values. Specifically, four variables, namely "um_cigs_30," "num_e_cigs_30," "num_mj_30," and "num_alc_30," exhibited more than 90% missing values. These variables were associated with follow-up questions that were not applicable if the response to a preceding question was "No." Additionally, variables "mom_spoke_pp1" and "mom_spoke_pp2" displayed 80% and 40% missing values, respectively, as they pertained to self-reported current smoker status during the 1st and 2nd postpartum visits. Notably, no participant had observations in both of these variables; it was either one or the other, and some participants had no observations in either. Another variable, "childasd," which related to Autism in the child, contained 57% missing values. However, all other variables exhibited 32% or fewer missing values.

Remarkably, eight observations were identified wherein 49% or more of the values were missing. Significantly, these eight observations also featured missing values in the answers that served as dependent variables. Due to the insufficiency of information in these cases, a decision was made to exclude them from further analysis.

Furthermore, an examination of the socio-demographic characteristics of participants revealed that parental ages spanned from 32 to 45 years, while children's ages ranged from 12 to 16 years. Notably, the dataset primarily consisted of females, with a single exception. It's pertinent to mention that among the parents, there were no individuals of Black or Asian race. A majority of parents (62%) identified themselves as White, while only 46% of children self-identified as White. In contrast, approximately 37% of children identified themselves as Black.

Characteristic	**N = 49**[1]
Hispanic/Latino	13 (32%)
American Indian/Alaskan Native	4 (8.2%)
Asian	0 (0%)
Native Hawaiian or Pacific Islander	8 (16%)
Black	0 (0%)
White	26 (53%)
Other Race	6 (12%)
Parent Age	37 (35, 39)
Parent Sex	
Male	1 (2.4%)
Female	40 (98%)
Another Language Spoken by Parent at Home	15 (37%)

Parent Employed	
No	12 (29%)
Part-Time	7 (17%)
Full-Time	22 (54%)
Highest Level of Education of Parent	
Some High School	3 (7.3%)
High School	3 (7.3%)
GED	5 (12%)
Some College	15 (37%)
2 Year Degree	3 (7.3%)
4 Year Degree	10 (24%)
Postgraduate Degree	2 (4.9%)

Table 1. Parents Demographics

These meticulous data preprocessing procedures served to enhance data quality, consistency, and readiness for subsequent analytical endeavors. The resulting dataset, characterized by standardized values and judicious handling of missing data, now stands poised for the exploration of research objectives centered on maternal smoking during pregnancy and its potential impact on adolescent outcomes.

Methods

Composite variables

In this section, we'll dive into how we calculated some essential composite variables to better understand the impact of prenatal and postnatal exposure to smoking on adolescent behavior. These variables will help us explore self-regulation, substance use, and externalizing behavior among adolescents in our dataset.

Prenatal Exposure Severity Variable (SDP)

To gauge smoking during pregnancy (SDP), we used a few key variables:

- **mom_smoke_16wk**, **mom_smoke_22wk**, and **mom_smoke_32wk** to figure out if moms smoked during specific timeframes.
- **cotimean_34wk** to measure the severity of smoking during pregnancy.

We didn't use **momcig** and **mom_numcig** because we weren't sure which time periods they referred to. But before we crunched the numbers, we needed to deal with missing values in **mom_smoke_16wk**, **mom_smoke_22wk**, and **mom_smoke_32wk**. Our fix? Replacing those missing values with "No" since it seemed reasonable to assume that missing values meant the participant didn't smoke during those times.

To calculate SDP, we created two new variables:

1. **smoking_status_prenatal** based on **mom_smoke_16wk**, **mom_smoke_22wk**, and **mom_smoke_32wk**.

2. **smoking_severity_prenatal** based on **cotimean_34wk**.

With these two variables, we could calculate SDP, giving us a snapshot of smoking behavior during pregnancy in terms of both status (Consistent Smoker, Inconsistent Smoker, Non-Smoker) and severity (Absent, High, Low) of smoking.

Smoking Status and Severity Prenatal		
Smoking Status Prenatal	**Smoking Severity Prenatal**	**Count**
Consistent Smoker	Absent	0
Inconsistent Smoker	Absent	0
Non-Smoker	Absent	8
Consistent Smoker	High	7
Inconsistent Smoker	High	0
Non-Smoker	High	0
Consistent Smoker	Low	1
Inconsistent Smoker	Low	3
Non-Smoker	Low	22

Table 2. Prenatal Smoking Status and Severity

Calculating Postnatal Exposure Severity Variable (ETS)

For the Postnatal Exposure Severity variable (ETS), we used:

- **mom_smoke_pp1**, **mom_smoke_pp2**, **mom_smoke_pp12wk**, and **mom_smoke_pp6mo** to assess postnatal smoking.
- **cotimean_pp6mo** to evaluate smoking severity postnatally.

Before crunching the numbers, we had to tackle missing values in these variables. We noticed that **mom_smoke_pp1** and **mom_smoke_pp2** had quite a few missing values, but upon closer inspection, it turned out that no participants had values in both variables; it was always one or the other. So, we merged them into a single variable, named **merged_mom_smoke_pp**, and filled in the missing values with the most common response, "No."

We applied a similar approach to deal with missing values in **cotimean_pp6mo** and **cotimean_pp6mo_baby**, replacing missing entries with "0."

Next, we created two new variables:

1. **smoking_status_postnatal** based on **merged_mom_smoke_pp**, **mom_smoke_pp12wk**, and **mom_smoke_pp6mo**.

2. **smoking_severity_postnatal** based on **cotimean_pp6mo** and **cotimean_pp6mo_baby**.

Using these variables, we calculated ETS, resulting in a summary table showing the distribution of postnatal smoking status and the severity of smoking exposure.

Smoking Status and Severity Postnatal		
Smoking Status Postnatal	**Smoking Severity Postnatal**	**Count**
Consistent Smoker	Absent	1
Inconsistent Smoker	Absent	1
Non-smoker	Absent	7
Consistent Smoker	High	5
Inconsistent Smoker	High	3
Non-smoker	High	3
Consistent Smoker	Low	0
Inconsistent Smoker	Low	4
Non-smoker	Low	17

Table 3. Postnatal Smoking Status and Severity

Calculating Variables for Adolescent Self-Regulation, Substance Use, and Externalizing

Now, let's focus on our dataset's variables to examine adolescent self-regulation, substance use, and externalizing behavior. We've already calculated the independent variables (SDP and ETS) and three dependent variables:

1. **Externalizing (EXT):** This variable combines responses related to externalizing problems. We had four missing values in **bpm_ext**, which we handled by imputing the median due to data skewness.

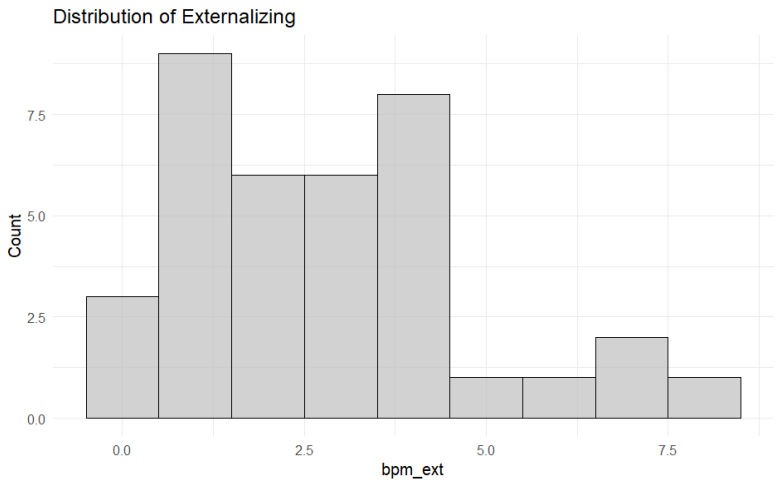

Figure 2. Distribution of Externalizing

2. **Self-Regulation (SR):** It's the average response on questions about Cognitive Reappraisal (**erq_cog**) and Expressive Suppression (**erq_exp**). We retained variables with some missing data, and for the six remaining missing values in **self_r**, we imputed the median.

Figure 3. Distribution of Self-Regulation, Cognitive Reappraisal, and Expressive Suppression

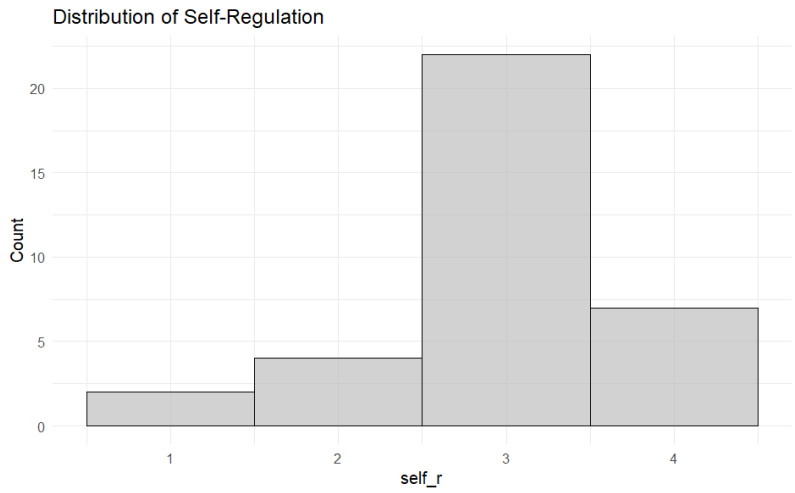

Figure 4. Distribution of Self-Regulation

3. **Substance Use (SU):** Calculated using variables about cigarette and e-cigarette use, marijuana, and alcohol. We handled missing data by considering them as negative responses for substance initiation questions and "0" for frequency-related variables.

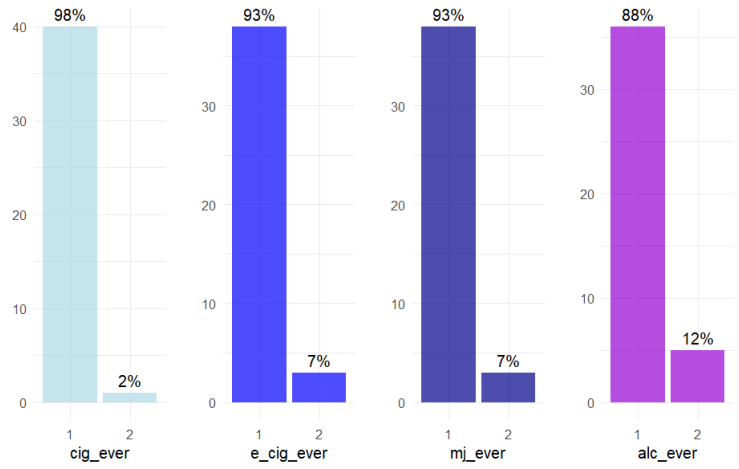

Figure 5. Distribution of Self-Regulation

These calculated variables, combined with SDP and ETS, form the basis for our analysis, helping us explore the impact of smoking exposure on adolescent development.

Analysis

In this section, we provide a descriptive overview of the independent and dependent variables, shedding light on the characteristics of the study population and the distribution of key variables.

Characteristic	**N = 41**
SDP	
Non-smoker & Absence of smoking	8 (19.5%)
Low smoking or non-smoker with some exposure	22 (53.7%)
Moderate smoking	3 (7.3%)
High smoking	1 (2.4%)
Very high smoking	7 (17.1%)
ETS	
Non-smoker & Absence of exposure	7 (17.1%)
Low exposure (inconsistent/low severity)	18 (43.9%)
Moderate exposure	8 (19.5%)
High exposure (consistent/low or inconsistent/high)	3 (7.3%)
Very high exposure (consistent & high)	5 (12.2%)
SR	2.88 [1.13, 4.50]
EXT	3.00 [0.00, 8.00]
SU	
Frequent user	1 (2.4%)
Non-user	34 (82.9%)
Occasional user	6 (14.6%)

Table 4. Descriptive table for independent and dependent variables

Smoking During Pregnancy (SDP): The majority of parents (53.7%) fall into the category of "Low smoking or non-smoker with some exposure," while 19.5% are "Non-smoker & Absence of smoking." Notably, approximately 20% of participants reported high levels of smoking during pregnancy.

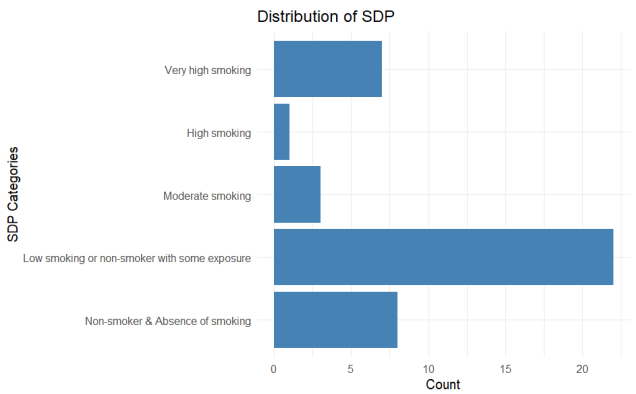

Figure 6. Distribution of SDP

Environmental Tobacco Smoke (ETS) Exposure: The largest proportion of participants (43.9%) experienced "Low exposure (inconsistent/low severity)" to ETS, while 17.1% had "Very high exposure (consistent & high)." Only 7.3% reported "High exposure (consistent/low or inconsistent/high)," and 17.1% were "Non-smoker & Absence of exposure."

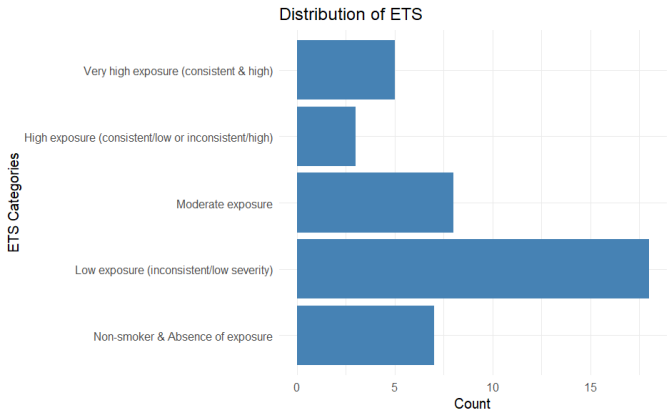

Figure 7. Distribution of ETS

Self-Regulation (SR): The self-regulation scores exhibit a distribution with a mean of 2.88 and a range from 1.13 to 4.50. The distribution has a slight left-skewness, indicating that a slightly larger number of participants have lower self-regulation scores.

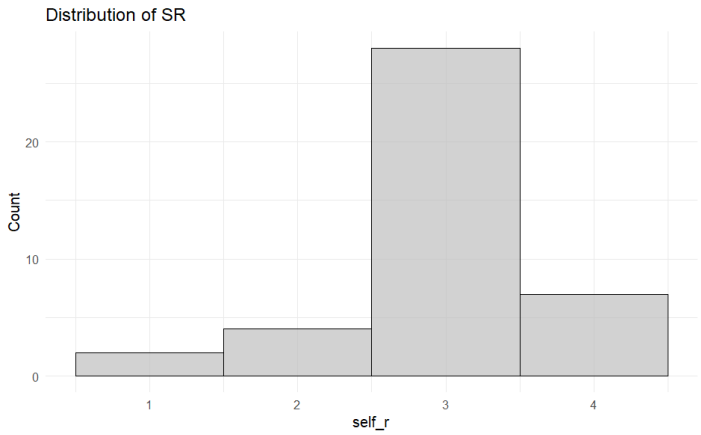

Figure 8. Distribution of SR

Externalizing Behavior (EXT): The distribution of externalizing behavior scores has a mean of 3.00, ranging from 0.00 to 8.00. Some missing values in this variable were imputed with the median due to data skewness.

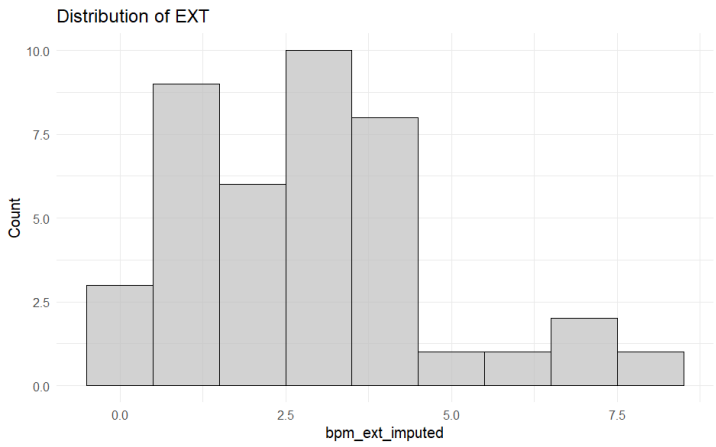

Figure 9. Distribution of EXT

Substance Use (SU): The majority of adolescents (82.9%) are categorized as "Non-users," while only a small fraction (2.4%) are "Frequent users." Approximately 14.6% fall into the category of "Occasional users." This suggests a relatively low prevalence of substance use among the studied population.

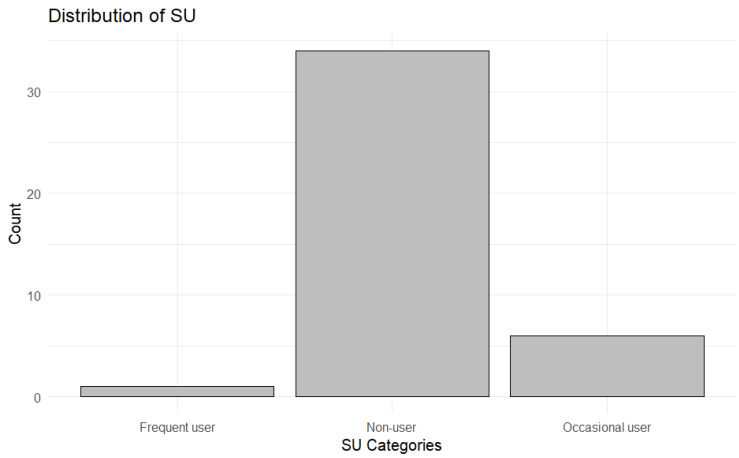

Figure 10. Distribution of SU

Several important insights can be drawn from the univariate analysis. Notably, a significant proportion of parents reported low or no smoking during pregnancy, which is promising from a public health perspective. However, it is concerning that even among those who claimed not to smoke (Non-smokers), some exhibited the presence of cotinine in their urine, indicating exposure to Environmental Tobacco Smoke (ETS). Additionally, the majority of adolescents had not experimented with any substances, suggesting a relatively low prevalence of substance use in this cohort. However, further analysis will be needed to explore the relationships between these variables in more detail.

To deepen our understanding of the data and investigate the relationships between variables, we conducted a correlation analysis. The correlation matrix reveals associations between key variables, shedding light on potential patterns and trends in the data.

Column1	SDP	ETS	SR	EXT	SU
SDP	1	0.6980129	0.1019456	0.2480115	0.145094
ETS	0.698013	1	0.1182923	0.2322279	0.3875211
SR	0.101946	0.1182923	1	0.2317548	0.1325622
EXT	0.248012	0.2322279	0.2317548	1	0.3160639
SU	0.145094	0.3875211	0.1325622	0.3160639	1

Table 5. Correlation matrix

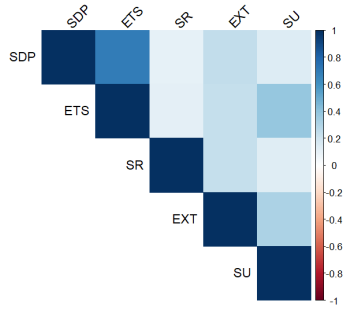

Figure 11. Correlation matrix

SDP and ETS: The correlation coefficient between SDP and ETS is approximately 0.698, indicating a strong positive association. This suggests that higher levels of smoking during pregnancy (SDP) are associated with higher levels of environmental tobacco smoke exposure (ETS) and vice versa.

SDP with Dependent Variables:

SDP vs. SR: A coefficient of 0.102 indicates a weak positive association, suggesting that higher levels of SDP are slightly associated with higher levels of self-regulation, although the relationship is weak.

SDP vs. EXT: A coefficient of 0.248 indicates a weak to moderate positive association, implying that higher levels of SDP are somewhat associated with higher levels of externalizing behaviors.

SDP vs. SU: A coefficient of 0.145 indicates a weak positive association, suggesting that higher levels of SDP are slightly associated with a higher likelihood of substance use.

ETS with Dependent Variables:

ETS vs. SR: A coefficient of 0.118 indicates a weak positive association.

ETS vs. EXT: A coefficient of 0.232 indicates a weak to moderate positive association.

ETS vs. SU: A coefficient of 0.388 indicates a moderate positive association. This suggests that higher levels of ETS are more strongly associated with a higher likelihood of substance use compared to SDP.

Self-Regulation (SR) vs. Externalizing (EXT) vs. Substance Use (SU):

SR vs. EXT: A coefficient of 0.232 indicates a weak to moderate positive association.

SR vs. SU: A coefficient of 0.133 indicates a weak positive association.

EXT vs. SU: A coefficient of 0.316 indicates a weak to moderate positive association.

These correlation findings provide valuable insights into the relationships between maternal smoking exposure, self-regulation, externalizing behaviors, and substance use among adolescents. Further analyses will explore these associations in greater depth to inform our research objectives.

Discussion

Our comprehensive exploratory data analysis has illuminated the intricate relationships among maternal smoking during pregnancy (SDP), environmental tobacco smoke exposure (ETS), and adolescent outcomes, including self-regulation, externalizing behaviors (EXT), and substance use (SU).

When examining adolescent self-regulation (SR), we observed a prevalence of lower SR scores. However, the link between SDP and SR appeared relatively weak, prompting the need for more extensive investigations to elucidate this connection further.

In the realm of adolescent externalizing behaviors (EXT), we observed considerable variability in EXT scores. While SDP exhibited a weak to moderate association with EXT, it remains essential to delve deeper into

potential mechanisms that may underlie the relationship between prenatal smoking exposure and externalizing outcomes.

Turning our attention to adolescent substance use (SU), our analysis revealed a noteworthy pattern—most adolescents were non-users of substances, suggesting a relatively low prevalence of SU. Interestingly, SDP showed a weak association with SU, whereas ETS demonstrated a stronger link. This finding emphasizes the importance of considering passive smoke exposure in households with children, given its potential impact on adolescent substance use.

From a public health perspective, our findings underscore the urgency of implementing interventions aimed at maternal smoking cessation and reducing ETS exposure during pregnancy to enhance child health outcomes. Moreover, it is imperative to explore the mediating factors and mechanisms that connect SDP to adolescent outcomes more comprehensively.

In terms of future research directions, longitudinal studies can provide valuable insights into the enduring effects of tobacco exposure on adolescent behavior. Additionally, investigating genetic factors and gene-environment interactions can offer a deeper understanding of individual variances in susceptibility to tobacco exposure. Qualitative research methods may capture the lived experiences and perspectives of pregnant individuals and adolescents concerning smoking and its consequences.

In conclusion, our exploratory data analysis has unveiled crucial insights into the intricate relationships between maternal smoking exposure and adolescent outcomes. However, further research is warranted to unravel the complexities of these associations and inform evidence-based interventions for the betterment of child and adolescent health.