Examining the Impact of Prenatal and Environmental Tobacco Exposure on Child Behavioral: An Exploratory Data Analysis

Yu Yan

2023-10-05

Abstract

Background: Exposure to smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) poses known risks to children behavioral development. This study examines the extent to which SDP/ETS influences adolescent self-regulation, substance use, and externalizing behavior.

Methods: Using data from a longitudinal study of 800 pregnant mothers and a follow-up subset of 100 mother-child pairs, we developed 'timing' and 'intensity' variables to quantify SDP/ETS exposure. The analysis explores correlations between these exposures and behavioral outcomes in adolescents in three domains: Self-Regulation, Externalizing behavior and Substance use problem.

Hypothesis: Earlier exposure and higher cumulative intensity to SDP/ETS is hypothesized to correlate with poorer self-regulation, earlier substance use initiation, and increased externalizing behavior in adolescents.

Conclusion: The findings aim to clarify the impact of early smoke exposure on adolescent behavior, providing valuable data for public health interventions and policies targeting smoking during pregnancy and childhood. With the exploratory analysis provided, there are associated relationships between Earlier exposure and higher cumulative intensity to SDP/ETS with bad behavioral outcomes in general. Statistical significance should be tested in the future.

Introduction

Our study, led by Dr. Lauren Micalizzi from the Department of Behavioral and Social Sciences, aims to understand how smoking during pregnancy and exposure to secondhand smoke affect children as they grow into teenagers. We are using data from a previous study of 800 pregnant women and a focused group of 100 of their children, tracking how early exposure to smoke relates to behaviors like self-control, substance use. The goal is to get a clear picture of how smoke exposure in early life might lead to behavioral challenges later on, which could help in creating better public health policies.

Missing Data and Outlier

We started doing missingness pattern checking of the data. We detected that there are 8 observations who have more than half of the columns as missing. We decided to remove those 8 observations at the beginning, so we end up have 41 observations. Then by looking at the 10 most missing columns from Table 1, we found out one interesting thing that almost all observations (41 observation in total), are lacking four particular columns, num_cigs_30 (missing 40), num_e_cigs_30(missing 39),num_mj_30(missing 38),num_alc_30(missing 37). By referring to the code book, these are the number of days in the last 30 days for teenagers to use cigratte, e_cigrattes, marijuana, and alcohol respectively. As the only four direct variables that we have relating to teenagers' substance use behavior, this implies that either the majority of teenagers didn't perform SU related activities, or the data is simply missing those variables.

Table 1: Missingness Count and Percentage of columns

Colnames	Missing_num	Missing_Pct
num_cigs_30	40	97.6
num_e_cigs_30	39	95.1
num_mj_30	38	92.7
num_alc_30	37	90.2
mom_smoke_pp1	33	80.5
childasd	21	51.2
mom_smoke_pp2	15	36.6
cotimean_pp6mo_baby	8	19.5
cotimean_pp6mo	8	19.5
pmq_parental_control	8	19.5

With this finding in mind, we calculated the number of teenagers that answered YES in those preceding columns, only a small amount of them answered YES. Considering the definition of those variables, only observation with a YES answer in those variables could potentially have a value to the number of four substances used. This becomes hard if we were to look into hypothesis of whether early exposure to SDP lead to earlier SU initiation or faster SU escalation since we do not really have a lot of samples to look for insights.

Table 2: Number of Teenagers answered YES Substance Use related questions

	Number
cig_ever	1
e_cig_ever	3
mj_ever	3
alc_ever	5

With this initial finding bearing in mind, we keep the analysis forward by identifying outliers. We identified that in the column of mom_cig, we have two very obvious outliers probably due to bad reporting of data. So we will recode the 'None' as 0 and the rest as 'NA' for later analysis.

Table 3: Irregular reporting of mom numcig

parent_id	mom_numcig
50102	2 black and miles a day
52702	44989
53802	20-25
54802	None

For other continuous variables like composite score from the three types of questionnaire (Brief Problem Moniter, Parental Knowledge, and Emotional Regulation), we perform univariate analysis looking at their distribution.

Univariate Analysis

By looking at the following three tables respectively for BPM,ERQ,and PMQ, we can have a holistic view of the distribution of main scores in the dataset, and how do parent and child scores differ.

The first one correspond to Brief Problem Moniter Scores. It is used to mainly record Behavioral questions. Separate questions are used for adults or adolescents to for rating Internalizing (INT), Attention (ATT), and Externalizing (EXT) problems. The parent and child grouping is differentiated as the 'score_group' indicate. Consistently for each of the three categories of problems, parent rating themselves all have the highest means and children rating themselves all have the lowest means.

Table 4: Summary of Brief Problem Moniter Scores

Var_name	mean	sd	score_group
bpm_att	3.000	2.625	P
bpm_att_a	1.474	1.955	C
bpm_att_p	2.056	2.216	P-C
bpm_ext	2.811	2.012	P
bpm_ext_a	1.237	1.567	C
bpm_ext_p	1.676	2.495	P-C
bpm_int	2.714	2.729	P
bpm_int_a	1.538	1.862	C
bpm_int_p	2.205	2.483	P-C

Note:

Score Group is interpreted as

[1] P:Parent; P-C:Parent on Children; C:Children

Table 5 describes the Emotional Regulation Scores summary, by two categories: (1) Cognitive Reappraisal and (2) Expressive Suppression. The parent and child grouping is differentiated as the 'score_group' indicate. Consistently for each of the two categories of problems, parent rating themselves all have a higher means than children rating themselves.

Table 5: Summary of Emotional Regulation Scores

Var_name	mean	sd	score_group
erq_cog	3	1	C
erq_cog_a	5	1	P
erq_exp	3	1	C
erq_exp_a	3	2	P

Note:

Score Group is interpreted as

[1] P:Parent; C:Children

Table 6 describes the Parental knowledge related Scores summary, by four categories: (1) Parental knowledge, (2) Child disclosure, (3) Parental solicitation and (4) Parental control. The parent and child grouping is differentiated as the 'score_group' indicate. Consistently for each of the four categories of problems, parent rating themselves all have a higher means than children rating themselves. And both parent and children reporting have Parental Control as the highest scoring categories in this group.

In conclusion, by looking at univariate distribution of each subcategory of the questionnaire, we do not observe any irregular patterns of the scores. This ensures the next step of our exploratory analysis.

Table 6: Summary of Parental Knowledge Scores

Var_name	mean	sd	score_group
pmq_child_disclosure	3.433	1.000	C
pmq_parental_control	4.345	0.932	C
pmq_parental_knowledge	3.990	0.788	C
pmq_parental_solicitation	2.977	1.341	C
ppmq_child_disclosure	3.676	0.672	P
ppmq_parental_control	4.584	0.954	P
ppmq_parental_knowledge	4.258	0.575	P
ppmq_parental_solicitation	4.182	0.729	P

Note:

Score Group is interpreted as

[1] P:Parent; C:Children

Data Tranformation

One of the main process of data transformation follows this section. As explained at the analysis plan in the beginning, we are creating both timing and intensity variables to evaluate the effects of SDP/EXT in a quantitative and visualizing way. We will using the time and behavior indicator variables such as 'mom_smoke_pp1' as indicators for SDP/EXT exposure at the corresponding time point. We plan to create timing variables in two ways. The first one is a binary variable divided in terms of prenatal vs postnatal. Using this variable we can see how in a big picture exposure first occurred at prenatal period or postnatal would impose effects. For prenatal, we have variables including mom_smoke from 16 weeks pregnant to 32 weeks, as well as lab recording of Urine cotinine in mothers at 34 weeks gestation. So any observation whose first exposure happened in these four period would be given a label of 'prenatal'. For postnatal, we have variables including mom smoke postpartum from visit 1 to 12 weeks, lab recording of Urine cotinine in both mothers and children at 6 month postpartum, and smoke exposure from mom or partner from begin of postpartum til 5 years. As a note, the exposure variables are retrospective of mothers at the new study. So any observation whose first exposure happened in these four period would be given a label of 'postnatal'.

The second way is to look deeper into each of the periods. In addition, as the aim is targeting, we are primarily interested in the effects of SDP and ETS, we could also incorporate this dividing in the second way of creating timing and intensity variables. Basically, for both natal periods, we subdivided their respective first exposure within that period as timing indicator for the observation and also create intensity indicator as well. For intensity, we coded as cumulative sum of all the related smoke or exposure variables within the specificity time frame (prenatal vs postnatal). So for prenatal, we named the timing variable as 'trimester', and the levels are set as 1,2,3 to indicate first, second, and third trimester for which the baby's initial smoke exposure. And intensity as 'trimester_int' to demonstrate level of intensity within the prenatal period.

For the lab Urine cotinine values, we decided a level of bigger than 0 as a indicator of smoke exposure for baby and a level bigger than 10 for mother as indication of smoker or exposed to second hand smoke.

For postnatal, variable establishment is quite similar except for the time point of 6 month postpartum. For this time point, we have all three of self-reported smoke, lab Urine cotinine and smoke exposure from mom or partner. Due to the nature of self-reported survey typed questions and its validity problems, when coding for postpartum timing, in the level of 6-month, we primarily look at the indicator from lab Urine cotinine values and mask the other two variables by it. The postpartum timing variable is named 'postpartum' and intensity 'postpartum_int'.

Moreover, considering the SDP vs ETS, we can say that prenatal variables (trimester,trimester_int) corresponds to smoke during pregnancy since at these stages, the mother is still in pregnancy and bay hasn't born yet. For ETS, we can say that postnatal variables (postpartum,postpartum_int) corresponds to environmental tobacco smoke since the baby is currently out of the mother's uterus and any surrounding environmental exposure to tobacco would be considered exposure. This is also why we decided that as long as the level of baby lab Urine cotinine is bigger than 0, there is exposure of ETS of the corresponding time, and adds 1 to the level of intensity.

Notable to mention that for the timing variables, we are considering initial exposure. So in cases there may be more than one time indicators in a specific time frame: eg: 'mom_smoke_16wk' and 'mom_smoke_22wk' both fall at the second trimester. As long as there's an indicator of the two, this observation has first SDP exposure at second trimester. Intensity is coded as sum of number of indicator in the cumulative time period.

Here's the summary of the variables that we generated at this stage.

This is a table of summary for the timing and intensity variables that we created. We can see in detail that take variable 'natal' as an example. By definition, this correspond to first exposure of SDP or EXT in Prenatal(33) VS Postnatal(7). Then referring back to the 'trimester' variable which is the continual timing variable for prenatal. There is a total of 33 observations (2 trimester, 11 observation), (3 trimester, 22 observations). Also from the summary and distribution of the variables that we created, we can see that our data is highly unbalanced. There are a lot of cells that do not have any observations. We may continue for the purpose of conducting exploratory analysis to identify potential patterns, but the current data quality may not be a good choice for performing any statisticl analysis.

With such creation of timing and intensity variables, we conduct exploratory analysis to address our question of interest: Examine the association between smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure and self-regulation, externalizing behavior, and substance use. We will conduct analysis for each of the three domains.

Prenatal VS Postnatal

With regard to the created variables, we then compute this table looking at how all the scores differ in terms of prenatal vs postnatal to look for general patterns. Each column consists of observations whose first exposure occurs at either prenatal or postnatal period. And they are exclusive.

Characteristic	**N**	**Postnatal**, N = 7[1]	**Prenatal**, N = 33[1]	**p-value**[2]
bpm_att	37	2.29	3.17	0.2
NA		0	3	
bpm_ext	37	2.29	2.93	0.4
NA		0	3	
bpm_int	35	2.57	2.75	0.8
NA		0	5	
erq_cog	36	3.64	3.09	0.2
NA		0	4	
erq_exp	36	2.75	2.75	>0.9
NA		0	4	
pmq_parental_knowledge	35	4.11	3.96	0.6
NA		0	5	
pmq_child_disclosure	36	3.63	3.39	0.6
NA		0	4	
pmq_parental_solicitation	35	3.00	2.97	>0.9
NA		0	5	
pmq_parental_control	33	3.87	4.45	0.2
NA		1	6	
erq_cog_a	38	6.07	5.26	0.085
NA		0	2	

erq_exp_a	38	3.25	3.50	0.6
NA		0	2	
bpm_att_p	35	1.86	2.14	0.7
NA		0	5	
bpm_ext_p	36	1.43	1.79	0.7
NA		0	4	
bpm_int_p	38	1.71	2.35	0.4
NA		0	2	
ppmq_parental_knowledge	36	4.17	4.28	0.6
NA		0	4	
ppmq_child_disclosure	36	3.94	3.61	0.2
NA		0	4	
ppmq_parental_solicitation	33	4.31	4.12	0.4
NA		0	7	
ppmq_parental_control	36	4.74	4.53	0.4
NA		0	4	
bpm_att_a	37	1.29	1.50	0.7
NA		0	3	
bpm_ext_a	37	1.86	1.13	0.4
NA		0	3	
bpm_int_a	38	0.86	1.74	0.088
NA		0	2	
swan_hyperactive	38	7.3	8.0	0.8
NA		0	2	
swan_inattentive	38	10.1	11.2	0.7
NA		0	2	

[1] Mean
[2] Welch Two Sample t-test

As we do not obtain any significant P-values and the score differences vary by type: some have larger score prenatal, some larger at postnatal. It does not seem to have any meaningful patterns at this stage, we shall proceed by looking at their respective sub timing and intensity in terms of bivariate comparisons.

Bivariate comparison

In this section, we will mainly perform bivariate comparisons between the selected scores to either timing or intensity variables that we created as before, either for prenatal and postnatal periods. Relating back to the questions that we are interested in, we want to explore associations between SDP/EXT to children behavior. More specifically, we were to look for meaning patterns such as whether earlier exposure to SDP/EXT would have different impacts on children behavior as compared to later exposure; as well as lower cumulative intensity vs higher cumulative intensity. We will conduct analysis for each of the three domains of main interests: Self-Regulation, Externalizing behavior and Substance use problem.

Self-Regulation

Based off resources, Self-regulation means the ability to understand and manage your own behavior and reactions, and it contains behaviors in the following four dimensions: executive function, emotion regulation, effortful control, vagal tone. Given what we have in the data set, we decided that the following variables are related to adolescent self-regulation issues: emotional regulation scores of the adolescent themselves; attention-relating scores on the Brief Problem Monitor questionnaire of both the adolescent own answering, and parents evaluation on their child; lastly, we think the SWAN response would also be relevant as

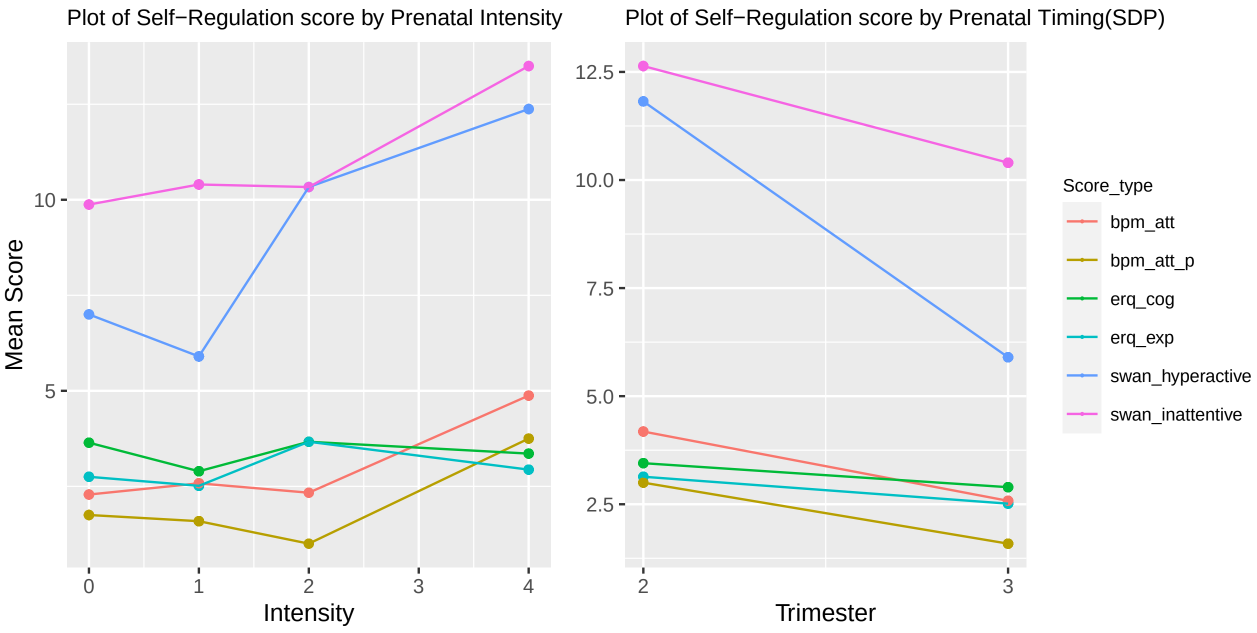
Table 7: Summary of Timing and Intensity Variables

Characteristic	**N = 41**
trimester	
2	11 (33%)
3	22 (67%)
NA	8
trimester_int	
0	8 (20%)
1	22 (54%)
2	3 (7.3%)
4	8 (20%)
postpartum	
1	2 (5.7%)
2	5 (14%)
3	3 (8.6%)
4	22 (63%)
5	2 (5.7%)
6	1 (2.9%)
NA	6
postpartum_int	
0	6 (15%)
1	13 (32%)
2	8 (20%)
3	2 (4.9%)
4	2 (4.9%)
5	3 (7.3%)
6	3 (7.3%)
7	2 (4.9%)
8	2 (4.9%)
natal	
Postnatal	7 (18%)
Prenatal	33 (83%)
NA	1

[1] n (%)

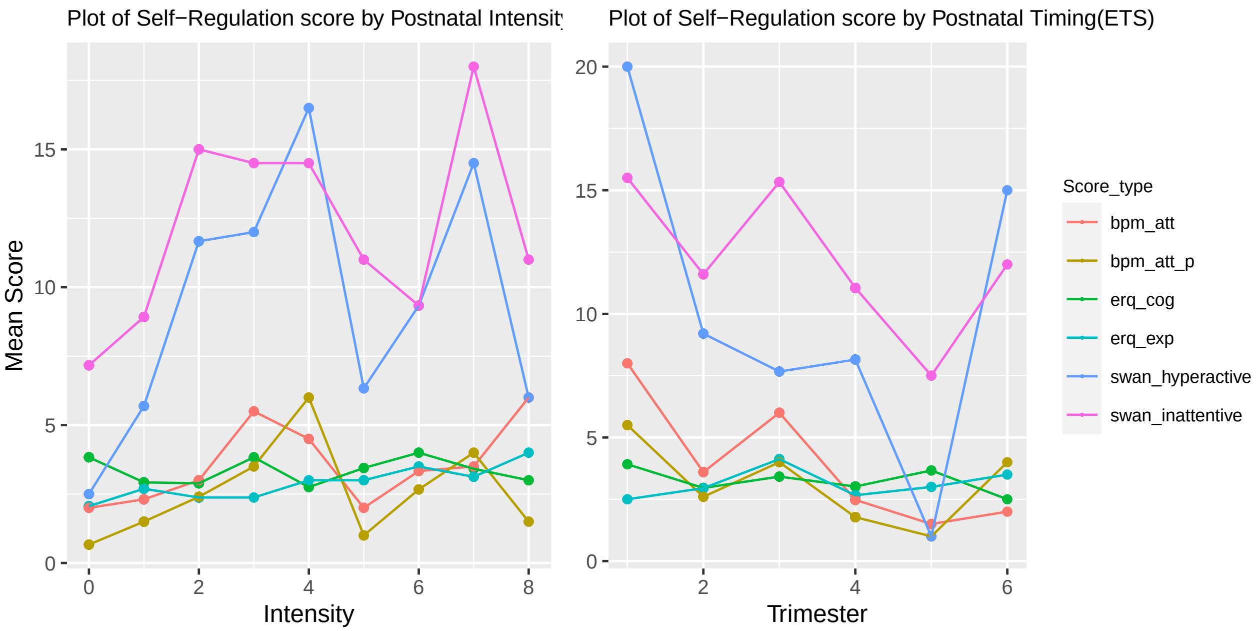
the conditions of ADHD reasonably indicate that the child has self-regulation problems such as easily get distracted. Here's the finding that we have:

The following two plots displays the respective trends of the selected scores with timing and intensity, we displayed two sets, one for prenatal and the other postnatal.

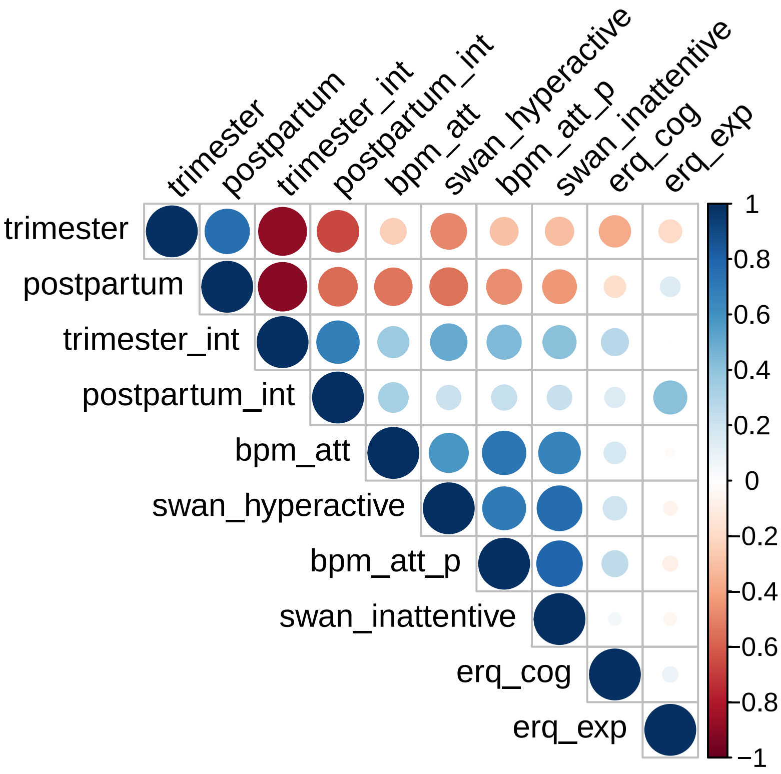


For prenatal associations, we found out that earlier first exposure time and higher cumulative intensity are associated with higher scores for all the scores that are plotted. Referring back to their definition, higher scores in general represent a bad performances on self-regulation tasks such as attention. Since during the prenatal period, the baby is still not born yet, so we can connect the conclusions to Smoke during pregnancy part where as for postnatal period, it is more connected to environmental tobacco smoke part.

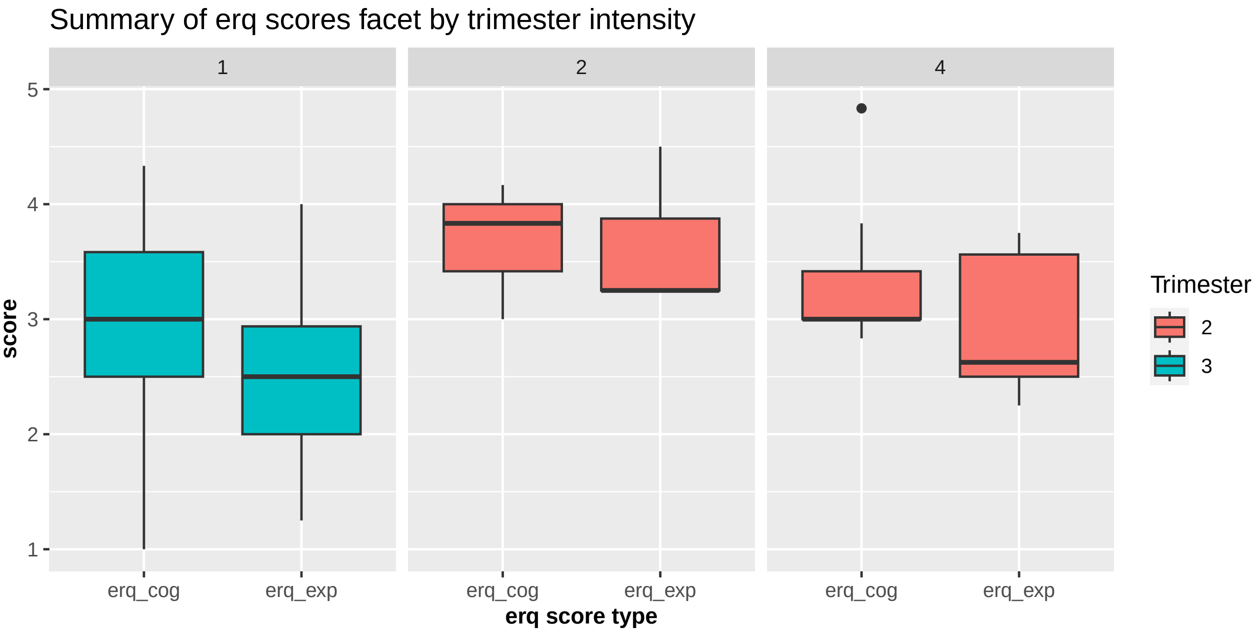
Similarly, we would like to assess the association between postnatal ETS exposure timing and intensity with the scores related to the self-regulation domain. The result plots are presented as below. Similar general trend can be observed for postnatal as compared to prenatal. The swan scores fluctuates more than the other scores. And higher intensity, earlier ETS exposure lead to higher scores, which correspond to bad behaviors relating to Self-Regulation.



We take a step further by looking at the correlation between each pairs of timing, intensity and scores. In this plot,from package(Wei and Simko 2021). Each box and circle within it represent a pair of variables aligned with the axis. An external color scheme on the right represent direction and value of correlation statistics. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. From the plot, we can see that timing parameters (trimester and postpartum) have in general red circles with the different scores, meaning positive correlation. Intensity variables in general have blue correlations. This finding is in line with the separate conclusions we draw from above.



Before coming to the final conclusion in this section, we want to highlight the differences between erq_cog and erq_exp scores. cog correspond to Cognitive Reappraisal while exp correspond to Expressive Suppression. By looking carefully through the original questions that was answered, we realize that their trends should be considered separately. Higher cog scores means better regulation while lower exp means better regulation. So we decided to explicitly look at these two variable in particular.



To compare, we can see that both mean erq score for trimester 2 is higher for intensity 2 in comparison to 4 and we do not have more data to compare the third trimester. So this is telling that controlling for the same initial exposure of SDP, higher intensity of SDP may lead to lower erq_cog and erq_exp scores.

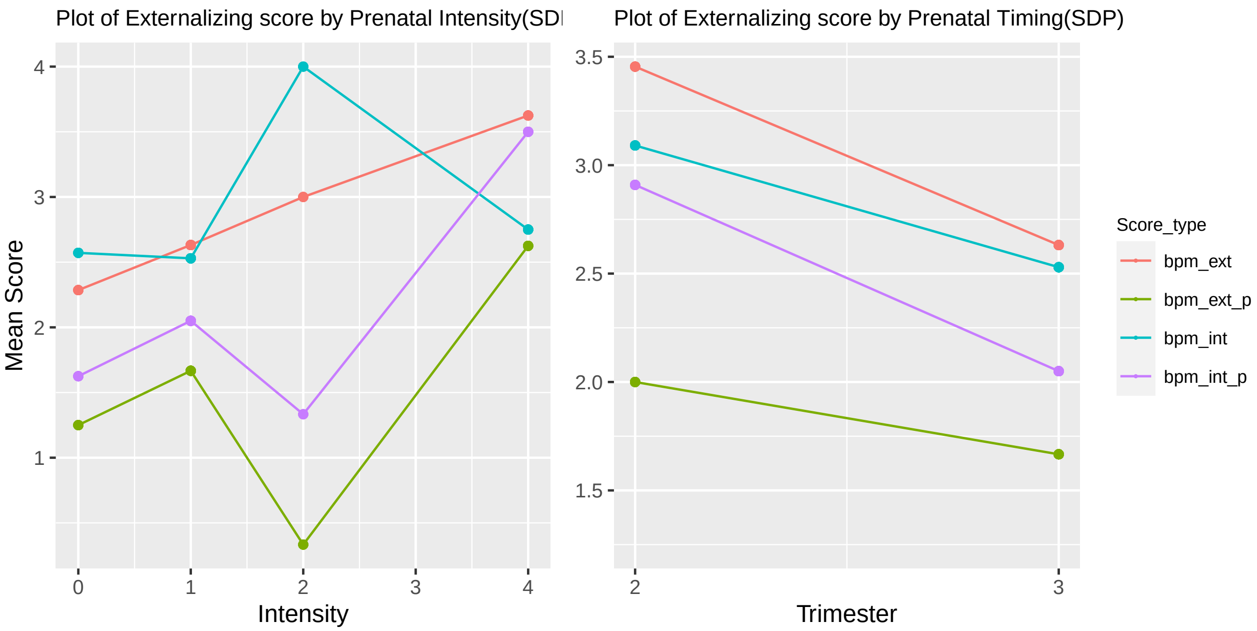
So our general conclusion in this section is as follows: Given the condition of data that we have, we focused on interpreting the effects of pre-natal SDP exposure in terms of time and intensity on adolescents' self-regulation problems. Although we do not obtain all statistical significant outcomes, due to the limited size of our data, we observe some consistent and generalized patterns: earlier first-time exposure and higher cumulative intensity of exposure may be potential risk factors of worse condition of adolescent's self-regulation behavior.

Externalizing Problems

Then we would like to focus of externalizing problems. The most intuitive indicator of externalizing problems would be bpm_score in terms of externalizing. Since we also have records about internalizing problems, we would like to also include that in this section for an additional reference. We follow the similar pattern of exploring as above: first see the general trend of prenatal exposure vs postnatal and then look deeper into how, in each phases, exposure time and intensity may impact adolescent externalizing behavior.

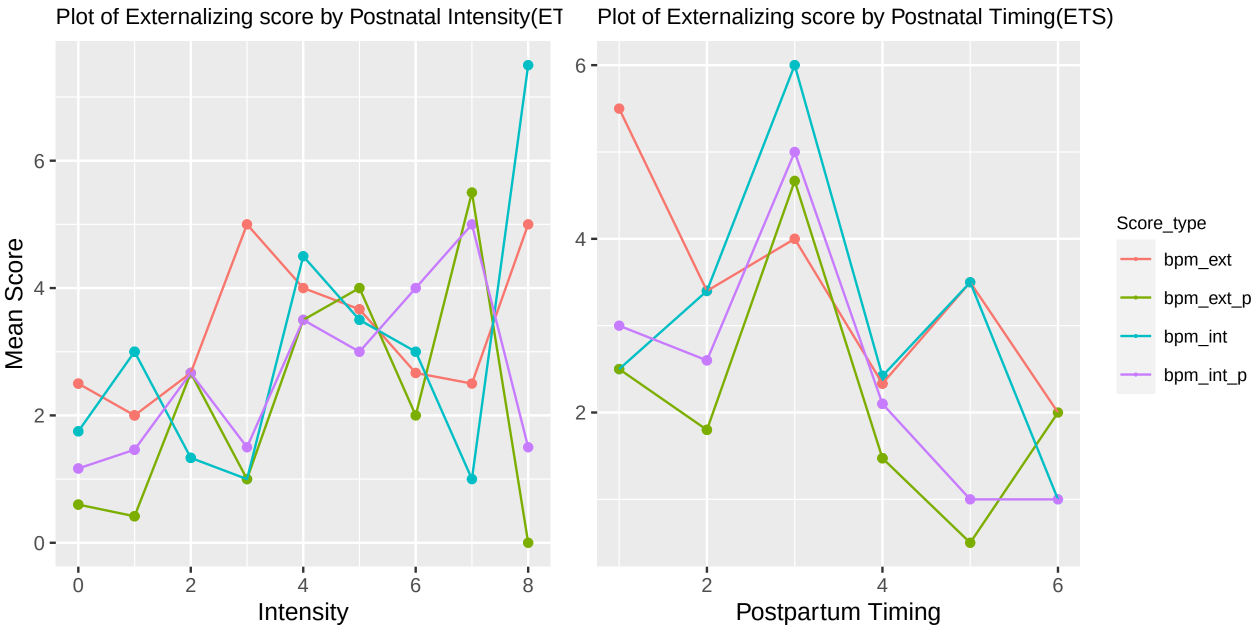
Through looking at the bmp scores of externalizing and internalizing between prenatal vs postnatal from the above table(bpm_ext,bpm_int,bpm_ext_p,bpm_int_p), although the P-value does not indicate statistical significance, again, this may be due to the quality of our data, we can see similar numerical pattern in all the scores, postnatal has generally lower scores than prenatal. We would like to explore the relationship between time and intensity.

Starting with prenatal timing and intensity to the Externalizing scores.



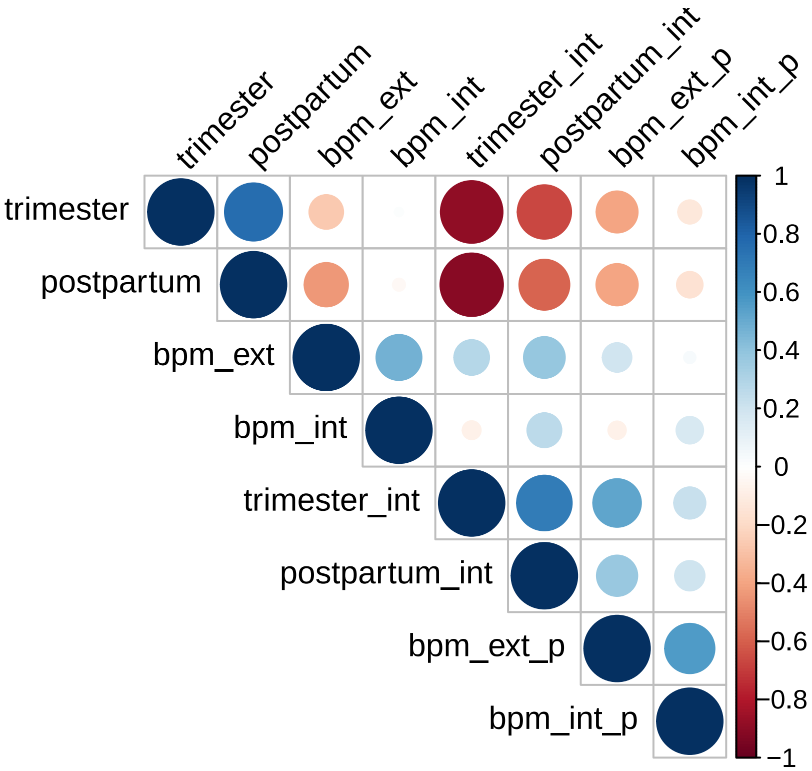
As expected, we observe associations between earlier exposure and higher cumulative intensity to higher levels of externalizing scores. Similarly as the previous self-regulation domain, we do observe some fluctuates in adjacent levels of intensities. And we wish we could have more timing points for prenatal to have a better visualization.

As for postnatal periods, despite more severe fluctuates, the general patterns remain the same: upward trend for intensity and downward trend for timing.



Both come down to the similar conclusion that earlier first-time exposure and higher cumulative intensity of exposure may be potential risk factors of worse condition of adolescent's Externalizing behavior.

We then move on to look at their correlations between each other.



From the graph, timing have positive correlation to the scores and intensity negative correlations to the scores. Another interesting finding is found between timing and intensity. They have positive correlation, meaning that children with later exposure tend to have higher cumulative exposure intensity. This is counter intuitive considering to the amount of time and probability of experience smoke exposure.

Substance Use Problem

The substance use section should contain analysis of how SDP/EST would impact adolescent's substance use behavior. From the data quality check section, we have found out that variables related to this problem is almost entirely missing, not to mention if we stratify them by timing and intensity. Here's the output

Table 9: Summary table of Substance use

	Number of YES	Cases with reporting
Cigarette	1	1
E-cigarette	3	2
Marijuana	3	3
Alcohol	5	4

From the summary table, we can see the among the total of 49 observations in our data, only 1,3,3,5 has reported YES to questions 'Have you ever' used cigarette, E-cigarette, Marijuana, and alcohol respectively and 1,2,3,4 observations reported a number of substance that they used. And this does not match to previous question, for example E-cigarette and alcohol. So we argue that with the current condition of data that we have, there's no point conducting analysis with substance use behavior and the main focus should be acquiring better data.

Conclusion and Limitation

Our exploratory data analysis focused on the potential effects of smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) on adolescents' self-regulation, substance use behaviors, and externalizing tendencies. We created the timing and intensity of SDP and ETS exposures to elucidate their impact on these behavioral dimensions. Despite data limitations, the analysis revealed noteworthy patterns:

Self-Regulation: There appears to be an association between early SDP/ETS exposure and diminished self-regulation abilities in adolescents, with greater exposure intensity exacerbating this effect. This trend persists across both prenatal and postnatal exposure periods.

Externalizing Behavior: Early and more intense exposure to SDP/ETS also seems to correlate with an increase in externalizing behaviors, mirroring the pattern observed for self-regulation.

These observations align with the prevailing understanding that smoke exposure during pregnancy is detrimental to child health. However, it is critical to approach these correlations cautiously and not infer causation without more rigorous statistical validation.

The analysis faces significant limitations:

The dataset is constrained in size, with numerous missing values, which impedes the detection of trends, especially within subgroups. The imbalance in data also presents challenges for subsequent analytical and modeling endeavors.

Reliance on self-reported questionnaire scores and a singular objective measure, urine cotinine, introduces potential reporting bias, questioning the outcome validity.

For substance use variables, the lack of reliable adolescent data, primarily due to missing or non-informative responses, necessitates alternative analytically approaches such as weighting or matching in future studies.

Further research with enhanced data quality is essential to substantiate these preliminary findings and to better understand the implications of SDP/ETS on adolescent development.

Reference

Wei, Taiyun, and Viliam Simko. 2021. "R Package 'Corrplot': Visualization of a Correlation Matrix." https://github.com/taiyun/corrplot.