Health Effects on Babies Born to Mothers Who Smoked During Pregnancy

Author Contributions

James Lu: Contributed R code, graphical analysis, conclusion / discussion, numerical analysis, and incidence analysis. I did the write ups for the numerical summary section, graphical analysis section, incidence report, and the conclusion / discussion section (minus the data limitations portion). I generated tables 1 and 2 as well as figures 1, 2, and 4.

Matin Ghaffari: Contributed R Code and explained them in the analysis for various numerical and graphical comparisons that I contributed in writing the interpretations of. Some in particular are for the code and explanations for the histogram, QQ plots, incidence. I also wrote the introduction, advanced analysis, discussion, and parts of the conclusion.

Introduction

This study examines the difference in birth weight between babies who were born to mothers who smoked during their pregnancy and mothers who did not smoke during pregnancy. Our data sample of these two populations of smoker pregnancies and non-smoker pregnancies consists of 1236 babies and mothers from all pregnancies in the Kaiser Health Plan in Oakland, California from 1960 and 1967. This data set comprises babies who are only boys, not twins, and lived for at least 28 days in order to control for these possible confounds and sources potential causes of noise within the observation of our data (*Maternal Smoking and Infant Health 16)*. In our current society with the abundant availability of information, modern science, and technology it is reasonable for the typical person to infer the severity of the effects when mothers smoke while pregnant since we often hear about the significant effect that certain foods, drugs, and lifestyle choices of pregnant women can have on their newborn's health. Especially when considering smoking, knowing the significant health effects that it already has even on a fully grown adult, we can infer how smoking can lead to much more severe health risks to a developing baby. Therefore, using this aforementioned data set, which comes from Child Health and Development Studies (CHDS), we would like to use graphical and numerical methods to summarize and interpret the information in the data set to gain insight on the observations and data figures important variables that are crucial to understand when addressing these such troubling problems.

Ultimately, we aim to provide insight and awareness to this issue by conducting various exploratory data methods in our analysis that I believe can help raise awareness on the gravity of this issue through the strong and clear relationships amongst the data variables that can also potentially bring to light new awareness or ideas for resolutions. Ultimately we will use these relationships and suggestions from these graphical, numerical, and incidence of the differences in the distribution of birth weights in newborns of smoking and non-smoking mothers. In the data set the variables we use are a continuous variable for the birth weight and a categorical variable of smoker which indicates either a 0 or a 1 if the mother is a smoker, which had the erroneous value of 9 for some observations that we had to remove the rows for. Furthermore, the framing issues we face is using this relationship of mother smoking status and baby weight and then having the necessary information necessary to understand the effect it has on the health of the baby.

1. Numerical Summary and Analysis of The Two Distributions for Birth Weight of Newborns of Smoking and Non-Smoking Mothers

Before getting into any complex statistical analysis it was important to summarize the data in each of the groups. To do this, we calculated some basic summary statistics to get an idea of the data that we might be working with. We began by finding the sample size known as the count of both groups followed by the minimum, 1st quartile, max, mean, median, mode, 3rd quartile, standard deviation, skewness, and kurtosis to try to see if there were any obvious differences between the two samples.

Table 1: A table outlining various numerical statistics using data of birth weights from mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy

	Non-Smoking Mothers	Smoking Mothers
Count	742	484
Minimum	55	58
1st Quartile	113	102
Mean	123.047	114.110
Median	123	115
Mode	129	115
3rd Quartile	134	126
Maximum	176	163
Standard Deviation	17.398	18.098
Skew	-0.187	-0.033
Kurtosis	4.026	2.975

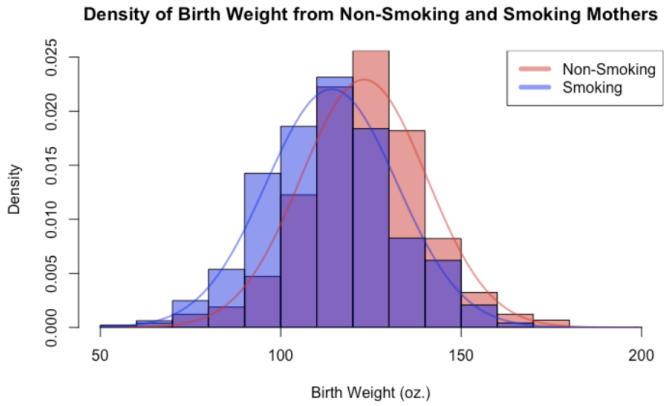
After calculating our basic summary statistics we found that the mean and median birth weight from smoking mothers were less than the mean and median birth weight from non-smoking mothers. We also found that the 1st and 3rd quartiles for smoking mothers were less than those from non-smoking mothers. The min, max, and standard deviation gave us a general idea of how the data might be spread out in each respective distribution. Comparing each distribution's skew and kurtosis to that of a standard normal (skew=0, kurtosis=3) we find that both distributions are relatively normal with the smoking mother baby weights being closer to a standard normal distribution.

There aren't any concrete conclusions that we can draw from these statistics alone, however we can see that generally the average birth weight from smoking mothers were smaller than the average birth weight from non-smoking mothers.

2. Graphical Methods of Analysis in Baby Birth Weight of Smoker and Non-Smoker Pregnancies

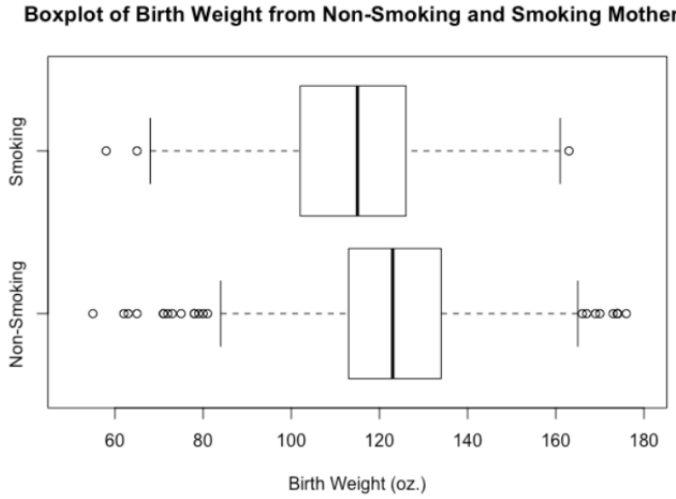
After gathering summary statistics we further understand the data within our samples by generating visualizations for the data which we accomplish with histograms, box plots, and Quantile-Quantile plots for our samples. Histograms of our samples allow us to better understand their distributions and frequency/density of their data, which we use in our analysis to compare the distributions of baby birth weights of non-smoking mothers to smoking mothers. The boxplot is another visualization that shows us other meaningful characteristics of the data, as it is helpful in showing skewness and how distributions might differ in regards to their median, range, outliers, and IQR. Furthermore, the QQ plot is another helpful way in understanding the distribution of data and can help clearly show the differences amongst them and if they potentially come from the same distribution.

Figure 1: Histogram that shows the density of birth weight from mothers who smoked during pregnancy and mothers who did not, with their corresponding density curve.



In Figure 1, we can see that both distributions look relatively normal as we predicted in our numerical analysis, however the difference in the range of the data is more clear. Both distributions appear to be unimodal and slightly asymmetric. It appears that the data from smoking mothers is shifted to the left of the data from non-smoking mothers, this observation implies that in this sample the birth weight from smoking mothers is generally less than the birth weight from non-smoking mothers.

Figure 2: Boxplot that visualizes the range, IQR, outliers, and median birth weights from mothers who smoked during pregnancy and mothers who did not.



In Figure 2 the difference in overall range, IQR, outliers, and median become much clearer. The boxplots also help us visualize the asymmetry of both distributions as the median line strays slightly from the middle of the IQR box which implies that both distributions diverge from the approximately symmetric normal distribution. Also we see many outliers for non-smokers as

there are many values beyond the upper and lower fence which shows how the median isn't as centered between Q1 and Q3 for non-smoker as opposed to the smoker sample who has only 3 outliers which helps us understand how the smoker's distribution is more symmetric and centered.

Figure 3: Quantile-Quantile for each of the two distributions helps visualize the normality of these distributions as well as the difference in the distributions and how the data is distributed in each samples' theoritical quantiles, the red lines are a reference to the QQ of a standard normal distribution.

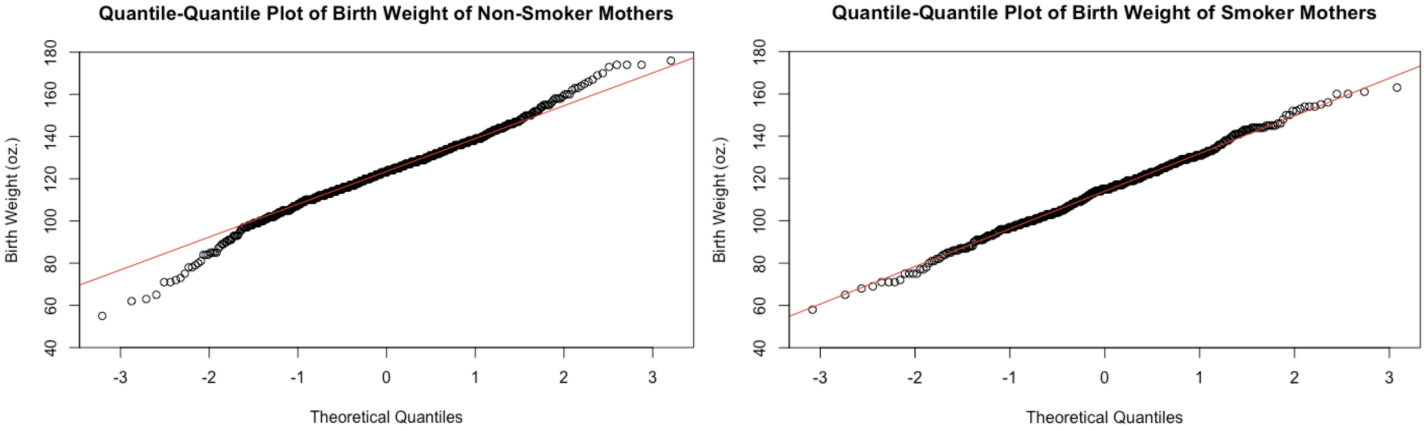
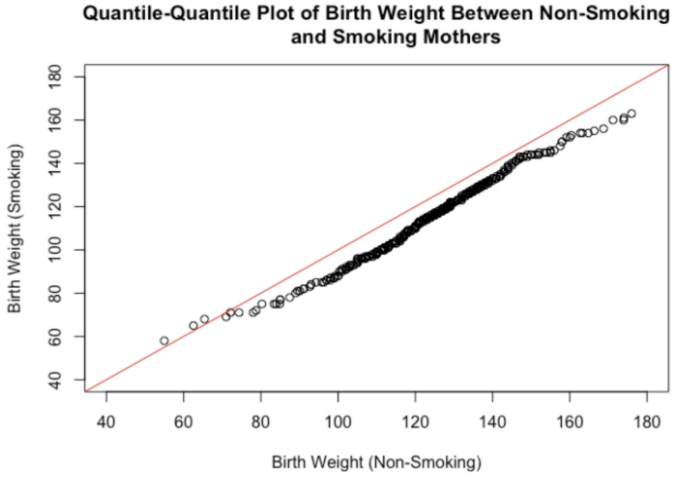


Figure 4: Quantile-Quantile of both distributions together so we can see how the two distributions differ from one another as seen with the shift from the reference line due to the off-centeredness and spread of non-smokers however this plot shows that they do come from the same approximately normal distribution since the points are quite linear.



In Figure 3 we see that both the distributions are approximately normal; however we see that the non-smoker distribution has more variance and skew in the 2nd quartile. In Figure 4, we can also see how the distributions differ, since even though they are mostly linear, suggesting they both come from an approximately normal distribution. It is however shifted from the standard normal reference line. When a shift occurs in a QQ plot, it implies that there is a shift in the distributions and in this scenario it is shifted towards non-smoking mothers, which also indicates the non-smoker higher median. Thus, with these QQ plots we found that the samples both come from the same approximately normal distribution, however the aforementioned spread and center of the non-smoker sample causes the QQ line to deviate from the standard normal line. These graphical methods that we generated also confirm that the two distributions have small amounts of skew and are mostly normal with non-smokers being slightly more skewed.

The basic histograms and boxplots were helpful in understanding the range and balance of the distributions while the QQ plot was very helpful in visualizing the normality and relationship between the two distributions. Overall the plots confirmed that the samples both come from a approximately normal distribution and showed that the distribution of non-smoking mothers were shifted to the right of the distribution of smoking mothers.

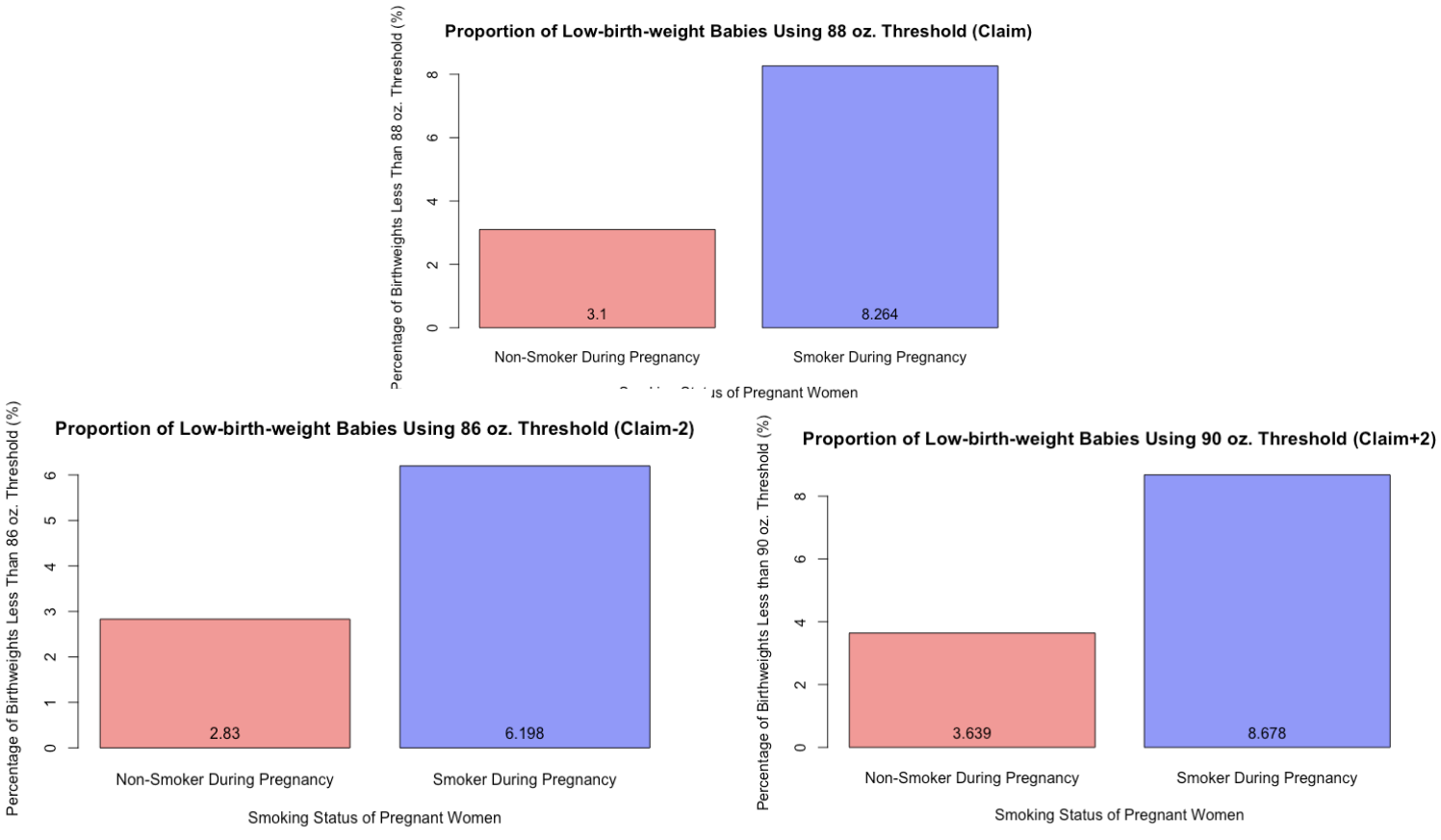
3. Incidence and Its Reliability as an estimate for Low-Birth-Weight Babies of Smoker and Non-Smoker Mothers

To compare the incidence and frequency of low-birth-weight (LBW) babies between mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy we can look at the proportion of LBW babies to the total number of babies in that distribution. After calculating these basic proportions we can test the robustness of our estimates by changing the threshold of what is considered a LBW baby. Currently our threshold is 88.1849 ounces however by modifying this threshold by ± 2, we can check if there are any unusual clusters around this value.

Table 2: A table outlining the incidence rate of low birth weight babies from mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy

Low Birth Weight Threshold	LBW Incidence Rate (Non-Smoking)	LBW Incidence Rate (Smoking)
86.1849 oz. (Claim - 2)	2.830%	6.198%
88.1849 oz. (Claim)	3.099%	8.264%
90.1849 oz. (Claim + 2)	3.634%	8.677%

Figure 5: Bar Plots outlining the incidence rate of low birth weight babies from mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy with varying thresholds in order to examine the robustness and sensitivity to change these estimates are.



After adjusting the low birth weight threshold we can observe that the incidence rate remained relatively stable and did not have any large deviations from the original incidence rate. Shifting the LBW threshold by plus and minus 2 allows us to test how robust our estimates of incidence rate are, and after conducting these calculations we found that there was little change between the different thresholds thus implying that our estimates for incidence rate are accurate.

Advanced Analysis

In an advanced analysis, we would like to implement other methodologies that will provide more definitive comparisons and compelling suggestions about our data that we may use to make more informative inferences with. These more ideal methodologies that we would like to implement include hypothesis testing and statistical tests for our two populations. Since we have a sample statistics for the mean and our population parameters are unknown, it would be best to perform a two sample t-test on the average birth weight of babies in smoker and non-smoker mothers (independent populations). However in order to run this test and get its proper results we must be sure to meet all of its assumptions to ensure our test statistic follows the degree of freedom of $T_{\text{min(sample 1 size - 1, sample 2 size - 1)}}$.

To begin with, the first assumption that we will satisfy is that the two populations of smoker pregnancies and non-smoker pregnancies are independent. This is because it is rational to assume that there aren't any relationships, links, or influences from one sample to the other in our scenario and as a result doesn't allow us to predict or understand the data points of one sample by just looking at the data points of the other. This is because there is no obvious pairing amongst mothers who smoke and don't in regards to their children's birth weight and we can't look at a certain non-smoker's baby and predict the weight of a smoker's baby, thus meeting the assumption that the populations are independent of another. Secondly we meet the second assumption which is the condition that the averages of the populations, regardless of their distribution, are approximately normal because of the central limit theorem. We meet this assumption since we see graphically previously in the analysis that there is limited skew and because of our large n values for both populations, which satisfy the central limit theorem balance condition, since we have much more than the 30 or more samples needed to balance out possible high skew. Meeting this assumption allows our sample mean to be approxinately normal. Lastly, we must meet the third assumption that the data in each sample is independent which we know because the study mentions the pregnancies were chosen randomly via a pseudo-randomization program, and because we meet the 10 percent rule which states that our sample size should be no more than 10% of the population which is reasonable since there is more than 10 * 742 and 10 * 484 total smoker and non-smoker pregnancies. Since we meet all these required assumptions, we conduct the t-test framework below so that we can more definitively support or reject our hypothesis. We do so below, since our results ultimately show that the data suggests a association between low birth weights in babies and mothers who smoke while pregnant as opposed to non-smoking pregnancies.

We can set up our 2-sample t-test for our two independent populations since we meet the aforementioned assumptions above, thus we use our 2 sample means and do the following:

1. Parameter: Let $\mu_d = \mu_{\text{non-smoker}} - \mu_{\text{smoker}}$ be the difference of averages (mean non-smoker baby birth weight - mean smoker baby birth weight) for all pregnancies in smokers and non-smokers.
2. Null Hypothesis (H_0): There is no difference in means of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
3. Alternative Hypothesis (H_1): There is a difference in birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
4. We perform A|B test through H_0: μ_d ($\mu_{\text{non-smoker}} - \mu_{\text{smoker}}$) = 0 vs. H_1: μ_d ($\mu_{\text{non-smoker}} - \mu_{\text{smoker}}$) > 0

5. Then we can either use the R t.test() method or we can compute manually the test statistic and degree of freedom through the sampling distribution: $t = \frac{\overline{x} - \overline{y} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}$ on $T_{\min(n_1-1, n_2-1)}$
    - Alternatively to the next step in R we can use this equation to compute the t-stat and degree of freedom to obtain a p-value by looking at the t-probability table.
6. After doing t.test(non-smoker baby weight data, smoker baby weight data, mu = 0, alternative = "greater") in R, we get the results that correspond to the hypothesis test from step 4, which tells us various values such as p-value = 2.2E-16, t-stat = 8.583, degree of freedom = 1003.2, and that the true difference in means is greater than zero.
7. Therefore we see that our t-test results in a p-value of 2.2E-16 which is much lower than the general significance level of 0.05. That being so allows us to reject the null hypothesis in favor of our alternative hypothesis since the very small p-value indicates that the data is strongly suggesting that smoking while pregnant may result in lower birth weight (not causation).

Conclusions and Discussion

This research paper set out to identify whether or not there was a difference in weight between babies whose mothers did not smoke during pregnancy and babies whose mothers did smoke during pregnancy and if that difference in weight had an effect on the health of the baby. After conducting numerical, graphical, and incidence analysis we found that the distribution of mothers who did not smoke had higher birth weights than that of mothers who did smoke during pregnancy. On average, we found that babies born from mothers who did not smoke were ~8 ounces heavier than those who were born to mothers who did smoke during pregnancy. We also found that mothers who smoked during pregnancy were 5% more likely to have a low birth weight baby than non-smoking mothers. The increased likelihood of LBW in smoking mothers is alarming as these babies are more likely to develop disease and health complications. The Children's Hospital of Philadelphia states that LBW babies suffer from low oxygen levels at birth, inability to maintain body temperature, difficulty feeding and gaining weight, infection, breathing problems, neurologic problems, gastrointestinal problems, and sudden infant death syndrome. Our findings generally coincide with other studies of this topic, in a study conducted by Kataoka et al. found that mothers who smoked during pregnancy had babies that were ~11.287 ounces lighter than mothers who did not smoke during pregnancy, however the small difference in our observations could be attributed to the differing origins of the datasets.

In our numerical analysis we found that the mean and median of the birth weight from non-smoking mothers were greater than that of the smoking mothers, we also found that our calculations of skewness and kurtosis implied that each of the distributions were approximately normal. The graphical analysis allowed us to confirm that the data was approximately normal, and gave us a visual representation of how one distribution differs from the other. Our QQ plots were especially helpful in determining normality and led us to the conclusion that the two sets of data come from the same distribution with one being shifted. The findings from our incidence study were important in understanding how smoking during pregnancy might affect the overall health of the baby.

Additionally, we discussed quite a few limitations of this data set that might raise possible issues with the validity of our numerical, graphical, and incidence analysis. One of these limitations is the source of the data. Since the data is from a single city and a single health plan, it is unreasonable to assume this would be representative of all babies since we lack randomization, which may possibly result in confounding factors connected to this particular city or health plan that could suggest a association that is from factors other than what we are studying, and thus possibly providing misleading results. Furthermore, with the data and analysis currently we were able to show an association between smoking and birth weight, however this may be invalid since

there are numerous other factors that can influence the birth weight of a baby that this analysis is failing to look at. For instance, the height of the father and mother may potentially have a significant influence other than just smoking on the birth weights of babies, however we don't have information on father heights or on baby health to determine if smoking and non-smoking birth weights are independent.

Works Cited

1) The Children's Hospital of Philadelphia. “Low Birthweight.” *Children's Hospital of Philadelphia*, The Children's Hospital of Philadelphia, 24 Aug. 2014, www.chop.edu/conditions-diseases/low-birthweight#:~:text=A%20baby%20with%20low%20birthweight%20may%20be%20at%20increased%20risk,staying%20warm%20in%20normal%20temperatures.
2) Kataoka, M.C., Carvalheira, A.P.P., Ferrari, A.P. *et al.* Smoking during pregnancy and harm reduction in birth weight: a cross-sectional study. *BMC Pregnancy Childbirth* 18, 67 (2018). https://doi.org/10.1186/s12884-018-1694-4
3) *Maternal Smoking and Infant Health*, vol. 1, no. 1, 1 Mar. 1995, pp. 1–27., doi:10.11120/msor.2001.01030053.