Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Caitlin Kennedy	April 15, 2016
Students Name	Date

Exposure to fine particulate matter from traffic in early life and childhood pneumonia-Survival analysis of an Atlanta birth cohort

Caitlin Kennedy

MSPH
Epidemiology and Environmental Health

_____
Matthew Strickland, PhD
Committee Chair

_____
Audrey Flak, PhD
Committee Member

_____
Paige Tolbert, PhD
Committee Member

Exposure to fine particulate matter from traffic in early life and childhood pneumonia- Survival analysis of an Atlanta birth cohort

Caitlin Kennedy

B.A. in Environmental Science and Spanish
Drew University
2013

Thesis Committee Chair: Matthew Strickland, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology and Environmental Health
2016

Abstract

Exposure to fine particulate matter from traffic in early life and childhood pneumonia: Survival analysis of an Atlanta birth cohort

By Caitlin Kennedy

Introduction: Pneumonia is one of the leading causes of global infant morbidity and mortality in children less than 5 years old. Environmental risk factors for pneumonia include indoor air pollution exposures as well as ambient traffic related exposures in urban areas. The objective of this study is to examine the association between residential particulate matter exposures and time to first pneumonia event during the first two years of life for children living in the greater Atlanta area.

Methods: Time to first clinical diagnosis of pneumonia was analyzed in a survival analysis using Cox proportional hazards regression. The outcome was defined as the first diagnosis of pneumonia after the first 28 days of life based on ICD-9 codes 480-486 in the medical record. Clinical data came from 22,520 children enrolled in the Kaiser Air Pollution and Pediatric Asthma Study (KAPPA), a historical birth cohort of children born between 2000 and 2010 and enrolled in the Kaiser Permanente Georgia HMO. Exposures to fine particulate matter ($PM_{2.5}$) from traffic emissions were modeled based on 2011 pollution estimates created using a research line-source dispersion model (RLINE) at 250-meter resolution.

Results: The effect estimate for the association between primary $PM_{2.5}$ exposure and first pneumonia event by age two was modestly elevated for a 1 microgram per cubic meter ($\mu g/m^3$) change in $PM_{2.5}$. The Hazard Ratio and 95% Confidence Interval for this association were estimated to be 1.17 (0.93, 1.47) in a no interaction, un-stratified model controlling for child sex, child race, maternal asthma status, maternal prenatal smoking status, maternal education, city region, and neighborhood socioeconomic status. Average $PM_{2.5}$ exposure was 1.17 $\mu g/m^3$ $PM_{2.5}$ with a standard deviation of 0.27 $\mu g/m^3$. Approximately 10% of the total 22,520 children were diagnosed with pneumonia during their first two years of life.

Conclusions: These results provide limited evidence for a harmful association between traffic-related primary $PM_{2.5}$ and pneumonia in the first two years of life. Our findings are consistent with the literature and suggest more research is needed to understand the relationship between chronic, early life exposures to fine particulates from vehicle emissions and early childhood pneumonia.

Exposure to fine particulate matter from traffic in early life and childhood pneumonia: Survival analysis of an Atlanta birth cohort

By

Caitlin Kennedy

B.A. in Environmental Science and Spanish
Drew University
2013

Thesis Committee Chair: Matthew Strickland, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology and Environmental Health
2016

Acknowledgements

This thesis would not have been possible without the help and support from Dr. Matthew Strickland and Dr. Audrey Flak. Thank you for all your time and consideration in providing guidance throughout this research experience. I feel fortunate to have had the opportunity to work with such talented and knowledgeable mentors.

Table of Contents

Chapter 1: Introduction and Background

Introduction. According to the World Health Organization, pneumonia is one of the leading causes of infant mortality and morbidity in young children less than five years old, accounting for about two million deaths globally each year (Lu et al., 2013). One important environmental risk factor attributable to these deaths at the global level is indoor particulate matter from biomass burning stoves used for cooking and heating in developing countries. Recent studies have suggested that traffic related ambient air pollution exposures are also an important environmental risk factor for early childhood pneumonia (Fuertes et al., 2014; Jedrychowski et al., 2013; Lu et al., 2013; MacIntyre et al., 2013; Penna et al., 1990; Rice et al., 2014; Vieira et al., 2012). This is a public health concern because the economic burden of treating pneumatic children amounts to approximately $109 million annually worldwide (WHO, 2016). The objective of this thesis is to examine the association between exposures to primary particulate matter from traffic emissions during the first year of life on the rate of first pneumonia event by age two in a cohort of children in Atlanta, Georgia.

Case definition. Pneumonia is an inflammatory disease of the lower respiratory tract. The inflammation is a result of accumulation of mucus and other miscellaneous environmental inflammatory products that impair gas exchange in the alveoli (WHO, 2016). As a result, oxygen intake is limited and breathing becomes difficult for the affected patient. Pneumonia is caused primarily by viruses and bacteria, but can also be attributed to fungi or parasites in some cases. The most common organisms known to cause pneumonia in children are the bacteria *streptococcus pneumoniae*, *repiratory syncytial virus* (RSV) and *haemophilus influenzae type B* (Hib) (WHO, 2016). Common

symptoms of childhood pneumonia include cough with sputum, fever, chest pain, difficulty breathing, muscular fatigue, chills, headaches, and nausea. The best available test for pneumonia diagnosis is a chest X- ray but other types of diagnostic screening tests are viable including blood test and sputum tests. For this study, pneumonia was defined as one clinical diagnosis of pneumonia based on ICD-9 codes 480-486 excluding neonatal pneumonia cases that occurred during the first 28 days of life.

Particulate Matter Exposure. Particulate matter is a complex mixture of liquid and solid droplets suspended in air (Kinney et al., 2011), comprised of particles of various size and chemical composition. Fine particulate matter, defined as particulates less than or equal to 2.5 micrometers in diameter ($PM_{2.5}$), is the main exposure of interest in this study. The US Environmental Protection Agency (EPA) regulates $PM_{2.5,}$ along with five other criteria air pollutants in accordance with an amendment to the Clean Air Act called the National Ambient Air Quality Standards (NAAQS) (EPA, 2015). The NAAQS established short term and long term standards for six total criteria air pollutants to target threshold levels for allowable emissions to better protect the health of the public. Particulate matter is regulated according to these standards based on size, not chemical composition. Smaller sizes of PM, such as $PM_{2.5}$ compared to PM_{10}, have been shown to be more harmful than larger sized particles because the smaller particles can evade the body's natural defense system and lodge deeper into the lungs and even circulate into the blood (Oberdorster 2001; Knibbs et al., 2011; Nelin et al., 2012; Schwartz et al., 2000). However, the chemical composition of PM is also important because some toxicants that adhere to PM, such as sulfates from diesel emissions, be more hazardous than other particulates and can induce more damage upon inhalation. Particles from incomplete

combustion of diesel engines and vehicular traffic sources are generally composed of soot, acid condensates, and sulfate particles and nitrate particles (Dockery, 1993). Despite regulation, health effects of exposure to PM have been documented in numerous studies (Ciccone et al., 1998; Dockery, 1993; Esposito et al., 2014; Estarlich et al., 2011; Fuertes et al., 2014; Gehring et al., 2013; Goldizen et al., 2015; Jedrychowski et al., 2013; Kinney et al., 2011; Knibbs et al., 2011; Laumbach & Kipen, 2012; Lu et al., 2013; MacIntyre et al., 2013; Morales et al., 2014; Morgenstern et al., 2008; Oberdorster, 2001; Pope et al., 2000; Pope et al., 2006; Raaschou-Nielsen et al., 2013; Rice et al., 2014; Schultz et al., 2012; Schwartz et al., 2000; Shaheen et al., 1994; Spira-Cohen et al., 2011).

Particulate matter and children's health. Children are especially vulnerable to lower respiratory infections because infant and child lungs are not fully developed. A study looking at critical windows of exposure in the mammalian respiratory system observed that childhood lung development is a complex process that cannot be based on studies of adults (Pinkerton & Joad, 2000). About 80% of lung alveoli in the adult lung are not fully developed at birth and they continue to branch and differentiate into adolescence to meet the needs of the increasing body mass from birth into adulthood. The three main anatomically distinct stages of prenatal lung development are termed pseudoglandular, canalicular, and saccular. Exposure to a toxicant during these critical and vulnerable stages of growth could affect cellular differentiation patterns, resulting in damage to immuno-respiratory development (Pinkerton & Joad, 2000). But the lungs continue to develop postnatally as well.

Postnatal exposures in early childhood can impact formation of new alveoli and differentiation of new cells, a crucial biological stage to meet the metabolic needs of a growing body. Toxicants such as environmental tobacco smoke, bio-activated compounds and oxidant gases have been shown to target epithelial cells undergoing maturation and or rapid proliferation, leading to greater susceptibility among exposed children (Pinkerton & Joad, 2000). In fact, an animal study looking at critical windows of development exposed rats to tobacco smoke both in utero and during early postnatal periods of development. The study found that rats exposed during both prenatal and postnatal development had increased markers for pulmonary disruption and oxidative stress compared to rats that were exposed during only prenatal or only postnatal periods but not both (Pinkerton & Joad, 2000). Children who spend more time outdoors or who live in direct proximity to sources of urban air pollution may be more vulnerable, especially if they were also exposed prenatally as well. Also, children accrue more exposures due to their dose to body weight ratio and increased inhalation rate (Esposito et al., 2014). A critical exposure during childhood could adversely affect the growth and function of the respiratory system, the effects of which could persist into adult life (Shaheen et al., 2012).

Literature Review. This research process began with a literature review of epidemiologic studies on air pollution exposures and respiratory health outcomes using Pub Med and Google Scholar. In this search I utilized keywords such as “ambient,”, “traffic,” “air pollution,” “childhood,” and “pneumonia”. Twelve articles met the search criteria for this analysis which required the exposure to be traffic related air pollutants

such as PM and the outcome to be childhood pneumonia. The other articles were examined for contribution to the background information on general health outcomes and biological pathways related to air pollution exposures. In all, 69 article abstracts were reviewed on topics related to childhood pneumonia and other health effects of PM and other air pollutants. See Appendix I for a full list of the references consulted or cited in this research study. This present analysis will focus on long-term exposures and the health effects related to more chronic traffic related exposures in children.

Based on the literature reviewed, several studies were selected for further review based on relevance to research on the effects of traffic related primary PM emissions on childhood pneumonia incidence. Although the literature is not extensive, several prospective studies have estimated childhood pneumonia outcomes in relation to chronic ambient traffic exposures (Esposito et al., 2014; Fuertes, et al., 2014; Jedrychowski et al., 2013; MacIntyre et al., 2013; Rice et al., 2014; Ryan et al., 2009). Among these, the studies by Fuertes, Jedrychowski, Rice and MacIntyre found modest, positive associations among cohort members, while other studies of air pollution and respiratory events in early childhood had mixed results. Studies by Ciccone (1998) and Jedrychowski (2013) found statistically significant elevated risk of pneumonia in children. In a cross sectional survey of 40,000 subjects Ciccione found the association between exposures to exhausts from heavy vehicle traffic in metropolitan areas on pneumonia in the first two years of life to be estimated at 1.84 (1.27, 2.65). In a cohort study of 214 children from birth to seven years, Jedrychowski found the association between PM and pneumonia and/or bronchitis combined to be elevated with an OR of 2.44 (1.12, 5.36).

On the other hand, several studies have found only moderately positive associations between ambient traffic related exposures and early childhood pneumonia outcomes (Esposito et al., 2014; Fuertes et al., 2014; Lu et al., 2013; MacIntyre et al., 2013). A prospective study of 777 children aged 2 to 18 years followed over 12 months found that an increase of 10 $\mu g/m^3$ of PM_{10} and NO_2 increased the onset of pneumonia only in children with a preexisting respiratory condition such as wheeze or asthma (continuous RR=1.08, 95%CI: 1.00,1.17). In a cross sectional study of 2,727 kindergarten children whose parents completed a health history questionnaire, the association between PM_{10} from traffic emissions and childhood pneumonia was elevated but not statistically significant (OR: 1.19, 95% CI: (0.84, 1.69)). In a meta-analysis of seven European birth cohorts including information on 15,980 children, the study conducted by Fuertes found more varied effect estimates in a random effects model (OR (95%CI): Cu=1.42(0.91, 2.22); Fe=1.40(1.05,1.87); K=2.03(0.91,4.52); Ni=1.11 (0.93, 1.32); S=1.85 (0.76,4.50); Si=1.88 (0.84,4.22); V=2.13(0.82, 5.54); Zn=1.37 (1.02, 1.85). This is important in the context of this study because the chemical composition of the particulate can vary by source and diesel exhaust particulates are more harmful than particles from other sources. The variation in associations may be due to factors such as geographic variation and differences in ambient pollution sources, study design, time period and ages of the study population, and the size of particulate matter measured. For a more detailed explanation of the scope of the current literature studies related to this topic, refer to the full literature review in Appendix I.

Chapter 2: Methods

KAPPA cohort data. Health data for this analysis came from the Kaiser Air Pollution and Pediatric Asthma Study (KAPPA), a historical birth cohort of children insured by Kaiser Permanente Georgia (KPGA) Health Maintenance Organization who were born between 2000 and 2010. There were 22,520 children included in the analysis, after exclusions were made from the total birth cohort of 24,608. Exclusions were made to eliminate cohort members for the following reasons: children who enrolled after day 29 of life (n=480); children whose enrollment ends before day 29 of life (n=9); children who were diagnosed with pneumonia in the first 28 days of life (n=131); children with no residential data in the first year of life (n=726); and children with at least one residence during the exposure period outside of the region for which we have RLINE pollution data (n=742).

Average ambient $PM_{2.5}$ from traffic at each child's residence was estimated for the first year of life, until the first pneumonia diagnosis event, or until censorship. If a child moved between birth and first year of life, exposure data from all residences during that year were incorporated into an average exposure estimate for that child instead using only birth residence as a proxy for total exposures. Approximately 10 percent, or 2,188 children, were diagnosed with pneumonia for the first time by their second birthday (excluding neonatal diagnoses of pneumonia). The outcome was defined as the child's first clinical diagnosis of pneumonia based on ICD-9 codes 480-486.

Exposure data. Exposure data for this analysis came from the SCAPE research project, a collaboration between Emory University and Georgia Institute of Technology. Primary $PM_{2.5}$ exposures were estimated using a research line source dispersion model for near

surface releases (RLINE). The RLINE model was created by the US EPA as a way to assess the impact of traffic emissions on people living near roadways (Zhai et al., 2015 & Community Modeling and Analysis System, 2015). The model incorporates traffic data from the Atlanta Regional Commission to represent emissions from each section of the roadway as well as surface meteorology data, such as wind patterns, and creates a smooth exposure surface to better understand transport and dispersion of pollutants (Community Modeling and Analysis System, 2015). Averages from the year 2011 were used to model primary particulate exposures at 250 meter grid resolution (1 estimate per grid for 2011). An annual exposure concentration from 2011 data was assigned to each child based on geocoded residential location(s) in the first year of life, incorporating exposure information from children who moved during the first year. For children whose first pneumonia diagnosis is after the first birthday and children who are censored after the first birthday, the calculated average was based on exposures between birth and the day before the first birthday. For children whose first pneumonia diagnosis is before the first birthday and children who are censored before the first birthday, the calculated average was based on average exposures until the day before diagnosis or censoring.

Description of covariates. The dataset for this analysis was created in November 2015 by Dr. Audrey Flak, and was originally sourced from RLINE data from GA Tech and health data from KAPPA Kaiser Permanente Georgia HMO. Covariates used in the adjusted analysis include neighborhood socioeconomic status (SES), child race, child sex, maternal asthma status, maternal education, maternal prenatal smoking status, city region and maternal age. Neighborhood SES was measured by major and minor demographic

cluster variables created from EASI Demographics data from 2010 census blocks and contained 25 variables related to age, income, family structure, housing value and type, education attainment and employment type (Demographic Clusters of Georgia, 2012 & Zhou, 2012). Neighborhood SES demographics were assigned to each child based on residence at birth only. Group A is considered to be of the highest SES while Group D is considered to be of the lowest SES. The modeling in the Cox regression analysis used minor cluster created as subgroups from the four major neighborhood SES groups. The sub-groupings allowed for greater control of neighborhood SES.

Child race has four categories: white (reference), black, other and unknown. Maternal asthma status has three categories: no (reference), yes, and missing. Highest level of education attained by the mother has four categories: some college (reference), less than 12^{th} grade, high school of GED, or missing. The category that represents whether or not the mother smoked during pregnancy has three categories: no (reference), yes and missing. The variable for city region describes where in the Atlanta area a child's primary residence is located: inside metropolitan Atlanta defined as inside the I-285 highway that surrounds the city, less than or equal to 10 miles from I-285, and more than 10 miles from I-285 (reference). The variables for child ethnicity and maternal smoking status during the child's first year of life were excluded due to a large proportion of children with missing data for those variables.

Statistical modeling. Cox proportional hazards (PH) regression was used to examine the association between first year of life exposures to primary $PM_{2.5}$ from traffic emissions and time to first pneumonia event by age two. The PH assumption was checked in

unadjusted and adjusted models using the following methods: Kaplan Meier log-log curves and adjusted log-log curves; goodness of fit (GOF) tests using Schoenfeld residual p values; extended Cox models which tested each variable's interaction with time, time squared and log of time separately. Variables were first tested separately in unadjusted models. Variables found to not violate the PH assumption based on 2 of 3 of the mentioned PH assumption tests did not need to go in the strata statement. Variables that did not satisfy the PH assumption in the unadjusted models were then tested in adjusted models using the same series of tests and methods described above. The four level variable for neighborhood socioeconomic status (major cluster) was used to test the PH assumption but the minor cluster variable was used in the modeling for more control of the neighborhood SES variable.

After assessing the proportional hazards assumption, collinearity, interaction and confounding were assessed. For collinearity, the cut point was determined to be a condition index of 10 and condition indices with a value below 10 were considered to not be suggestive of collinearity issues, based on the approach proposed by Kleinbaum & Klein (2012). Interaction was assessed using likelihood ratio tests which compare the negative two log likelihood (-2lnL) values for the reduced model (with limited covariates) and full model (with all covariates) and an alpha value of 0.05. Interaction between $PM_{2.5}$ exposure and sex, maternal asthma, prenatal smoking status, and city region was examined based on a priori criteria and to be consistent with studies in the literature. Likelihood ratio tests were carried out to compare the -2lnL value of the reduced with the full models under the null hypothesis that the interaction terms were equal to zero. P values that were not significant at the 0.05 alpha level suggested that

there was not enough evidence to conclude that there is significant interaction between the exposure and covariate (Kleinbaum and Klein, 2012). All p values for the interaction terms of the likelihood ratio tests were not significant at the 0.05 alpha level and thus the interaction terms examined could be dropped from the model.

Confounding was assessed based on the all-possible-subsets approach (Kleinbaum & Klein, 2012). This approach determines the Hazard Ratio (HR) and confidence interval (CI) for the "Gold Standard" (GS) model, the full model with all covariates being assessed. Then the HR and CI of each possible subset of covariates (the reduced models) are compared to the HR of the gold standard model in separated adjusted analyses. Subsets of covariates whose HR does not differ from the gold standard HR by more than 10% after dropping a certain subset of covariates are considered to adequately control for confounding and can be dropped from the model, especially if the subset improves precision by making the CI ratio or width more narrow. Results of the confounding assessment are shown in the results section (Table 3). We will also be able to account for sibling clustering in the analysis by using the robust sandwich estimator implemented using the "covs(aggregate)" statement in Proc PHREG. Data was analyzed using SAS 9.4 (Cary, NC).

Chapter 3: Results

Descriptive statistics. The RLINE modeled mean $PM_{2.5}$ exposure was 1.17 $\mu g/m^3$ with a standard deviation of 0.27$\mu g/m^3$, a minimum exposure concentration of 0.31$\mu g/m^3$ and a maximum of 2.66 $\mu g/m^3$. About half the children in the KAPPA cohort were males (Table 1). The majority of the children insured by Kaiser were born into a neighborhood classified as having the highest SES (Group A Neighborhood SES). The most common child race was white, accounting for 39.5% of the children in the cohort, followed by children of black race, accounting for 34.9% of the cohort. The majority of mothers did not have asthma (79%) and did not smoke while pregnant (78.6%). About sixty percent of the mothers had some college education, but almost 30 percent of data on maternal education is missing. Because approximately one third of the data for this variable is missing, we might expect some of the mothers with missing data to have attained some higher education, making 60% the lower bound of the true proportion of mother's who attended some college. Most of the children (46.7%) lived in residences greater than 10 miles outside the metropolitan Atlanta at birth, while only about 10.5 percent of the population lived in residences inside metropolitan Atlanta at birth.

Variation in frequency of pneumonia diagnoses by year was considered to examine trends in annual diagnosis (Figure 1). There were 2,188 pneumonia diagnoses, accounting for about 10% of the total children followed during the study period. For this study only the first pneumonia diagnosis was counted for each child; there would have been a higher frequency of total diagnoses during the follow up period if all pneumonia diagnoses were counted for each child. The year with the highest frequency of first pneumonia diagnoses was 2009 (n=237), followed by 2002 (n=226). The year with the

lowest frequency of first pneumonia diagnoses was 2012 (n=22) followed by 2000 (n=50). Less children were included in the first and last years of follow up, resulting in fewer diagnoses in these years. Pneumonia diagnosis also varied by season, with the highest proportion of diagnoses (38%) occurring in the winter (Figure 2). Only about 10% of children were diagnosed with pneumonia during the summer.

Model Selection. The selected model for this analysis was a no interaction, un-stratified Cox regression model that controlled for child sex, child race, maternal asthma, prenatal smoking, maternal age, maternal education, city region, and neighborhood socioeconomic status. This model was selected based on the results of PH tests, interaction likelihood ratio tests, confounding all possible subset assessment and collinearity assessment. There were no interaction terms with exposure that were significant at the alpha 0.05 level and thus there was no evidence of significant interaction, allowing us to drop the interaction terms from the model. Using the three assessment methods to test the PH assumption, it was confirmed that no variables violated the assumption; therefore, the final model was an un-stratified model. The results of the all-possible subsets test for evaluation of confounding found that city region, child race and neighborhood socioeconomic status had a meaningful impact on the exposure disease relationship by changing the HR effect estimate by more than 10% from the gold standard (Table 3). Thus it was important not to drop these variables from the final model

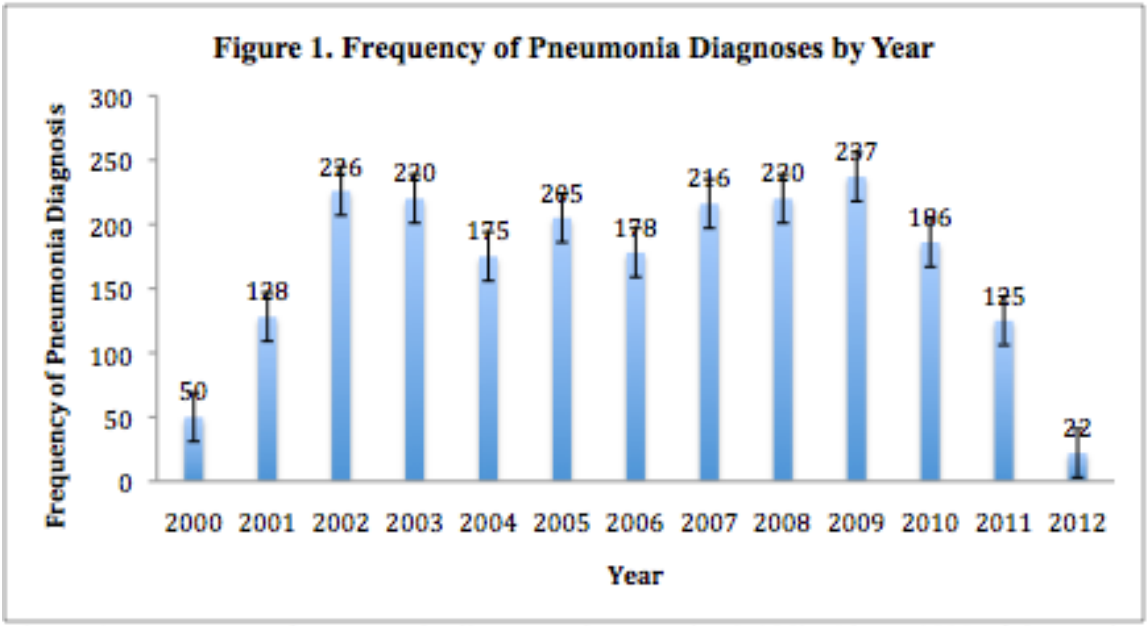

Figure 1. Frequency of pneumonia diagnoses by year (n=2,188). About 10% of the total children in the birth cohort were diagnosed with their first pneumonia event during the two year follow up. The year with the highest frequency of first pneumonia diagnoses was 2009 and the year with the least frequency of first diagnoses was 2012.

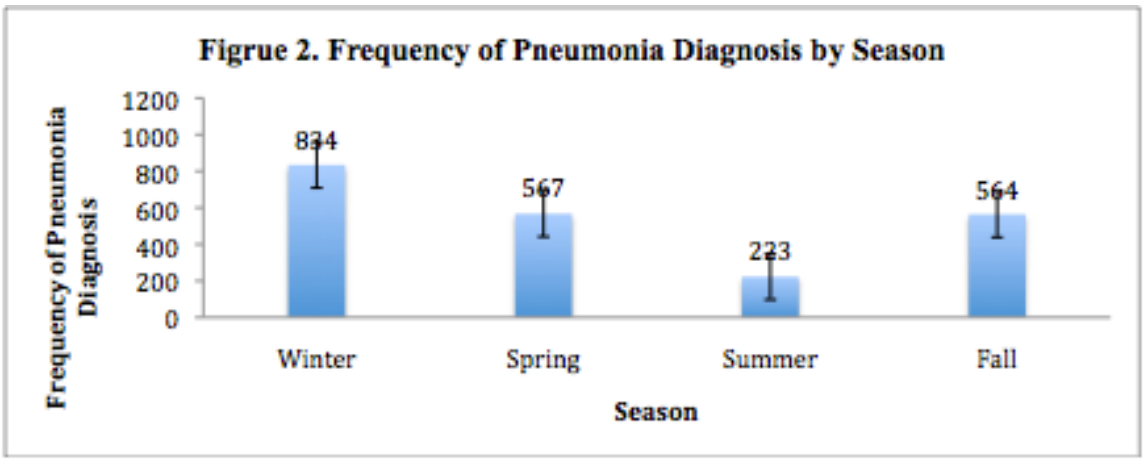

Figure 2. Pneumonia diagnosis frequencies by season. Winter is defined as December (n=392), January (n=221), February (n=221). Spring is defined as March (n=225), April (n=192), May (n=150). Summer is defined as June (n=69), July (n=61), August (n=93). Fall is defined as September (n=129), October (n=151), November (n=284). Summary statistics for each month were calculated by adding the total pneumonia diagnosis frequencies from each month over the course of the follow up period. According to the summary, 38% of the pneumonia diagnoses occur during the winter.

because they were found to be important confounders of the measured exposure disease association. Based on a priori criteria in the literature and to better ensure exchangeability among covariate groups, it was decided to keep the other potential confounders in the model even though dropping them individually did not change the hazard ratio more than 10% from the gold standard estimate. Advantages and disadvantages to using the hazard ratio estimate of effect for this analysis will be discussed in later sections.

Modeling Results. This study followed a cohort of 22,520 children from the KAPPA study until the first pneumonia event, or until censoring. Children were censored either because they were not diagnosed with pneumonia by age two or due to HMO enrollment attrition. Figure 6 shows the Kaplan Meier curve that describes time until first pneumonia diagnosis in children over the total follow up period. The independent variable represents the age of the child in terms of days and the dependent variable is probability of a pneumonia diagnosis. Of the 22,520 cohort members, there were 2,188 children diagnosed with pneumonia for the first time by age two and 20,332 were censored.

The estimated hazard ratio for the association between $PM_{2.5}$ exposure and pneumonia incidence by age two was positive and modest, although non-significant (HR (95% CI): 1.17 (0.93, 1.47) for a 1 $\mu g/m^3$ change in $PM_{2.5}$ (Table 2)). The selected model was a no interaction, un-stratified Cox regression survival model controlling for child sex, child race, maternal asthma status, prenatal smoking, maternal age (dichotomized at the mean), maternal education, city region, and neighborhood socioeconomic status. By contrast, the unadjusted effect estimate for $PM_{2.5}$ exposure on the outcome was close to

the null value of one and the confidence interval included the null (HR (95% CI): 0.98 (0.84, 1.15).

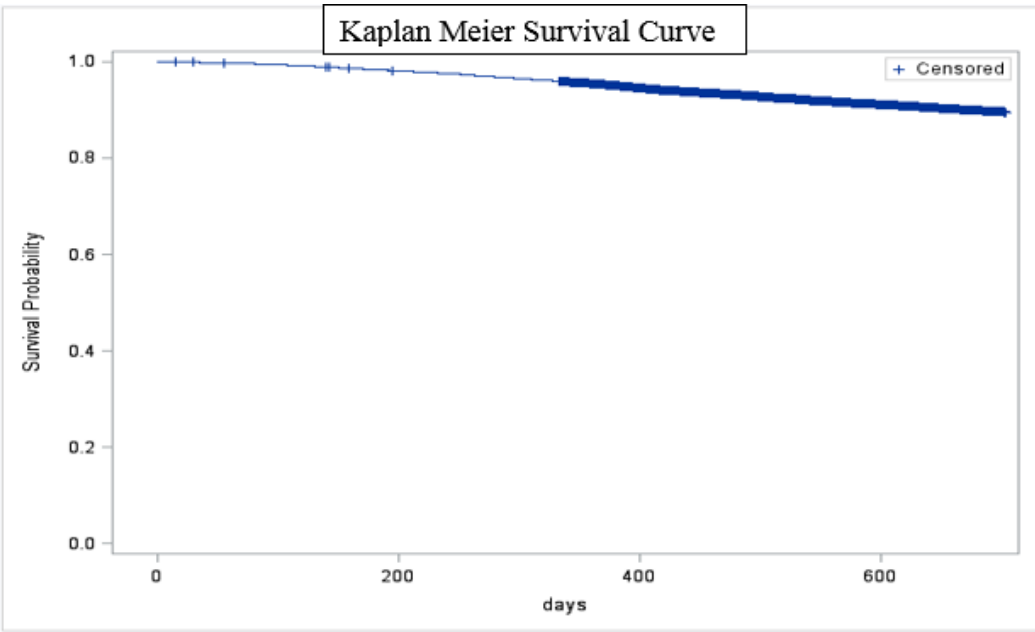

Figure 3. Kaplan-Meier survival curve of survival probability for pneumonia diagnosis in children over the total follow up period. There were a total of 22,520 children in the cohort, of which 2,188 (10%) were diagnosed with pneumonia and 20,332 were censored.

Results suggest that low maternal education was the strongest risk factor for childhood pneumonia (adjusted model HR (95%CI): 1.57 (1.11, 2.22) comparing a child whose mother did not finish high school to a child whose mother attended at least some college (Table 2)**.** The sex of the child had a significant protective effect on incidence of pneumonia diagnosis in females compared to males (adjusted model HR (95% CI): 0.83 (0.76, 0.90) (Table 2). When stratified by sex, there was no meaningful difference in the incidence of pneumonia (adjusted model HR (95% CI): 1.16 (0.82, 1.64) for females compared to 1.15 (0.85, 1.55) in males). The majority of neighborhood SES clusters had

higher rates of pneumonia compared to the neighborhood of the highest SES, but the HRs did not increase proportionately with decreasing neighborhood SES status.

The rate of first pneumonia diagnosis was 1.15 times higher for black children compared to white children and this result was statistically significant in the adjusted model (95% CI: 1.02, 1.28). Children of mothers who have asthma are significantly more likely to have a pneumonia diagnosis compared to children whose mothers do not have asthma (HR: 1.20, 95% CI: 1.06, 1.37). A mother who indicated that she smoked during or prior to pregnancy was 1.18 times more likely to have a child who had a pneumonia event compared to children whose mothers did not smoke during or before pregnancy (95% CI: 0.90, 1.56). Children who live in metropolitan Atlanta are 0.93 times less at risk of developing pneumonia compared to children who live greater than 10 miles outside the perimeter. This estimate of association is close to the null and the confidence interval is non-significant (95% CI: 0.75, 1.15). Similarly children who live less than or equal to 10 miles from metropolitan Atlanta are 0.94 times less at risk of developing pneumonia compared to children who live greater than 10 miles outside the perimeter. This estimate of association is close to the null and the confidence interval is non significant (95% CI: 0.83, 1.06). Thus city region may be an important confounder, but it does not have a strong effect on predicting the pneumonia outcome.

Table 1. Kaiser HMO birth cohort summary statistics for children born between 2000-2010 in the Atlanta, GA greater metropolitan area.

Characteristic	N	Mean (Std. Dev)
RLINE modeled PM2.5 Exposure[1]	22,520	1.17 (0.27)
Maternal age (years)	20,258	31.64 (4.96)

Category	Frequency	Percent
Child Sex		
Reference=Males	11,452	50.9
Females	11,068	49.2
Major Demographic Cluster [2]		
Reference=Group A	14,065	62.5
Group B	2,231	9.9
Group C	1,093	4.9
Group D	5,128	22.8
Child's Race		
Reference=White	8,901	39.5
Black	7,853	34.9
Other	2,728	12.1
Unknown	3,038	13.5
Maternal Asthma Status		
Reference=No	17,797	79.0
Yes	2,461	10.9
Missing	2,262	10.0
Highest Maternal Education		
Reference= Some college	13,312	59.1
Less than 12th grade	280	1.2
High School or GED	2,575	11.4
Missing	6,353	28.2
Prenatal Smoking Status		
Reference=No	17,693	78.6
Yes	455	2.0
Missing	4,372	19.4
City Region [3]		
Reference: >10 mi OTP	10,514	46.7
Metro Atlanta (inside I-285)	2,353	10.5
Less than or equal 10 mi OTP	9,653	42.9

[1] Particulate matter 2.5 ($PM_{2.5}$) is measured in $\mu g/m^3$

[2] Demographic clusters were created from EASI Demographics data (2011) available at the census block group level of the 2010 census geographies. The information for this variable incorporates 25 variables related to age, income, family structure, housing value and type, education attainment and employment type. The highest SES level is Group A and the lowest SES level is Group D. Minor demographic clusters are subgroups created from the major demographic clusters and allow for greater control of SES in modeling.

[3] Defined based on distance from the metro Atlanta perimeter. OTP means Outside the Perimeter as defined by interstate 285

Table 2. Results of un-stratified, no interaction, adjusted Cox regression survival analysis. A modest positive association was estimated between $PM_{2.5}$ exposure and childhood pneumonia incidence by age 2 for a change of 1 microgram per cubic meter. Low maternal education was the strongest risk factor while neighborhood SES, child race and city region were other notable risk factors.

Parameter	Estimate (Std. Error)	Hazard Ratio	P value	Confidence Interval
RLINE PM2.5 Exposure [1]	0.16 (0.12)	1.17	0.18	(0.93,1.47)
Maternal Age (years) Dichotomized at mean of 32	-0.05 (0.05)	0.95	0.27	(0.87, 1.04)
Child Sex	Reference=Males			
Females	-0.19 (0.04)	0.83	<0.0001	(0.76, 0.90)
SES Demographic Cluster [2]	Reference=A.1			
A.2	0.08 (0.09)	1.08	0.41	(0.90, 1.30)
A.3	0.08 (0.08)	1.08	0.31	(0.93, 1.26)
B.1	0.02 (0.14)	1.02	0.90	(0.78, 1.33)
B.3	0.18 (0.12)	1.20	0.13	(0.95, 1.51)
B.4	0.56 (0.34)	1.75	0.10	(0.90, 3.39)
C.1	0.01 (0.35)	1.01	0.97	(0.51, 2.00)
C.2	0.39 (0.13)	1.47	0.00	(1.14, 1.92)
C.3	0.20 (0.21)	1.23	0.34	(0.81, 1.87)
C.4	0.24 (0.41)	1.27	0.56	(0.57, 2.86)
D.1	0.05 (0.10)	1.05	0.64	(0.87, 1.27)
D.3	-0.10 (0.16)	0.90	0.51	(0.66, 1.23)
D.4	0.00 (0.12)	1.00	0.99	(0.80, 1.26)
D.5	-0.11 (0.18)	0.89	0.53	(0.63, 1.27)
D.6	-0.46 (0.31)	0.63	0.14	(0.35, 1.15)
D.7	0.36 (0.35)	1.44	0.31	(0.72, 2.87)
Child Race	Reference=White			
Black	0.13 (0.06)	1.15	0.02	(1.02, 1.28)
Other	-0.14 (0.08)	0.87	0.07	(0.74, 1.01)
Unknown	0.01 (0.09)	1.01	0.88	(0.84, 1.22)
Maternal Asthma Status	Reference=No			
Missing	0.03 (0.10)	1.03	0.76	(0.85, 1.26)
Yes	0.18, (0.07)	1.20	0.01	(1.06, 1.37)
Highest Maternal Education	Reference= Some college			
Less than 12th grade	0.45 (0.18)	1.57	0.01	(1.11, 2.22)
High School of GED	0.00 (0.07)	1.00	0.99	(0.87, 1.15)
Missing	0.15 (0.07)	1.16	0.04	(1.01, 1.34)
Prenatal Smoking Status	Reference= No			
Missing	-0.18 (0.10)	0.84	0.07	(0.69, 1.01)
Yes	0.17 (0.14)	1.18	0.24	(0.90, 1.56)
City Region [3]	Reference: >10 mi OTP			
Metro Atlanta (inside I-285)	-0.07 (0.11)	0.93	0.50	(0.75, 1.15)
Less than or equal 10 mi OTP	-0.06 (0.06)	0.94	0.29	(0.83, 1.06)

[1] Particulate matter 2.5 ($PM_{2.5}$) is measured in $\mu g/m^3$

[2] Demographic clusters were created from EASI Demographics data (2011) available at the census block group level of the 2010 census geographies. The information for this variable incorporates 25 variables related to age, income, family structure, housing value and type, education attainment and employment type. The highest SES level is Group A and the lowest SES level is Group D. Minor demographic clusters are subgroups created from the major demographic clusters and allow for greater control of SES in modeling.

[3] Defined based on distance from the metro Atlanta perimeter. OTP means Outside the Perimeter as defined by interstate 285

Table 3. Adjusted Hazards Ratio Estimates from the all possible subsets procedure [1] to evaluate confounding in the no interaction, un-stratified model. This is a strategy used in model development to assess the covariates as potential confounders.

Model	Hazard Ratio	Confidence Interval	CI Width	CI Ratio
Gold Standard	1.17	(0.93, 1.47)	0.54	1.58
Unadjusted	0.98	(0.84, 1.15)	0.31	1.37
Drop Neighborhood SES	1.08	(0.88, 1.33)	0.45	1.51
Drop Child Sex	1.17	(0.93, 1.47)	0.54	1.58
Drop Child Race	1.10	(0.88, 1.38)	0.50	1.57
Drop Maternal Asthma Status	1.18	(0.94, 1.48)	0.54	1.57
Drop Maternal Education	1.18	(0.93, 1.48)	0.55	1.59
Drop Prenatal Smoking Status	1.17	(0.93, 1.47)	0.54	1.58
Drop City Region	1.11	(0.90, 1.36)	0.46	1.51
Drop Maternal Age	1.17	(0.93, 1.47)	0.54	1.58
Drop Child Sex and Maternal Asthma Status	1.18	(0.94, 1.48)	0.54	1.57
Drop Child Sex and Maternal Education	1.18	(0.94, 1.48)	0.54	1.57
Drop Child Sex and Prenatal Smoking Status	1.17	(0.94, 1.48)	0.54	1.57
Drop Child Sex and Maternal Age	1.17	(0.93, 1.47)	0.54	1.58
Drop Maternal Asthma Status and Education	1.18	(0.94, 1.49)	0.55	1.59
Drop Maternal Asthma Status and Prenatal Smoking Status	1.18	(0.94, 1.48)	0.54	1.57
Drop Maternal Asthma Status and Maternal Age	1.18	(0.94, 1.48)	0.54	1.57
Drop Maternal Education and Prenatal Smoking Status	1.18	(0.94, 1.48)	0.54	1.57
Drop Maternal Education and Maternal Age	1.18	(0.93, 1.48)	0.55	1.59
Drop Prenatal Smoking Status and Maternal Age	1.17	(0.93, 1.47)	0.54	1.58

[1] All possible subsets approach modified from Kleinbaum and Klein (2012)

Chapter 4: Discussion and Conclusion

These preliminary results suggest evidence of a positive yet modest association between primary traffic-related $PM_{2.5}$ exposures during the first year of life and the time to first pneumonia event by age two in the cohort of children enrolled in the Atlanta Kaiser Air Pollution and Pediatric Asthma (KAPPA) study within the Kaiser Permanente Georgia HMO. The advantage of this cohort is that it is robust (n=22,520) and well defined. The measure of association for this analysis was an adjusted hazard ratio, using a no-interaction, un-stratified final model. In the fully adjusted model, an increase of 1 $\mu g/m^3$ of exposure to primary $PM_{2.5}$ from traffic in the first year of life was associated with a 1.17 times multiplicative increase in the hazard.

Survival analysis instead of a logistic regression was chosen for several reasons. First, a hazard ratio allows us to determine the time that each child is at risk of a pneumonia outcome. However, when thinking about the hazard ratio as an estimate of effect, it is important to consider that the measured effect depends on the exact time of pneumonia diagnosis in order to measure a constant rate in the cohort, which may or may not be valid in real applications. A second advantage of a survival analysis is that a hazard ratio as a measure of effect can increase statistical power by including children who were lost to follow up in the analysis as censored observations but who would have otherwise been excluded in a logistic regression analysis.

Several covariates were considered in this analysis and several were found to be important confounders of the exposure disease relationship. Including neighborhood SES in the model made an important impact on the observed association; the hazard ratio changed more than 10% from the Gold standard model when this variable was dropped in

the confounding assessment indicating it was an important confounder that needed to stay in the model (Table 3). Similar conclusions were made for city region and child race based on this procedure. Including city region and neighborhood SES in the model was important to account for unmeasured confounders that could vary by location among the children enrolled in the study such as variation in exposure types (traffic related exposures compared to agricultural exposures which vary spatially) and differences in access to health care that might vary based on income level and place of residence. Including the city region and neighborhood SES covariates thus offered better exchangeability between children whose residences were in spatially variable locations of Atlanta. However, even when these variables are included in the fully adjusted model, confidence limits still include the null indicating that while they may be important confounders, they do not have a significant effect on the measure of association between exposure and disease.

We have done the best we can to control for important confounders, based on what other studies had controlled for in the literature. Several covariates examined in other related literature studies, but not in this study, include paternal asthma status, preexisting constitutional respiratory conditions of the child (wheeze, bronchitis, other genetic complications), whether or not the child was breastfed, whether or not there were pets in the home. No information on these covariates was collected in the information provided by the KAPPA cohort. Residual confounding from unmeasured factors may result in biased effect estimates. However, we do not expect this to meaningfully impact our objective of estimating the association between traffic related air pollution exposures and childhood pneumonia outcomes.

A limitation of this analysis might appear to be that exposure averages for each child were modeled based on RLINE data from 2011 exposure averages. In other words, a child who was born in 2001 has an annual average exposure assigned from ten years after their birth. However, in this analysis we assume that the residential level exposure data from traffic pollution remains constant over time and that there are no new major roads put in during the follow up period that would drastically change temporal exposures. Thus an exposure estimate from 2011 would be considered appropriate for children born in all years of this study's follow up period. Misclassification of personal exposures could occur if children go to day care and do not spend most of their day at home. This would be an issue if the objective of this study were to examine the effects of personal exposures on pneumonia outcomes, but that is not the objective of this study. This study is examining the effect of residential level exposures on pneumonia, modeled using RLINE techniques that we assume to be an accurate classification of exposure concentration for this study. The outcome of interest in this study is pneumonia as defined based on medical diagnosis using ICD-9 codes 480-486. The advantage of defining the outcome based on clinical diagnosis instead of self-report minimizes chance of recall bias.

Selection bias and loss to follow up must be considered as well. From the start, selection into the KAPPA cohort requires children to have health insurance, a feature that not all children in Atlanta maintain. Thus our cohort does not represent a random sample of the general population of all children in Atlanta. Furthermore we must consider whether the children who were lost to follow up are different from the children who were not lost to follow up. If children were differentially lost to follow up before their second

birthday or before first pneumonia event for reasons related to exposure or pneumonia diagnosis, then this would lead to selection bias. Results of this study are generalizable to children whose parents can afford health insurance and who also live in urban areas like Atlanta. However, the results of this analysis may still be important for children across the US and other cities of the developed world because children as a group are expected to be similarly biologically vulnerable to the effects of air pollution during this critical stage of development. Thus, the children enrolled in the study are not expected to be differentially affected by air pollution compared to the children not included in the study.

This study contributes to our knowledge of whether chronic, residential traffic related $PM_{2.5}$ exposure could affect early life childhood pneumonia. The results of previous studies show mixed associations between early life exposures to particulate matter (PM_{10} and $PM_{2.5}$) and respiratory outcomes (pneumonia included) in children and young adults. Our result of a moderate positive, but non significant association between exposure and outcome was consistent with comparable studies in the literature (Esposito et al., 2014; Fuertes, et al., 2014; Jedrychowski et al., 2013; MacIntyre et al., 2013; Rice et al., 2014; Ryan et al., 2009).

Future research should examine the effects of early life exposures to fine and ultrafine particulates because smaller particulates can have more of a human health effect especially for vulnerable populations like urban children. Smaller sized particulates such as $PM_{2.5}$ and ultrafine particulates (not yet regulated by the EPA) can evade the body's natural defense systems and lodge more deeply into the lungs and circulatory system which can lead to detrimental effects especially in developing children. Compared to short term studies related to acute exposures and hospital admittance, research on long

term exposures to smaller sized particulate matter is needed to understand how chronic, low-levels of traffic exposures impact early life lung development. This study makes an important contribution to the literature because of the extended follow up period to examine associations during the first 2 years of life, the use of hazard ratios to determine relative effects of exposure on a chronic outcome, the use of RLINE dispersion modeling techniques to determine effects of exposures on outcomes with greater reliability near the source, and enhanced robustness of the large Kaiser cohort available for analysis.

Chapter 5: References

1. Bateson, T. F., & Schwartz, J. (2007). Children's Response to Air Pollutants. *Journal of Toxicology and Environmental Health, Part A, 71*(3), 238-243.

2. Chauhan, A. J. (2003). Air pollution and infection in respiratory illness. *British Medical Bulletin, 68*(1), 95-112.

3. Ciccone, G., Forastiere, F., Agabiti, N., Biggeri, A., Bisanti, L., Chellini, E., . . . Viegi, G. (1998). Road traffic and adverse respiratory effects in children. SIDRIA Collaborative Group. *Occupational and Environmental Medicine, 55*(11), 771-778.

4. Colley, J., Holland, W., & Corkhill, R. (1974). Influence Of Passive Smoking And Parental Phlegm On Pneumonia And Bronchitis In Early Childhood. *The Lancet, 304*(7888), 1031-1034.

5. Community Modeling and Analysis System (2015). "R-LINE: A Research LINE-source dispersion model for near surface releases." Accessed April 3, 2016 from https://www.cmascenter.org/r-line/

6. Demographic Clusters of Georgia (2012). Georgia Department of Public Health: Office of Health Indicators for Planning (OHIP).

7. Dietert, R. R., Etzel, R. A., Chen, D., Halonen, M., Holladay, S. D., Jarabek, A. M., . . . Zoetis, T. (2000). Workshop to Identify Critical Windows of Exposure for Children's Health: Immune and Respiratory Systems Work Group Summary. *Environ Health Perspect Environmental Health Perspectives, 108*(S3), 483-490.

8. Dockery, D. (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine*, *329*(24), 1753-1759.

9. Eenhuizen, E., Gehring, U., Wijga, A. H., Smit, H. A., Fischer, P. H., Brauer, M., . . . Hoek, G. (2012). Traffic-related air pollution is related to interrupter resistance in 4-year-old children. *Eur Respir J European Respiratory Journal, 41*(6), 1257-1263.

10. National Ambient Air Quality Standards Table. Environmental Protection Agency web site. https://www.epa.gov/criteria-air-pollutants/naaqs-table. Updated March 29, 2016. Accessed February 6, 2016.

11. Esposito, S., Galeone, C., Lelii, M., Longhi, B., Ascolese, B., Senatore, L., . . . Principi, N. (2014). Impact of air pollution on respiratory diseases in children with recurrent wheezing or asthma. *BMC Pulmonary Medicine BMC Pulm Med, 14*(1), 130.

12. Estarlich, M., Ballester, F., Aguilera, I., Fernández-Somoano, A., Lertxundi, A., Llop, S., . . . Iñiguez, C. (2011). Residential Exposure to Outdoor Air Pollution during Pregnancy and Anthropometric Measures at Birth in a Multicenter Cohort in Spain. *Environ Health Perspect Environmental Health Perspectives, 119*(9), 1333-1338.

13. Fuertes, E., Macintyre, E., Agius, R., Beelen, R., Brunekreef, B., Bucci, S., . . . Heinrich, J. (2014). Associations between particulate matter elements and early-life pneumonia in seven birth cohorts: Results from the ESCAPE and TRANSPHORM projects. *International Journal of Hygiene and Environmental Health, 217*(8), 819-829.

14. Gauderman, W. J., Mcconnell, R., Gilliland, F., London, S., Thomas, D., Avol, E., . . . Peters, J. (2000). Association between Air Pollution and Lung Function Growth in Southern California Children. *Am J Respir Crit Care Med American Journal of Respiratory and Critical Care Medicine, 162*(4), 1383-1390.

15. Gauderman, W. J., Avol, E., Gilliland, F., Vora, H., Thomas, D., Berhane, K., . . . Peters, J. (2004). The Effect of Air Pollution on Lung Development from 10 to 18 Years of Age. *New England Journal of Medicine N Engl J Med, 351*(11), 1057-1067.

16. Gauderman, W. J., Urman, R., Avol, E., Berhane, K., Mcconnell, R., Rappaport, E., . . . Gilliland, F. (2015). Association of Improved Air Quality with Lung Development in Children. *New England Journal of Medicine N Engl J Med, 372*(10), 905-913.

17. Gehring, U., Gruzieva, O., Agius, R. M., Beelen, R., Custovic, A., Cyrys, J., . . . Brunekreef, B. (2013). Air Pollution Exposure and Lung Function in Children: The ESCAPE Project. *Environ Health Perspect Environmental Health Perspectives*.

18. Goldizen, F. C., Sly, P. D., & Knibbs, L. D. (2015). Respiratory effects of air pollution on children. *Pediatric Pulmonology Pediatr Pulmonol., 51*(1), 94-108.

19. Götschi, T., Heinrich, J., Sunyer, J., & Künzli, N. (2008). Long-Term Effects of Ambient Air Pollution on Lung Function. *Epidemiology, 19*(5), 690-701.

20. Hamra, G. B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J. M., . . . Loomis, D. (2014). Outdoor Particulate Matter Exposure and Lung Cancer: A Systematic Review and Meta-Analysis. *Environ Health Perspect Environmental Health Perspectives*.

21. Jakkula, M., Cras, T., Gebb, S., Hirth, P., Tuder, R., Voelkel, N., Abman, S. (2000). Inhibition of angiogenesis decreases alveolarization in the developing rat lung. *Am J Physiol Lung Cell Mol Physiol.* 279:600-607.

22. Jedrychowski, W., Galas, A., Pac, A., Flak, E., Camman, D., Rauh, V., & Perera, F. (2005). Prenatal Ambient Air Exposure to Polycyclic Aromatic Hydrocarbons and the Occurrence of Respiratory Symptoms over the First Year of Life. *Eur J Epidemiol European Journal of Epidemiology, 20*(9), 775-782.

23. Jedrychowski, W. A., Perera, F. P., Spengler, J. D., Mroz, E., Stigter, L., Flak, E., . . . Jacek, R. (2013). Intrauterine exposure to fine particulate matter as a risk factor for increased susceptibility to acute broncho-pulmonary infections in early childhood. *International Journal of Hygiene and Environmental Health, 216*(4), 395-401.

24. Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., . . . Giovis, C. (2004). A review and evaluation of intraurban air pollution exposure models.

J Expo Anal Environ Epidemiol Journal of Exposure Analysis and Environmental Epidemiology, 15(2), 185-204.

25. Johnston, I. D., Strachan, D. P., & Anderson, H. R. (1998). Effect of Pneumonia and Whooping Cough in Childhood on Adult Lung Function. *New England Journal of Medicine N Engl J Med, 338*(9), 581-587.

26. Kajekar, R. (2007). Environmental factors and developmental outcomes in the lung. *Pharmacology & Therapeutics, 114*(2), 129-145.

27. Kinney, P., Gicjuru, G., Volavka-Close, N., Ngo, N., et al. (2011). Traffic impacts on PM2.5 air quality in Nairobi, Kenya. *Environmental Science and Policy, 14,* 369-378.

28. Kleinbaum, D., and Klein, M. (2012). Survival Analysis: A self learning text on statistics for biology and health. Third edition. Springer.

29. Knibbs, L., Hunter, T., Morawska, L. (2011). A review of commuter exposure to ultrafine particles and its health effects. *Atmospheric Environment 45,* 3224-3227.

30. Korytina, G. F., Yanbaeva, D. G., Babenkova, L. I., Etkina, E. I., & Victorova, T. V. (2005). Genetic polymorphisms in the cytochromes P-450 (1A1, 2E1), microsomal epoxide hydrolase and glutathione S-transferase M1, T1, and P1 genes, and their relationship with chronic bronchitis and relapsing pneumonia in children. *Journal of Molecular Medicine J Mol Med, 83*(9), 700-710.

31. Laumbach, R. J., & Kipen, H. M. (2012). Respiratory health effects of air pollution: Update on biomass smoke and traffic pollution. *Journal of Allergy and Clinical Immunology, 129*(1), 3-11.

32. Liu, L., Poon, R., Chen, L., Frescura, A., Montuschi, P., Ciabattoni, G., . . . Dales, R. (2008). Acute Effects of Air Pollution on Pulmonary Function, Airway Inflammation, and Oxidative Stress in Asthmatic Children. *Environ Health Perspect Environmental Health Perspectives, 117*(4), 668-674.

33. Lu, C., Deng, Q., Yu, C. W., Sundell, J., & Ou, C. (2013). Effects of ambient air pollution on the prevalence of pneumonia in children: Implication for National Ambient Air Quality Standards in China. *Indoor and Built Environment, 23*(2), 259-269.

34. Macintyre, E. A., Gehring, U., Mölter, A., Fuertes, E., Klümper, C., Krämer, U., . . . Heinrich, J. (2013). Air Pollution and Respiratory Infections during Early Childhood: An Analysis of 10 European Birth Cohorts within the ESCAPE Project. *Environ Health Perspect Environmental Health Perspectives*.

35. Mcdonald, J. D., Barr, E. B., White, R. K., Chow, J. C., Schauer, J. J., Zielinska, B., & Grosjean, E. (2004). Generation and Characterization of Four Dilutions of Diesel Engine Exhaust for a Subchronic Inhalation Study. *Environmental Science & Technology Environ. Sci. Technol., 38*(9), 2513-2522.

36. Morales, E., Garcia-Esteban, R., Cruz, O. A., Basterrechea, M., Lertxundi, A., Maria D Martinez López De Dicastillo, . . . Sunyer, J. (2014). Intrauterine and early postnatal

exposure to outdoor air pollution and lung function at preschool age. *Thorax, 70*(1), 64-73.

37. Morgenstern, V., Zutavern, A., Cyrys, J., Brockow, I., Koletzko, S., Krämer, U., . . . Heinrich, J. (2008). Atopic Diseases, Allergic Sensitization, and Exposure to Traffic-related Air Pollution in Children. *Am J Respir Crit Care Med American Journal of Respiratory and Critical Care Medicine, 177*(12), 1331-1337.

38. Mostofsky, E., Schwartz, J., Coull, B. A., Koutrakis, P., Wellenius, G. A., Suh, H. H., . . . Mittleman, M. A. (2012). Modeling the Association Between Particle Constituents of Air Pollution and Health Outcomes. *American Journal of Epidemiology, 176*(4), 317-326.

39. Nelin, T. D., Joseph, A. M., Gorr, M. W., & Wold, L. E. (2012). Direct and indirect effects of particulate matter on the cardiovascular system. *Toxicology letters*, *208*(3), 293-299.

40. Niessen, L. (2009). Comparative impact assessment of child pneumonia interventions. *Bulletin of the World Health Organization Bull World Health Org, 87*(6), 472-480.

41. Nordling, E., Berglind, N., Melén, E., Emenius, G., Hallberg, J., Nyberg, F., . . . Bellander, T. (2008). Traffic-Related Air Pollution and Childhood Respiratory Symptoms, Function and Allergies. *Epidemiology, 19*(3), 401-408.

42. Oberdorster, G. (2001). Pulmonary effects of ultrafine particles. *Journal of Occupational Environmental Health, 74*, 1-8.

43. Peden, D. B. (2000). Development of Atopy and Asthma: Candidate Environmental Influences and Important Periods of Exposure. *Environmental Health Perspectives, 108*, 475.

44. Penna, M. L., & Duchiade, M. P. (1990). Air pollution and infant mortality from pneumonia in the Rio de Janeiro metropolitan area. *Bulletin of the Pan American Health Organization*, *25*(1), 47-54.

45. Perera, F. P., Jedrychowski, W., Rauh, V., & Whyatt, R. M. (1999). Molecular epidemiologic research on the effects of environmental pollutants on the fetus. *Environ Health Perspect Environmental Health Perspectives, 107*(Suppl 3), 451-460.

46. Perera, F., Rauh, V., Whyatt, R., Tang, D., Tsai, W., Bernert, J., . . . Kinney, P. (2005). A Summary of Recent Findings on Birth Outcomes and Developmental Effects of Prenatal ETS, PAH, and Pesticide Exposures. *NeuroToxicology, 26*(4), 573-587.

47. Peters, J. M., Avol, E., Navidi, W., London, S. J., Gauderman, W. J., Lurmann, F., . . . Thomas, D. C. (1999). A Study of Twelve Southern California Communities with Differing Levels and Types of Air Pollution. *Am J Respir Crit Care Med American Journal of Respiratory and Critical Care Medicine, 159*(3), 760-767.

48. Pinkerton, K. E., & Joad, J. P. (2000). The Mammalian Respiratory System and Critical Windows of Exposure for Children's Health. *Environmental Health Perspectives, 108*, 457.

49. Polgar, G., and Weng, T. (1979). The functional development of the respiratory system: From the period of gestation to adulthood. *American Review of Respiratory Disease, 120*(3), 625-695.

50. Pope, C. A. (2000). Epidemiology of Fine Particulate Air Pollution and Human Health: Biologic Mechanisms and Who's at Risk? *Environmental Health Perspectives, 108*, 713.

51. Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., . . . Hoek, G. (2013). Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The Lancet Oncology, 14*(9), 813-822.

52. Rice, M. B., Rifas-Shiman, S. L., Oken, E., Gillman, M. W., Ljungman, P. L., Litonjua, A. A., . . . Gold, D. R. (2014). Exposure to traffic and early life respiratory infection: A cohort study. *Pediatric Pulmonology Pediatr Pulmonol., 50*(3), 252-259.

53. Riedl, M., & Diaz-Sanchez, D. (2005). Biology of diesel exhaust effects on respiratory function. *Journal of Allergy and Clinical Immunology, 115*(2), 229.

54. Rioux, C. L., Gute, D. M., Brugge, D., Peterson, S., & Parmenter, B. (2010). Characterizing Urban Traffic Exposures Using Transportation Planning Tools: An Illustrated Methodology for Health Researchers. *Journal of Urban Health J Urban Health, 87*(2), 167-188.

55. Ritz, B., Wilhelm, M., & Zhao, Y. (2006). Air Pollution and Infant Death in Southern California, 1989-2000. *Pediatrics, 118*(2), 493-502.

56. Rojas, R., Romieu, I., Perez-Padilla, R., Mendoza, L., Fortoul, T., & Olaiz, G. (2006). Lung Function Growth in Children with Long-Term Exposure to Air Pollutants in Mexico City. *Epidemiology, 17*(Suppl).

57. Romieu, I., Samet, J. M., Smith, K. R., & Bruce, N. (2002). Outdoor Air Pollution and Acute Respiratory Infections Among Children in Developing Countries. *Journal of Occupational and Environmental Medicine, 44*(7), 640-649.

58. Romieu, I., Ramirez-Aguilar, M., Sienra-Monge, J. J., Moreno-Macias, H., Rio-Navarro, B. E., David, G., . . . London, S. (2006). GSTM1 and GSTP1 and respiratory health in asthmatic children exposed to ozone. *European Respiratory Journal, 28*(5), 953-959.

59. Rudan, I. (2008). Epidemiology and etiology of childhood pneumonia. *Bulletin of the World Health Organization Bull World Health Organ, 86*(5), 408-416.

60. Ryan, P. H., Bernstein, D. I., Lockey, J., Reponen, T., Levin, L., Grinshpun, S., . . . Lemasters, G. (2009). Exposure to Traffic-related Particles and Endotoxin during Infancy Is Associated with Wheezing at Age 3 Years. *Am J Respir Crit Care Med American Journal of Respiratory and Critical Care Medicine, 180*(11), 1068-1075.

61. Schultz, E. S., Gruzieva, O., Bellander, T., Bottai, M., Hallberg, J., Kull, I., . . . Pershagen, G. (2012). Traffic-Related Air Pollution And Lung Function In Children At 8

Years Of Age- A Birth Cohort Study. *B17. How Bad Is Traffic Pollution? Health Effects And Interventions*.

62. Schwartz, J., & Neas, L. M. (2000). Fine Particles Are More Strongly Associated than Coarse Particles with Acute Respiratory Health Effects in Schoolchildren. *Epidemiology, 11*(1), 6-10.

63. Selevan, S. G., Kimmel, C. A., & Mendola, P. (2000). Identifying Critical Windows of Exposure for Children's Health. *Environmental Health Perspectives, 108*, 451.

64. Shaheen, S. O., Barker, D. J., Shiell, A. W., Crocker, F. J., Wield, G. A., & Holgate, S. T. (1994). The relationship between pneumonia in early childhood and impaired lung function in late adult life. *Am J Respir Crit Care Med American Journal of Respiratory and Critical Care Medicine, 149*(3), 616-619.

65. Smarr, M. M., Vadillo-Ortega, F., Castillo-Castrejon, M., & O'Neill, M. S. (2013). The use of ultrasound measurements in environmental epidemiological studies of air pollution and fetal growth. *Current Opinion in Pediatrics, 25*(2), 240-246.

66. Soto-Martinez, M., & Sly, P. D. (2009). Review Series: What goes around, comes around: Childhood influences on later lung health?: Relationship between environmental exposures in children and adult lung disease: The case for outdoor exposures. *Chronic Respiratory Disease, 7*(3), 173-186.

67. Spira-Cohen, A., Chen, L. C., Kendall, M., Lall, R., & Thurston, G. D. (2011). Personal Exposures to Traffic-Related Air Pollution and Acute Respiratory Health among Bronx Schoolchildren with Asthma. *Environ Health Perspect Environmental Health Perspectives, 119*(4), 559-565.

68. Šrám, R. J., Binková, B., Dejmek, J., & Bobak, M. (2005). Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. *Environ Health Perspect Environmental Health Perspectives, 113*(4), 375-382.

69. Stocks, J., Hislop, A., & Sonnappa, S. (2013). Early lung development: Lifelong effect on respiratory health and disease. *The Lancet Respiratory Medicine, 1*(9), 728-742.

70. Vieira, S. E., Stein, R. T., Ferraro, A. A., Pastro, L. D., Pedro, S. S., Lemos, M., . . . Saldiva, P. H. (2012). Urban Air Pollutants Are Significant Risk Factors for Asthma and Pneumonia in Children: The Influence of Location on the Measurement of Pollutants. *Archivos De Bronconeumología (English Edition), 48*(11), 389-395.

71. Wang, M., Gehring, U., Hoek, G., Keuken, M., Jonkers, S., Beelen, R., . . . Brunekreef, B. (2015). Air Pollution and Lung Function in Dutch Children: A Comparison of Exposure Estimates and Associations Based on Land Use Regression and Dispersion Exposure Modeling Approaches. *Environ Health Perspect Environmental Health Perspectives*.

72. Pneumonia. World Health Organization web site. http://www.who.int/mediacentre/factsheets/fs331/en//. Updated November 2015. Published 2016. Accessed February 6, 2016.

73. Zeltner, T. B., & Burri, P. H. (1987). The postnatal development and growth of the human lung. II. Morphology. *Respiration Physiology, 67*(3), 269-282.

74. Zhai X, Sampath P, Mulholland JA, et al. (2015). "Comparison and calibration of Research-Line (RLINE) model results with measurement-based an CMAQ-based source impacts." *Poster presented at Community Modeling and Analysis System (CMAS) Conference*.

75. Zhou Y (2012). Creating Demographic Clusters of Georgia, 2011, Georgia Department of Public Health, Office of Health Indicators for Planning.

Appendix I: Literature Review

Although the literature is not extensive, several studies have specifically looked at childhood pneumonia outcomes in relation to chronic ambient traffic exposures. Most studies in the literature examine hospitalizations due to short term exposures to peak air pollution levels, typically in urbanized areas. Acute exposures are important to understand, but that is not the focus of this study. This present analysis will focus on longer term exposures and the health effects related to more chronic low level exposures in children. The literature on the effects of chronic traffic related exposures and childhood pneumonia is not extensive, so this study is a vital contribution to the literature on this topic. The literature basis for this study will now be discussed, focusing on several available meta-analysis, prospective cohort, and cross sectional studies related to traffic related ambient air pollution exposures and early childhood pneumonia outcomes in developed countries.

For example, a metaanalysis of population based prospective birth cohort studies from Milan, Italy recruited 777 children aged 2 to 18 years from the local pediatric clinic between November and December 2011 (Esposito et al., 2014). The children were stratified based on history of recurrent wheezing or asthma, based on the occurrence of at least three clinical diagnosed lower respiratory tract illnesses with wheezing in a 6 month period or pediatrician diagnosed asthma. In this study, asthma was defined as “the presence of episodes of cough, breathlessness or dyspnea”. The other arm of the cohort was made up of healthy children born at term with no history of wheezing/asthma, negative for presence of lower respiratory tract disease at baseline, who were randomly selected and enrolled in the study during the same period as those who attended the clinic

for mionor surgery procedures. Outcomes were reported based on standardized parental reports over a period of 12 months beginning in January 2012. The researchers regularly contacted the participants by phone to make sure they were completing the surveys so that recall bias could be minimized in this study. Exposure data were combined with information collected from centralized air pollution monitors, providing daily concentrations of NO_2 and PM_{10}. Comparisons between children in each strata were made using contingency tables with a chi square test for categorical variables or a wilcoxon's rank sum test for continuous variables that were not normally distributed. ORs were computed using unconditional multiple logistic regression models that included terms for age, sex, number of siblings, parental education, and presence of smokers at home to examine the effect of incremental increases in pollutant concentration on the incidence of respiratory symptoms over a 12 month period in both study groups. A disadvantage of this study is that they did not stratify by age to look at differences in effect across different age groups. According to the results of this study, living close to a busy street was increased the risk for asthma episodes (OR:1.79, CI: (1.13,2.84), on the other hand living near a park reduced the risk of asthma (OR: 0.50, CI:(0.31, 0.80). Futhermore, an increase in 10 micrograms per meter cubed of PM_{10} and NO_2 increased the incidence of pneumonia, although barely (RR=1.08, CI: (1.00-1.17) for PM_{10} and RR=1.08, CI: (1.01, 1.17). Note that these results are barely significant and may be collinear with each other, so interpretation of these results should be considered cautiously. Also, children with a history of wheezing or asthma reported significantly ($p<0.001$) more episodes of pneumonia compared to healthy controls.

Another meta-analysis examined the effects of air pollution on childhood pneumonia outcomes in population based prospective birth cohorts from various European countries (Fuertes et al., 2014). Pneumonia was defined by a parental report of at least one clinically diagnosed pneumonia incident between birth and two years of age. The study used standardized land use regression models and logistic regression models adjusted for host and environmental covariates and total mass of $PM_{2.5}$ and PM_{10}. The analysis found that pneumonia was only associated with zinc from non tailpipe emissions derived from PM_{10} (OR:1.47, CI (0.99, 2.18) per 20 μg/m3 increase). As noted by the confidence interval that includes the null value, this effect estimate is non significant. However it was interesting to note that the study found stronger associations with first year of life exposures compared to second year of life exposures for the same unit increase in emissions. However, the results were not significant.

A third meta-analysis also found strong associations between elevated exposure to air pollution and increased incidence of pneumonia outcomes in children during the first 2 years of life (MacIntyre et al., 2013). Pneumonia was defined by parent report of physician diagnosed pneumonia in relation to annual average pollutant levels during the first 2 years of life and utilized land use regression models and logistic binomial regression models for the analysis. Unlike the results from the previous meta-analysis that found strong but insignificant associations, this meta-analysis was able to find strongly significant associations between exposure to air pollution and childhood pneumonia. The combined adjusted OR was found to be up and away from the null and statistically significant for all pollutants except $PM_{2.5}$. For example the OR for a 10-μg/m3 increase in NO_2 was found to be 1.30 with a confidence interval of (1.02,1.65).

Furthermore, the OR for a 10-μg/m3 increase in PM10 was found to be 1.76 with a confidence interval of (1.00, 3.09). Notably the confidence intervals for the associations of these two pollutants are barely significant, but significant nonetheless. Possibly with repeated samples or increased sample size the associations and effects would have been stronger and more statistically significant. The OR for $PM_{2.5}$ was 2.58 for a 5 μg/m3 increase but the confidence interval was insignificant because it includes the null value of 1, and also is comparably wider and less precise than the other confidence intervals ($PM_{2.5}$ CI: (0.91, 7.27)). It is interesting to note that this study also found stronger associations when the data was restricted to outcomes in the first year of life compared to the second year of life. These results are consistent with the results from the previously mentioned meta-analysis. Also notable, the stratified meta-analysis suggested slightly stronger effects in females and in those from middle SES groups. Lastly, pneumonia effect estimates were stronger and statistically significant for children who had moved from their original home to another home compared to children who remained at the same residence (OR: 1.62, CI: (1.20, 2.18).

Several other articles examined the association between traffic related ambient air pollution and early childhood pneumonia outcomes using prospective birth cohort study designs. One example is a 2013 study by Jedrychowski et al. in the Krakow area of Poland. In this prospective birth cohort, a total of 214 children were followed from birth until seven years of age. The mothers of these children were recruited from ambulatory prenatal clinics. Upon recruitment, these mothers were asked to fill out questionnaires detailing the health and exposure information. Health information was then continuously collected during annual follow up interviews during the seven year follow up period to

determine incidence of doctor diagnosed pulmonary outcomes. The exposure of interest was $PM_{2.5}$ and the outcome was defined as incidence of recurrent episodes of bronchitis and pneumonia during the seven year follow up. The results were analyzed in SAS using multivariable logistic models adjusted for potential confounders such as prenatal and postnatal ETS, city residence (a proxy for postnatal urban exposure), child's sensitization to domestic aeroallergens as well as asthma. According to the results, the adjusted OR for incidence of recurrent broncho-pulmonary infections (five or more spells of bronchitis and or pneumonia) recorded in the follow up was found to be significantly correlated with pneumonia outcomes in a dose response manner ($PM_{2.5}$ OR=2.44 and CI (1.12,5.36). It is interesting to note that both the physician diagnosed bronchitis and pneumonia cases were lower in the younger age group than the older age group. This is contradictory to our hypothesis that younger children will have higher incidence than older children (given same proximity to road way exposures). Furthermore, the study results suggest that children with asthma showed a two fold higher number of bronchitis episodes and more than a three fold higher number of pneumonia episodes. The incidence of these episodes was significantly correlated with the level of prenatal $PM_{2.5}$ exposures. This particular result is consistent with our second hypothesis.

Another relevant prospective cohort study by Rice et al from 2015 examined the association between traffic related exposures and childhood pneumonia outcomes. This study took place between 1999-2002 in Boston, Massachusetts and recruited women in their first trimester during their first prenatal visit. The children of these women were followed from birth until early childhood (median age was 3.3 years). The outcome was respiratory infection defined as a maternal report of greater than or equal to 1 doctor

diagnosed pneumonia, bronchitis croup or other respiratory infection from birth until the early childhood wellness visit (median age 3.3 years). The questionnaires for this information were distributed to mothers at study enrollment, when the child reached 6 months, when the child was 1 year, 2 years and then again at the early childhood wellness visit. The air quality emissions data were estimated based on geocoded subject addresses using ArcGIS at study enrollment and then again at the time of delivery. Estimates of traffic density were quantified by the Massachusetts Department of Transportation. The results were analyzed in SAS using relative risk regression models adjusted for potential confounders that were selected a priori. In fully adjusted models the risk ratio for respiratory infection were 1.30 (CI: 1.08, 1.55) for living less than 100m from the road; 1.15 (CI: 0.93, 1.41) for living between 100 to 200m from the road; and 0.95 (CI: 0.84, 1.07) for living between 200 to 1000m from the road. The reference group was living greater than 1000m away from the roadway. Results also indicate that for each interquartile range increase in distance further from roadway, there was an associated 8% (CI: 0.87, 0.98) lower risk of childhood pneumonia outcome. It was also noted that each interquartile range increase in traffic density was associated with a 5% (CI: 0.98, 1.13) increased incidence of respiratory infection. These results were not statistically significant. However, the risk for respiratory infection was 1.31 (1.08, 1.60) times higher for those living less than 100m compared to those living greater than or equal to 1000m from a major road. Notably, this result is statistically significant. Thus not only does proximity to roadway matter but also traffic volume is an important predictor of the association between traffic emissions and childhood pneumonia outcomes.

The results of one last prospective birth cohort study are included in this review. This study took place from 2001-2006 and examined a birth cohort from coming from Cincinnati, Ohio during the child's first year of life until three years of age (Ryan et. al, 2009). The children were recruited from parental report of two or more wheezing episodes in the past 12 months as of the 36 months clinic visit routine check up consultation. The children were clinically evaluated annually at ages 1,2,3 years receiving an SPT and physical examination. At each visit, parents of the child were asked to fill out a questionnaire on child's health and environmental exposures in the previous year including environmental exposures, ETS exposures and pet interactions. We are not looking at all of these potential covariates as potential confounders or effect modifiers in our model because available information is not the same for our cohort. The models were analyzed using land use regression to examine the correlation and distribution of average daily exposure to ECAT, adjusting for race, household income, sex, parental history of asthma, day care attendance, report of an upper and or lower respiratory condition in the past 12 months. According to the results of this analysis, persistent wheezing by 36 months was significantly associated with exposure to increased levels of traffic related particles before one year of age (OR: 1.75 (CI:1.07, 2.87). It is interesting to note that a co- exposure to endotoxin had a synergistic effect with traffic exposure on persistent wheeze after adjustment for significant covariates (OR:5.85 CI: (1.89, 18.13). However note that the precision for this estimate is rather low, reflected by the wide confidence interval. Furthermore it was found that persistent allergic wheeze to high emissions was a strong but not statistically significant predictor of the outcome (OR:2.11, CI (0.97, 4.61) compared with children without persistent allergic wheeze, after controlling for

endotoxin, sex, parental asthma, race, lower respiratory conditions, and breastfeeding. The interaction product term created between the exposure and endotoxin was not found to be statistically significant and did not remain significant in the final model.

Several cross- sectional studies have also examined the association between early life traffic related exposure and childhood pneumonia outcome. One such study took place in China (Lu et al., 2013). Emissions data was collected between 2008 and 2011 and the participants were recruited between September 2011 and January 2012. Children between the ages of 3 and 6 years of age were recruited based on 4988 questionnaires that were randomly distributed to parents in 29 local kindergartens in 5 community districts. A total of 2727 completed questionnaires were returned within one week to give a response rate of 59%. Of these, 2706 responses from eligible children aged 3-6 were included in the study. The outcome was defined as prevalence of pneumonia in children aged 3-6 in related to traffic related exposures to PM_{10}, SO_2, and NO_2. The results were analyzed based on 2 stage hierarchical regression techniques, involving multivariate logistic regression to determine the personal variables associated with pneumonia diagnosis. The adjusted prevalence odds ratios were calculated for the resulting final models. The study concluded that the overall prevalence of pneumonia was 38.2% in the children aged 3-6 years. The prevalence of pneumonia was significantly associated with exposure to NO_2, with an OR=1.16 and CI: (1.12, 1.20). However the pneumonia prevalence was not significantly associated with either PM_{10} or SO_2 (ORs are 1.19 and 1.08 respectively and the CIs are (0.84, 1.26) and (0.92, 1.26) respectively. Moreover, the increase in one episode day (based one day's average exceeding the NAAQS standards) per year for NO_2 was suggested to lead to an increase pneumonia prevalence by 3.8%

with a rather imprecise CI of (2.4, 53.1). Interestingly, it was noted that the morbidity for males was 40.7% which was relatively higher than the morbidity for females which was 35.3%. There was no meaningful difference in incidence of pneumonia among the categories of ages between 3 and 6. However, children living in urban areas had the highest prevalence of pneumonia compared to children living in more rural areas away from the source of exposure.

A second cross sectional study conducted in the metro area of Rio de Janeiro, Brazil in 1980 examined the effect of exposure to traffic emissions during the first year of life. The outcome definition was infant mortality from pneumonia in 1980. The analysis used multiple linear regression (the stepwise method) controlling for potential covariates such as areas of residence and incomes. Data regarding the number of deaths among children under one year by cause and area of residence were provided by the State Secretariat of Health. This is the agency responsible for collecting and codifying death certificates in Brazil. However it is important to note that the criteria for coding pneumonia may vary between studies if countries use various internal coding systems (similar to ICD9 codes) to diagnose cause of death. The coding used in this analysis was not reported. The results suggest that the best model explains 83% of the overall variation in infant mortality, but only 5.27% of the total variation can be explained by pollution. There were no odds ratio or CI estimate associated with this correlation coefficient. This could lead to exposure misclassification bias if such a small percentage of the variation is explained by the air pollution mixture. This investigation reveals an association between air pollution (measured in terms of annual average concentration of suspended particles) and infant pneumonia mortality at the level of aggregate data for each administrative

region in the metro area of Rio de Janeiro. This association persists even when controlling for proportion of households that a total income below two minimum wages, thus controlling for SES factors.

Another notable cross sectional study of traffic related childhood pneumonia outcomes was also conducted in Brazil, this one specifically recruiting patients from Sao Paulo during August- October 2009 (Vieira et. al, 2012). Children ages 6-10 were recruited from a primary care health clinic operated by the public administration so that the patients recruited would be representative of the general population. These children, who were selected from a specific residential district city of Sao Paulo with intense traffic, were hypothesized to have higher risk for respiratory morbidity (given appropriate monitoring) due to higher exposures. Chi square tests for linear tendency were used to examine the association between exposure and wheezing, asthma and pneumonia outcome (as reported by parental questionnaire.) Logistic regression models were used for univariate and multivariate analysis (adjusted for SES factors, family factors, and environmental factors.) Results found that respiratory morbidity was high with 43 of 64 children (67.2%) reporting having wheezing at any time, 27 of 64 (42.2%) wheezing in the last month, 17 of 64 (26.6%) having had asthma at any time and 21 (32.8%) having had pneumonia at any time. It is interesting to note that in a sub analysis of this population, it was suggested that in children between the ages of 5 and 10, the risk for respiratory disease increases 20% for every increase of 28.3 μg/m3 in the NO_2 concentration. Overall higher exposures of NO_2 and O_3 were associated with increased risk for asthma and pneumonia in children, especially for children living in areas of intense pollution.

There is a need to further examine the association between traffic related PM exposures on early childhood pneumonia outcomes. The literature that is available on this topic has several notable limitations. This could be due to inaccurate model prediction of exposure or could be due to bias due to recall of parental report on the health questionnaires. Another limitation of several of the articles is that different pollutants may have been measured (PM_{10}, or secondary PM instead of primary $PM_{2.5}$) or different time periods of a child's exposure or outcome could have been measured, leading results that are difficult to generalize to this study. Our study is unique because there are no identifiable studies in the literature that look at first year of life chronic traffic related exposures and first year of life pneumonia outcomes on a birth cohort located in Atlanta.

Appendix II: SAS code

```
libname h 'H:\Thesis';

*Create a temporary dataset;
Data thesis;
Set h.pneumonia_dataset;
RUN;

*Examining the data;
Proc Contents data=thesis;
RUN;
*There are a total of 22520 observations (children) and 36 variables;

Proc Freq data=thesis;
Tables gender MJRCLUSTER MNRCLUSTER child_ethnicity child_race
       mom_asthma mom_educ prenatal_smoking year1_smoking/missing;
RUN;

proc print data=thesis (obs=50);
var child_studyID pneumonia_dx_date birth_date days status rline_pm25
rline_co rline_nox gender
      mjrcluster mnrcluster child_race maternal_age mom_asthma mom_educ
prenatal_smoking;
RUN;

proc freq data=thesis;
tables mjrcluster;
run;

proc print data=thesis (obs=50);
var pneumonia_dx_date birth_date days;
format pneumonia_dx_date 8. birth_date 8.;
run;

proc logistic data=thesis descending;
model status=rline_pm25;
run;

proc gchart data=thesis;
vbar rline_pm25;
run;

proc gchart data=thesis;
vbar rline_co rline_nox;
run;quit;

Proc Means data=thesis;
var days average_household_income maternal_age median_house_value
median_household_income
    median_yr_built per_capita_income percent_adult_poverty
percent_child_poverty percent_families_poverty
    percent_lt_highschool percent_unemployment pneumonia_dx_date
population_density rline_CO rline_NOX rline_PM25;
RUN;
```

```
*To calculate the total number of children per year (by birth year);
proc freq data=thesis;
tables birth_date;
RUN;

proc freq data=thesis;
tables birth_date*status;
where status=1;
RUN;

*To calculate the total number of children per year (by pneumonia
diagnosis date);
Proc freq data=thesis;
tables pneumonia_dx_date;
RUN;

Proc freq data=thesis;
tables pneumonia_dx_date*status;
where status=1;
RUN;


*Create histograms to examine the distribution of several variables;
proc sgplot data=thesis;
histogram days;
label days='Days until censoring or pneumonia';
RUN;

proc sgplot data=thesis;
histogram average_household_income;
label average_household_income='Average Household Income';
RUN;

proc sgplot data=thesis;
histogram maternal_age;
label maternal_age='Maternal Age (years)';
RUN;

proc sgplot data=thesis;
histogram percent_families_poverty;
label percent_families_poverty='Percent of families below the Poverty
Line';
RUN;

proc sgplot data=thesis;
histogram percent_lt_highschool;
label percent_lt_highschool='Percent of Mothers with Less than
Highschool Education';
RUN;

proc sgplot data=thesis;
histogram percent_unemployment;
label percent_unemployment='Percent Unemployed Mothers';
RUN;

proc sgplot data=thesis;
```

```
histogram percent_child_poverty;
label percent_child_poverty ='Percent of children living below the
poverty line';
RUN;

proc sgplot data=thesis;
histogram rline_CO;
label rline_CO='Modeled Carbon Monoxide Concentration from RLINE';
RUN;

proc sgplot data=thesis;
histogram rline_NOX ;
label rline_NOX ='Modeled Nitrous Oxide Concentration from RLINE';
RUN;

proc sgplot data=thesis;
histogram rline_pm25;
where status=1;
RUN;

proc sgplot data=thesis;
histogram rline_PM25;
label rline_PM25 ='Modeled PM2.5 Concentration from RLINE';
RUN;

*SG panels for first year of life maternal smoking;
proc sgpanel data=thesis;
panelby year1_smoking/columns=3 missing novarname;
reg x=days y=rline_PM25;
label days='Days until censoring or pneumonia';
label rline_PM25 ='Modeled PM2.5 Concentration from RLINE';
Title 'RLINE modeled PM2.5 concentration (micrograms per meter cubed)as
a Function of Days until censoring or pneumonia for maternal smoking in
the first year of life';
RUN;

*SG panels for mothers education;
proc sgpanel data=thesis;
panelby mom_educ/columns=4 missing novarname;
reg x=days y=rline_PM25;
label days='Days until censoring or pneumonia';
label rline_PM25 ='Modeled PM2.5 Concentration from RLINE';
Title 'RLINE modeled PM2.5 concentration (micrograms per meter cubed)
as a Function of Days until censoring or pneumonia for mothers level of
education';
RUN;

*SG plots for mothers asthma status;
proc sgpanel data=thesis;
panelby mom_asthma/columns=3 missing novarname;
reg x=days y=rline_PM25;
label days='Days until censoring or pneumonia';
label rline_PM25 ='Modeled PM2.5 Concentration from RLINE';
Title 'RLINE modeled PM2.5 concentration (micrograms per meter cubed)
as a Function of Days until censoring or pneumonia for mothers asthma
status';
RUN;
```

```
*SG plots for maternal prenatal smoking;
proc sgpanel data=thesis;
panelby prenatal_smoking/columns=3 missing novarname;
reg x=days y=rline_PM25;
label days='Days until censoring or pneumonia';
label rline_PM25 ='Modeled PM2.5 Concentration from RLINE';
Title 'RLINE modeled PM2.5 concentration (micrograms per meter cubed)
as a Function of Days until censoring or pneumonia for prenatal
smoking';
RUN;

*****************************************************************;
*Developing a Cox regression model- unadjusted log log curves;

*regular KM survival curves for gender;
proc lifetest data=thesis method=KM PLOTS=(s,lls);
time days*status(0);
strata gender;
RUN;
*Interpretation of gender: Carry out a log rank test to determine if
there is a significant difference between the 2 curves and assess the
PH assumption using log log curves;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 18.96 and the pvalue is <.0001. Thus we
have evidence to reject the null and conclude there is a significant
difference between the two curves. Since the Wilcoxon test, which
weights early failures more, is also significant (p<0.0001), this
indicates that there is a difference in the early part of the curves as
well. The log log curves show no gross violation of the PH assumption
*/

*regular KM survival curves for demographic cluster (four categories);
proc lifetest data=thesis method=KM PLOTS=(s,lls);
time days*status(0);
strata MJRCLUSTER;
RUN;
*Interpretation of demographic clusters in four categories: Carry out a
log rank test to determine if there is a significant difference between
the curves and assess the PH assumption using log log curves;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 8.0437 and the p value is 0.0451
(borderline significant). Thus we have evidence to reject the null
(barely) and conclude there is a significant difference between the two
curves. However since it is only borderline significant it is highly
recommended that we check the assumption using other methods as well.
The PH assumption may not be violated (based on a very  subjective,
loose interpretation of parallelism. Verification using other methods
especially for MJRCLUSTER is highly recommended*/

*Create dummy variables for the major clusters to dichotomize it into
Group A vs B,C,D and dummy variables for child_race, mom_asthma,
mothers education, prenatal smoking, and first year of life smoking;
*Also dichotomize all continuous variables at the mean;
Data thesis_2;
Set thesis;

*Gender "m"=1, "f"=0;
```

```
if gender='M' then genderDV=1;
else genderDV=0;

*Mjrcluster "A"=1 reference category encompasses all other categories;
if mjrcluster='A' then DemCluster=1;
else DemCluster=0;

*Child race "white"=1, and the reference category encompasses all other
categories;
if child_race='White' then race=1;
else race=0;

*Mothers asthma "No"=1, and the reference category encompassess all
other categories;
if mom_asthma='no' then momasthma=1;
else momasthma=0;

*Mothers education "some college"=1, and the reference category
encompasses all other categories;
if mom_educ='some college+' then momedu=1;
else momedu=0;

*Prenatal smoking "No"=1, and the reference category encompasses all
other categories;
if prenatal_smoking='no' then prenatalsmk=1;
else prenatalsmk=0;

*For ITP_only10 "Inside"=1, and the reference category encompasses all
the other categories;
if ITP_only10=1 then Inside=1;
else Inside=0;

*Dichotomize mothers age at the mean;
if maternal_age ge 32 then maternal_ageDV=1;
else maternal_ageDV=0;


/*if average_household_income ge 77219 then
average_household_incomeDV=1;
else average_household_incomeDV=0;

if median_house_value ge 203817 then median_house_valueDV=1;
else median_house_valueDV=0;

if median_household_income ge 64497 then median_household_incomeDV=1;
else median_household_incomeDV=0;

if median_yr_built ge 1989 then median_yr_builtDV=1;
else median_yr_builtDV=0;

if per_capita_income ge 28635 then per_capita_incomeDV=1;
else per_capita_incomeDV=0;

if percent_adult_poverty ge 9.83 then percent_adult_povertyDV=1;
else percent_adult_povertyDV=0;

if percent_child_poverty ge 15.98 then percent_child_povertyDV=1;
```

```
else percent_child_povertyDV=0;

if percent_families_poverty ge 9.31 then percent_families_povertyDV=1;
else percent_families_povertyDV=0;

if percent_lt_highschool ge 11.38 then percent_lt_highschoolDV=1;
else percent_lt_highschoolDV=0;

if percent_unemployment ge 9.23 then percent_unemploymentDV=1;
else percent_unemploymentDV=0;

if pneumonia_dx_date ge 16962 then pneumonia_dx_dateDV=1;
else pneumonia_dx_dateDV=0;

if population_density ge 2194 then population_densityDV=1;
else population_densityDV=0;*/

 if rline_CO ge 138.2 then rline_CODV=1;
else rline_CODV=0;

if rline_NOX ge 45.6 then rline_NOXDV=1;
else rline_NOXDV=0;

if rline_PM25 ge 1.2 then rline_PM25DV=1;
else rline_PM25DV=0;
RUN;

*Cross tabs to check coding of the categorical variables;
Proc freq data=thesis_2;
tables DemCluster*mjrcluster;
RUN;

Proc freq data=thesis_2;
tables child_race*race;
RUN;

proc freq data=thesis_2;
tables mom_asthma*momasthma;
RUN;

Proc freq data=thesis_2;
tables mom_educ*momedu;
RUN;

proc freq data=thesis_2;
tables prenatal_smoking*prenatalsmk;
RUN;

Proc freq data=thesis_2;
tables ITP_only10*inside;
RUN;


*Regular KM survival curves for demographic cluster;
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata DemCluster;
```

```
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 0.7164 and the p value is 0.3973. Thus
we do not have evidence to reject the null and conclude there is a no
significant difference between the two curves. The log log curves seem
to be overlaping completely which is a gross violation of the PH
assumption*/

*regular KM survival curves for child's race (four categories);
proc lifetest data=thesis method=KM PLOTS=(s,lls);
time days*status(0);
strata child_race;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 12.29 and the p  value is 0.0065. Thus
we have evidence to reject the null and conclude there is a significant
difference between the curves. Based on a very very loose definition of
parallelism, we could say that there is no gross violation of the PH
assumption based on the log log curves. However, we would especially
recommend verifying this conclusion of the PH assumption using other
methods as well*/

proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata race;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is .434 and the p value is 0.5100. Thus we
do not have evidence to reject the null and conclude there is not a
significant difference between the curves. The curves seem to overlap,
suggesting a no violation of the PH assumption based on the log log
curves. However, we would especially recommend verifying this
conclusion of the PH assumption using other methods as well*/

*regular KM survival curves for whether or not the mother had asthma
(original coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata mom_asthma;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 10.1672 and the p value is 0.0062. Thus
we have evidence to reject the null and conclude there is a significant
difference between the curves. There seems to be a violation of the PH
assumption because the curves cross in several places and are overall
not parallel. Needs to be verified using other methods*/

*regular KM survival curves for whether or not the mother had asthma
(Dummy variables coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata momasthma;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 5.10 and the p value is 0.0239. Thus we
have evidence to reject the null and conclude there is a significant
difference between the curves. There seems to be a violation of the PH
```

```
assumption because the curves cross in several places and are overall
not parallel. Needs to be verified using other methods*/

*regular KM survival curves for mothers highest level of education
(regular coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata mom_educ;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 8.28 and the p value is 0.0405
(borderline significant). Thus we have evidence to reject the null and
conclude there is a significant difference between the curves. Although
the curves cross in several places there does not appear to be a gross
violation of the PH assumption*/

*regular KM survival curves for mothers highest level of education
(dummy variable coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata momedu;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 1.38 and the p value is 0.2397. Thus we
do not have evidence to reject the null and conclude there is no
significant difference between the curves. Although the curves cross in
several places there does not appear to be a gross violation of the PH
assumption*/

*regular KM survival curves for whether or not the mother smoked during
pregnancy (regular coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata prenatal_smoking;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 3.09 and the p value is 0.2133. Thus we
do not have evidence to reject the null and conclude there is not a
significant difference between the curves. Based on a very lose
interpretation of the log log curves there does not seem to be a gross
violation of the PH assumption. However, it is highly recommended to
verify this conclusion using other methods of checking the PH
assumption as well, especially for this variable*/

*regular KM survival curves for whether or not the mother smoked during
pregnancy (dummy variable coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata prenatalsmk;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is .0562 and the p value is 0.8126. Thus we
do not have evidence to reject the null and conclude there is not a
significant difference between the curves. Based on a very lose
interpretation of the log log curves there does not seem to be a gross
violation of the PH assumption. However, it is highly recommended to
verify this conclusion using other methods of checking the PH
```

```
assumption as well, especially for this variable*/

*regular KM survival curves for whether or not the person lives in the
atlanta perimeter (regular coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata ITP_only10;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is .1001 and the p value is 0.9512. Thus we
do not have evidence to reject the null and conclude there is no
significant difference between the curves. There Seems to be a gross
violation of the PH assumption*/

*regular KM survival curves for whether or not the person lives in the
atlanta perimeter (regular coding);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata Inside;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is .0963 and the p value is 0.7563. Thus we
do not have evidence to reject the null and conclude there is no
significant difference between the curves. There Seems to be a gross
violation of the PH assumption*/

*regular KM survival curves for maternal_age (dichotomized at the
mean);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata maternal_ageDV;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is 1.52 and the p value is 0.2176. Thus we
do not have evidence to reject the null and conclude there is no
significant difference between the two curves. There is no gross
violation of the PH assumption based on assessment of the log log
curves*/

*regular KM survival curves for rline_pm25 (dichotomized at the mean);
proc lifetest data=thesis_2 method=KM PLOTS=(s,lls);
time days*status(0);
strata rline_pm25DV;
RUN;
/*The null hypothesis is that both curves are the same. The chi square
value for the log rank test is ,1546 and the p value is .6942. Thus we
do not have evidence to reject the null and conclude there is not a
significant difference between the two curves. There is no gross
violation of the PH assumption based on assessment of the log log
curves*/

*****************************************************************;

*Consider separately all predictors again (one at a time). Using
Schoenfeld residuals GOF tests, evaluate whether each variable
satisfies the PH assumption;
```

```
*GOF testing in a model that only includes ITP_only10 ;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=ITP_only10;
output out=SR_thesis ressch=SR_ITP10;
id family_ID;
RUN;

data ITP10_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=ITP10_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_ITP10;
with timerank;
RUN;
*p-value for the GOF correlation between SR_ITP10 and survival time
(days)= 0.8263.;
*Suggests that PH assumption is not violated for ITP_only10;

*GOF testing in a model that only includes maternal_age;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= maternal_age ;
output out=SR_thesis ressch=SR_maternalage;
id family_id;
RUN;

data maternalage_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=maternalage_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_maternalage;
with timerank;
RUN;
*p-value for the correlation between SR_maternalage and survival time
(days)= 0.1677;
*Suggests that PH assumption is not violated for maternal age;

*GOF testing in a model that only includes rline_PM25;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=  rline_PM25;
output out=SR_thesis ressch=SR_rlinepm25;
id family_id;
RUN;
```

```
data rlinepm25_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=rlinepm25_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_rlinepm25;
with timerank;
RUN;
*p-value for the correlation between SR_rlinepm25 and survival time
(days)= 0.4827;
*Suggests that PH assumption is not violated for rline_pm25;

*GOF testing in a model that only includes gender;
proc phreg data=thesis covs(aggregate);
class gender;
model days*status(0)=  gender;
output out=SR_thesis ressch=SR_gender;
id family_id;
RUN;

data gender_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=gender_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_gender;
with timerank;
RUN;
*p-value for the correlation between SR_gender and survival time
(days)= 0.4653;
*Suggests that PH assumption is not violated for gender;

*GOF testing in a model that only includes child_race;
proc phreg data=thesis covs(aggregate);
class child_race;
model days*status(0)=  child_race;
output out=SR_thesis ressch=SR_child_race;
id family_id;
RUN;

data child_race_events;
set SR_thesis;
if status=1;
RUN;
```

```
proc rank data=child_race_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_child_race;
with timerank;
RUN;
*p-value for the correlation between SR_child_race and survival time
(days)= 0.2185;
*Suggests that PH assumption is not violated for child_race;

*GOF testing in a model that only includes mom_asthma;
proc phreg data=thesis covs(aggregate);
class mom_asthma;
model days*status(0)=  mom_asthma;
output out=SR_thesis ressch=SR_mom_asthma;
id family_id;
RUN;

data mom_asthma_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=mom_asthma_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_mom_asthma;
with timerank;
RUN;
*p-value for the correlation between SR_mom_asthma and survival time
(days)= 0.0038;
*Suggests that PH assumption is violated for mom_asthma;

*GOF testing in a model that only includes prenatal_smoking;
proc phreg data=thesis covs(aggregate);
class prenatal_smoking;
model days*status(0)=  prenatal_smoking;
output out=SR_thesis ressch=SR_prenatal_smoking;
id family_id;
RUN;

data prenatal_smoking_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=prenatal_smoking_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;
```

```
Proc corr data=ranked_events nosimple;
var SR_prenatal_smoking;
with timerank;
RUN;
*p-value for the correlation between SR_prenatal_smoking and survival
time (days)= 0.2630;
*Suggests that PH assumption is not violated for prenatal_smoking;

*GOF testing in a model that only includes mom_educ;
proc phreg data=thesis covs(aggregate);
class mom_educ;
model days*status(0)= mom_educ;
output out=SR_thesis ressch=SR_mom_educ;
id family_id;
RUN;

data mom_educ_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=mom_educ_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_mom_educ;
with timerank;
RUN;
*p-value for the correlation between SR_mom_educ and survival time
(days)= 0.5188;
*Suggests that PH assumption is not violated for mom_educ;

*GOF testing in a model that only includes mjrcluster;
proc phreg data=thesis covs(aggregate);
class mjrcluster;
model days*status(0)=mjrcluster;
output out=SR_thesis ressch=SR_mjrcluster;
id family_id;
RUN;

data mjrcluster_events;
set SR_thesis;
if status=1;
RUN;

proc rank data=mjrcluster_events out=ranked_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked_events nosimple;
var SR_mjrcluster;
with timerank;
RUN;
*p-value for the correlation between SR_mjrcluster and survival time
```

```
(days)= 0.1137;
*Suggests that PH assumption is not violated for mjrcluster;


*********************************************************************;

/* Consider separately all predictors once more. Using an extended Cox
model that contains each predictor and a product term of the form V*t,
where V denotes a given predictor and t denotes days (a continuous
variable), evaluate whether each predictor variable satisfies the PH
assumption*/

*Extended cox model when only maternal_ageDV and its interaction with
time is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = maternal_ageDV maternal_ageDV_t;
maternal_ageDV_t = maternal_ageDV*days;
id family_id;
Run;
*The wald p value for the interaction term = 0.4240. Suggests PH
assumption is not violated;

*Extended cox model when only ITP_only10 and its interaction with time
is evaluated;
proc phreg data=thesis_2 covs (aggregate);
model days*status(0) = ITP_only10 ITP_only10_t;
ITP_only10_t = ITP_only10*days;
id family_id;
Run;
*The wald p value for the interaction term = 0.8123. Suggests PH
assumption is not violated;

*Extended cox model when only ITP_only10 Dummy variable and its
interaction with time is evaluated;
proc phreg data=thesis_2 covs (aggregate);
model days*status(0) = inside inside_t;
inside_t = inside*days;
id family_id;
Run;
*The wald p value for the interaction term = 0.2883. Suggests PH
assumption is not violated;

*Extended cox model when only rline_PM25DV and its interaction with
time is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = rline_PM25DV rline_PM25DV_t;
rline_PM25DV_t = rline_PM25DV*days;
id family_id;
Run;
*The wald p value for the interaction term = 0.2707. Suggests PH
assumption is not violated;

*Extended cox model when only gender and its interaction with time is
evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = genderDV genderDV_t/rl;
genderDV_t = genderDV*days;
```

```
id family_id;
Run;
*The wald p value for the interaction term = 0.4692. Suggests PH
assumption is not violated;

*Extended cox model when only child_race and its interaction with time
is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = race race_t;
race_t = race*days;
id family_id;
Run;
*The wald p value for the interaction term=.3244. Suggests the PH
assumption is not violated;

*Extended cox model when only mjrcluster and its interaction with time
is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = demcluster demcluster_t;
demcluster_t = demcluster*days;
id family_id;
Run;
*The wald p value for the interaction term=.1355. Suggests the PH
assumption is not violated;

*Extended cox model when only mom_asthma and its interaction with time
is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = momasthma momasthma_t;
momasthma_t = momasthma*days;
id family_id;
Run;
*The wald p value for the interaction term=.0843. Suggests the PH
assumption is not violated;

*Extended cox model when only prenatal_smoking and its interaction with
time is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = prenatalsmk prenatalsmk_t;
prenatalsmk_t = prenatalsmk*days;
id family_id;
Run;
*The wald p value for the interaction term=.2780. Suggests the PH
assumption is not violated;

*Extended cox model when only mom_educ and its interaction with time is
evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0) = momedu momedu_t;
momedu_t = momedu*days;
id family_id;
Run;
*The wald p value for the interaction term=.9602. Suggests the PH
assumption is not violated;


*Using Time Squared;
```

```
*Extended cox model when only maternal_ageDV and its interaction with
time squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=maternal_ageDV maternal_ageDV_Time2/ rl;
maternal_ageDV_Time2 = maternal_ageDV*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.8083. Suggests the PH
assumption is not violated;

*Extended cox model when only ITP_only10 and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=ITP_only10 ITP_only10_Time2/ rl;
ITP_only10_Time2 = ITP_only10*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.8373. Suggests the PH
assumption is not violated;

*Extended cox model when only ITP_only10 DV and its interaction with
time squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=inside inside_Time2/ rl;
inside_Time2 = inside*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.3918. Suggests the PH
assumption is not violated;

*Extended cox model when only rline_PM25DV and its interaction with
time squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=rline_PM25DV rline_PM25DV_Time2/ rl;
rline_PM25DV_Time2 = rline_PM25DV*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.2985. Suggests the PH
assumption is not violated;

*Extended cox model when only gender and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=genderDV genderDV_Time2/ rl;
genderDV_Time2 = genderDV*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.6586. Suggests the PH
assumption is not violated;

*Extended cox model when only mjrcluster and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=demcluster demcluster_Time2/ rl;
demcluster_Time2 = demcluster*(days**2);
id family_id;
run;
```

```
*The wald p value for the interaction term=0.0703. Suggests the PH
assumption is not violated;

*Extended cox model when only child race and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=race race_Time2/ rl;
race_Time2 = race*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.3454. Suggests the PH
assumption is not violated;

*Extended cox model when only mom asthma and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=momasthma momasthma_Time2/ rl;
momasthma_Time2 = momasthma*(days**2);
id family_id;
run;
*Wald p value for the interaction term=0.0457. Borderline significant
but Suggests the PH assumption is not violated;

*Extended cox model when only mom educ and its interaction with time
squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=momedu momedu_Time2/ rl;
momedu_Time2 = momedu*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.8616. Suggests the PH
assumption is not violated;

*Extended cox model when only prenatal smoking and its interaction with
time squared is evaluated;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)=prenatalsmk prenatalsmk_Time2/ rl;
prenatalsmk_Time2 =prenatalsmk*(days**2);
id family_id;
run;
*The wald p value for the interaction term=0.2244. Suggests the PH
assumption is not violated;



*Using ln time;

*Extended cox model when only maternal_ageDV and its interaction with
natural log time is evaluated;
 proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=maternal_ageDV maternal_ageDV_lnt;
maternal_ageDV_lnt= maternal_ageDV*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.1565. Suggests the PH
assumption is not violated;
```

```
*Extended cox model when only ITP_only10 and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=ITP_only10 ITP_only10_lnt;
ITP_only10_lnt= ITP_only10*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.6333. Suggests the PH
assumption is not violated;

*Extended cox model when only ITPonly10DV and its interaction with
natural log time is evaluated;
 proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=inside inside_lnt;
inside_lnt= inside*log(days);
id family_id;
run;
*The wald p value for the interaction term=.1547. Suggests the PH
assumption is not violated;

*Extended cox model when only rline_PM25DV and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=rline_PM25DV rline_PM25DV_lnt;
rline_PM25DV_lnt= rline_PM25DV*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.1802. Suggests the PH
assumption is not violated;

*Extended cox model when only gender and its interaction with natural
log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=genderDV genderDV_lnt;
genderDV_lnt= genderDV*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.2031. Suggests the PH
assumption is not violated;

*Extended cox model when only mjrcluster and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=demcluster demcluster_lnt;
demcluster_lnt= demcluster*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.2363. Suggests the PH
assumption is not violated;

*Extended cox model when only child race and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=race race_lnt;
race_lnt= race*log(days);
id family_id;
run;
```

```
*The wald p value for the interaction term=0.3764. Suggests the PH
assumption is not violated;

*Extended cox model when only mom asthma and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=momasthma momasthma_lnt;
momasthma_lnt= momasthma*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.1245. Suggests the PH
assumption is not violated;

*Extended cox model when only mom educ and its interaction with natural
log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=momedu momedu_lnt;
momedu_lnt= momedu*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.5989. Suggests the PH
assumption is not violated;

*Extended cox model when only prenatal smoking and its interaction with
natural log time is evaluated;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0)=prenatalsmk prenatalsmk_lnt;
prenatalsmk_lnt= prenatalsmk*log(days);
id family_id;
run;
*The wald p value for the interaction term=0.6212. Suggests the PH
assumption is not violated;

**********************************************************************;

/*In a no interaction Cox model that contains the Rline exposure
variables and controls for the variables that satisfy the PH
assumption, evaluate the PH assumption for what variables do not
satisfy the PH assumption (separately),using:
a) Adjusted ln-ln survival curves. Note: be sure to choose appropriate
values for the covariates when estimating these!
b) A linear interaction term with time. That is, use an extended cox
model that contains all predictors and a product term of the form V*t,
where V denotes what does not satisfy the PH assumption and t=month.
c) Schoenfeld residuals

*a) Adjusted ln-ln survival curves. Note: be sure to choose appropriate
values for the covariates when estimating these!;
*Use the dataset to obtain plots (survival and log- log survival) for
those who have pneumonia and those who do not, adjusted for variables
that satisfy the PH assumption.

*Since we are looking for ADJUSTED survival and log- log survival
plots, we need to first find the mean values of the things we are
adjusting for;

proc means data=thesis_2;
```

```
var rline_pm25DV rline_CODV rline_noxDV maternal_ageDV;
RUN;

*We now need to create a new dataset that only contains these mean
values;
data thesis_meanvalues;
input rline_pm25DV rline_CODV rline_noxDV maternal_ageDV ;
datalines;
0.4651421  0.3950710    0.3997336    0.4221137
;
run;

*Variables that satisfy the PH assumption appear in the model statement
while variable that does not, is in the strata statement;

*MJRCLUSTER;
proc phreg data = thesis_2 plots (overlay=row) = survival
covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV;
strata demcluster;
id family_id;
baseline covariates = thesis_meanvalues out = outputdataset survival =
s loglogs = lls;
title 'Stratified Cox Procedure - Stratified on Major Demographic
Cluster';
title2 'Adjusted for mean values variables that satisfy the PH
assumption';
Run;title;

proc sgplot data = outputdataset;
title 'Log-log survival plots for Major Demographic Cluster groups';
title2 'Adjusted for variables that satisfy the PH assumption';
step x = days y = lls / group = demcluster;
run;title;

*CHILD_RACE;
proc phreg data = thesis_2 plots (overlay=row) = survival covs
(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV;
strata RACE;
id family_id;
baseline covariates = thesis_meanvalues out = outputdataset survival =
s loglogs = lls;
title 'Stratified Cox Procedure - Stratified on Childs Race';
title2 'Adjusted for mean values variables that satisfy the PH
assumption';
Run;title;

proc sgplot data = outputdataset;
title 'Log-log survival plots for Childs Race groups';
title2 'Adjusted for variables that satisfy the PH assumption';
step x = days y = lls / group =race;
run; title;

*MOM_ASTHMA;
```

```
proc phreg data = thesis_2 plots (overlay=row) = survival
covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV;
strata MOMASTHMA;
id family_id;
baseline covariates = thesis_meanvalues out = outputdataset survival =
s loglogs = lls;
title 'Stratified Cox Procedure - Stratified on Mothers Asthma Status';
title2 'Adjusted for mean values variables that satisfy the PH
assumption';
Run;title;

proc sgplot data = outputdataset;
title 'Log-log survival plots for Mothers Asthma Status groups';
title2 'Adjusted for variables that satisfy the PH assumption';
step x = days y = lls / group = MOMASTHMA;
run; title;

*PRENATAL_SMOKING;
proc phreg data = thesis_2 plots (overlay=row) = survival
covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV;
strata PRENATALSMK;
id family_id;
baseline covariates = thesis_meanvalues out = outputdataset survival =
s loglogs = lls;
title 'Stratified Cox Procedure - Stratified on Prenatal Smoking
Status';
title2 'Adjusted for mean values variables that satisfy the PH
assumption';
Run;title;

proc sgplot data = outputdataset;
title 'Log-log survival plots for Prenatal Smoking Status groups';
title2 'Adjusted for variables that satisfy the PH assumption';
step x = days y = lls / group = PRENATALSMK;
run;title;

*Mom educ;
proc phreg data = thesis_2 plots (overlay=row) = survival
covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV;
strata MOMEDU;
id family_id;
baseline covariates = thesis_meanvalues out = outputdataset survival =
s loglogs = lls;
title 'Stratified Cox Procedure - Stratified on Mothers Education';
title2 'Adjusted for mean values variables that satisfy the PH
assumption';
Run;title;

proc sgplot data = outputdataset;
title 'Log-log survival plots for Mothers Education groups';
title2 'Adjusted for variables that satisfy the PH assumption';
```

```
step x = days y = lls / group = MOMEDU;
run;title;



*b) A linear interaction term with time. That is, use an extended cox
model that contains all predictors and a product term of the form V*t,
where V denotes the variable that does not satisfy the PH assumption
and t denotes days.*g(t) = t;

*MJRCLUSTER;
proc phreg data = thesis_2 covs (aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV demcluster demcluster_t;
demcluster_t = demcluster*days;
id family_id;
run;
*CONCLUSION:WALD P VAULE FOR THE INTERACTION TERM WITH TIME =0.1350.
SUGGESTS PH IS NOT VIOLATED;

*CHILD_RACE;
proc phreg data = thesis_2 covs (aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV race race_t;
race_t=race*days;
id family_id;
run;
*CONCLUSION:WALD P VAULE FOR THE INTERACTION TERM WITH TIME =0.3242.
SUGGESTS PH IS NOT VIOLATED;

*MOM_ASTHMA;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV momasthma MOMASTHMA_t;
MOMASTHMA_t= MOMASTHMA*days;
id family_id;
run;
*CONCLUSION:P VALUE FOR THE INTERACTION TERM WITH TIME IS 0.0837.
SUGGESTS PH IS NOT VIOLATED;

*PRENATAL_SMOKING;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV prenatalsmk prenatalsmk_t;
prenatalsmk_t = prenatalsmk*days;
id family_id;
run;
*CONCLUSION:P VALUE FOR THE INTERACTION TERM WITH TIME IS 0.2793.
SUGGESTS PH IS NOT VIOLATED;

*MOM_EDUC;
proc phreg data = thesis_2 covs(aggregate);
model days*status(0) = rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV momedu momedu_t;
momedu_t = momedu*days;
id family_id;
run;
```

```
*CONCLUSION:P VALUE FOR THE INTERACTION TERM WITH TIME IS 0.9642.
SUGGESTS PH IS NOT VIOLATED;



*C)Schoenfeld residuals:  Using Schoenfeld residuals GOF tests,
evaluate whether the variables that do not satisfy the PH assumption
satisfies the PH assumption given the other variables in the model;

*MJRCLUSTER;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV demcluster;
output out=SR2_thesis2 ressch=SR_rline_pm25DV SR_rline_CODV
SR_rline_noxDV SR_maternal_ageDV SR_demcluster;
id family_id;
RUN;

data DEMCLUSTER_events2;
set SR2_thesis2;
if status=1;
RUN;

proc rank data=DEMCLUSTER_events2 out=ranked2_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked2_events nosimple;
var SR_rline_pm25DV SR_rline_CODV SR_rline_noxDV SR_maternal_ageDV
SR_demcluster;
with timerank;
RUN;
*P VALUE FOR THE GOF TEST IS 0.1166. SUGGESTS PH IS NOT VIOLATED;

*CHILD RACE;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV race;
output out=SR2_thesis2 ressch=SR_rline_pm25DV SR_rline_CODV
SR_rline_noxDV SR_maternal_ageDV SR_race;
id family_id;
RUN;

data Race_events2;
set SR2_thesis2;
if status=1;
RUN;

proc rank data=Race_events2 out=ranked2_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked2_events nosimple;
var SR_rline_pm25DV SR_rline_CODV SR_rline_noxDV SR_maternal_ageDV
SR_race;
```

```
with timerank;
RUN;
*P VALUE FOR THE GOF TEST IS 0.2980. SUGGESTS PH IS NOT VIOLATED;

*MOM ASTHMA;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV momasthma;
output out=SR2_thesis2 ressch=SR_rline_pm25DV SR_rline_CODV
SR_rline_noxDV SR_maternal_ageDV SR_momasthma;
id family_id;
RUN;

data asthma_events2;
set SR2_thesis2;
if status=1;
RUN;

proc rank data=asthma_events2 out=ranked2_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked2_events nosimple;
var SR_rline_pm25DV SR_rline_CODV SR_rline_noxDV SR_maternal_ageDV
SR_momasthma;
with timerank;
RUN;
*P VALUE FOR THE GOF TEST IS 0.0849 . SUGGESTS PH IS NOT VIOLATED;

*PRENATAL SMOKING;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV prenatalsmk;
output out=SR2_thesis2 ressch=SR_rline_pm25DV SR_rline_CODV
SR_rline_noxDV SR_maternal_ageDV SR_prenatalsmk;
id family_id;
RUN;

data prenatalsmk_events2;
set SR2_thesis2;
if status=1;
RUN;

proc rank data=prenatalsmk_events2 out=ranked2_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked2_events nosimple;
var SR_rline_pm25DV SR_rline_CODV SR_rline_noxDV SR_maternal_ageDV
SR_prenatalsmk;
with timerank;
RUN;
*P VALUE FOR THE GOF TEST FOR Prenatal smoking is 0.2252. SUGGESTS PH
NOT VIOLATED;
```

```
*MOTHERS EDUCATION;
proc phreg data=thesis_2 covs(aggregate);
model days*status(0)= rline_pm25DV rline_CODV rline_noxDV
maternal_ageDV momedu;
output out=SR2_thesis2 ressch=SR_rline_pm25DV SR_rline_CODV
SR_rline_noxDV SR_maternal_ageDV SR_momedu;
id family_id;
RUN;

data momedu_events2;
set SR2_thesis2;
if status=1;
RUN;

proc rank data=momedu_events2 out=ranked2_events ties=mean;
var days;
ranks timerank;
RUN;

Proc corr data=ranked2_events nosimple;
var SR_rline_pm25DV SR_rline_CODV SR_rline_noxDV SR_maternal_ageDV
SR_momedu;
with timerank;
RUN;
*P VALUE FOR THE GOF TEST FOR Prenatal smoking is 0.8903. SUGGESTS PH
NOT VIOLATED;


*FINAL CONCLUSION: After assessment using the Adjusted log log curves,
there are no variables that do not satisfy the
PH assumption now is;


**************************************************************************;

/* Next, we will examine the possible interaction of each factor with
PM2.5 exposure (then similarily for CO and NOX as
well), one at a time in unadjusted analyses. There might be important
effect modification to consider.*/
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure;
*B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
*C) Using the log rank results and the cumulative "failure probability"
in the plots comment on the possibility of sig effect modification for
each of the covariates. How might you critique the results?;

*Examination of interaction between PM2.5 and gender;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
```

```
data thesis_gender;
set thesis_2;
RUN;

proc sort data=thesis_gender;
by gender;
RUN;

proc lifetest data=thesis_gender plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by gender;
RUN;
*There is no effect modification by gender because the p values for
each strata are both non significant.;

*Examination of interaction between PM2.5 and mom_educ;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
data thesis_momedu;
set thesis_2;
RUN;

proc sort data=thesis_momedu;
by momedu;
RUN;

proc lifetest data=thesis_momedu plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by momedu;
RUN;
*There is no effect modification by mothers edu because the p values
for each strata are both non significant.;

*Examination of interaction between PM2.5 and maternal_ageDV;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
data thesis_maternal_ageDV;
set thesis_2;
RUN;

proc sort data=thesis_maternal_ageDV;
by maternal_ageDV;
RUN;

proc lifetest data=thesis_maternal_ageDV plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
```

```
by maternal_ageDV;
RUN;
*There is no effect modification by maternal age because the p values
for each strata are both non significant.;

*Examination of interaction between PM2.5 and mjrcluster;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
data thesis_demcluster;
set thesis_2;
RUN;

proc sort data=thesis_demcluster;
by demcluster;
RUN;

proc lifetest data=thesis_demcluster plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by demcluster;
RUN;
*There is no effect modification by demographic cluster because the p
values for each strata are both non significant.;

*Examination of interaction between PM2.5 and prenatal_smoking;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
data thesis_prenatalsmk;
set thesis_2;
RUN;

proc sort data=thesis_prenatalsmk;
by prenatalsmk;
RUN;

proc lifetest data=thesis_prenatalsmk plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by prenatalsmk;
RUN;
*There is no effect modification by prenatal smoking because the p
values for each strata are both non significant.;

*Examination of interaction between PM2.5 and child_race;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
```

```
those with and without exposure;
data thesis_race;
set thesis_2;
RUN;

proc sort data=thesis_race;
by race;
RUN;

proc lifetest data=thesis_race plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by race;
RUN;
*There is no effect modification by childs race because the p values
for each strata are both non significant.;

*Examination of interaction between PM2.5 and mom_asthma;
*A) Provide stratified KM failure plots (ie. without adjustment in a
regression model), showing for each level of each potential effect
modifier (considered separately), the comparison between those with and
without exposure.
B) Conduct appropriate log rank tests within the strata that compare
those with and without exposure;
data thesis_momasthma;
set thesis_2;
RUN;

proc sort data=thesis_momasthma;
by momasthma;
RUN;

proc lifetest data=thesis_momasthma plots=survival(failure);
time days*status(0);
strata rline_PM25DV;
by momasthma;
RUN;
*There is no effect modification by mothers asthma because the p values
for each strata are both non significant.
However the p value for momasthma=0 strata is barely non significant so
there could be potential interaction;


*****************************************************************;

/*Regardless of your conclusions from earlier questions, fit a
stratified cox PH model with PM2.5 as the E variable and control
simultaneously for the other variables as both V and W variables, where
all two way(E*W) EM terms are considered. Assume that no variables
violate the PH assumption based on adjusted and unadjusted analysis so
no variables need to be stratified upon in the strata statement;
A) State the form of the hazard function you have used to fit this
model
B) Obtain collinearity diagnostics (based on the inverse of the
information matrix) for this model. Proceed from this point to examine
collinearity for these data, modifying the model and obtaining
additional collinearity diagnostics as you consider appropriate
```

```
C) What do you conclude about collinearity for these data and what if
anything do you recommend be done to remedy any collinearity problem
found?

*A) State the form of the hazard function you have used to fit the
model for PM2.5;
*First redefine dummy variables to include all levels of the category,
not just a dichotomized version;

data thesis_3;
set thesis_2;

if gender='F' then genderDV=1;
else genderDV=0;
*females=1, the reference is M. Males is the largest category;

if mjrcluster='B' then mjrclusterB=1;
else mjrclusterB=0;
if mjrcluster='C' then mjrclusterC=1;
else mjrclusterC=0;
if mjrcluster='D' then mjrclusterD=1;
else mjrclusterD=0;
*The reference is A. A is the largest category and the one that makes
the most sense because it is the highest SES and
thus we expect the HR to be the smallest for this group;

if child_race='Black' then child_raceB=1;
else child_raceB=0;
if child_race='Other' then child_raceO=1;
else child_raceO=0;
if child_race='Unknown' then child_raceU=1;
else child_raceU=0;
*The reference is White. This is the largest category;

if mom_asthma='missing' then mom_asthmamissing=1;
else mom_asthmamissing=0;
if mom_asthma='yes' then mom_asthmaYes=1;
else mom_asthmaYes=0;
*The reference is No. This is the largest category and the one that
makes the most sense because children whose mothers
do not have asthma would be potentially less at risk;

if mom_educ='<12th grade' then mom_LT12=1;
else mom_LT12=0;
if mom_educ='HS/GED' then mom_HS=1;
else mom_HS=0;
if mom_educ='missing' then mom_missingedu=1;
else mom_missingedu=0;
*The reference is some college+. This is the largest category and the
category that makes the most sense because children
whose mothers are more educated might be less at risk;

if prenatal_smoking='missing' then prenatalsmokingmissing=1;
else prenatalsmokingmissing=0;
if prenatal_smoking='yes' then prenatalsmokingYes=1;
else prenatalsmokingYes=0;
*The reference is No. This is the largest category and the category
```

```
that makes the most sense because children whose
mothers did not smoke prenatally might be less at risk;

if ITP_only10=1 then Location1=1;
else location1=0;
if ITP_only10=2 then Location2=1;
else location2=0;
*The reference is 0= >10 miles outside the perimeter. This makes the
most sense because it is the largest category
and would potentially be where children are the least exposed to air
pollution so their risk might be less there;

RUN;


*Cross tabs to check coding of the categorical variables;
Proc freq data=thesis_3;
tables gender*genderDV
mjrcluster*mjrclusterB*mjrclusterC*mjrclusterD
child_race*child_raceB*child_raceO*child_raceU
mom_asthma*mom_asthmamissing*mom_asthmaYes
mom_educ*mom_LT12*mom_HS*mom_missingedu
prenatal_smoking*prenatalsmokingmissing*prenatalsmokingYes
ITP_only10*location1*location2/ list;
RUN;


*Developing the model;
data thesis_PM25;
set thesis_3;
PMZ1=rline_pm25*genderDV;
PMZ2=rline_pm25*mom_asthmamissing;
PMZ3=rline_pm25*mom_asthmaYes;
PMZ4=rline_pm25*prenatalsmokingmissing;
PMZ5=rline_pm25*prenatalsmokingYes;
PMZ6=rline_pm25*location1;
PMZ7=rline_pm25*location2;
RUN;

*A)form of the hazard function=
hg(t,x)=h0g(t)*exp[(B*rline_pm25)+(B11*genderDV)+(B12*child_raceB)+(B13
*child_raceO)+(B14*child_raceU)+(B15*mom_asthmamissing)+(B16*mom_asthma
Yes)+(B17*mom_LT12)+(B18*mom_HS)+(B19*mom_missingedu)+(B111*prenatalsmo
kingmissing)+(B112*prenatalsmokingYes)+(B113*location1)+(B114*location2
)+ (B115*maternal_ageDV) +
(B21*PMZ1)+(B22*PMZ2)+(B23*PMZ3)+(B24*PMZ4)+(B25*PMZ5)+(B26*PMZ6)+(B27*
PMZ7)];
*No need to stratify on any variables after unadjusted and adjusted
analysis;

proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1 PMZ2 PMZ3 PMZ4 PMZ5 PMZ6 PMZ7/RL;
```

```
id family_id;
RUN;
*The standard errors when the covs option is added: 0.18440, 0.09549,
0.07777, 0.14080, 0.12094, etc;

proc phreg data=thesis_PM25;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1 PMZ2 PMZ3 PMZ4 PMZ5 PMZ6 PMZ7/RL;
RUN;
*The standard errors when the covs option is not added: 0.18721,
0.09481, 0.07653, 0.13891, 0.12055, etc;
*The standard errors change when the covs option is added. When it is
not there, the standard errors are a little larger
overall compared to when they are there.;

*****************************************************************;

*B) Obtain collinearity diagnostics (based on the inverse of the
information matrix) for this model. Proceed from this
    point to examine collinearity for these data, modifying the model
and obtaining additional collinearity diagnostics
      as you consider appropriate;

*Full model;
%include 'S:\course\Epi740\MACRO\collin_2011.sas';
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1 PMZ2 PMZ3 PMZ4 PMZ5 PMZ6 PMZ7/RL;
id family_id;
RUN;
%collin(covdsn=info, output=outcr);
RUN;
*There are 6 CIs that are over 10, indicating at least one collinearity
issue;

*Try rerunning the model without PMZ6 (Highest CI and has a VDP of
0.8821)and also must drop PMZ7 because both are
dummy variables for ITP_only10 (city region);
%include 'S:\course\Epi740\MACRO\collin_2011.sas';
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1 PMZ2 PMZ3 PMZ4 PMZ5 /RL;
id family_id;
RUN;
%collin(covdsn=info, output=outcr);
```

```
RUN;
*There are 4 (maybe 5) CIs that are over 10, indicating at least one
collinearlity issue;

*Try rerunning the model without PMZ4 (Highest CI and has a VDP of
0.8460) and also must drop PMZ5 because both are
dummy variables for prenatal_smoking;
%include 'S:\course\Epi740\MACRO\collin_2011.sas';
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1 PMZ2 PMZ3/RL;
id family_id;
RUN;
%collin(covdsn=info, output=outcr);
RUN;
*There are 3 CIs that are over 10, indicating at least one
collinearlity issue;

*Try rerunning the model without PMZ3 which has the highest VDP of the
product terms at 0.7606 and also must drop PMZ2
because they are dummy variables for mothers asthma;
%include 'S:\course\Epi740\MACRO\collin_2011.sas';
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1/RL;
id family_id;
RUN;
%collin(covdsn=info, output=outcr);
RUN;
*There are 1 CIs that are over 10 (but just barely over 10 with a CI of
10.998), indicating at least one potential
collinearlity issue;

*Try rerunning the model without PMZ1 (the product term for
gender)which a high product term VDP of 0.9865;
%include 'S:\course\Epi740\MACRO\collin_2011.sas';
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV /RL;
id family_id;
RUN;
%collin(covdsn=info, output=outcr);
RUN;
*There are no remaining CIs that are over 10, indicating no further
collinearity issue once all the product terms drop;
```

```
*Thus the final model includes the terms following terms and does not
include product terms:
rline_PM25, gender, ITP_only10, mom_educ, maternal_age,
mjrcluster,prenatal_smoking, child_race and mom_asthma;

*****************************************************************;

*Based on the interaction analysis above, there was an indication that
mom_asthma might be an important effect modifier;
*Run a liklihood ratio test to determine if mom_asthma terms are
significant in the model, even if they have collinearity
issues;

*Full model;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ2 PMZ3/RL;
id family_id;
RUN;
*The -2logL for this full model is 43253.082;

*Reduced model;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
RUN;
*The -2logL for this reduced model is 43254.358;

data p_momasthma;
pvalue=1-probchi(43254.358-43253.082,2);
RUN;

proc print data=p_momasthma;
RUN;

*Test statistic: -2lnL(reduced)-(-2lnL(full))= 43254.358-
43253.082=1.276~chisquare with 2 df under the null;
*p value under the null is 0.59 (nonsignificant);
*Null hypothesis: PMZ2=PMZ3=0;
*The test statistic of 1.276 is non significant at the 0.05 level, thus
we do not have evidence to conclude that
there is a significant interaction effect based on this chunk test and
that the interaction terms can be dropped from
the model;


*Furthermore, for the sake of a priori criteria based on the
literature, there is evidence that gender, mothers asthma
status, prenatal smoking status, and city region may also be important
```

```
effect modifiers with the exposure;
*We will do a chunk test to examine if there is significant effect
modification by these variables in our sample;

*Gender interaction LRT test;
*Full model for gender;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ1/RL;
id family_id;
RUN;
*The -2lnL for the full model is 43252.445;

*Reduced model for gender;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
RUN;

data p_gender;
pvalue=1-probchi(43254.358-43252.445,1);
RUN;

proc print data=p_gender;
RUN;
*The -2lnL for the reduced model is 43254.358;
*Test statistic: -2lnL(reduced)-(-2lnL(full))= 43254.358-
43252.445=1.913~chisquare with 1 df under the null;
*The p value under the null is 0.17 (nonsignificant);
*Null hypothesis: PMZ1=0;
*The test statistic of 1.913 is non significant at the 0.05 level, thus
we do not have evidence to conclude that
there is a significant interaction effect based on this chunk test and
that the interaction terms can be dropped from
the model;

*Prenatal_smoking LRT test;
*Full model for prenatal smoking;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ4 PMZ5/RL;
id family_id;
RUN;
*The -2lnL for the full model is 43253.618;
```

```
*Reduced model for prenatal smoking;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
RUN;

data p_smoking;
pvalue=1-probchi(43254.358-43253.618,2);
RUN;

proc print data=p_smoking;
RUN;
*The -2lnL for the reduced model is 43254.358;
*Test statistic: -2lnL(reduced)-(-2lnL(full))= 43254.358-
43253.618=0.72~chisquare with 2 df under the null;
*The p value under the null is 0.69 (nonsignificant);
*Null hypothesis: PMZ4=PMZ5=0;
*The test statistic of 0.72 is non significant at the 0.05 level, thus
we do not have evidence to conclude that
there is a significant interaction effect based on this chunk test and
that the interaction terms can be dropped from
the model;

*City region (ITP_only10) LRT test;
*Full model for city region;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV
PMZ6 PMZ7/RL;
id family_id;
RUN;
*The -2lnL for the full model is 43253.421;

*Reduced model for city region;
proc phreg data=thesis_PM25 covs(aggregate)covout outset=info;
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
RUN;

data p_cityregion;
pvalue=1-probchi(43254.358-43253.421,2);
RUN;

proc print data=p_cityregion;
RUN;
*The -2lnL for the reduced model is 43254.358;
```

```
*Test statistic: -2lnL(reduced)-(-2lnL(full))= 43254.358-
43253.421=0.937~chisquare with 2 df under the null;
*The p value under the null is 0.63 (nonsignificant);
*Null hypothesis: PMZ6=PMZ7=0;
*The test statistic of 0.937 is non significant at the 0.05 level, thus
we do not have evidence to conclude that
there is a significant interaction effect based on this chunk test and
that the interaction terms can be dropped from
the model;

*Thus there are no interaction terms that contribute significantly to
the model, suggesting the final model will be
a no interaction model;

*********************************************************************;

*Using all possible subsets change in estimate approach, carry out
confounding/precision assessment, summarizing the
results in a table. Make a conclustion about which model is "best";

*This is assumed to be the GS full model;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'HR for rline_PM25 in the GS model' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 in the gold standard model, HR= 1.17, +-10% range
of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54;*CI ratio is 1.58;
*This is the gold standard;

*Try rerunning the model without mnrcluster;
proc phreg data=thesis_PM25 covs(aggregate);
model days*status(0)=rline_pm25 genderDV child_raceB child_raceO
child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop MNRCLUSTER' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop mnrcluster, HR= 1.08, +-10% range of
GS=(1.05, 1.29);
*Confidence interval: (0.88, 1.33);
*CI width is 0.45;*CI ratio is 1.51;

*Try rerunning the model without genderDV;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster child_raceB child_raceO
child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
```

```
id family_id;
contrast 'Drop genderDV' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop genderDV, HR= 1.17, +-10% range of
GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54; *CI ratio is 1.58;

*Try rerunning the model without child_race;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV mom_asthmamissing
mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop Child_race' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop child_race, HR= 1.10, +-10% range of
GS=(1.05, 1.29);
*Confidence interval: (0.88, 1.38);
*CI width is 0.50;*CI ratio is 1.57;

*Try rerunning the model without mom_asthma;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop mom_asthma' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop mom_asthma, HR= 1.18, +-10% range of
GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try rerunning the model without mom_educ;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
prenatalsmokingmissing prenatalsmokingYes location1 location2
maternal_ageDV/RL;
id family_id;
contrast 'Drop mom_educ' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop mom_educ, HR= 1.18, +-10% range of
GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.48);
*CI width is 0.55;*CI ratio is 1.59;

*Try rerunning the model without prenatal_smoking;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
```

```
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop prenatal_smoking' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop prenatal_smoking, HR= 1.17, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54;*CI ratio is 1.58;

*Try rerunning the model without city region;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes maternal_ageDV/RL;
id family_id;
contrast 'Drop city region' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop city region, HR= 1.11, +-10% range
of GS=(1.05, 1.29);
*Confidence interval: (0.90, 1.36);
*CI width is 0.46;*CI ratio is 1.51;

*Try rerunning the model without maternal_ageDV;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2/RL;
id family_id;
contrast 'Drop maternal_ageDV' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop maternal_ageDV, HR= 1.17, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54; *CI ratio is 1.58;


*Try dropping two variables at a time during all possible subsets
testing;

*Try dropping gender and asthma;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster child_raceB child_raceO
child_raceU
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop gender and asthma' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop gender and asthma, HR=1.18, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
```

```
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping gender and education;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster child_raceB child_raceO
child_raceU mom_asthmamissing mom_asthmaYes
prenatalsmokingmissing prenatalsmokingYes location1 location2
maternal_ageDV/RL;
id family_id;
contrast 'Drop gender and education' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop gender and education, HR= 1.18, +-
10% range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping gender and smoking;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster child_raceB child_raceO
child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop gender and smoking' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop gender and smoking, HR= 1.17, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping gender and age;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster child_raceB child_raceO
child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 /RL;
id family_id;
contrast 'Drop gender and age' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop gender and age, HR= 1.17, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54;*CI ratio is 1.58;

*Try dropping asthma and education;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU
 prenatalsmokingmissing prenatalsmokingYes location1 location2
maternal_ageDV/RL;
id family_id;
contrast 'Drop asthma and education' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop asthma and education, HR= 1.18, +-
```

```
10% range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.49);
*CI width is 0.55;*CI ratio is 1.59;

*Try dropping asthma and smoking;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU
mom_LT12 mom_HS mom_missingedu location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop asthma and smoking' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop asthma and smoking, HR= 1.18, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping asthma and age;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2/RL;
id family_id;
contrast 'Drop asthma and age' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop asthma and age, HR= 1.18, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping education and smoking;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
 location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'Drop education and smoking' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop education and smoking, HR= 1.18, +-
10% range of GS=(1.05, 1.29);
*Confidence interval: (0.94, 1.48);
*CI width is 0.54;*CI ratio is 1.57;

*Try dropping education and age;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
 prenatalsmokingmissing prenatalsmokingYes location1 location2 /RL;
id family_id;
contrast 'Drop education and age' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop education and age, HR= 1.18, +-10%
```

```
range of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.48);
*CI width is 0.55;*CI ratio is 1.59;

*Try dropping smoking and age;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu location1 location2/RL;
id family_id;
contrast 'Drop smoking and age' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 when we drop smoking and age, HR= 1.17, +-10%
range of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54;*CI ratio is 1.58;


*****************************************************************;


*This is assumed to be the GS full model and best model;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
contrast 'HR for rline_PM25 in the GS model' rline_PM25 1/estimate=exp;
RUN;
*Effect of rline_PM25 in the gold standard model, HR= 1.17, +-10% range
of GS=(1.05, 1.29);
*Confidence interval: (0.93, 1.47);
*CI width is 0.54;*CI ratio is 1.58;
*This is the gold standard;

*Estimate of the HR in the unadjusted model;
proc phreg data=thesis_PM25 covs(aggregate);
model days*status(0)=rline_pm25 /RL;
id family_id;
contrast 'HR for rline_PM25 in the unadjusted model' rline_PM25
1/estimate=exp;
RUN;
*Effect of rline_PM25 in the unadjusted model, HR= 0.9824, +-10% range
of GS=(1.05, 1.29);
*Confidence interval: (0.84, 1.15);
*CI width is 0.31;*CI ratio is 1.37;
*This is the gold standard;

*KM curves for the best model;
proc lifetest data=thesis_PM25 method=KM plots=s;
time days*status(0);
title 'No interaction, KM curves for PM2.5';
title2 'Non stratified but adjusted for SES cluster, Gender, Childs
Race, Mothers Asthma, Mothers Education, Maternal Smoking, Maternal Age
```

```
and City region';
RUN;
TITLE;

*HR by gender. The HR is very slightly highest for females than for
males in this analysis;
proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
Where genderDV=1;
contrast 'HR for rline_PM25 for females' rline_PM25 1/estimate=exp;
RUN;
*HR effect estimate for females: 1.1628, CI=(0.8234 1.6422);

proc phreg data=thesis_PM25 covs(aggregate);
class mnrcluster (param=ref ref='A.1');
model days*status(0)=rline_pm25 mnrcluster genderDV child_raceB
child_raceO child_raceU mom_asthmamissing mom_asthmaYes
mom_LT12 mom_HS mom_missingedu prenatalsmokingmissing
prenatalsmokingYes location1 location2 maternal_ageDV/RL;
id family_id;
Where genderDV=0;
contrast 'HR for rline_PM25 for males' rline_PM25 1/estimate=exp;
RUN;
*HR effect estimate for males: 1.1454, CI=(0.8487 1.5458);
```