In this report we will look back at all the progress we made following the first quarter report. We will also work on a step by step document for the R data management and analysis. But first, let's revise what we have accomplished in the first quarter;

First quarter accomplishments

- Reviewed the evidence from the literature for biological plausibility and significance of association
- Reviewed the burden of disease model
- Reviewed the exposure model (LUR model)
- Summarized the exposure estimates and provided plots (tables and graphs needs updating)
- Reviewed US Census system and provided a summary of available data (tables need updating)
- Discovered that exposure concentrations were at the census centroid

Main R Scripts written during the first quarter:

- NO2 summary statistics.R
    - This scripts provides summary statistics for N02 exposures using the "Processed data set" for both years 2000, 2010. The data produced was used to populate (table 6) for the first quarter report.
    - To estimate population weighted averages of exposure for year 2000, two other data set were used (NHGIS population data + website exposure values)
- Other scripts include those used to read different census data downloaded from NHGIS and scripts used for preliminary exploratory data analysis comparing exposure to the census data.

Work in progress for the second quarter

Weekly report 3; comparing census data with exposure data

In this week we compared the available census and exposure data sets, summarized the exposure values for each data set and examined the nature of missing observations. We then made a final decision on which data sets to use for the year 2000 (website data set) and year 2010 (processed data set), and used census data downloaded from NHGIS.

Data sets used (see Attachment)

- Processed data set (for NO2 exposure values at census block with 2010 geographical framework)
- Annual Average "website" data set (for NO2 exposure values at census block with 2000 geographical framework)
- NHGIS census data sets for years 2000 & 2010 which included the available variables:
    - Total population (block level)
    - Race by age group by gender (block level)
    - Urban vs Rural (block level)
    - Household and family income (block group level)
    - Personal income (block group level)
    - Educational attainment >25 years of age by gender (block group level)

Table 1. - Summary for the data sets used

	Processed		AA (website)	
	2000	2010	2000	2010
True # of blocks	8,164,718	11,007,989	8,164,718	11,007,989
Observations	11,007,989	11,007,989	8,151,762	8,151,762
Blocks not covered	N/A	114,332	12,956	N/A
Blocks with population not covered	N/A	0	628	N/A
Mean NO2 ppb	9.69	6.32	9.94	6.47
Mean NO2 ug/m3	18.22	11.88	18.69	12.16

We also concluded that to go forward with the analysis we will be using two data sets for the exposure values, since each data set uses different geographical framework for the census blocks:

- For year 2000 the Annual Average "website" data set
- For year 2010 the Processed data set

We also examined the nature of missing observations, for the year 2010 processed data set we had 114,332 missing observations, however, none of the observations had any population counts within them and no further steps will be taken for the 2010 data set. For the year 2000 "website" data set we had 12,956 missing observations of which 628 blocks had population counts, and the sum of the population without exposure values mounted to 18,212. We also contacted Mathew regarding the missing values, he will update the data set by Wednesday.

Main R scripts written:

- Comparing NO2 with Census data 2000
    - In this script we cleaned the data set, changed column names, and merged exposure data set with NHGIS census data for 2000
    - We then compared the blocks with missing exposure values and summarized them.

Weekly report 4; LUR and Incidence papers summary, Burden estimation, Abstract submission

In this week, we searched for other land use regression models to use for future studies, we also looked into alternate methods to examine traffic related pollution exposure. We also examined incidence rates for asthma in the US. We completed a preliminary estimate of the burden of asthma and submitted an abstract. (See attachment for summaries of papers.)

Summary of LUR papers:

- A census in the US near-roadway population: Public health and environmental justice considerations. (*Gregory M. Rowangould)*
    - Estimate size and distribution of the population living near high volume roads in the US

    - Distance to highway & traffic density used to model exposure
    - Investigate race and income (median household income) disparities near roadway populations.
    - Examine coverage of national ambient monitors network
    - Unit of analysis was the census block for the years 2000 and 2010
- Spatial Modeling of PM10 and NO2 in the Continental US 1985-2000 (Jaime E. Hart et al.)
- Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression (*Michael T. Young)*
- A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology (*Paul D. Sampson)*
- A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States (*Bernardo S. Beckerman)*

Conclusion of summary:

- All the exposure models except (*Gregory's)* paper did not cover the desired years 2000 & 2010.
- I searched for NO2 and PM models only, will have to look for other pollutants.
- In our study we want to estimate **traffic** air pollution by using regression models that assume the true values of exposure to be the values obtained from EPA monitors for ambient air pollution (model builders try to fit their models (using R^2) to EPA monitor data), however, there is evidence that EPA monitors do not describe well air pollution sourced or related to traffic, since they are not positioned near traffic in most counties (*Gregory M. Rowangould)*
- We don't know if current regression models are underestimating or overestimating the true NO2 levels near roadways, one way to examine this is to first measure NO2 levels with monitors near or around residential areas then examine their classification into either exposed to high traffic or low traffic then compare them with the results from the LUR models.
- The use of a categorizing method (blocks are categorized as either far or close to traffic) would possibly be a better representation of the true exposure to traffic related air pollution, however, we don't have a concentration response function for this method.
- Summary of individual studies is available in the annex

Summary of Incidence and prevalence papers:

We currently have one paper with asthma incidence estimates, we also have a couple of reports for asthma prevalence.

Incidence:

Asthma Incidence among Children and Adults: Findings from the Behavioral Risk Factor Surveillance System Asthma Call-back Survey—United States, 2006–2008

Rachel A. Winer, Xiaoting Qin, Theresa Harrington, Jeanne Moorman & Hatice Zahran

Aim:

- Estimate annual asthma incidence rate and determine if they differ by age group, sex, and race/ethnicity.

Method:

- Analyzed the Behavioral Risk Factor Surveillance System (BRFSS) data and the Asthma Call Back Survey (ACBS). Note: BRFSS includes all 50 states, DC, Puerto Rico, US Virgin Islands, and Guam, however not all participated in the ACBS (check figure 1)
- Definition of incidence case: children diagnosed with asthma by healthcare provider within 12 months prior to survey participation.

Results:

- Number of Participating states were 24 states & DC in 2006, 34 states and DC in 2007 and 2008. (these include Alaska & Hawaii which we did not include in our analysis)
- Asthma incidence rate among at-risk children in areas participating in the ACBS was 12.5 per 1000 during 2006-2008.
- Stratified Asthma incidence rate among at-risk children in areas participating in the ACBS was 13.6, 9.9 & 14.4 per 1000 for the years 2006, 2007, 2008 respectively.
- Weighted population estimates of child incident cases were 435,847 in 2006, 440,575 in 2007 and 623,291 in 2008 **in ACBS participating states**.
- Year to year incidence can't be compared due to different number of participating states each year (Check figure 1)

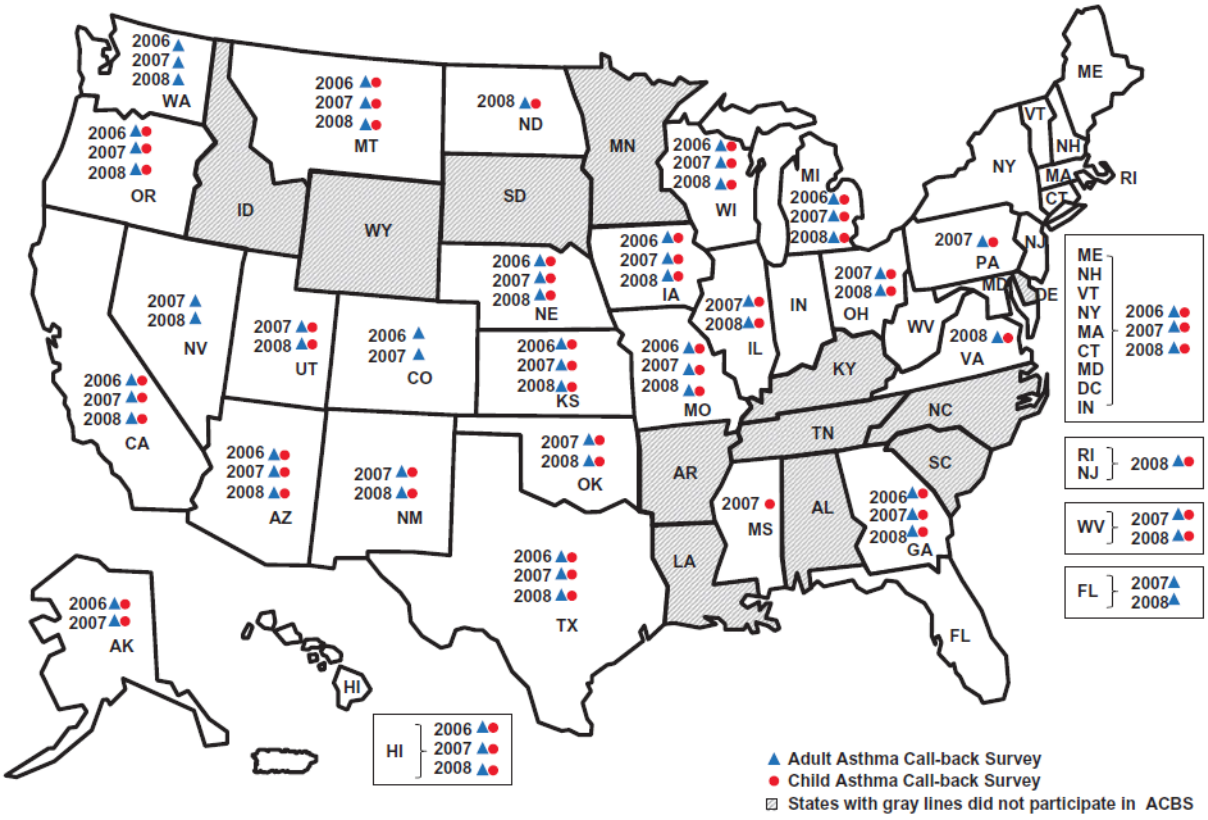

FIGURE 1.—States participating in Adult and Child Asthma Call-back Survey (ACBS), 2006–2008.

TABLE 2.—Estimated annual asthma incidence rate per 1000 at-risk children by year, sex, age group, and race/ethnicity—Behavioral Risk Factor Surveillance System (BRFSS) child Asthma Call-back Survey (ACBS), 2006–2008.

Annual incidence rate over rate period, 2006–2008 (n = 164,327)[a]	Number[b]	Rate[c]	95% CI
Total	592	12.5	(10.5,14.4)
Year			
2006 (22 states + DC, n = 38,004)[d]	172	13.6	(9.6,17.6)
2007 (30 states + DC, n = 60,031)	186	9.9	(7.4,12.4)
2008 (31 states + DC, n = 66,292)	234	14.0	(10.2,17.7)
Sex			
Male	313	13.6	(10.5,16.6)
Female	277	11.5	(8.7,14.2)
Age (years)			
0–4	265	23.4	(17.8,28.9)
5–11	191	11.1	(8.1,14.2)
12–17	131	4.4	(2.9,5.9)
Race[e]/ethnicity			
NH White	408	12.8	(10.3,15.3)
NH Non-White	110	15.8	(10.0,21.7)
Hispanic	68	9.9	(6.0,13.8)

Prevalence:

We have a couple of reports produced by the CDC:

1- National Surveillance of Asthma: United States, 2001-2010. (https://www.cdc.gov/nchs/data/series/sr_03/sr03_035.pdf)

2- Summary Health Statistics for U.S. Children: National Health Interview Survey (1999-2010). (https://www.cdc.gov/nchs/products/series/series10.htm)

Year	Ever Asthma	Current Asthma
1999	10.8%	5.3%
2000	12.4%	5.5%
2001	12.7%	5.7%
2006	13.6%	9.4%
2007	13.9%	9.5%
2008	13.1%	9.1%
2009	13.9%	9.7%
2010	13.7%	9.5%
2011	14.2%	9.6%

Table 3 – Prevalence of asthma among children in the US (age adjusted rates for years 2006-2011) .
Source: US census, Summary Health Statistics for U.S. Children: National Health Interview Survey, National Surveillance of Asthma: United States, 2001–2010

Burden Estimation (Results)

	2000	2010
Total population[1]	279,583,437	306,675,006
Total children[1]	71,807,328	73,690,271
Prevalence among children (using Ever Asthma)[2,3]	12.4%	13.7%
Number of prevalent cases (Using Ever Asthma) [2,3]	9,119,531	10,095,567
Prevalence among children (using Current Asthma[2,3])	6%	10%
Number of prevalent cases (Using Current Asthma) [2,3]	3,949,403	7,000,576
Asthma Incidence among children[4]	12.5 / 1,000	12.5 / 1,000
Total incident cases of asthma	897,592	921,128
cases of asthma attributable to NO2	238,651	164,463
cases of asthma attributable to NO2 (with prevalence adjustment)	209,058	141931
Mean NO2 ppb[5]	9.94	6.32
Mean NO2 ug/m3[5]	18.69	11.88

Table 4 – Summary of burden results. Sources: [1]US census, [2]Summary Health Statistics for U.S. Children: National Health Interview Survey, [3]National Surveillance of Asthma: United States, 2001–2010, [4]*Rachel A. Winer et al 2012,* [5]*Bechle et al 2015,* [6]*Khreis et al 2016.*

Looking forwards

Before moving forward with the analysis there are a couple of points we need to look into:

- Re-examine the calculated number of asthma incident cases (compare with literature)
- Re-think of what prevalence to use (Ever is more conservative)

We can also adjust are analysis by incorporating available information mainly the incidence rates for different strata of the children population, information we have is available in table 5. We can also search for state specific incidence rates in the literature.

Variable	Incidence rate per 1000	95% CI	Difference within strata
Age group			
0-4	23.4	(17.8-28.9)	Significant
5-11	11.1	(8.1-14.2)	
12-17	4.4	(2.9-5.9)	
Sex			
Male	13.6	(10.5-16.6)	Not significant
Female	11.5	(8.7-14.2)	
Race/Ethnicity			
NH White	12.8	(10.3-15.3)	Not significant
NH Non-White	15.8	(10.0-21.7)	
Hispanic	9.9	(6.0-13.8)	

Table 5 – Strata specific incidence rate. Source: *Rachel A. Winer et al 2012*

Attachment

Summary of Exposure Papers

(1) A census in the US near-roadway population: Public health and environmental justice considerations

Gregory M. Rowangould

Aim:

- Estimate size and distribution of the population living near high volume roads in the US
- Investigate race and income (median household income) disparities near roadway populations.
- Examine coverage of national ambient monitors network

Method:

- Census blocks are classified by (traffic density & proximity to road) using 2008 traffic data and 2000/2010 census, population exposure is assigned based on the % of the census block within the buffer area of the road.
- For income level estimation, only block group level is available, therefore, blocks are assigned to block group values.
- Average annual daily traffic volume (AADT) for 2008 was obtained through highway performance monitoring system (HPMS), classifications starts from 25,000 AADT up to 200,000 AADT with increments of 25,000 AADT
- Proximity is measured using series of distance of buffers (100m intervals up to 500m)
- Census blocks with no population or traffic density are discarded

Results:

- 52..2% of populated census blocks have **some level** of traffic density
- 19% live near **high volume roads**
- Non-white & low income household have greater traffic volume
- 84% of counties show some level of exposure disparity.
- Most counties with residents living near high volume roads do not have regulatory monitors (EPA).

Discussion:

- Traffic density is a continuous variable, no need to classify traffic density & proximity categories
- The link between traffic density and emission concentrations is less understood

(2) Spatial Modeling of PM10 and NO2 in the Continental US 1985-2000

Jaime E. Hart et al.

Aim:

- National model of annual PM10 and NO2 for the years 1985-2000

Method:

- Mixture of smoothing and generalized additive models using annual monitoring data and GIS covariates
- Covariates included : distance to road, elevation, % low-intensity residential, % high-intensity residential, industrial, commercial or transportation land use,

Results:

- R^2 = 0.49 for PM10
- R^2 = 0.88 for NO2

Discussion:

- The model describes in accuracy **ambient** air pollution
- However, the years are not appropriate for our study

(3) Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression

Michael T. Young,,† Matthew J. Bechle,‡ Paul D. Sampson,⊥ Adam A. Szpiro,§ Julian D. Marshall,‡ Lianne Sheppard,§,∥ and Joel D. Kaufman*

Aim:

- Build a predictive model of NO2 on a national level using a mixture of satellite and land-use regression/spatial smoothing

Method:

- The use of satellite data, land use data, GIS data and EPA monitor data
- GIS covariates include: proximity to roads, commercial areas, airports, railroads, railyards, … and many other covariates (>800)
- They used dimension reduction to reduce the number of covariates, testing out (2, 3, 4 and 5) components.
- Satellite images have a resolution of 13*24 km^2!!
- Used kriging

Results

- Predicted NO2 in a 25*25 km^2 national grid for years 1990, 2006.

Discussion

- Spatial resolution is course (13*24 km), satellite data cannot provide a better resolution.

(4) A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology

Paul D. Sampson a,, Mark Richards b , Adam A. Szpiro c , Silas Bergen c , Lianne Sheppard d , Timothy V. Larson e , Joel D. Kaufman d*

Aim:

- Build a national model to predict PM2.5 at a fine scale for year 2000

Discussion:

- Does not cover target years

(5) A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States

Bernardo S. Beckerman,,† Michael Jerrett,† Marc Serre,‡ Randall V. Martin,§ Seung-Jae Lee,∥ Aaron van Donkelaar,§ Zev Ross,⊥ Jason Su,† and Richard T. Burnett*

Aim:

- Build a national model to predict PM2.5 using data from 1999-2008

Discussion:

- Does not cover target years

Conclusions and thoughts:

- All the exposure models except (*Gregory's)* paper did not cover the desired years 2000 & 2010.
- I searched for NO2 and PM models only, will have to look for other pollutants.
- In our study we want to estimate **traffic** air pollution by using regression models that assume the true values of exposure to be the values obtained from EPA monitors for ambient air pollution (model builders try to fit their models (using R^2) to EPA monitor data), however, there is evidence that EPA monitors do not describe well air pollution sourced or related to traffic, since they are not positioned near traffic in most counties (*Gregory M. Rowangould)*
- We don't know if current regression models are underestimating or overestimating the true NO2 levels near roadways, one way to examine this is to first measure NO2 levels with monitors near or around residential areas then examine their classification into either exposed to high traffic or low traffic then compare them with the results from the LUR models.
- The use of a categorizing method (blocks are categorized as either far or close to traffic) would possibly be a better representation of the true exposure to traffic related air pollution, however, we don't have a concentration response function for this method.

R walk through

All data sets and scripts have been collected in one folder, names of files have been updated for ease of navigation. In the following I will discuss how to estimate the burden of asthma due to NO2 exposure in a step by step manner using R, by following the provided R script (Burden_NO2_children.R).

Data sets

We will be using two data sets for each year:

- **Census data set** from the NHGIS with age and sex stratification (Steven Manson, Jonathan Schroeder, David Van Riper, 2017)
  - NHGIS_2000_Sex_Age.csv
  - NHGIS_2010_Sex_Age.csv
- **NO2 data set** from Bechle (Bechle, Millet, & Marshall, 2015)
  - NO2_2000
  - NO2_2010

R scripts Used:

- Burden_NO2.R

Steps of Burden estimations

The steps of should be followed using the available R script provided, the analysis will be done separately for each year (i.e. year 2000 data set than year 2010)

Step 1: Read the census dataset into R, (census dataset should be the specific data set provided).

Step 2: Clean the data set, removed unwanted columns and renamed the remaining columns.

Step 3: Validate the population size, drop age groups > 18 years, and validate the child population size.

Step 4: Load the NO2 data set, merge the two data sets. (Note: you will now have a new data set named burden, we will then remove the two old data sets and some columns for reducing the size of the file)

Step 5: Estimate the burden of disease in the following order:

- Estimate the number of asthma cases
- Transform NO2 levels from ppb to ug/m^3
- Estimate RR of new exposure (RRnew), Attributable fraction (AF), Attributable cases (AC) with lower and upper limits
- Estimate the counts of cases attributable to NO2 exposure with upper and lower limits

Summary of R scripts used in all of the analysis

NO2 summary statistics.R

- This scripts provides summary statistics for NO2 exposures using the "Processed data set" for both years 2000, 2010.
- They provided the plots and tables for full progress report 1.

Comparing NO2 with Census data 2000.R

- This scripts provides an exploration/summary of the missing NO2 levels for the year 2000 by comparing the "website" data set with NHGIS census data.

Comparing NO2 Processed with Annual average

- This scripts compares two NO2 data sets, the "website" data set and the "Processed" data set.
- It summarizes the mean NO2 levels between the two for both years 2000 and 2010

NHGIS……block level…….R

- These are a group of scripts that read the NHGIS data sets for year 2000/2010 and cleans them by removing unwanted columns and renaming the columns as provided in the codebook
- Files with sum prepares data sets with total population count only
- Files with sex and age prepares data sets with population counts stratified by sex and age

Data set file names

Exposure files:

- Exposure files will be named using the following convention:
  - Pollutant_Year_CW
  - CW stands for cross walk; if the file is using the 2010 census block geographical boundaries, if not CW will not be added. i.e.
  - NO2_2000; the pollutant is NO2, Year is 2000, and the geographical census block boundaries is for the same year (2000).
- An exception to this rule is the two files named
  - NO2_2000_2010_Summary
  - NO2_2000_2010_Summary_Census
  - These data sets were used for the **NO2 summary statistics.R script**
  - These files will only be used for this script and not in future scripts.