Project 5 (National):

A National Study to Assess Susceptibility, Vulnerability and Effect Modification of Air Pollution Health Risks

Principle Investigator: Francesca Dominici
Co-Principle Investigators: Joel Schwartz, Michelle Bell
ABSTRACT – Project 5

b. Title: A National Study to Assess Susceptibility, Vulnerability, and Effect Modification of Air Pollution Health Risks
c. Investigators: PI: Francesca Dominici; Co-PIs: Joel Schwartz, Michelle Bell, Antonella Zanobetti
d. Institution: Harvard School of Public Health
e. Project Period and Location: 2010-2015, Boston MA
f. Project Cost: $ 1 Million

Project Summary:

1) Objectives: The proposed research responds directly to four of the research questions posed in the EPA solicitation: 1) What are the explanations for regional and temporal differences in air pollution risk? 2) What subpopulations are at increased risk? 3) How can the health effects of PM be better understood in a multi-pollutant context? 4) What is the exposure-response relationship for these sources and mixtures? To address these we will conduct a National study aimed at identifying factors that explain the heterogeneity of health risks associated with air pollution exposure. We hypothesize that such factors include medical and social conditions, conditions that modify exposure, and differences in pollution composition that modify exposure toxicity. Moreover, we hypothesize that the relevant factors vary among different health outcomes. Our research will be fully interactive with the other Center projects. Our previous results (e.g. diabetic susceptibility) have guided their analyses, and their results have generated specific hypotheses that we will test. We have 3 objectives. In Aims 1 and 2, we will conduct national studies of short- and long-term exposures to individual pollutants, sources, and mixtures. A main focus of our Center is to study established cohorts (NAS, Framingham, and Viva) in Massachusetts and surrounding states using novel, validated approaches to assess exposure. In Aim 3, we will complement those cohort studies, by establishing a cohort of 2.3 million Medicare enrollees residing in the same region and following its members prospectively for cause-specific hospital admissions and mortality for the period 2000-2014, and also by studying all live births in Eastern MA, geo-coded to exact address and followed for adverse birth outcomes.

2) Experimental Approach: We will use the largest available collection of national datasets including: 1) daily time series data of mortality and hospital admissions for thousands of zip codes in the US; 2) a cohort of 12 million Medicare enrollees followed prospectively for cause-specific hospital admission and death during the period 2000 to 2014; 3) a comprehensive collection of individual and area-level modifying factors; 4) exposures to individual pollutants, mixtures, and sources developed by the Biostatistics and Exposure cores.

3) Expected Results: Identifying factors that explain heterogeneity of risks will help to identify: 1) the populations that are more susceptible/vulnerable to air pollution; and 2) the emission sources, pollutants and pollutant mixtures that are more toxic. The characterization of susceptibility factors, such as age, gender, race, and pre-existing health conditions will inform research on biological mechanisms. The identification of the most harmful air pollution sources and components can aid decision-makers in developing targeted air quality regulations.

Supplemental Keywords: National studies, air pollution, heterogeneity, vulnerability, susceptibility
1. OBJECTIVES/HYPOTHESES

1.1. Objectives and Hypotheses: The proposed research directly responds to the following research questions posed in the RFA: 1) What are the explanations for regional and temporal differences in air pollution risk? 2) What subpopulations are at increased risk? 3) How can the health effects of PM be better understood in a multi-pollutant context? 4) What is the exposure-response relationship for these sources and mixtures? We will conduct a National study aimed at identifying individual and area-level factors that explain the geographical and temporal heterogeneity of the air pollution health risks across the US. This project will use the largest national datasets on air pollution, health outcomes, potential confounders and modifiers to develop innovative statistical methods and to provide evidence addressing the following specific hypotheses: 1) Long-term and short-term exposures to individual air pollutants, sources types, and pollutant mixtures are associated with risk of increased hospital admissions and mortality in a population of elderly individuals (over 65 yr); 2) Associations vary by individual pollutant, source type and pollutant mixture; 3) Health risks of both short- and long-term exposures to air pollution vary greatly across the US; and 4) This variation depends on exposure efficiency (e.g. infiltration), susceptibility, vulnerability, and pollution characteristics. Our research will be fully interactive with the other Center’s Projects. The development of our research hypotheses will be also guided by the preliminary findings other Center’s Projects. Our Specific Aims are:

A1. Develop statistical methods and conduct national studies to estimate mortality and hospitalization risks associated with short-term exposures to individual pollutants, source types and air pollution mixtures. We will: 1) estimate location-specific health risks associated with short-term exposures to individual pollutants (PM$_{2.5}$ mass and components and ozone) and source types; 2) characterize vulnerability and susceptibility to air pollution by individual- and area-level socio-economic, behavioral and co-morbidity factors; and 3) investigate whether location-specific health risks associated with short-term exposure to individual pollutants are modified by air pollution mixtures.

A2. Develop statistical methods and conduct national studies to estimate mortality and hospitalization risks associated with long-term exposures to individual pollutants, source types and air pollution mixtures. We will: 1) estimate location-specific health risks associated with long-term exposure to individual pollutants (PM$_{2.5}$ and components and ozone) and source types; 2) characterize vulnerability and susceptibility to air pollution by individual- and area-level socio-economic, behavioral and co-morbidity factors; and 3) investigate whether location-specific health risks associated with long-term exposure to individual pollutants are modified by air pollution mixtures.

A3. Conduct two cohort studies in Massachusetts and surrounding states to estimate health risks associated with long-term exposures to individual pollutants, sources, and air pollution mixtures. We will estimate spatially resolved long-term exposures to individual pollutants (PM$_{2.5}$ and components and ozone) using models described in the Exposure and Biostatistics Cores. We will use administrative data to construct the following two study populations:

A3.1. A cohort of approximately 700,000 live births in Eastern Massachusetts registered between January 1, 1996 and December 31, 2007 (residence of the mothers geo-coded). We will estimate the risks of adverse birth outcomes associated with long-term exposure to individual pollutants (PM$_{2.5}$ and components and ozone), source types and mixtures adjusted by individual- and area-level risk factors.
A3.2. A cohort of 2.3 million Medicare enrollees residing in the Massachusetts and surrounding states (residence known at the zip-code level) followed prospectively for 14 years. We will estimate the mortality and morbidity risks associated with long-term exposure to individual pollutants (PM2.5 and components and ozone), source types and mixtures adjusted by individual and area-level risk factors for the period 2000 to 2014.

1.2. Background/Preliminary Results: In this Project we will estimate adverse health risks associated with short- and long-term exposure to individual pollutants, source types and mixtures on national and regional scales. Our focus will be to develop specific hypotheses and use national data sets to explain regional and temporal differences in air pollution risks. A wide array of potential effect modifiers can be investigated, including socio-economic status, co-morbidity, stress, climate conditions, and components of the air pollution mixture. To maintain focus in our analyses, our research strategy will include the development of a clear set of research hypotheses. Specifically, the development of our research hypotheses will be guided by: 1) a comprehensive and systematic review of the literature on effect modification (e.g., see Table 5.1); and by 2) preliminary findings of the other Center’s projects. In fact, each Center’s project will share a common set of objectives, exposure metrics and health outcomes and will test them using different, synergistic methods thereby providing the unique opportunity to look for consistency of findings across studies.

1.2.1. Previous Work and Contributions for A1: In our national studies we have found evidence of geographical and seasonal heterogeneity in the association between short-term exposure to PM10 and PM2.5 total mass, ozone, and risks of hospitalization and mortality.1-13 As summarized in Table 5.1 (next page), several studies have indicated that associations between particles or ozone and health may be modified by various factors, such as AC use, although results are inconsistent. For example, we have found that the heterogeneity of the health risks can be explained in part by: 1) PM composition,8,12,14,15 (2) exposure differences due to home ventilation as inferred by air conditioning use or seasonal mean temperature;14,16-18 and 3) susceptibility factors such as diabetes and atrial fibrillation.19 We have also shown that pollutant mixtures, as indexed by air masses with specific back trajectories, influence the health effects of particles.20 Earlier work indicates that some individuals may be more susceptible and more vulnerable to air pollution than others due to existing health conditions, race, or gender.19,21-27

In Aim 1, we will develop specific research hypotheses and use updated national data sets (2000-2014) to fully characterize effect modification of risks associated with short-term exposure to individual pollutants and source types. This work will extend our previous efforts in the following ways. First, a set of specific hypotheses will be developed as a result of the collaboration and synergy with the work proposed by the other Center’s projects. Second, we will use a substantially larger number of years, thus increasing the precision of the location-specific estimates of the health risks and therefore the statistical power to detect effect modification, particularly for components. Third, we will study a larger number of geographical locations across the US. This will enhance our ability to detect differences in health risks due to differences in sources and/or composition of the air pollution. Fourth, in addition to time series Poisson regression analyses, we will develop a new approach for statistical analysis that will allow us to investigate both individual- and area-level effect modifiers. Fifth, we have assembled and plan to study an extensive list of effect modifiers that have not previously been studied such as: 1) area-level smoking and obesity prevalence; 2) area-level proportion of green space; 3) area-level average of Volatile Organic Compounds (VOCs); 4) several individual-level chronic
disease diagnoses, including a diagnosis of depression; and 5) air pollution mixtures. Sixth, we will explore exposure-response and distributed lags for sources, mixtures, and components.

**Table 5.1:** Summary of the literature on effect modifiers of short-term associations between air pollution (PM\(_{10}\), PM\(_{2.5}\) and O\(_3\)) and mortality and morbidity

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Pollutant (Duration, # of Communities)</th>
<th>Health Outcome</th>
<th>Effect Modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bell et al. (2009)(^9)</td>
<td>PM(_{2.5}) (1999-2005, 106)</td>
<td>Hospital admissions for CVD and respiratory</td>
<td>AC(^<em>), 20 PM(_{2.5}) chemical components† (Ni(^</em>), V(^<em>), EC(^</em>)) Education, income, race, urbanization</td>
</tr>
<tr>
<td></td>
<td>PM(_{2.5}) (1987-2000, 100)</td>
<td>Non-accidental mortality</td>
<td>AC(^<em>), 20 PM(_{2.5}) chemical component† (Ni(^</em>)), Education, income, race, urbanization</td>
</tr>
<tr>
<td>Bell et al. (2008)(^9)</td>
<td>PM(_{2.5}) (1999-2005, 168)</td>
<td>CVD hospital admissions</td>
<td>AC(^*)</td>
</tr>
<tr>
<td></td>
<td>PM(_{10}) (1987-2000, 84)</td>
<td>Respiratory hospital admissions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-accidental mortality</td>
<td></td>
</tr>
<tr>
<td>Franklin et al. (2008)(^7)</td>
<td>PM(_{2.5}) 1999-2006, 25</td>
<td>Non-accidental, CVD, Stroke, and Resp deaths</td>
<td>Mean seasonal temperature, chemical components (S, Al, As, Ni), season, region</td>
</tr>
<tr>
<td>Zanobetti and Schwartz (2009)(^11)</td>
<td>PM(_{2.5}), CM 1999-2005, 112</td>
<td>Cause specific deaths</td>
<td>Climate zone, season</td>
</tr>
<tr>
<td>Samoli et al. (2008)(^8)</td>
<td>PM(_{10}) (1987-1997, 134)</td>
<td>Non-accidental mortality</td>
<td>Age(^<em>), health status, percent unemployed(^</em>), proportion of cardiorespiratory(^*), NO(<em>2)-PM(</em>{10}) ratio, O(_3) temperature, humidity</td>
</tr>
<tr>
<td>Dominici et al. (2007)(^5)</td>
<td>PM(_{10}) (1987-2000, 60)</td>
<td>All-cause mortality</td>
<td>Ni(^<em>), V(^</em>)</td>
</tr>
<tr>
<td>Franklin, et al. (2007)(^7)</td>
<td>PM(_{2.5}) 1997-2002, 27</td>
<td>All-cause mortality</td>
<td>Age(^<em>), sex, geographic location(^</em>), annual PM(_{2.5}) concentration, AC</td>
</tr>
<tr>
<td>Lippmann et al. (2006)(^2)</td>
<td>PM(_{10}) (1987-1994, 60)</td>
<td>Mortality</td>
<td>16 key components† (Ni(^<em>), V(^</em>))</td>
</tr>
<tr>
<td>Zeka et al. (2006)(^2)</td>
<td>PM(_{10}) (1989-2000, 20)</td>
<td>Non-accidental mortality</td>
<td>Sex, race, age(^<em>), education, location of death, season(^</em>), 4 contributing causes of death(^*)</td>
</tr>
<tr>
<td>Janssen et al. (2002)(^17)</td>
<td>PM(_{10}) (1985-1994, 14)</td>
<td>Hospital admissions for COPD, CVD and pneumonia</td>
<td>AC(^<em>), 9 combustion sources† (PM(_{10}) from highway vehicles(^</em>), PM(_{10}) from highway diesels(^*))</td>
</tr>
<tr>
<td>Zanobetti et al. (2001)(^2)</td>
<td>PM(_{10}) (1988-1994, 1)</td>
<td>Hospital admissions for COPD, CVD and pneumonia</td>
<td>Age, diabetes(^*), age by diabetes interaction</td>
</tr>
<tr>
<td>Zanobetti et al. (2000)(^2)</td>
<td>PM(_{10}) (1986-1993, 4)</td>
<td>Non-accidental mortality</td>
<td>Race, sex(^*), education</td>
</tr>
<tr>
<td>Zanobetti et al. (2000)(^2)</td>
<td>PM(_{10}) (1988-1994, 1)</td>
<td>Hospital admissions for COPD, CVD and pneumonia</td>
<td>Age, sex, race, 9 presence of co-morbidities† (respiratory infections(^<em>), heart disease(^</em>), heart failure(^<em>), asthma(^</em>))</td>
</tr>
<tr>
<td>Bell et al. (2008)(^8)</td>
<td>Ozone (1987-2000, 98)</td>
<td>Non-accidental mortality</td>
<td>Education, income(^<em>), race, urbanization, transportation(^</em>), population, AC(^<em>), 5 air quality factors, 2 weather factors† (temperature(^</em>))</td>
</tr>
<tr>
<td>Medina-Ramon et al. (2008)(^19)</td>
<td>Ozone (1989-2000, 48)</td>
<td>Non-accidental mortality</td>
<td>Sex(^<em>), age(^</em>), race, education, city characteristics, 10 primary and secondary causes of death† (atrial fibrillation(^*)), AC</td>
</tr>
<tr>
<td>Bell et al. (2007)(^1)</td>
<td>Ozone (1987-2000, 98)</td>
<td>Non-accidental mortality</td>
<td>PM(<em>{10}), PM(</em>{2.5})</td>
</tr>
</tbody>
</table>

Abbreviations: CVD (cardiovascular disease), COPD (chronic obstructive pulmonary disease), AC (air conditioning), Ni (Nickel), V (Vanadium), EC (elemental carbon), O\(_3\) (Ozone), CM (coarse mass)
* Significant modifier; † Significant factors provided in parenthesis

1.2.2. Previous work and contributions for Aim 2: In our earlier studies we found evidence of geographical differences in mortality risks associated with long-term exposure to PM\(_{2.5}\) in the Medicare population.\(^{32,33,34}\) However, compared to acute risks of short-term exposure to PM\(_{2.5}\), less is known about heterogeneity of chronic risks associated with long-term exposure to PM\(_{2.5}\). Further, little is known about regional variation, and potential explanations of such variation, in
the health risks associated with long-term exposure to PM$_{2.5}$ components which relate to different sources and atmospheric processes. In this Aim, we will develop statistical methods and conduct a study on **12 million** Medicare enrollees followed for hospitalization and death over a 14-year period (2000 to 2104). Our study population includes only the enrollees that did not change residence during the study period. Because data from the Medicare database is available for 100% of the zip codes in the US, we will use a **nearest-monitor approach** to estimate long-term exposure (e.g., yearly averages) to individual pollutants. To the best of our knowledge, this will be the largest study to date to estimate health risks associated with long-term exposure to PM$_{2.5}$ mass, its components, and ozone. We will develop computationally efficient statistical models that will allow for spatially varying risks of long-term exposures to air pollution. We will investigate which individual and/or area-level effects might explain geographical variation in the chronic effects of long-term exposure to individual pollutants and individual sources. The list of research hypotheses and potential effect modifiers will be the same as that assembled for Aim 1, thus further increasing our ability to look for consistency of findings.

1.2.3. **Previous work and contributions for Aim 3**: In our earlier work, we identified links between air pollution exposure and health outcomes for older persons, including risk of hospitalizations. Many of these studies were based on Medicare datasets. 4,8,15,33-37 Our previous research also identified associations between air pollution exposure and pregnancy outcomes, including risk of low birth weight in the New England area 38-40 and other regions. 41 The overarching research objective of the Center is the investigation of health effects of individual pollutants, sources, and pollutant mixtures in MA and the surrounding States (“the Region”) using populations at different stages of life. Specifically, the NAS, Framingham, and Viva Projects will focus on neurocognitive, neurovascular and other outcomes and they will use a detailed and rich exposure assessment data set developed by the Exposure and the Biostatistics Cores. In **Aim 3**, we will complement these studies by constructing two cohorts and using the same exposure assessment methods. The first will be a study of over **2.3 million Medicare enrollees** followed prospectively from January 1, 2000 to December 31, 2014. For a small subpopulation (approximately 1,000 individuals) we will have detailed information on individual-level risk factors. For every zip code of the study Region, we will have very extensive information on area-level confounders. In addition, to complement the Viva Project, we will also conduct a study of approximately **700,000 live births** in Eastern MA followed prospectively from January 1 1996 to December 31 2007.

1.2.4. **Description of how each objective is related to the Center’s integrated approach**: This project will address several research questions posed by the RFA, using a **national approach** (Aims 1 and 2), and also focusing on **MA and surrounding states** (Aim 3). First, the two administrative cohorts (Medicare and live births) used in Aim 3 of this project, and the NAS, Framingham and Viva cohorts studied in Projects 2, 3 and 4 will use the same estimated exposures developed by the Exposure and the Biostatistics Cores. In each of these cohorts, we will also investigate a similar list of individual-level and area-level modifiers. Second, both the Viva study and Aim 3.1 of this Project will focus on births in MA. Third, both the Framingham study and Aims 1, 2, and 3.2 of this Project will focus on cardiovascular outcomes. Fourth, findings from the Framingham study during the first three years will inform this project on key susceptibility factors and cardiovascular outcomes that merit further investigation on a national scale. In summary, the NAS, Framingham, Viva and two proposed studies described in Aim 3 of this project will focus on different populations, spanning different life stages but using the same exposure and similar health outcomes, thus providing the unique opportunity to compare and contrast results across
studies. Discrepancies in results, if detected, will stimulate further research aimed at resolving these issues. The integration of these projects is assured not only by our regular meetings, but also by the participation of investigators from this project in the cohort studies, including as PI and Co-PIs. Moreover, Dr. Schwartz, the Co-PI of this project is PI of the NAS and Co-PI of the Framingham study, ensuring tight collaboration. As an example of the Center's integrated approach, each project will help test the hypothesis that exposure to particulate matter (PM) generated by vehicle emissions increases the risk of cardiovascular disease (CVD) more than air pollution from any other source. Project 5 will conduct a national analysis of short and long-term exposures, focusing on cardiovascular health outcomes in the elderly. Project 4 (Viva) will be a regional cohort study of links between pre- and post-natal exposures on blood pressure, cardiovascular fitness, and cognitive outcomes. Two additional cohort studies will investigate associations between exposure and changes in CVD biomarkers in adults (Project 3 Framingham) and people over the age of 74 (Project 2 NAS). Project 1 will use vehicle emissions PM in animal-based laboratory experiments, where cardiovascular performance of mice will be measured.

2. APPROACH/ACTIVITIES: The analytical approaches for each aim are detailed below.

2.1. Description of the datasets: Our team has decades of experience in the administration, management and analysis of national databases on health, environmental agents, confounders and effect modifiers. Below is the data inventory (by Agency) of the national data sources that are already available to us (Table 5.2).

- **The US Environmental Protection Agency (EPA):** 1) Air-Data: Daily ambient pollution concentrations for criteria pollutants (PM$_{10}$, PM$_{2.5}$, O$_3$, NO$_2$, CO) for all available Federal Information Processing Standards (FIPS) throughout the United States (Figure 5.1, right). PM mass concentration data are available as 24-hour integrated samples, usually daily, but sometimes on a 1-in-3 or 1-in-6 day sampling schedule. When more than one monitor is available in one county, daily ambient levels will be averaged over the county using a method previously described; and 2) PM Speciation data: PM$_{2.5}$ constituent data consist of particle mass together with more than 50 elemental and ion fractions. However, for each US region we will select elements with levels more than 50% above the minimum detectable limits of the analytical methods. Data are available for 24-hour integrated samples at all the available FIPS throughout the US, normally on a 1-in-3 day or 1-in-6 day schedule.

- **National Oceanic and Atmospheric Administration (NOAA):** Daily temperature, dew-point temperature, relative humidity, wind speed, and precipitation data from more than 8,000 monitoring stations. Relative humidity and allied measures are available from a sub-sample of 500 monitoring stations.

- **Mortality data from the National Center for Health Statistics (NCHS):** Individual cause-specific mortality data for the entire US. Individual-level data include county of residence, state and county of death, date of death, race, gender, education and age of decedent.

- **Centers for Medicare and Medicaid services (CMS):** 1) Cohort: All Medicare enrollees by year of enrollment, including age, gender, race, state and county of residence and 5- and 9-digit zip-code identifiers (40,000,000 people per year); 2) Mortality: Date of death for those who die within the Medicare cohort; 3) Hospitalizations: Hospitalization records for all Medicare enrollees, including date of hospitalization, length of stay in hospital, International Classification of Diseases (ICD) primary and secondary diagnostic and procedure codes associated with the hospitalization, and the costs billed to Medicare for the hospitalization.
Table 5.2: Current data resources (Data will be updated through 2014)

<table>
<thead>
<tr>
<th>Agency</th>
<th>Data</th>
<th>Approximate No. of Records</th>
<th>Years covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPA</td>
<td>AirData Speciation VOC</td>
<td>Usually available on 1-3 or 1-6 day schedule</td>
<td>1980-2007, 1999-2007</td>
</tr>
<tr>
<td>NOAA Meteorology</td>
<td>3 million hourly records per year</td>
<td></td>
<td>1948-2007</td>
</tr>
<tr>
<td>NCHS</td>
<td>Mortality</td>
<td>2,400,000 records per year</td>
<td>1962-2008</td>
</tr>
<tr>
<td>Census Bureau</td>
<td>Population and community characteristics</td>
<td>432,000 zip codes per census</td>
<td>1980, 1990 and 2000</td>
</tr>
<tr>
<td>NLCD</td>
<td>Green space 1992, 2001</td>
<td>29,000 zip code per year</td>
<td>1992 and 2001</td>
</tr>
<tr>
<td>AHS</td>
<td>Available by county by year</td>
<td></td>
<td>1985-2000</td>
</tr>
<tr>
<td>CDC</td>
<td>BRFSS</td>
<td>Available by metropolitan area</td>
<td>1990-2008</td>
</tr>
</tbody>
</table>

- **The Medicare Current Beneficiary Survey (MCBS):** A representative sample of approximately 13,000 persons from the Medicare cohort. The data include survey responses covering: smoking, income, household size, marital status, education, general health, history and current status with regard to a number of medical conditions, including cancer, heart disease, emphysema, asthma, pneumonia, and respiratory infections. (Figure 5.1, left).

- **PAM (Photochemical Assessment Monitoring Stations) for Volatile Organic Compounds (VOCs):** the dataset includes around 58 compounds, for all available FIPS throughout the United States.

Figure 5.1: (left) map indicating the distribution of the MCBS enrollees by zip code; (right) Long-term averages of PM$_{2.5}$ ambient values (μg/m$^3$) plotted at each monitor station.
• **The U. S. Census Bureau:** The census provides socio-demographic data at the state, county, 5 digits zip-code tabulation area, and the census block level.

• **National Land Cover Dataset (NLCD):** This dataset provides data on land cover derived from satellite photographs taken in the years 1992 and 2001. The data include the following variables: total space, green space, open water area, proportion of green space, proportion open water.

• **North American Regional Re-analysis (NARR):** Data every 3 hours on the height of the planetary boundary layer (HPBL) in meters, for grid points in MA from 1994 to the end of 2008. Data for the full country is available, and will be updated to future years.

• **Centers for Disease Control: Behavioral Risk Factor Surveillance System (BRFSS):** Data on smoking, diabetes, and obesity are available on a metropolitan area basis.43

• **American Housing Survey (AHS):** Annual county-level data on prevalence of AC, percent of housing within 300 m of green space, percent multifamily housing, and percent of housing during the period 1994 to 2002.

2.2. **Exposure assessment and spatial misalignment:** The selection of the geographical unit of the analysis is affected by the tradeoff between minimizing exposure measurement error and maximizing statistical power. We anticipate that the geographical unit of the analysis will range from the zip code, where the exposure measurement error is small, to the county or metropolitan areas, where the exposure measurement error is larger. The statistical methods we will develop in Aims 1-3 will account for measurement error caused by the spatial misalignment of the data. In our research plan, the smallest level of aggregation at which the health data are available defines the geographical unit of the analysis. The place of residence of the Medicare enrollees is provided at the zip code level. Zip codes can be subsequently translated into zip code tabulation areas (ZCTAs). The ZCTA are polygonal boundaries that can be used for mapping and defining non-overlapping areas of residence. Using the ZCTAs we can use the nearest-monitor approach to assign exposure. In Aims 1, 2 and 3 we will use the **nearest-monitor approach** to assess **short- and long-term exposures to individual pollutants.** Specifically, we will match people to air pollution monitors based on the distance from the monitor to the centroid of their ZCTA of residence. Specifically, we will define geographical locations as all the ZCTAs falling within a circle of a fixed radius centered on a monitor. With this approach, we will be able to modify the study population by adjusting the radius of the circle and thereby adjusting the average distance the study population resides from a given monitor. We will explore the sensitivity of the estimated exposure to the selection of the geographical location. As a special case, we can aggregate the data to create county-average exposures to make comparisons with previous studies. In **Aim 1,** we will also **assess short-term exposure to individual PM sources in 10 locations** (e.g., counties) using the EPA Positive Matrix Factorization (PMF) source apportionment method.44,45 The 10 locations will be selected based on availability of PM$_{2.5}$ chemical composition data and population density. For the same 10 locations, we will **assess short-term exposure to mixtures** as a follows: 1) we will group days based upon their multi-pollutant profiles (see the Exposure Core for details on the temporal clustering algorithm); 2) we will investigate whether the “cluster type/mixture type” is an effect modifier of the health risk associated with short-term exposure to individual pollutants. In **Aim 2,** we will **assess long-term exposure to sources in 30 locations** (e.g. counties) using the PMF source apportionment method. As in Aim 1, the 30 locations will be selected based on availability of PM$_{2.5}$ chemical composition data and population density. For the same 30 locations, we will **assess long-term exposure to mixtures** as
follows: 1) we will calculate the annual concentrations for each individual for all the geographical locations included in the National Study; 2) we will perform a cluster analysis to group US geographical locations with similar multi-pollutant profiles (see the Exposure and Biostatistics Cores for details on the spatial clustering algorithm); 3) we will investigate whether the “cluster type/mixture type” is an effect modifier of the health risk associated with long-term exposure to individual pollutants. In Aim 3, we will also use **spatially resolved long-term exposures to individual pollutants and source types** using several air pollution data sets available in MA and surrounding states (e.g., Boston Supersite, Satellite data, Harvard School of Public Health spatial sites) using spatio-temporal models described in the Exposure and Biostatistics Cores.

2.3. **Approach to Aim 1:** In this aim, we focus on the following main analyses. First, we will conduct multi-site time series and case-crossover analyses\textsuperscript{46-49} to estimate the health risks associated with short-term exposure to individual pollutants (PM\textsubscript{2.5} mass and components and ozone) for several hundred geographical locations (e.g., counties) across the US and for the period of 2000-2014. Second, we will investigate the roles of air pollution mixtures, age, gender, race, socio-economic variables and chronic diseases in modifying these air pollution risks. Third, for 10 locations in the US, we will also estimate health risks associated with short-term exposures to sources.

2.3.1. **Statistical methods for Aim 1:** Here we detail the statistical models that will be used to: 1) estimate location-specific mortality and morbidity risks associated with short-term exposure to individual pollutants; 2) assess effect modification. In this aim, in addition to the common approach of fitting Bayesian hierarchical models for multi-site time series analyses,\textsuperscript{49,50} we will also fit the following conditional logistic regression model to allow the investigation of effect modification by both individual and area-level characteristics:

\[
\text{logit}(E[Y_{idjt}]) = b_0 + u_{jt} + \left( b_1 + v_{jt} \right) \times PM_{2.5djt} + b_2 \times M_i + \left( b_3 + w_{jt} \right) \times O_{3djt} \times M_i + \text{covariates} \tag{1}
\]

where \(Y_{idjt}\) indicates a death (or an admission) for individual \(i\) on day \(d\) in community \(j\) and in year \(t\). The variable \(M_i\) denotes the individual-level effect modifiers. The variance of \(u_{jt}\) will capture differences in baseline rates as well as differences in populations across communities and years, and the variance \(v_{jt}(w_{jt})\) will capture the variation of the short-term effects of individual pollutants across communities and years. All models will control for day of the week using indicator variables. Because the unit of observation for a case-crossover analysis is the individual, it is possible to examine individual-level factors as effect modifiers via interaction terms as planned. At the second level, we will investigate whether area-level characteristics will explain some of the variance of the random effects. We will assume:

\[
E(u_{jt}) = \gamma_0 \sum_p \gamma_p X_{jpp}; E(v_{jt}) = \lambda_0 \sum_p \lambda_p X_{jpp}; E(w_{jt}) = \alpha_0 \sum_p \alpha_p X_{jpp}; \tag{2}
\]

where \(X_{jpp}\) are the area-level covariates (e.g. such as land use, average temperature, age-sex-race-cause specific annual mortality rates). Because separate slopes for each calendar year are likely to be noisily estimated, we will fit a penalized likelihood to smooth out jumps from year to year within community. This will be done by adding a penalty to the conditional likelihood proportional to the sum of the squares of the differences between the coefficients of air pollution in adjacent years. The constant of proportionality for the penalty will be estimated using Akaike’s Information Criterion (AIC). To control for seasonal confounding, we will apply a
time-stratified approach by choosing control days only within the same month of the same year that death occurred. We will determine whether the impact of significant modifiers changes over time by looking at both varying coefficient models and stratifying the interaction terms by decades.

2.3.2. Nonlinearity of air pollution-mortality associations: In the model above, we have assumed that a linear exposure-response relation is adequate. We will relax that assumption by fitting a penalized piecewise linear spline at each lag, with a penalized distributed lag for each piece. Again, we will use AIC to choose the penalties, and thus will be able to both examine the evidence for nonlinear relations in the seasonally stratified models as well as estimate the nonlinear distributed lag, where present. This is a generalization of a nonlinear distributed lag model we have already published, with the added benefit of season-specific models, which will reduce the extent of nonlinearity. See also our related work.

2.3.3. Description of the effect modifiers: The focus of this Project will be to conduct a National study aimed at identifying factors that explain the heterogeneity of the air pollution health risks. We hypothesize these factors include medical and social conditions, conditions that modify exposure, and compositional factors that modify the toxicity of the exposure. Moreover, we hypothesize the relevant factors will vary by outcome. The effect modifiers examined and their purpose are outlined below:

2.3.3.1. Individual-level modifiers: age, race, gender, education, co-morbidity (e.g., chronic obstructive pulmonary disease (COPD), cardiovascular outcome and diabetes).

2.3.3.2. Area-level climatic factors: seasonal averages of daily temperature. These will be used as a surrogate for ventilation, since windows will be closed during cold weather, but are also more likely to be closed during hot weather, particularly in locations where air conditioning (AC) prevalence is high.

2.3.3.3. Area-level environmental and socioeconomic characteristics: population density, county-level median household income (separately by race for analyses stratified on race), percent below poverty, percent with less than a college education, and county level income inequality as measured by the Gini coefficient. From the NLCD we will consider measures of green space, altitude, and impermeable surfaces.

2.3.3.4. Area-level epidemiologic characteristics: annual age-race-gender-specific baseline mortality rates for specific diseases, e.g., diabetes, as an indicator of background risk of that disease. We will use the Behavioral Risk Factor Surveillance System (BRFSS) and the MCBS, which are both nationally representative, to calculate the prevalence of several risk factors including smoking, mean body mass index (BMI) and diagnosed diabetes and hypertension.

2.3.4. Health risks associated with short-term exposure to source types in 10 US locations: In each location, using data on PM$_{2.5}$ chemical composition from the STN, we will estimate short-term exposures to PM source using the EPA Positive Matrix Factorization (PMF) source apportionment method. The source types that will be investigated are traffic, soil, coal and oil combustion, among others. We will select the 10 counties with the largest population density. Data on PM$_{2.5}$ chemical composition will be available every third day in most of the locations, and therefore short-term exposure to sources will be calculated only for the days where data is available. Since source emissions and their profiles have been gradually decreasing over the last two decades, source apportionment analyses will be done by year. From this analysis we will estimate location-specific mortality and morbidity risks associated with the source types within
each location. Therefore we will be able to assess the relative toxicity of the identified sources within each location.

2.3.5. **Pollution mixtures as a modifier of the acute health risks in 10 US locations:** Within each location we will apply a temporal clustering algorithm to group days that have the same air pollution mixture profile (see Exposure Core for details). We will then investigate whether the cluster type/mixture type modifies the health risks associated with short-term exposure to individual pollutants.

2.3.6. **VOC as a modifier of the acute health risks in all US:** We will use VOC data from monitoring sites to stratify days based on low versus high levels of VOC concentrations. Here we hypothesize that VOCs are potential effect modifiers because they are both precursors to secondary particles, and interact with ozone to form new chemicals with potentially greater toxicity. We will estimate the health risks associated with short-term exposure to individual pollutants (e.g., PM$_{2.5}$) separately for the days with low VOC concentrations and for days with high VOC values. Then we will test whether these health risks are different between these two groups.

2.4. **Approach to Aim 2:** From the Medicare enrollment files, we will construct a cohort of about 12 million people with individual-level information on age, sex, race and zip-code of residence for the period 2000–2014. To account for residential history, we will include in the analysis only enrollees who stayed in the same residence during the entire study period. The cohort is dynamic, including all new enrollees to Medicare during this interval. First, we will estimate location-specific health risks associated with long-term exposures to individual pollutants. Second, we will investigate effect modification, vulnerability and susceptibility using the list of individual and area-level modifiers described in Aim 1. Third, we will estimate location-specific health risks associated with long-term exposures to sources in 30 US locations. Fourth, to investigate effect modification by air pollution mixture, we will cluster geographical locations based on similar air pollution profile mixture (see Exposure and Biostatistics Cores for more details on the clustering algorithm), and investigate whether the health risks of long-term exposure to individual pollutants vary across clusters.

2.4.1. **Description of the study population:** Because we have information on the zip code of residence of every Medicare enrollee, we will use a nearest-monitor approach for exposure assessment to the individual pollutants. That is, we will assign the same exposure to all the enrollees who live less than a few miles from an ambient monitor for PM$_{2.5}$ and O$_3$, and less than a few miles from a STN monitor for PM$_{2.5}$ chemical components, thus maximizing accuracy of ambient exposure estimates (see also section 2.2.). For example, if we assume a 5-mile distance between the zip code centroids where the subject resides and the monitor location, long-term exposure to PM$_{2.5}$ mass can be calculated for 5,762 zip codes. In these 5,762 zip codes we anticipate 5,119 deaths per 1,000 person-years during the period 2000 to 2014. If we assume a 5-mile distance between the zip code centroids and the STN monitor location, long-term exposure of each of the PM$_{2.5}$ components can be calculated for over 1,563 zip codes. In these 1,563 zip codes we anticipate 1,810 deaths per 1,000 person-years during the period 2000 to 2014 (Table 5.3). We will use individual-level data on race, gender, and co-morbidity measures from Medicare billing claims to characterize susceptibility.
Table 5.3: Characteristics of study population for zip codes whose centroids are within a few miles (1 to 8 miles) from each monitor (PM$_{2.5}$ and STN)

<table>
<thead>
<tr>
<th>Distance (miles)</th>
<th>Zip codes</th>
<th>Deaths (x1000)</th>
<th>PM$_{2.5}$ sites</th>
<th>Zip codes</th>
<th>Deaths (x1000)</th>
<th>STN Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>8,417</td>
<td>7,077</td>
<td>1,412</td>
<td>7,746</td>
<td>7,045</td>
<td>234</td>
</tr>
<tr>
<td>6</td>
<td>6,658</td>
<td>5,853</td>
<td>1,349</td>
<td>6,248</td>
<td>6,066</td>
<td>231</td>
</tr>
<tr>
<td>5</td>
<td>5,762</td>
<td>5,119</td>
<td>1,293</td>
<td>5,426</td>
<td>5,319</td>
<td>223</td>
</tr>
<tr>
<td>4</td>
<td>4,821</td>
<td>4,326</td>
<td>1,212</td>
<td>4,593</td>
<td>4,546</td>
<td>217</td>
</tr>
<tr>
<td>3</td>
<td>3,675</td>
<td>3,401</td>
<td>1,065</td>
<td>3,710</td>
<td>3,698</td>
<td>196</td>
</tr>
<tr>
<td>2</td>
<td>2,591</td>
<td>2,436</td>
<td>826</td>
<td>2,730</td>
<td>2,705</td>
<td>161</td>
</tr>
<tr>
<td>1</td>
<td>1,226</td>
<td>1,173</td>
<td>414</td>
<td>1,365</td>
<td>1,354</td>
<td>85</td>
</tr>
</tbody>
</table>

2.4.2. Statistical methods for Aim 2: We will use the connection between Cox proportional hazard models and Poisson regression\textsuperscript{56} to fit a proportional hazard model to the individual-level data in our cohort of Medicare enrollees. For each month and location, we will compute the average concentrations of each individual pollutant in the previous year and assign this exposure to all subjects at that location (running window). As a sensitivity analysis, we will also compute average exposure for the preceding two and three years, and use these estimates as model predictors. To properly account for the spatio-temporal data structure, we will investigate the residual correlation structure. Greven et al (2009) in a previous analysis of the Medicare data found that, after inclusion of location-specific hazard functions in the model, no additional modeling of the spatial and temporal structure was necessary.\textsuperscript{57} However, we will revisit this point in our setting and include an appropriate correlation structure, or a spatial random field and temporal smooth function, if necessary. To investigate whether the health risks of long-term exposure to individual pollutants differ across geographical locations, we will extend this base model to let the air pollution effect $\beta(u)$ be a function of spatial location $u$. We will include a random slope $\beta(u)$ for each geographical location (zip code or county) $u$. The log-linear model then becomes a generalized linear mixed model\textsuperscript{58}

$$\log E(Y_{uta}) = \log(T_{uta}) + \log(h_u(a)) + (x_{ut} - \bar{x}_t)\beta(u) + \bar{x}_t\gamma, \quad \beta(u) \sim N(0, \tau^2),$$

where $Y_{uta}$ is the number of deaths at age-race stratum $a$ in month $t$ at location $u$, $T_{uta}$ is the total time that people of age $a$ were at risk of dying at location $u$ in month $t$, and $h_u(a)$ is the baseline hazard of dying for people of age $a$ at location $u$ (modeled using splines). We will consider two exposures: 1) $x^*_u = (x_{ut} - \bar{x}_t)$ where $x_{ut}$ is the air pollutant exposure during the previous year up to month $t$ at location $u$; and 2) $\bar{x}_t$ is the average of $x_{ut}$ across location (national trend). This approach can be denoted as a spatio-temporal cohort study (see also \textsuperscript{34}). Under this approach we will estimate two regression coefficients. The first regression coefficient ($\gamma$) estimates the association between the national trend in the air pollutant and the national trend in mortality (admission) rates, and is likely to be confounded by other factors that vary slowly on the national level, such as smoking prevalence. The second coefficient, $\beta(u)$, estimates the association between: 1) deviations of location-specific trends in the air pollutant from the national trend, and 2) deviations of location-specific trends in mortality rates from the national trend. In Greven et al 2009, we showed that the local association is less affected by confounding due to unmeasured time-varying variables, individual-level risk factors or location-level characteristics.\textsuperscript{57} We will further assume that $\beta(u)$ varies across space due to factors that differ spatially, such as climate,
socio-economic factors, susceptibility of the population, or other pollutants. Results from the random slope model thus can be compared to levels of covariates that are potential effect modifiers at the county or zip code level (see section 2.3.3. for details on the list of effect modifiers).

2.4.3. Effect modification in epidemiological studies of long-term effects: We will consider the same list of individual-level and area-level effect modifiers described in section 2.3.3. To estimate effect modification, we will extend the model by including an interaction term between the air pollutant $x_{ut} = (x_{ut} - \bar{x}_t)$ and the effect modifier $z_{ut}$. We will determine whether the variable $z_{ut}$ is an important effect modifier by assessing the statistical significance of the interaction term. We will also report the percent reduction in the variance of the random effect that is obtained when the interaction term $x_{ut} \times z_{ut}$ is included in the regression model (3). For example, if the effect modifier explains a large proportion of the spatial heterogeneity in the pollutant effect, the estimates of $\beta(u)$ will be much closer to a constant (i.e., a spatially homogeneous effect) and its variance $\tau^2$ will decrease after inclusion of the interaction term. Our approach thus provides a means to systematically assess not only statistical significance of an interaction, but also epidemiological significance of an effect modifier in explaining observed spatial heterogeneity in pollutant effects.

2.4.4. Dealing with limited information on the individual-level risk factors in the full cohort: Prospective studies of air pollution and mortality need to account for the following potential sources of bias: 1) individual-level lifestyle factors; 2) area-level characteristics; and 3) unmeasured area-level confounders that vary with pollution and mortality. The Medicare cohort provides individual-level age, gender, and race but not information on lifestyle factors. However, an advantage of the Medicare cohort is the full availability of individual-level measures of any co-morbidity assessed by hospitalization. Re-analyses of the American Cancer Society Study (ACS) and the Six Cities Studies (SCS)\textsuperscript{59-62} have reported that health risks associated with long-term exposure to particulate matter on mortality were not sensitive to adjustment for lifestyle factors.\textsuperscript{63} In addition, we have constructed race-age-gender adjusted mortality rates from Medicare and yearly average concentrations of PM$_{2.5}$ for the period 1999-2002 and for the same geographical locations included in the ACS and SCS. With these data we have successfully replicated the results of these two landmark cohort studies.\textsuperscript{32} To partially account for individual-level SES variables, we will include into the regression model area-level SES variables from the 2000 US census. In addition, we will be able to construct a very extensive list of area level confounders (at zip code levels) using the data described in section 1.2.5. To account for unmeasured confounders that vary smoothly in space, we will include a smooth function $s(u, \lambda)$ of county location $u_i$ (longitude and latitude) with degrees of freedom. A smaller $\lambda$ value corresponds to a smoother $s(u_i, \lambda)$ and less control for spatial confounding. A value of $\lambda = 0$, corresponds to no adjustment for spatial confounding. Specifically, we will fit a Generalized Additive Model (GAMM)\textsuperscript{64} within each age-race stratum:

$$\log E(Y_{uta}) = \log(T_{uta}) + \log(h_u(a)) + (x_{ut} - \bar{x}_t)\beta(u) + \bar{x}_t\gamma + s(u, \lambda)$$

We will report results by age-race stratum and aggregated across strata. We estimate the log-relative risk $\beta(u)$ for different values of $\lambda$. We plan to extend the regression model described in (4) to allow for more flexible spatial decompositions. For example, we plan to build on the approach proposed by Jansen et al. (2009)\textsuperscript{65} aimed at transforming irregularly-spaced spatial data into different spatial scales using wavelets. Transforming air pollution data, covariates, and
mortality data into such spatial scales will thereby allow us to estimate a different health risk at each spatial scale. If the estimated scale-specific associations differ by spatial scale, then there is evidence of confounding or measurement error. As a result a robust estimate of the health risk associated with long-term exposure to individual pollutants can be obtained by averaging the scale-specific associations only over those scales that are less likely to be confounded.

2.4.5. Susceptibility: We will construct cohorts of individuals who have been diagnosed for a chronic disease and will follow them prospectively for hospitalization and death for the period 2002-2014. The susceptible cohorts will include individuals who were hospitalized for a given disease at any time during the study period. We will estimate the mortality and cause-specific hospital admission risks associated with long-term exposures to individual pollutants and sources in each of these cohorts of susceptible individuals, using the methods described above. As summarized in Table 5.4, among all the individuals residing in a zip code whose centroids are within 3 miles of the PM$_{2.5}$ monitor, we will be able to construct a cohort of 4,826,000 individuals that have been hospitalized for CVD. Among this CVD cohort we have estimated 150,000 deaths and 359,999 hospitalizations during the first year. Table 5.4, also shows the sizes of the susceptible cohorts by disease (COPD, CVD and Diabetes) and the hospital admission rates and mortality rate for a follow-up period of one year.

Table 5.4: The number of individuals in each cohort (N) is in thousands

<table>
<thead>
<tr>
<th>Zip code centroid within 3 mi of PM$_{2.5}$ monitor</th>
<th>Zip code centroid within 3 mi of PM$_{2.5}$ STN monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>N</td>
</tr>
<tr>
<td>COPD</td>
<td>1,439</td>
</tr>
<tr>
<td>CVD</td>
<td>4,826</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1,619</td>
</tr>
</tbody>
</table>

The mortality rates and hospitalization rates are per 1,000 individuals and are rates for the first year in the cohort. Calculations are made for individuals residing in a zip code whose centroid is within 3 miles of the PM$_{2.5}$ monitor (left table) and of a PM$_{2.5}$ STN monitor (right table).

2.4.6. Estimating health risks of long-term exposure to sources in 30 US locations: We will estimate the health effects of long-term exposure to sources on the sub-population of the 12 million enrollees with a place of residence in the largest 30 US locations. Specifically, as discussed in section 2.3.4 and further detailed in the Exposure core, in each location, using data on PM$_{2.5}$ chemical composition from the STN, we will estimate long-term exposures to PM sources using the EPA Positive Matrix Factorization (PMF) source apportionment method. The source types that will be investigated are traffic, soil, and coal combustion, among others. Since source emissions and their profiles have been gradually decreasing over the last two decades, source apportionment analyses will be done by year. From this analysis we will estimate location-specific mortality and morbidity risks associated with several source types within each location. For each year and for each of the 30 locations we will have available yearly average exposures to six source types. These will be the new exposure variables that will be used in the regression model detailed in equation (3).

2.4.7. Pollution mixtures as a modifier of the chronic health risks: First, we will calculate the annual concentrations for PM$_{2.5}$ mass and species as well as gaseous pollutants for all the geographical locations (e.g. zip codes or counties) included in the study (see Table 5.3). Second, we will conduct cluster analyses to group the geographical locations with similar multi-pollutant profiles (see Exposure core for details). Third, we will use the regression model detailed in
equation (3) to estimate location-specific health risks associated with long-term exposure to PM$_{2.5}$ total mass separately for each group of locations. Fourth, we will investigate whether these risks differ across clusters. By investigating between-cluster variability of the health risks, we will assess the relative toxicity of the different pollutant mixtures encountered in various regions of the US.

2.5. Approach to Aim 3: In A3, we plan to conduct two studies (live births and Medicare enrollees) in MA and surrounding states. The geographical area studied in this Aim substantially overlaps with the study regions of all the other cohorts (NAS, Framingham, Viva) and therefore we will use the same exposure metrics (short and long-term exposures to individual pollutants, sources and mixtures) as the other cohort studies (Projects 2, 3 and 4).

2.5.1. Approach to A3.1: We will study approximately 700,000 singleton live births obtained from the Massachusetts Birth Registry for the period between January 1, 1996 and December 31, 2007. The address of the mother at the time of birth was geocoded by a private firm and was reassessed by us for accuracy and completeness. In this study we will investigate the effect of prenatal exposures to individual pollutants, sources and mixtures on birth outcomes (pre-term births and low birth weights) adjusted by mother’s covariates and area-level risk factors. We will also plan to identify and explain disparities in low birth weight and preterm births as a function of the individual and area-based level SES, the physical environment, and their interaction. The Massachusetts birth registry has particularly extensive individual covariates including parental education and maternal smoking not available in many other states, such as California. This study will complement the investigations of the Viva cohort study. The individual-level data on infants and pregnancy data, as well as the individual geocoded addresses of residence at time are significant advantages of this study design.

2.5.2. Approach to A3.2: We will study a cohort of 2.3 million Medicare enrollees residing in MA and the surrounding states who will be followed prospectively for the period 2000-2014. Individual-level information on cause-specific hospitalizations, age, gender, and race are available for the entire sample (see Figure 5.2). Place of residence will only be available at zip code level. However, for a subset of approximately 1,000 subjects in the cohort we will have extensive information on individual-level risk factors, available from the MCBS-linked-Medicare files.

Figure 5.2: (left) Medicare enrollees by zip code for the subpopulation having detailed covariate information from MCBS, (right) Medicare enrollees by zip code. These individual-level risk factors will be available longitudinally for the entire 14 years of follow-up. We plan to develop and apply data-augmentation methods to impute individual-level covariates for the individuals that are in the Medicare cohort but that are not part of the MCBS.$^{56,67}$ In addition, we will use several area-level confounders including SES, diabetes, smoking diseases (see section 2.1.) for 100% of the zip codes where the Medicare enrollees reside from the BRFSS.
3. EXPECTED RESULTS: The proposed research is a natural extension of our previous work, and directly responds to the following research questions posed in the RFA: 1) What are explanations for regional and temporal differences in air pollution risk? 2) What subpopulations are at increased risk of adverse health outcomes from exposure to PM (components, size fractions) and/or air pollutant mixtures? 3) How can the health effects of PM (its components/size fractions and sources) be better understood in a multipollutant context? 4) What is the exposure-response relationship for these sources and mixtures? The focus of this Project is to identify factors that explain the heterogeneity of the air pollution health risks. We hypothesize these factors include medical and social conditions, conditions that modify exposure, and compositional factors that modify the toxicity of the exposure. Moreover, we hypothesize the relevant factors will vary by outcome. Our research will be fully interactive with the other Center projects. Our previous results (e.g. diabetic susceptibility) have guided their more focused analyses, and their results are generating specific hypotheses we will examine. The strengths of this study include the use of linked national datasets, including the opportunity of conducting a cohort study of 12 million individuals followed for deaths and hospitalizations for 14 years in all US. This will provide large statistical power to investigate complex questions such as susceptibility of sub-populations and exposure-response curves, as well as regional variation and factors that potentially contribute to such variation. We have already identified some factors modifying the response to particles and ozone, including particle composition. We will extend this to a) examine a fuller set of potential modifiers; b) expand our investigation of particle components to better understand more complex pollution mixtures and sources, including exposure-response. Answering these questions is critical for agency decision making and risk assessment, because it will only identify the key aspects of air pollution in terms of risk, it will also help identify the distribution of risk in the population. The latter (susceptible populations) is a requirement of the Clean Air Act and a key focus of EPA’s efforts on environmental justice. Moreover, it is likely that the distribution of susceptibility, exposure, and pollution characteristics is not independent. Understanding susceptibility will feed directly into our cohort studies (for validation) and help identify potential biological mechanisms.

4. GENERAL PROJECT INFORMATION: Our team includes experts in analysis of national epidemiological studies on air pollution and health. Drs. Dominici (PI) and Bell (co-PI) have a track record of collaborating on air pollution projects involving analysis of large databases, including work on PM and its sources. Dr. Bell has also worked on numerous epidemiological studies of air pollutants, including research on the chemical constituents of particles and the links between exposure and pregnancy outcomes. Dr. Schwartz is a prominent environmental epidemiologist, with expertise in the effects of weather, air pollution and lead exposure on health. His work has established that exposure to fine combustion particles in the air, at concentrations well below current standards, is associated with increased cardiovascular hospital admissions and deaths. He has examined cardiovascular effects of lead exposure in adults and its cognitive and growth effects in children. Dr. Zanobetti has done extensive work in collaboration with Dr. Schwartz on the adverse effects of air pollution on health, including conducting statistical analysis in the APHEA and NMMAPS projects. She has over 10 years of experience in environmental epidemiology and has extensive experience with the statistical models to be used in this study. Dr. Dominici has established collaborations with the other investigators.
REFERENCES


14. Franklin M, Koutrakis P, Schwartz J. The role of particle composition on the association between PM$_{2.5}$ and mortality. *Epidemiol.* In press:


