

Exposure modeling

By: Brian Reich
North Carolina State University

with considerable help from

Eric Kalendra (NCSU), John Langstaff (EPA), and Ana Rappold (EPA)

E-mail: reich@stat.ncsu.edu

- ▶ **Personal exposure** is the amount of pollutant that enters a subject's lungs.
- ▶ **Ambient concentration** is the amount of pollutant in the air outside.
- ▶ The most powerful way to study the effect of a pollutant on health is

$$y_i \sim \text{Bernoulli}(f[e_i])$$

- ▶ y_i is the response for subject i
 - ▶ $f[e_i]$ is some function of personal exposure e_i
 - ▶ $i = 1, \dots$, number of people in the US.
-
- ▶ For many reasons this study will never take place.

Today we'll discuss four alternatives and their pros and cons:

- ▶ Small chamber study
- ▶ Small panel study
- ▶ Fine-scale spatial modeling
- ▶ Stochastic model for personal exposure

Multi-city time series studies

These studies (i.e., those discussed by Howard Chang) typically do not have data for individuals. Rather, they relate an aggregate response with an aggregate measure of exposure.

- ▶ Response: y_{ts} is the number of events on day t in region s .
- ▶ Predictor: x_{ts} is a measure of the spatial average ambient air pollution in region s on day t .
- ▶ Model: $\log [E(y_{ts})] = x_{ts}\beta + \text{confounders}$.

Is this a good predictor?

The predictor is a sophisticated trimmed mean of the monitors in the region. Give a list of reasons why an individual's exposure may differ from this average ambient concentration.

- ▶
- ▶
- ▶
- ▶
- ▶
- ▶
- ▶

Ecological fallacy

- ▶ Ecological inference fallacy: An error associated with making inferences about individuals based solely on aggregate data.
- ▶ Example 1: Robinson (1950) showed that state literacy rate and proportion of the population born outside the US had a positive correlation. However, individuals born outside the US were less likely to be literate.
- ▶ Example 2: In 2004 Bush won the 15 poorest states. However, only 36% of voters with income less than \$15K voted for Bush.
- ▶ Is this a problem for multi-city time series studies? (Jon Wakefield, others)

Using spatial modeling

- ▶ One way to approximate an individual level analysis is to define the regions to be as small as possible.
- ▶ Eric Kalendra (NCSU) is studying this (and many other things) for his thesis.
- ▶ He has geo-coded mortality data for North Carolina for 2001-2002.
- ▶ He is studying the effects of aggregation at different scales.

Predicting the spatial surface

- ▶ Having specific locations for the events may not be useful unless the exposure is also available at a fine spatial scale.
- ▶ There are not enough pollution monitors to make a detailed interpolation for the entire state.
- ▶ Therefore he is using the EPA's "fusion model".
- ▶ The fusion model combines monitored data with numerical model output (CMAQ) to estimate the spatial map of ambient pollution.

- ▶ CMAQ is a deterministic model that estimates ambient pollution on a grid (say $12\text{km} \times 12\text{km}$).
- ▶ The inputs are meteorology and emission data.
- ▶ Estimates are made using a series of differential equations that govern the creation and transport of pollution.
- ▶ Advantage: it's available at all spatial and temporal locations.
- ▶ Disadvantage: it's not monitoring data.

- ▶ The data fusion model combines data from monitors with CMAQ using a hierarchical Bayesian model. Here is a simplified version of their model for a single day.
- ▶ Let $\mu(s)$ be the true value of the pollution at location (lat/long point) s .
- ▶ Monitor data model: $y(s) = \mu(s) + \text{error}$.
- ▶ CMAQ data: z_i estimates the average pollution in grid cell D_i .
- ▶ CMAQ data model: $z_i = a + b * \frac{1}{|D_i|} \int_{D_i} \mu(s) ds + \text{error}$.

- ▶ The true pollution surface $\mu(s)$ has a Gaussian process prior with spatial covariance.
- ▶ a and b represent additive and multiplicative bias. They can be allowed to vary with space and/or time.
- ▶ Fitting this model using MCMC gives a posterior for $\mu(s)$.
- ▶ From this, you can estimate $\mu(s)$ at any point or its average in a grid cell.
- ▶ This could either be repeated every day or generalized using space/time models.

Effects of aggregation

- ▶ Estimated percent increase in mortality per increase of 10 units of PM2.5 using standard Poisson regression. (Results given as $est_{(se)}[z]$.)

	Region	County	Tract
PM2.5 Monitor	0.018 _(0.007) [2.7]	0.015 _(0.006) [2.5]	0.015 _(0.006) [2.7]
PM2.5 Fusion	0.021 _(0.006) [3.5]	0.027 _(0.005) [5.0]	0.032 _(0.005) [6.0]

- ▶ Tracts and counties give the same results using only monitoring data.
- ▶ However, using fusion data the tract-level analysis gives a more significant result.

How representative is ambient pollution

Give a list of reasons why an individual's exposure may differ from the ambient concentration outside their house.

- ▶
- ▶
- ▶
- ▶
- ▶
- ▶

Although we often do not observed personal exposure, there is a lot of data that can be used to estimate the distribution of exposure in the population, e.g.,

- ▶ Demographic data
- ▶ Housing information
- ▶ Personal activity diaries
- ▶ Diurnal cycles of ambient concentration
- ▶ Weather information

Simulating exposure via APEX

- ▶ APEX (Air Pollution EXposure model) generates a large number of hypothetical individuals with demographic characteristics to represent the population.
- ▶ SHEDS (Stochastic human exposure and dose simulator) is another similar model.
- ▶ Each person's daily activity is randomly assigned by drawing the personal activity diary of a person with similar characteristics from a large database of diaries.
- ▶ Hourly personal exposure is computed based on the personal activity and the hourly PM concentration.
- ▶ A complete distribution of APEX can be found in the handouts.

Simulating exposure via APEX

- ▶ APEX returns the simulated exposure for M hypothetical individuals each day. For example, on day t we have $\hat{E}_1(t), \dots, \hat{E}_M(t)$.
- ▶ The simulated exposures are used to represent the population's *exposure distribution*.
- ▶ These exposure distributions have both variability and uncertainty.
- ▶ Variability: the true variation in exposure across people, $\text{Var}(E_1(t), \dots, E_M(t))$.
- ▶ Uncertainty: reflection of our incomplete knowledge of the exposure distribution, e.g., $\text{Var}_{MC}(\text{Var}(\hat{E}_1(t), \dots, \hat{E}_M(t)))$.

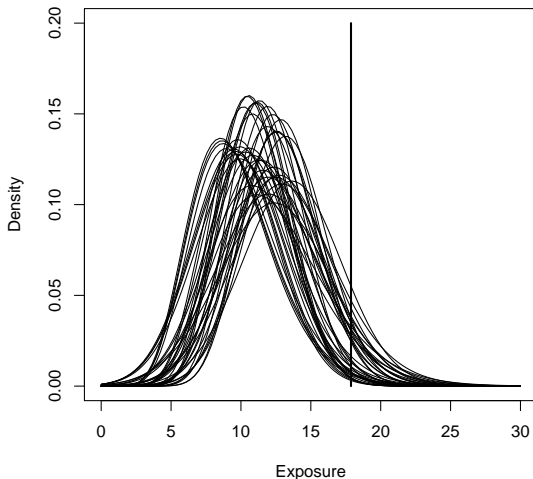
Sources of uncertainty:

- ▶ Randomization of diaries.
- ▶ Uncertainty in the met and ambient concentration inputs.
- ▶ Uncertainty in the physical constants, like air exchange rates.
- ▶ More...

To account for these sources of uncertainty, we ran the model several times, drawing different inputs from their priors each replication.

SHEDS output for a typical day

This plot shows several replications of the exposure distribution (curves) compared to that day's ambient concentration (line).



Relating the simulated exposure distribution with mortality

- ▶ We start by modeling each individual, and then build to a model for the whole population.
- ▶ For a single individual i , let $y_i(t)$ be 1 if they died on day t , 0 otherwise.
- ▶ Temporarily ignore confounders.
- ▶ For a rare event, we can approximate

$$y_i(t) \sim \text{Poisson} \left(\exp \left(a + \hat{E}_i(t)b \right) \right)$$

where b is the relative risk for PM.

Relating the simulated exposure distribution with mortality

- ▶ We don't observe $y_i(t)$, we only observe the total number of events on day t , $y(t) = \sum_{i=1}^M y_i(t)$.
- ▶ Assuming responses are conditionally independent, $y(t) \sim \text{Poisson} \left[\sum_{i=1}^M \exp \left(a + \hat{E}_i(t)b \right) \right]$.
- ▶ This gives a model for the response as a function of the simulated exposure.
- ▶ This "convolution model" is straight-forward to fit using ML or MCMC.

Relating the simulated exposure distribution with mortality

- ▶ The analysis can be simplified by assuming the exposure distribution is Gaussian.
- ▶ Assume the exposure distribution is Normal($\bar{E}(t)$, $s^2(t)$) with density $f(E)$.
- ▶ Then

$$\begin{aligned} E(y(t)) &= \sum_{i=1}^M \exp(a + \hat{E}_i(t)b) \\ &\approx \int \exp(a + Eb) f(E) dE \\ &= \exp(a + b\bar{E}(t) + b^2 s^2(t)/2) \end{aligned}$$

Relating the simulated exposure distribution with mortality

$$\log E(y(t)) \approx a + b\bar{E}(t) + b^2 s^2(t)/2$$

- ▶ If all subjects have the same exposure, then $s^2(t) = 0$ and $\log E(y(t)) \approx a + b\bar{E}(t)$. Further, if exposure is linear in the ambient concentration, this is equivalent to the usual model.
- ▶ Also, if there is not effect and $b = 0$, then this is the same as the usual model.
- ▶ So accounting for exposure is most important with large exposure variability a strong health effect.

Accounting for uncertainty

We don't know $\bar{E}(t)$ or $s^2(t)$ exactly, so we should account for this uncertainty. The final model we fit was:

- ▶ $y(t) \sim \text{Poisson} [\exp (a + b\bar{E}(t) + b^2s^2(t))]$
- ▶ $\bar{E}(t) \sim N (\mu(t), \sigma^2(t))$.
- ▶ $s^2(t) \sim \text{InvG} (e_1(t), e_2(t))$.
- ▶ Where $\mu(t)$, $\sigma^2(t)$, $e_1(t)$, and $e_2(t)$ are known constants estimated from the multiple simulator runs on day t .
- ▶ For example, we ran the model N times and took $\sigma^2(t)$ to be the variance of the N exposure distribution means.

Comparing the effects of concentration and exposure

- ▶ The daily exposure distributions for $PM_{2.5}$ and three diameters of fine particles are simulated using SHEDS.
- ▶ We analyze mortality using both the ambient concentrations and the simulated exposure distributions.
- ▶ We included the usual confounders for time trend, meteorology, etc.
- ▶ We let the lag be an unknown random variable with $U(\{0, \dots, 7\})$ prior.

Median and 95% intervals for the relative risks, $\exp(b)$

Diameter	Ambient Concentration	Exposure Distribution
<i>DIC</i> (p_D)	2170.9 (10.0)	2172.2 (10.6)
0.02 μm	1.008 (0.946, 1.073)	1.020 (0.796, 1.319)
0.05 μm	1.038 (0.978, 1.115)	1.060 (0.948, 1.207)
0.20 μm	0.945 (0.885, 1.003)	0.945 (0.882, 1.019)
$PM_{2.5}$	0.975 (0.921, 1.032)	0.958 (0.875, 1.056)

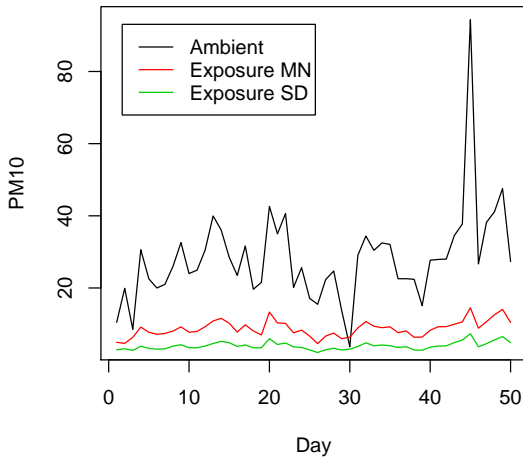
- ▶ For the most part, the relative risks are the similar for both regressions.
- ▶ The intervals are generally wider when accounting for uncertainty in the exposure.

Simulation study

- ▶ How much power could be gain using simulated exposure?
- ▶ To answer this question we conduce a brief simulation study.
- ▶ We use real APEX data, simulated for 2538 days in Chicago.
- ▶ We compare this using ambient mean for the same days from the NMMAPS data set.
- ▶ I put the data and code on my website
<http://www4.stat.ncsu.edu/~reich/Code/>

APEX versus monitoring data

Chicago data, first 50 days

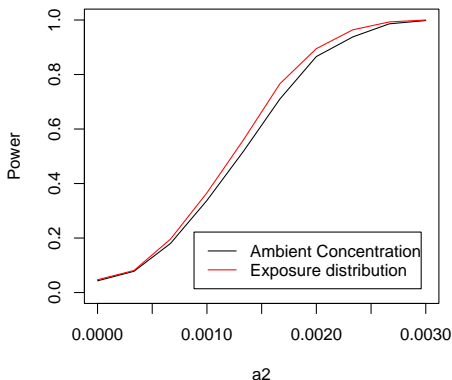


Generating data

- ▶ Each day we generate binary responses for $M = 10,000$ people.
- ▶ We first generate their exposure from the APEX distribution, $E_i(t) \sim N(\bar{E}(t), s^2(t))$.
- ▶ We then generate the response $y_i(t) \sim \text{Bern}[\exp(a_1 + a_2 E_i(t))]$.
- ▶ From the individual response, we calculate $y(t) = \sum_{i=1}^M y_i(t)$.

- ▶ We analyze $y(t)$ for each data set using two different Poisson regression model.
 - ▶ $\log(E(y(t))) = a_1 + a_2[\text{ambient concentration day } t]$
 - ▶ $\log(E(y(t))) = a_1 + a_2\bar{E}(t) + a_2^2s^2(t)/2$
- ▶ We simulated 500 data sets for several values of a_2 .
- ▶ For each a_2 , we calculate the power for each regression model.

Power of various methods



The largest gain in power is with $a_2 = 0.0017$. The power is 0.77 with APEX compared to 0.71 without APEX.

Summary of the exposure model approach

- ▶ Accounting for variability did not have a large effect in this study.
- ▶ It is likely that we would see more impact in a study of a more heterogeneous population.
- ▶ The simulation showed that moderate gains in power are possible.
- ▶ Extending this approach to a nationwide study would be great. This poses several computational challenges. One approach would be to develop a statistical emulator for APEX/SHEDS.
- ▶ How to combine APEX/SHEDS with chamber and panel data?

Advantages of each kind of study

- ▶ Chamber study:
- ▶ Panel study:
- ▶ Fine-scale spatial epi study:
- ▶ Fine-scale using SHEDS/APEX:

Disadvantages of each kind of study

- ▶ Chamber study:
- ▶ Panel study:
- ▶ Fine-scale spatial epi study:
- ▶ Fine-scale using SHEDS/APEX:

Clinical Human Studies

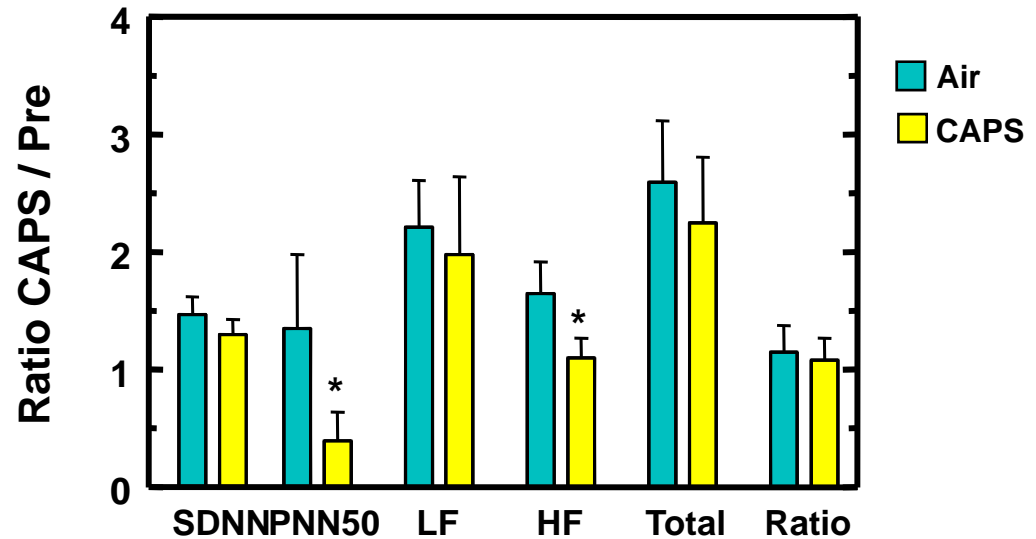
Ana Rappold, Statistician
Clinical Research Branch (CRB),
NHEERL, US EPA
SAMSI, October 2009

At CRB scientists use three complementary approaches to investigate the health risks posed by exposure of humans to environmental pollutants:

- Human Exposure Studies: Controlled and Panel Studies**
- Pharmacokinetic and Dosimetry Studies**
- In Vitro Studies**

Controlled Human Exposure Studies

- CRB Scientists use controlled human exposure studies to investigate the effect of pollutants on the respiratory, cardiovascular, and neurological systems
- CRB researchers have access to specialized chambers for controlled human exposure studies to gaseous air pollutants, particles, acid aerosols, organic compounds



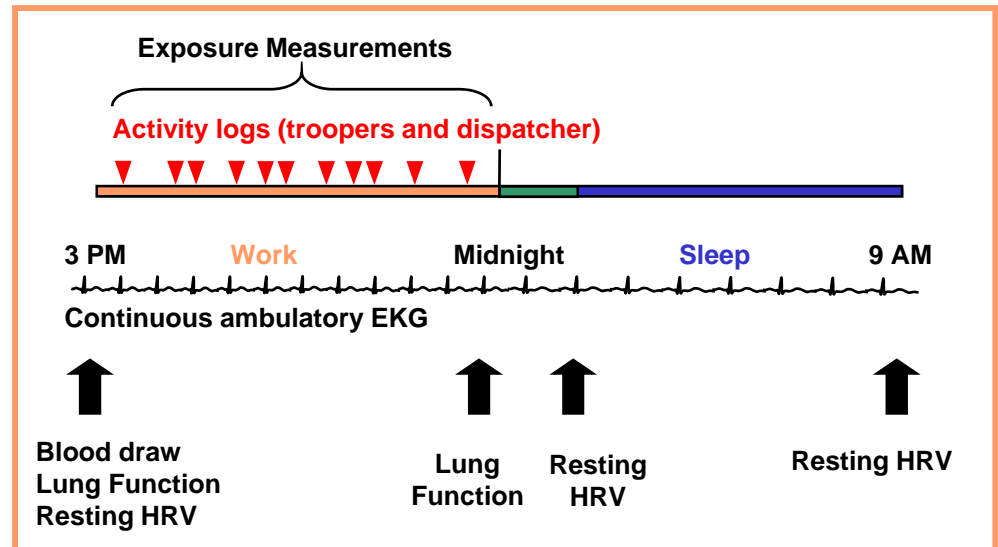
Panel studies

Panel Studies are used to investigate the effect of pollutants on the respiratory, cardiovascular, and neurological systems in susceptible population (Diabetics, Asthmatics)

Panel studies use personal and/or indoor monitoring, rooftop, and central cite monitors to approximate daily exposure

Biological Endpoints Measured

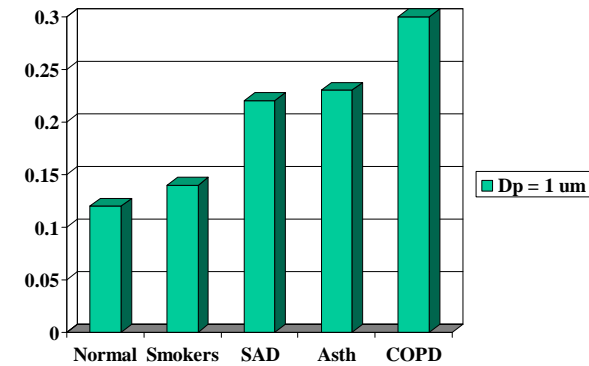
- Pulmonary function
- Impairment of host defense systems
- Cellular assays of inflammation and injury
- Cardiac rhythm and variability in heart rate
- Changes in vascular components
- Neurobehavioral and sensory function



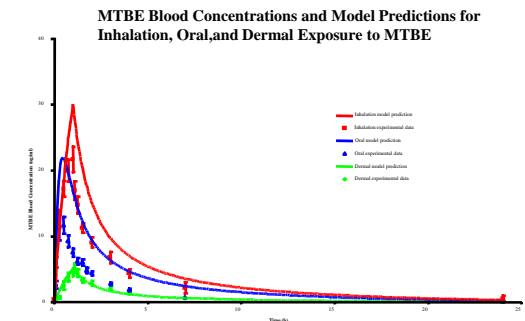
Pharmacokinetic and Dosimetry Research



Pharmacokinetics (PK) research describes the absorption, distribution, metabolism, and elimination of pollutants.



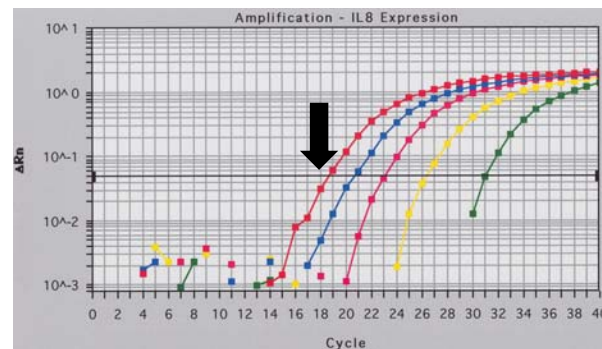
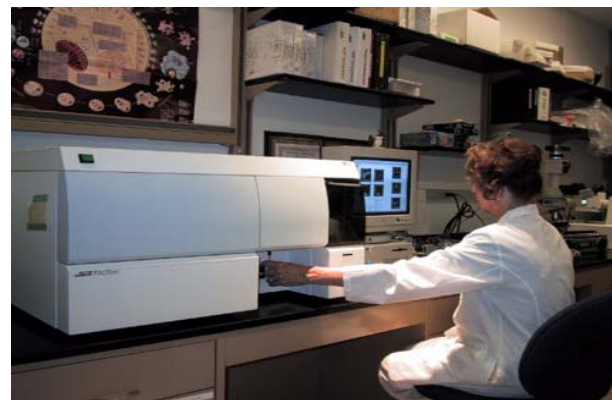
Dosimetry research characterizes the deposition and fate of pollutants taken up in the respiratory tract.



In vitro studies

Biochemical, molecular and immunological changes are measured in cells and fluids obtained from the respiratory tract of humans following exposure to pollutants.

Cells are also exposed to pollutants using in vitro exposure chambers and cutting-edge molecular biology techniques are used to study the mechanisms by which pollutants induce adverse health effects.



Asthmatic Panel Study-MASAES

- **Study is designed to determine if moderate to severe asthmatics respond more severely to air pollution than those with mild asthma**
- **The hypothesis is that a degree of inflammation due to the exposure to particulate matter is larger in moderate to severe asthmatics than in mild asthmatics and is linked to the degree of allergic inflammation present in the airway.**

Typical Panel Study Design

Subject Population

- 30 Moderate to Severe Asthmatics
- 30 Mild Asthmatics
- 30 Normal Controls

All subjects are followed for 5, 24 hour segments, approximately 1 week apart.

Biological Measurements

Pulmonary Function, Blood, Exhaled Nitric Oxide, Exhaled Breath Condensate, Heart Rate Variability, Induced Sputum

Surveys

Daily Activity and Medication Usage Survey
24 hour Exposure Survey
Residence Survey, Asthma History Survey

Exposure Assessment

Personal monitors (15 subjects), Rooftop Monitors, State and Federal Monitors
PM10, PM2.5, O₃, SO₂, NO₂, and CO
Activity monitor

Heart Rate Variability (Brook, R.D)

Epidemiologic studies provide strong and consistent evidence implicating PM2.5 to the adverse effects occurring in the cardiovascular system

Risk increase is small but it affects all of population

American Heart Association recognizes air pollution as a significant contributor to cardiovascular disease

The effect has been demonstrated for acute and chronic exposures

Chamber exposure studies provide experimental evidence

Panel studies provide evidence on susceptible population

Biological Pathways

Biological mechanisms are established based on human, animal, basic science, and molecular level experiments

Three main pathways are identified in Brooks review:

- 1) neural pathways (parasympathetic withdraw and sympathetic activation of the nervous system)
- 2) release of pro oxidative and pro inflammatory mediators from the lungs inducing secondary cardiovascular response
- 3) soluble PM components entering the blood circulation directly

All three involve oxidative stress

Exposure Mechanism

There are many unanswered questions:

- **PM is a complex mixture that varies – are components be equally harmful**
- **Are regional differences in the effect size related to regional exposure differences**
- **Does the source of components matter**
- **Do the characteristics of particles matter**
- **Characterization of exposure - central site monitoring, personal monitoring, indoor, outdoor monitoring?**

Typical Project

PM0.1, PM2.5 and PM10 from the chambers, personal and rooftop monitors are analyzed for metal components by

- Xray Fluorescence (XRF)**
- Inductively coupled plasma mass spectrometry (ICP-MS)**

Statistical Analysis:

- 1) Temporal modeling of 20 or so dimensional outcome**
- 2) Temporal modeling of the relationship between personal and rooftop outcomes**
- 3) Linking biological markers to metal components**