Statistical Methods for Time Series Analyses of Air Pollution and Health

Francesca Dominici
Department of Biostatistics
Bloomberg School of Public Health
fdominic@jhsph.edu

Outline

- **Background**
- **Statistical methods for estimating air pollution risks from time series data**
- **Hierarchical models for combining risks across locations**
- **The Medicare Study**

Background

From crisis to questions

- **We began with crisis---London 1952 and have moved to questions:**
  - **Are there adverse effects of today’s air pollution?**
  - **How large are these risks?**
  - **What is the cost-benefit ratio for control?**

Background

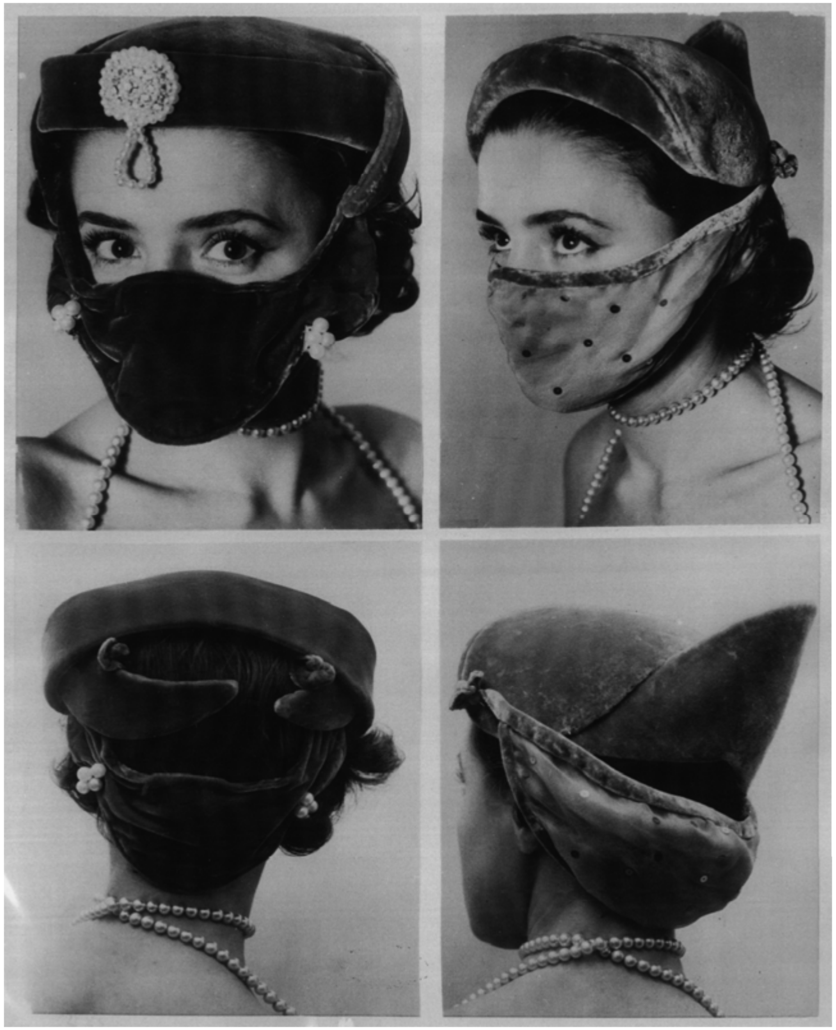
- London Fog 1952: five fold increase in death rates associated with air pollution episode of four days
- Several epidemiological studies in the 70's and 80's
- The state of the science 1980's did not establish a sufficiently robust link between air pollution and death and called for additional investigations

Daytime in London, 1952

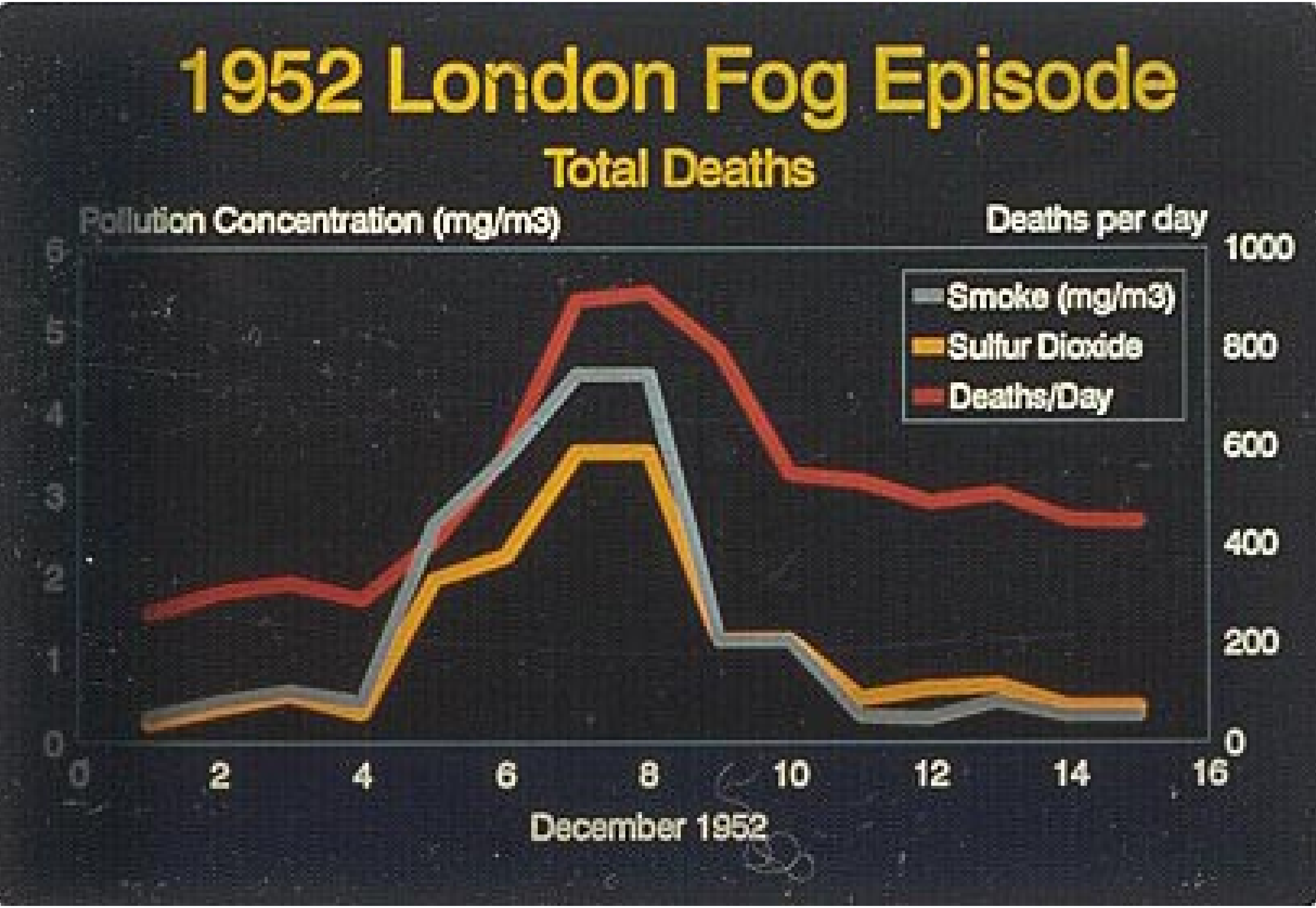
Particulate levels – 3,000 μg/m^3

Source: National Archives

Designer Smog Masks - London 1950's



Source: DL Davis. *When Smoke Ran Like Water* (2002)



1952 London Fog Episode
Total Deaths
Pollution Concentration (mg/m3)
Deaths per day
Smoke (mg/m3)
Sulfur Dioxide
Deaths/Day
0
1
2
3
4
5
6
0
200
400
600
800
1000
0
2
4
6
8
10
12
14
16
December 1952

13 DECEMBER 2002 VOL 298 SCIENCE www.sciencemag.org

AIR POLLUTION

Counting the Cost of London's Killer Smog

LONDON—In December 1952, an acrid yellow smog settled on this city and killed thousands of people. The catastrophe, known as the "Big Smoke," was a turning point in efforts to clean up polluted air in cities across the Western world. It has taken half a century, though, for some of the fog to clear around the death toll from the roiling sulfurous clouds. New research suggests that the U.K. government might have underestimated the number of smog-related deaths by a factor of 3.

Experts agree that the foul fog, which descended on London for a weekend in December 1952, killed roughly 4000 people that month alone. But researchers are now sparring over the cause of death of another 8000 Londoners in January and February 1953. Fresh analyses, debated at a conference here earlier this week to mark the 50th anniversary of the Big Smoke, suggest that these people succumbed to delayed effects of the smog or

CREDIT: SYNGENTA

50th Anniversary Meeting

- 4,000 first week
- 8,000 over next 2 months
- Pollution or flu or both?

Can air pollution kill at order of magnitude lower doses?

- Air pollution: many constituents
  - Particles (<2.5 microns penetrate to deep lung)
  - Ozone
  - Gases: NO_2, SO_2, CO
  - Many others

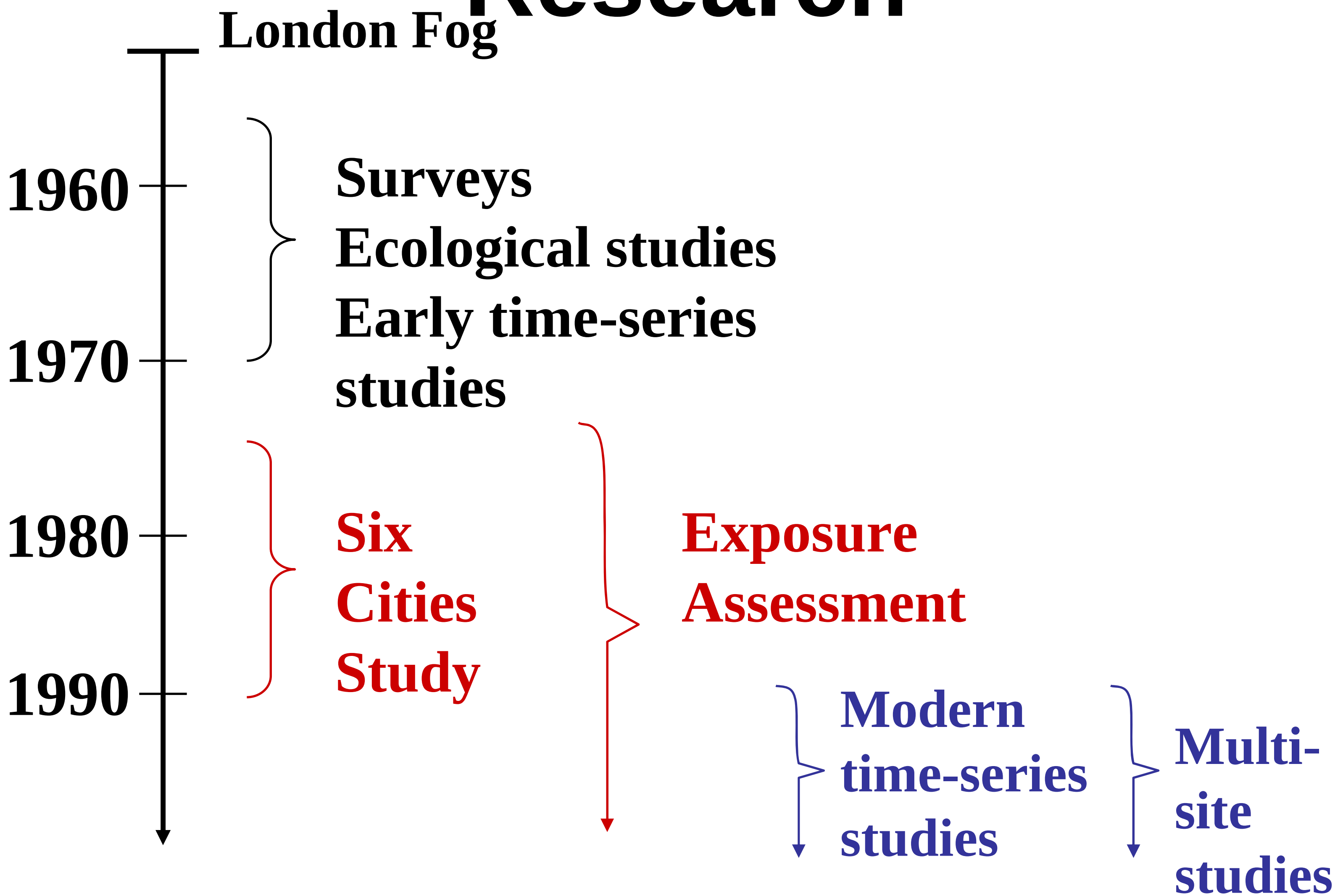
- Focus on particles because of epidemiologic data

APS Power Plant, Four Corners Area

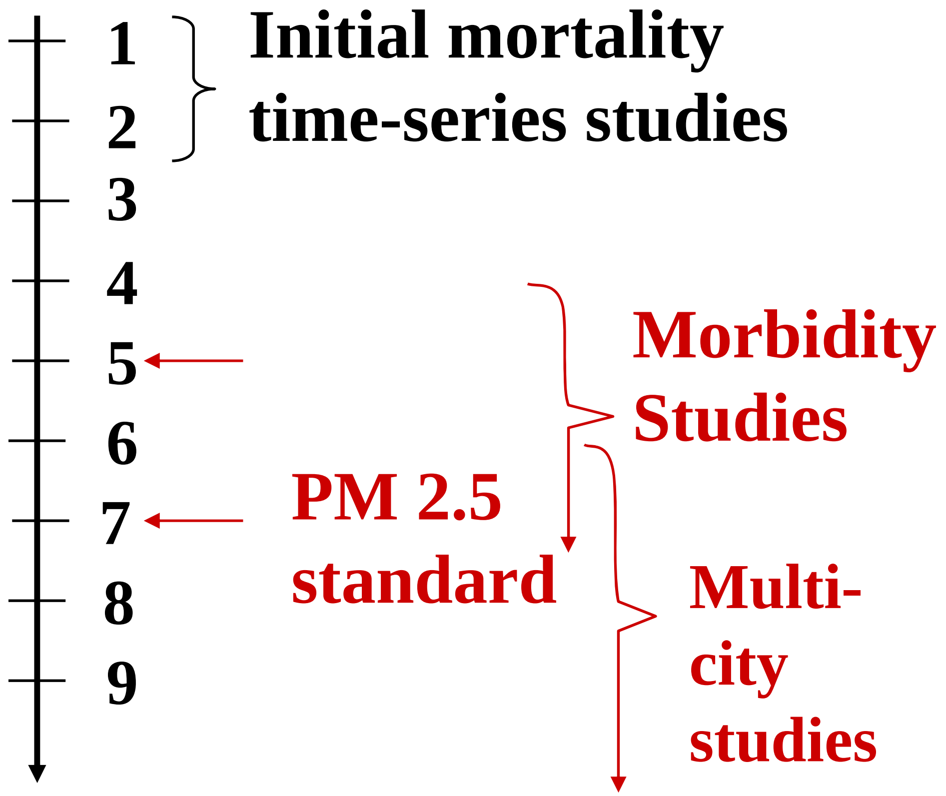
How do we study the risks of air pollution and health?

- **Epidemiological approaches:** **time-series, cross-sectional, cohort and panel**
- **Clinical studies:** **controlled exposures of normal and susceptible persons**
- **Laboratory research:** **animal exposures, *in vitro* approaches, genetic approaches**

50 Years of Air Pollution Research



PM Research in the 1990s



Issues

- Validity of findings
- Mortality displacement
- Confounding
- PM effect?
- Measurement error
- Toxic particles

Social, Political, and Regulatory Context

- EPA's process for review of the National Ambient Air Quality Standard (NAAQS) creates a sensitive political context
- Approx $100 million annually are spent in the United States alone to address uncertainties in the understanding of the health effects of particulate matter
- Several expert committees are created such as: Clean Air Act Scientific Advisory Board (CASAC) of the EPA, committees of National Academy of Sciences, World Health Organization and many others

Findings in the 90's

Cohort Studies

- **Harvard Six Cities Study (**1993): the risk of death in high polluted areas were **26%** larger than areas with lower pollution levels
- **American Cancer Society Study (1995):** the risk of death in high polluted areas were **17%** larger than areas with lower pollution levels
- **American Cancer Society (2002):** 10 units elevation in fine particulates concentrations was associated with a **4%, 6%** and **8%** in all-causes, cardiopulmonary and lung cancer mortality

Multi-Site Time Series Studies

- **APHEA Study (2000): 0.6%** increase in total mortality per 10 units elevation in particulate concentrations
- **NMMAPS Study (2002): 0.21%** increase in total mortality per 10 units elevation in particulate concentrations

Public Health Significance

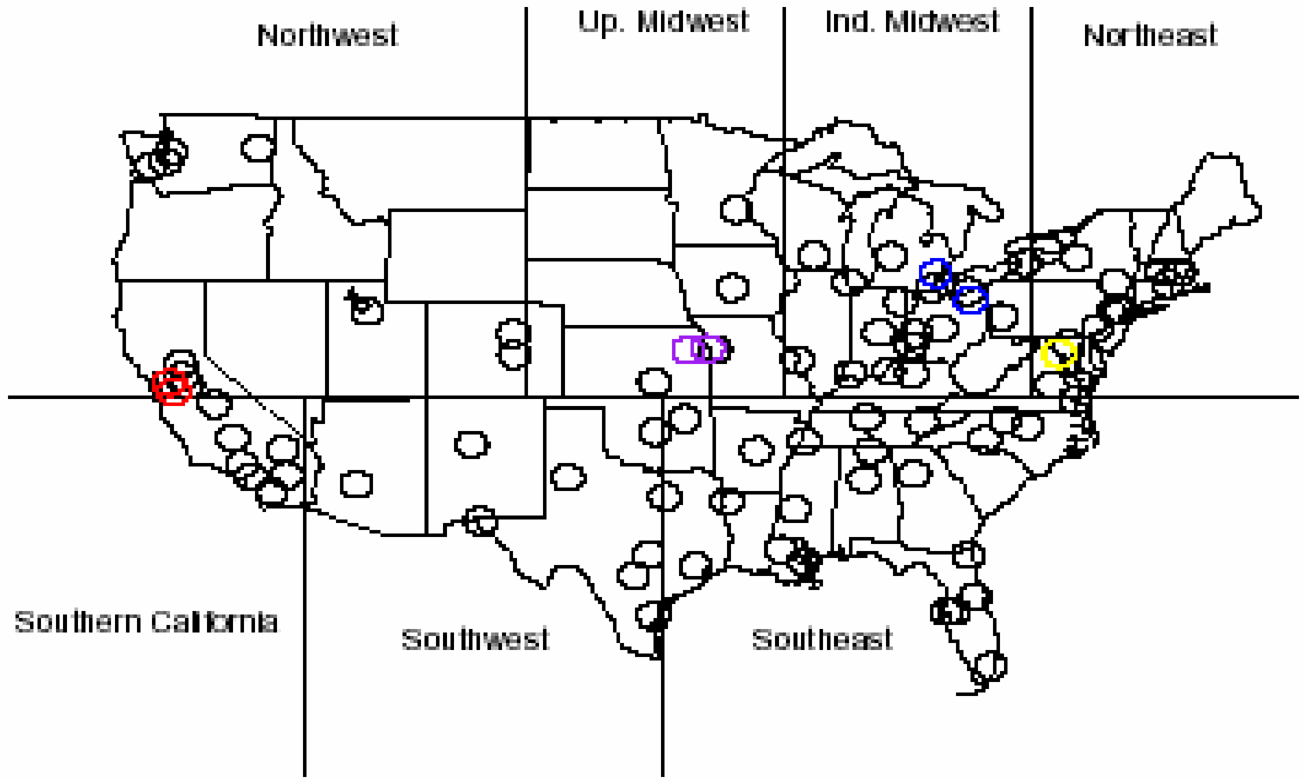
In US, EPA estimates on order of 10,000 particle-attributable deaths per year if cohort relative risks represent a causal effect

Smoking – 400,000 smoking attributable deaths per year

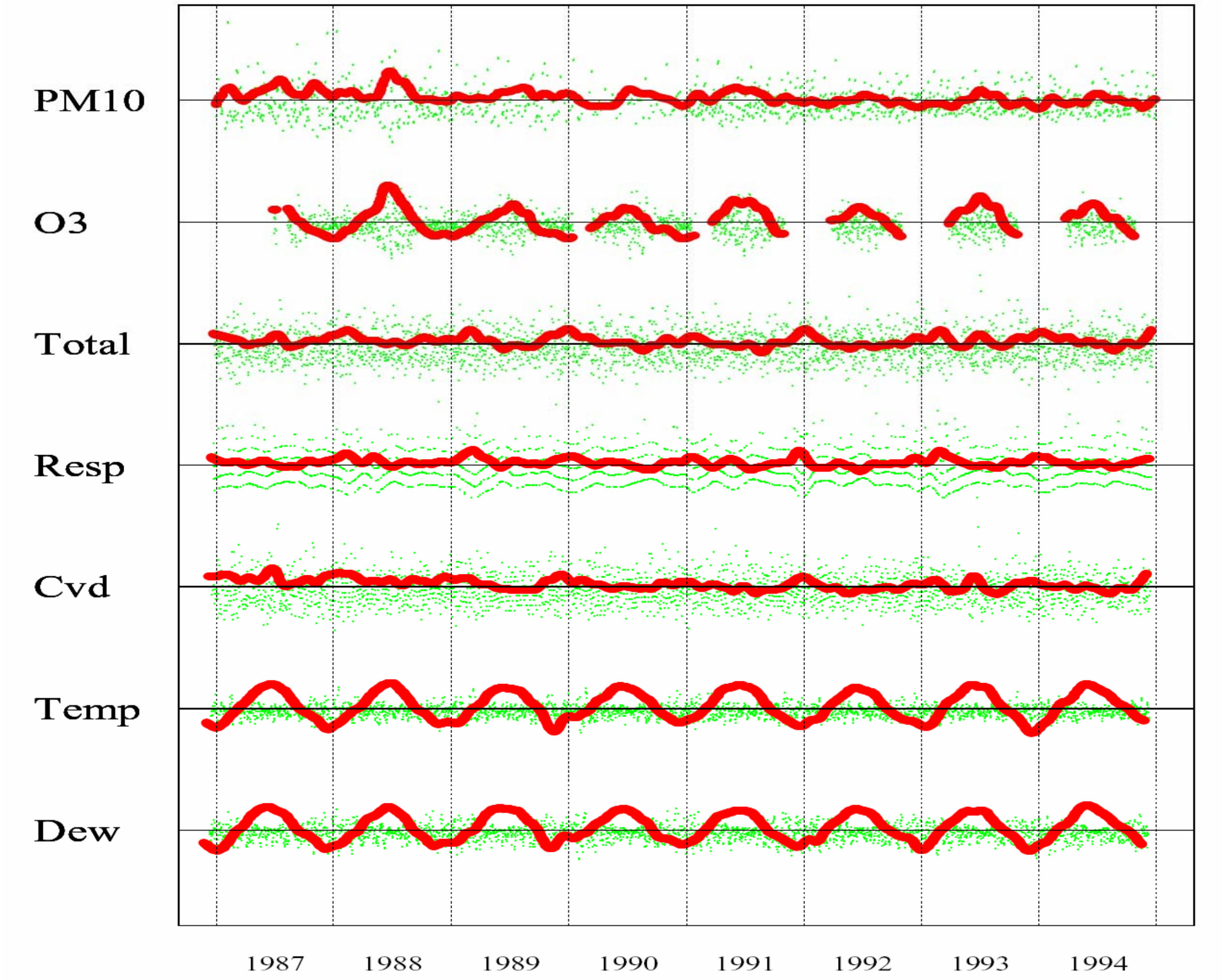
National Morbidity and Mortality Air Pollution Study (NMMAPS)

- HEI funded collaboration of Johns Hopkins and Harvard Universities; Jon Samet, PI
- 90 largest U.S cities covering roughly 40% of annual deaths
- 1987- 1994; now updated through 2000
- Mortality and hospitalizations (14 cities)

NMMAPS Locations



Data for Baltimore, Maryland



Multi-city Time Series Studies of Acute Effects

- Compare higher to lower polluted days within the same community
- Avoid problem of unmeasured differences among populations
- Key confounders
  - Longer-term trends in population characteristics, medical practice, etc
  - Seasonal effects of infectious diseases and weather
  - Day of month, week, holidays

Statistical Methods for Estimating Air Pollution Effects from Time Series Analyses

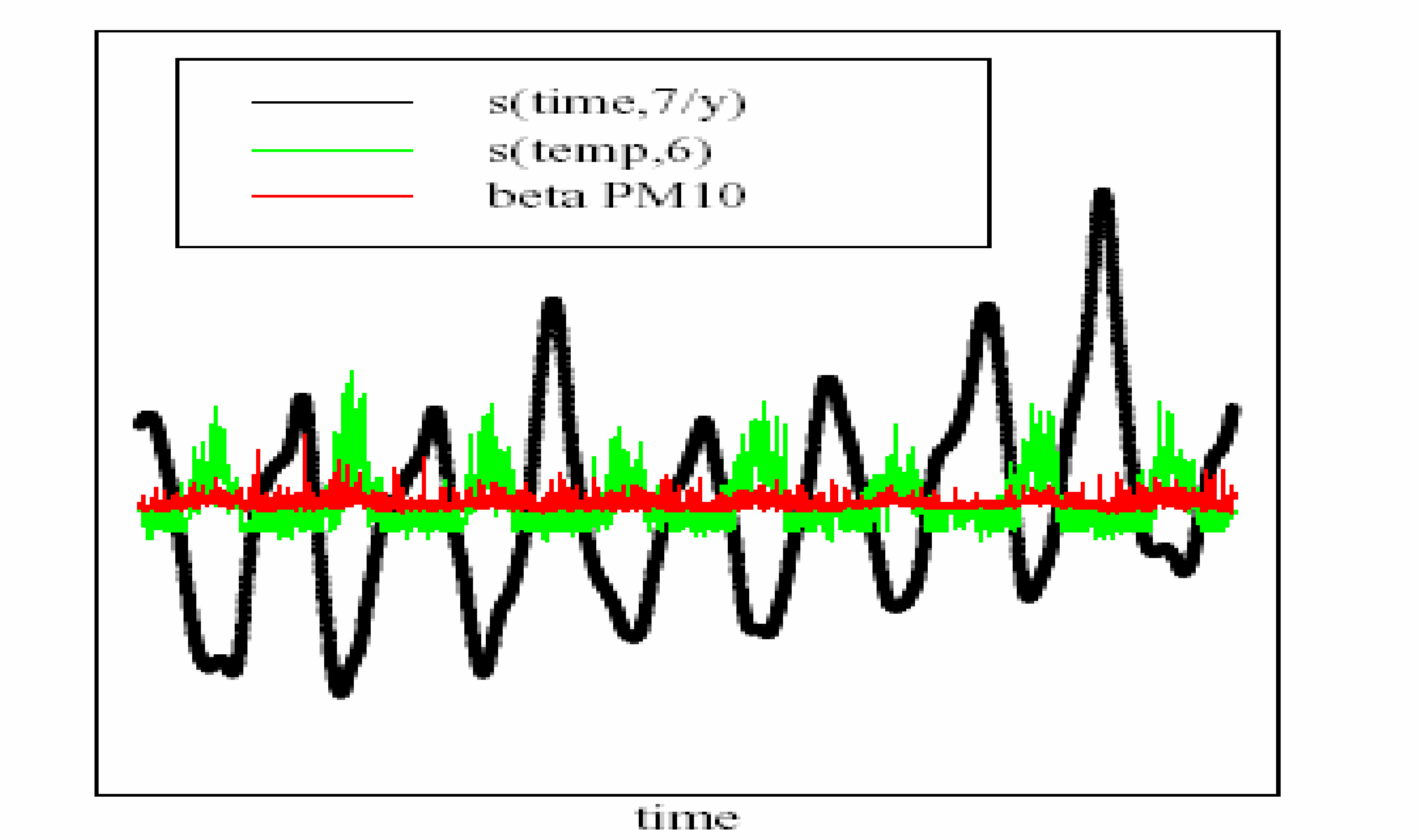
Statistical Methods

- **Within city. Semi-parametric regressions for estimating associations between day-to-day variations in air pollution and mortality controlling for confounding factors**
- **Across cities. Hierarchical Models for estimating:**
  - **national-average relative rate**
  - **national-average exposure-response relationship**
  - **exploring heterogeneity of air pollution effects across the country**

Challenges

- **For any given city, we try to estimate a small pollution effect relative to confounding effects of trend, season and weather**
- **Strong role of other time-dependent factors**
- **High correlation between non linear predictors**
- **Sensitivity of findings to model specifications**

Air pollution signal order of magnitude smaller than confounders



Estimates of model predictors in the GAM model Pittsburgh (1987-1994)

Outcome and Exposure

- Outcome:
  - daily number of deaths for all causes cardiovascular-respiratory and other causes mortality
  - daily number of hospital admissions for respiratory symptoms, asthma, or COPD
- Exposure:
  - Air pollution level on the same day, or the day before
  - Average air pollution levels in the past few days
  - Time-scale decomposition of air pollution time-series

Confounding

- The association between air pollution and mortality is potentially confounded by:
  - Weather: mortality is higher at low and high temperature
  - Seasonality: mortality generally peaks in winter because of influenza epidemics
  - Long-term trend: improvement in medical practice lower mortality over time
- All these phenomena cannot be attributed to air pollution

Measured and Unmeasured Confounding

- **Measured confounders: time-varying covariates that are associated with pollution and mortality, as for example weather variables**
- **Unmeasured confounders: other time-varying factors that generally are not measured. For example, influenza epidemics and trends in survival that affect mortality and are temporally associated with variations in air pollution**
- **Goal: estimate associations between day-to-day variations in air pollution and day-to-day variations in mortality taking into account measured and unmeasured confounders (reducing confounding bias)**

Semi-parametric regression model

- Y_t^c is the mortality count on day t at location c
- PM_{10t}^c is the level of particulate matter on day t at location c
- β^c is the percentage increase in mortality for $10\ \mu g/m^3$ increase in PM_{10}
- $s(\text{temp}, df)$ are smooth functions

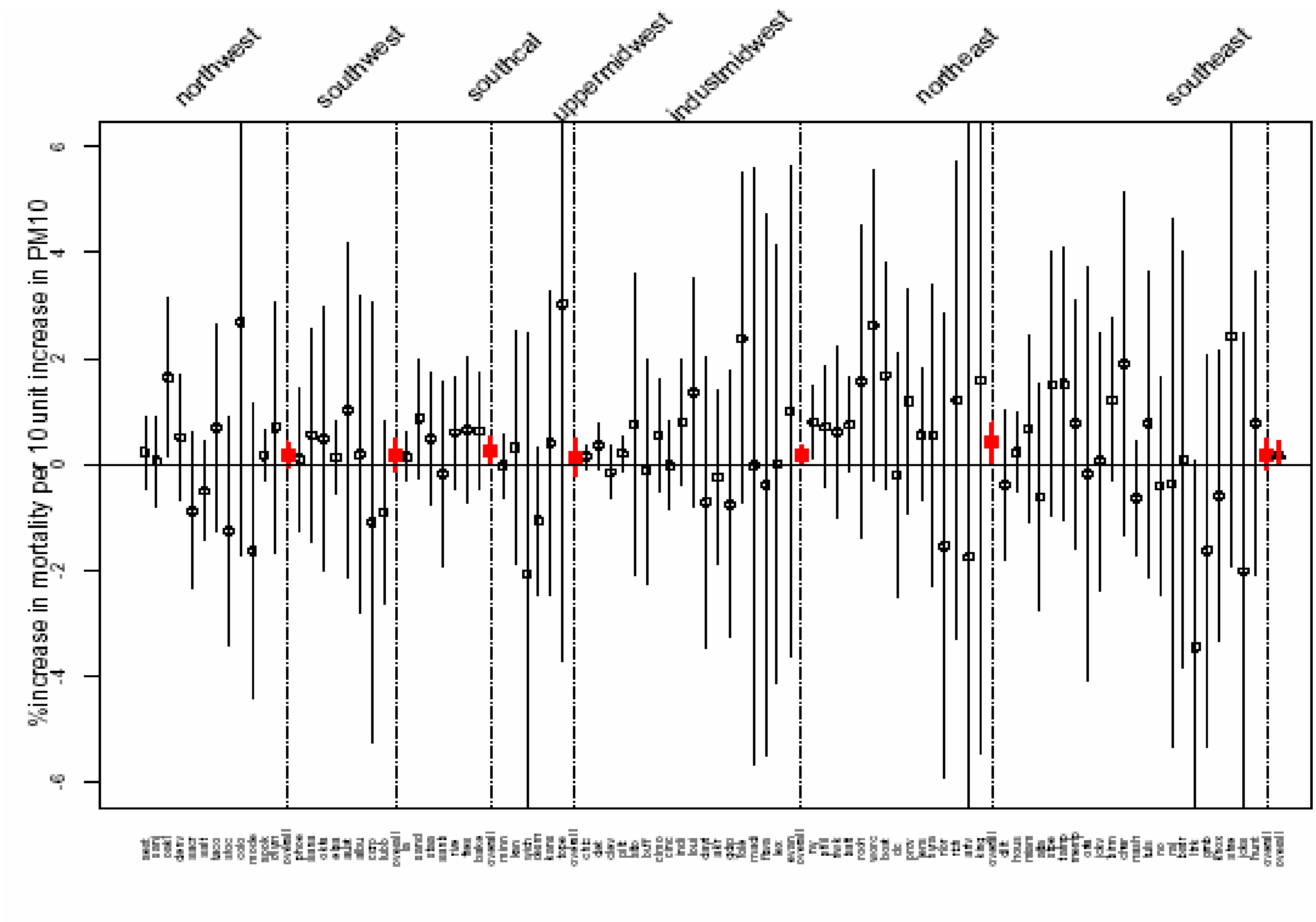
$$\begin{aligned}\log E[Y_t^c] &= \text{age-specific intercept} + \beta^c PM_{10t}^c + s(\text{time}, 7/\text{year}) + \\ &+ s(\text{temp}, 6) + s(\text{dewpoint}, 3) + \text{age} \times s(\text{time}, 8) + \ldots.\end{aligned}$$

$$\log E[Y_t^c] = \beta^c PM_{10t}^c + \text{confounders}$$

- estimate $\hat{\beta}^c$ and its statistical variance v^c within each location
- (GAM) with smoothing splines or
- (GLM) with natural cubic splines are methods of choice

Kelsall Zeger Samet AJE (1997) Dominici Zeger Samet RSSA (2000)

City-specific and regional estimates



Adjustment for confounders

How do we choose the degree of adjustment for confounding factors that leads to air pollution effect estimates with optimal statistical properties?

Adjusting for confounders

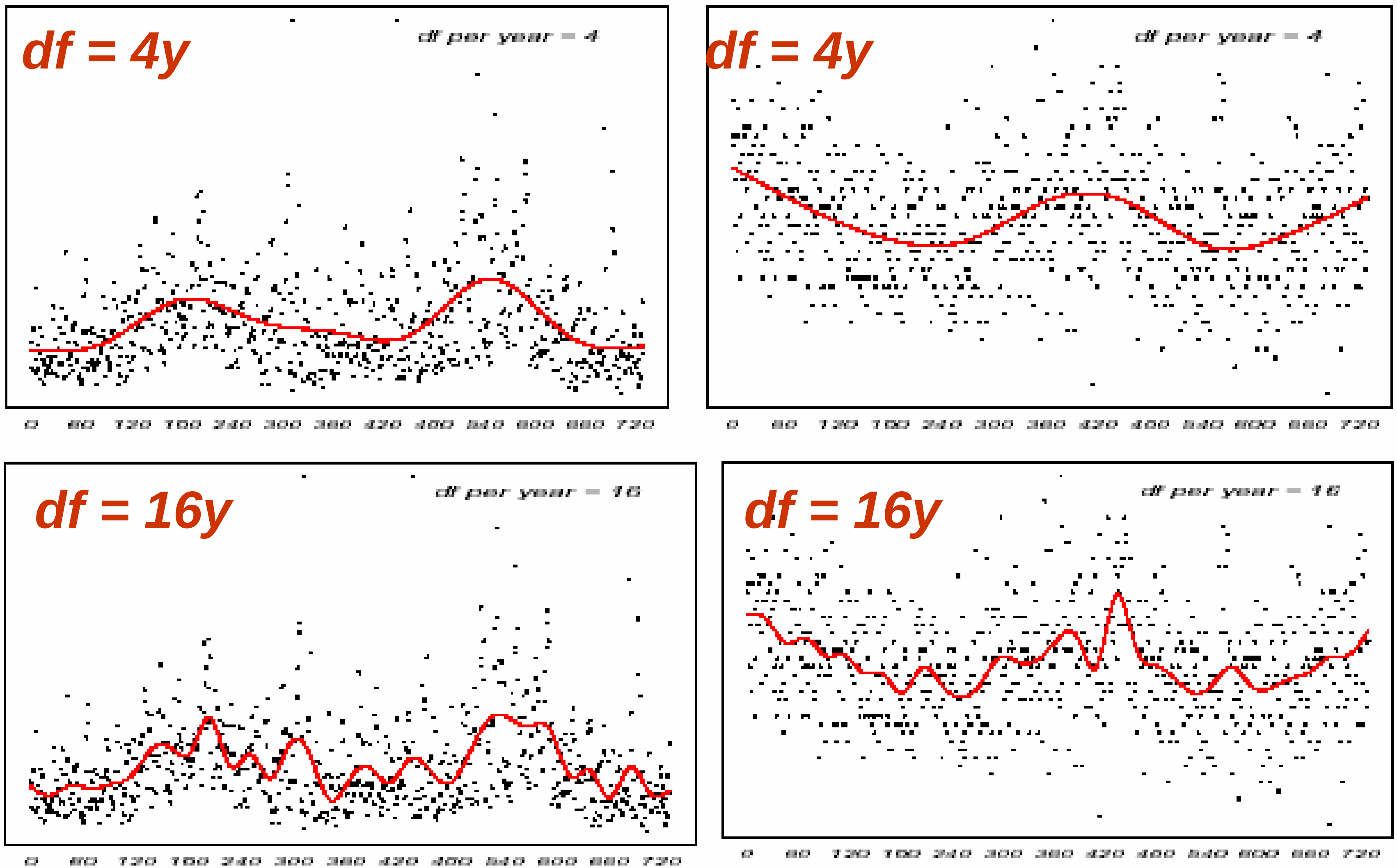
- The choice of the degree of adjustment for confounding factors is critical and still unresolved
- Need to select degrees of freedom such that:
  - are large enough to remove confounding
  - are not <u>too</u> large to wash out the air pollution signal

Choosing the number of df in the smooth function of time

- **Choosing too small a *df***
  - **Over-smoothing**
  - **Leaves temporal cycles in the residuals**
  - **Confounding bias might occur**
- **Choosing too large a *df***
  - **Under-smoothing**
  - **Removes all temporal variability in residuals**
  - **Wash out the pollution effect**

Adjusting for unmeasured confounding

PM10 (1987-88) | Mortality (1987-88)



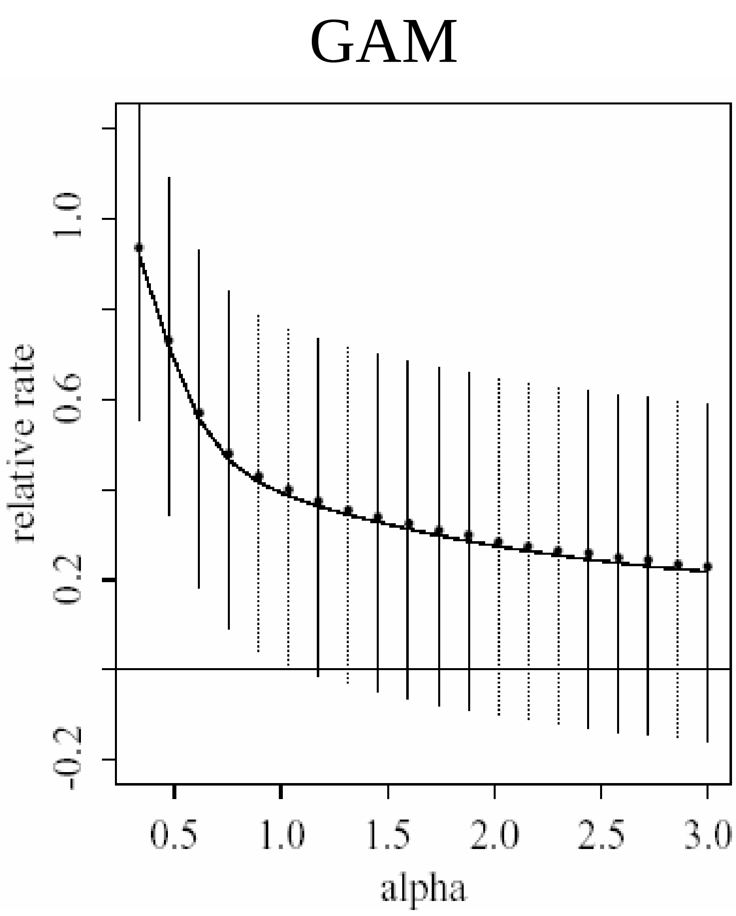
Pittsburgh NMMAPS data

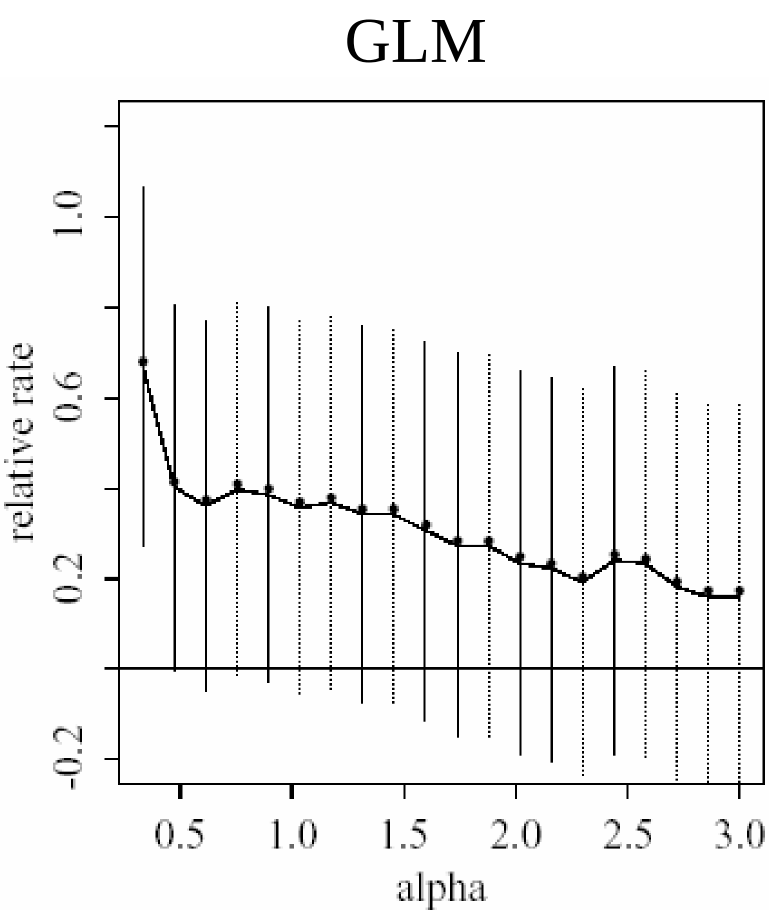
A calibration approach

$$\log E[Y_t^c] \;=\; \beta^c(\alpha) PM_{10t}^c + s_1(\text{time}, 7 \times \alpha) + s_2(\text{temp}, 6 \times \alpha)$$

Multiply all degrees of freedom by a calibration parameter α, which measures the degree of adjustment for confounding factors

Relative rates estimates as function of the degree of adjustment for confounding factors





How do we choose the degree of adjustment for confounding factors?

- **Report city-specific or national average air pollution effects as function of the degree of adjustment for confounding factors**
  - **It provides an important piece of evidence**
  - **Appropriate for multi-site time series studies**
  - **A fully transparent analysis**

A calibration approach

Semi-parametric regression model
a calibration approach

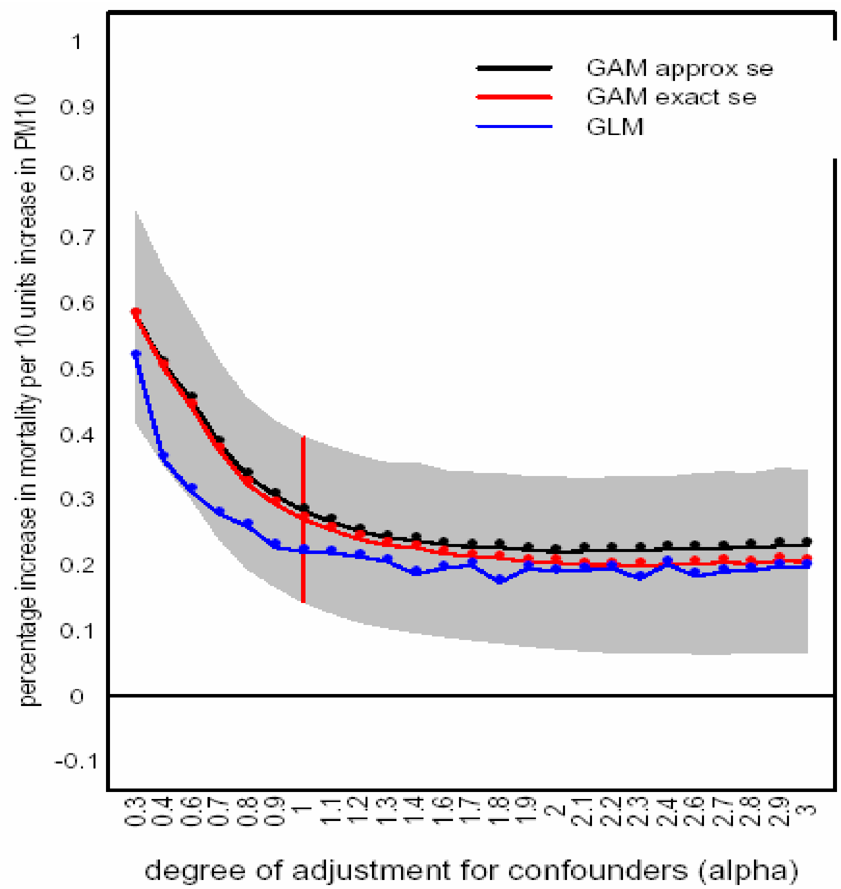
$$\log E[Y_t^c] \ = \ \beta^c(\alpha) PM_{10t}^c + \sum_j s_j(h_t, \lambda_j \times \alpha)$$

- h_t are the time-varying confounders such as time, trend and weather
- λ_j are degrees of freedom
- estimate $\hat{\beta}^c(\alpha)$ and its statistical variance $v^c(\alpha)$ within each location
- Goal: pool $\hat{\beta}^c(\alpha)$ across locations and estimate a national-average relative rate as function of α.

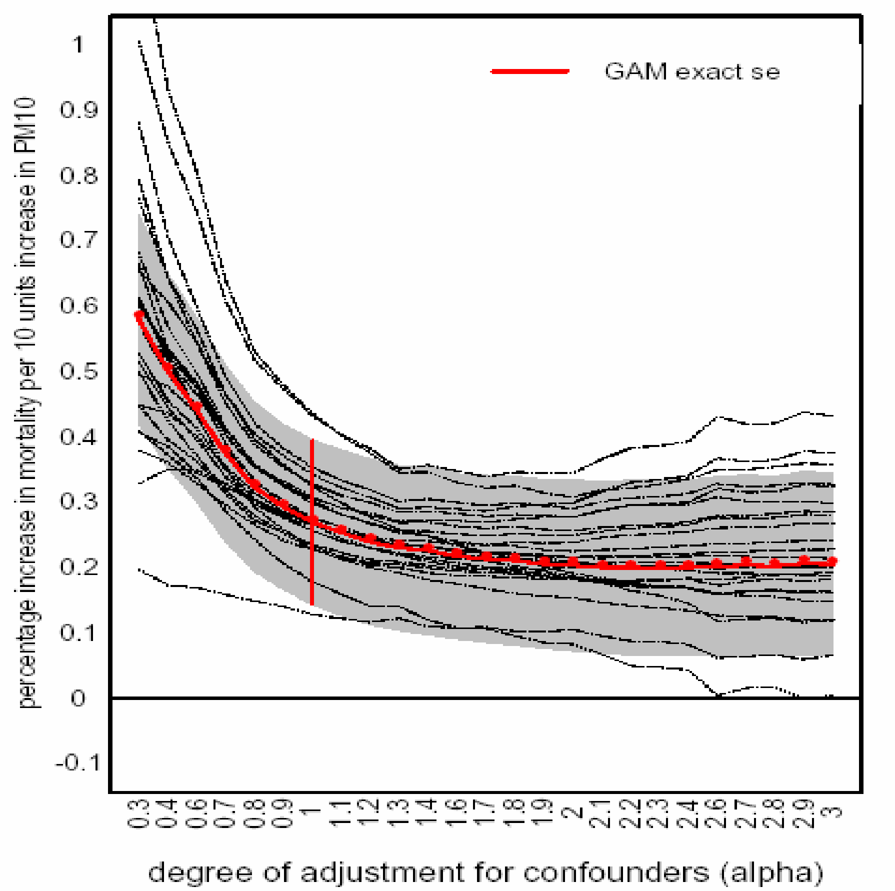
Dominici, McDermott, Trevor Hastie (2004)

National average estimates versus degrees of adjustment for confounders

National-average estimates



Bayesian city-specific estimates



Dominici, McDermott, Trevor Hastie (2004)

Relative risk estimates from time-series studies:

- Should not be the main source for quantifying risks and benefits in policy setting
  - Miss most of the lagged effects (positive, neutral or negative)
  - Are not directly amenable to years of life lost calculations
- Are important to establish causation and to assess relative magnitudes of effects across groups, causes of death, periods, seasons,…

Single-city studies are of limited value:

- For any given city, we are estimating a small pollution effect relative to the confounding effects of trend, season and weather
- High correlation between non-linear predictors
- Sensitivity of findings to model specifications

Relative risk estimates from TS studies likely under-estimate the total health effects because:

- They capture a small number of lags
- Acute effects only
- To control for known confounders, measured (temp) and unmeasured (influenza), must control for longer-term effects
- This results in "over-matching", ignoring effect of long-term exposure that is common to the population within a city

What is the Value of a Single City Analysis?

- Not much on own, valuable as a component of multi-city analysis
  - In most cities, too imprecise
  - Sensitive to difficult modeling decisions
  - Subject to publication bias
  - Little evidence of heterogeneity; better single-city estimates available from multi-city analyses

Hierarchical Models for Estimating a National Average Relative Rate of Mortality

Pooling

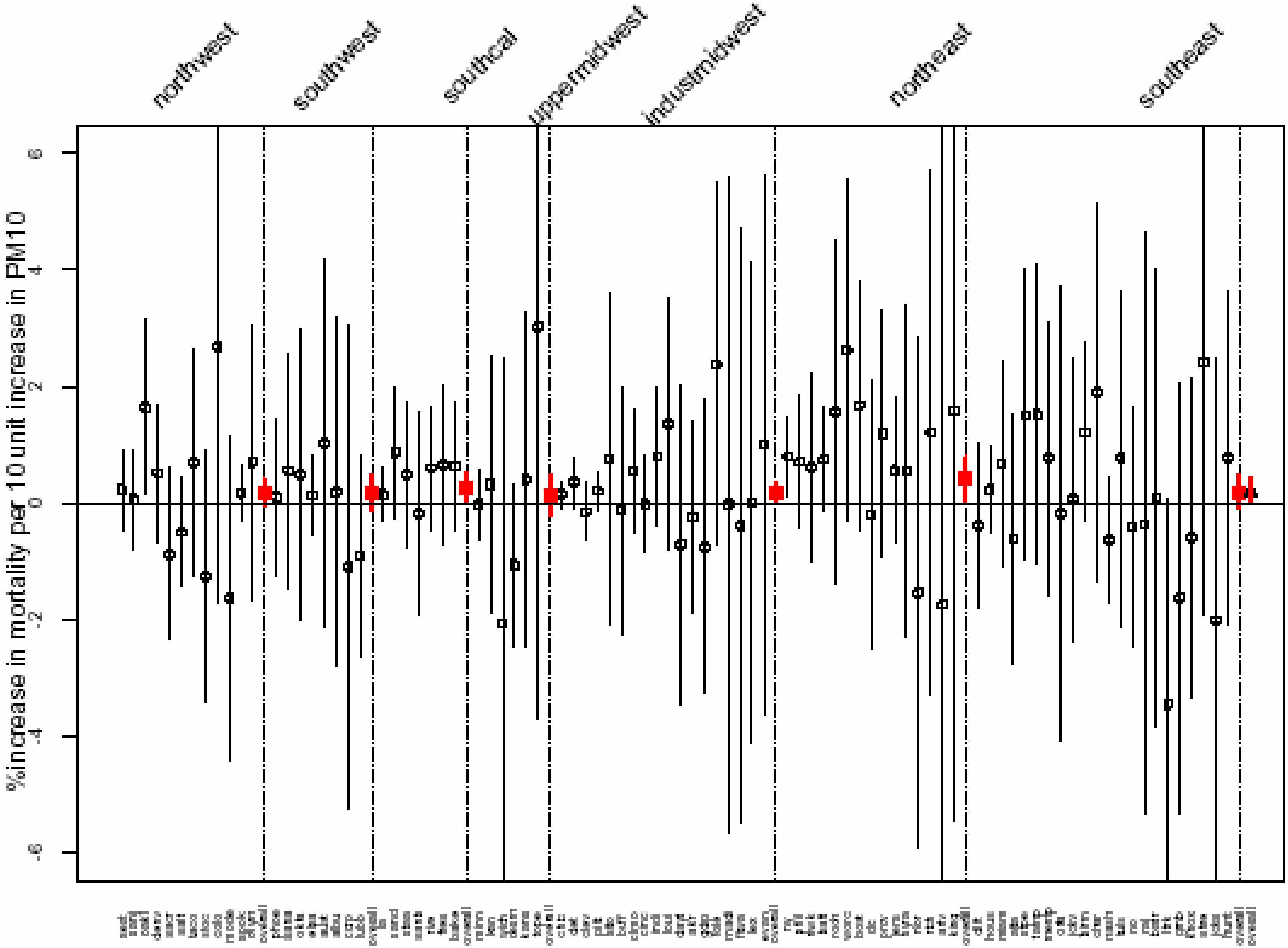
City-specific relative rates are pooled across cities to:

1. **estimate a *national-average* air pollution effect on mortality;**
2. **explore geographical patterns of variation of air pollution effects across the country**

Pooling

- **Implement the old idea of *borrowing strength across studies***
- **Estimate heterogeneity and its uncertainty**
- **Estimate a national-average effect which takes into account heterogeneity**

City-specific and regional estimates



Spatial Model for Relative Rates

- β_r^c = log relative rate of mortality associated with PM_{10} in city c in region r
- α_r = average log relative rate in region r
- α^* = national average log relative rate

$$\hat{\beta}_R^c = \alpha^* + \underbrace{(\hat{\beta}_r^c - \beta_r^c)}_{\text{statistical error}} + \underbrace{(\beta_r^c - \alpha_r)}_{\text{heterogeneity within region}} + \underbrace{(\alpha_r - \alpha^*)}_{\text{heterogeneity across regions}}$$

Three Models

- "**Three stage**"- as in previous slide
- "**Two stage**"- ignore region effects; assume cities have exchangeable random effects
- **Two stage with "spatial" correlation -** city random effects have isotropic exponentially decaying autocorrelation function

Estimating a national-average relative rate

Two-stage normal-normal hierarchical model

$$\hat{\beta}^c \sim N(\beta^c, v^c)$$

$$\beta^c = \beta^\star + \sum_j \beta_j^\star (x_j^c - \bar{x}^c) + N(0, \tau^2)$$

where

- $\beta^\star$ is the national-average relative rate
- τ^2 is the variance across cities of the true city-specific relative rates β^c (also called heterogeneity)
- $\text{corr}(\beta^c, \beta^{c'}) = \exp\left(-\phi \text{distance}(c, c')\right)$
- Goal: marginal posterior distributions of $\beta^\star$ and τ^2

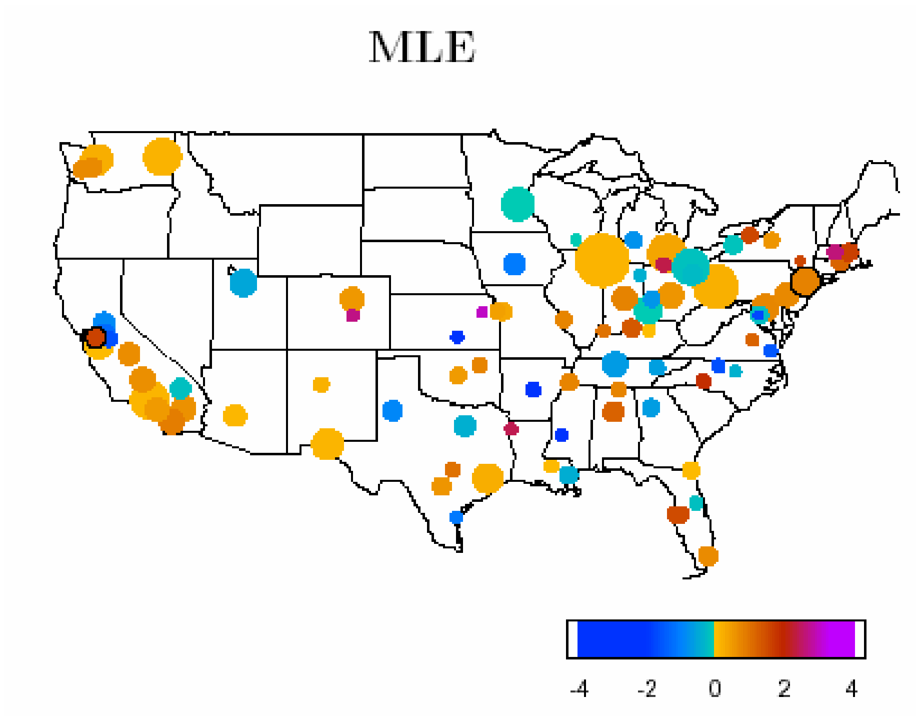
Dominici, Zeger, Samet RSSA 2000

Samet, Dominici, Zeger et al. NEJM 2000

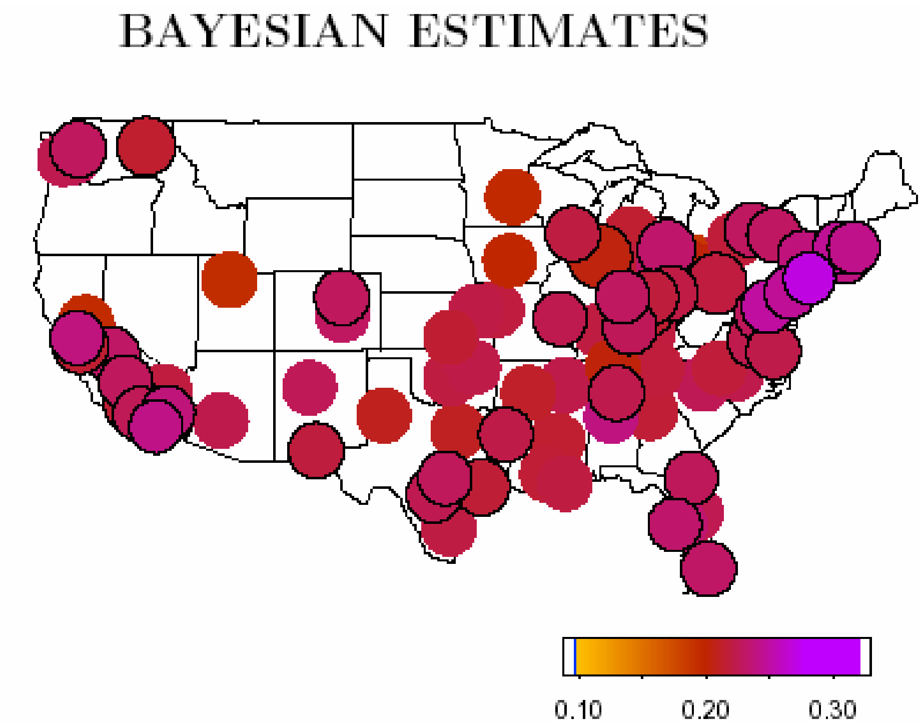
Epidemiological Evidence from NMMAPS

Maximum likelihood and Bayesian estimates of air pollution effects

Use only city-specific information

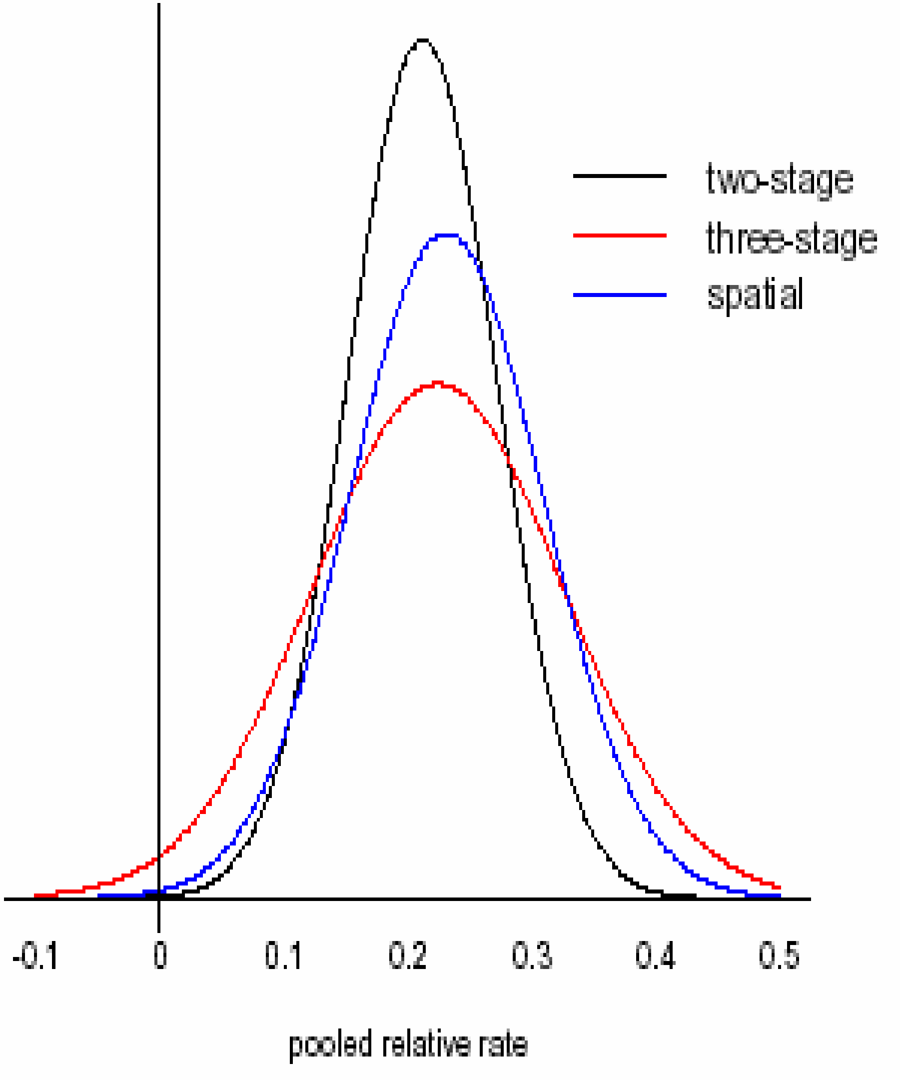


Borrow strength across cities

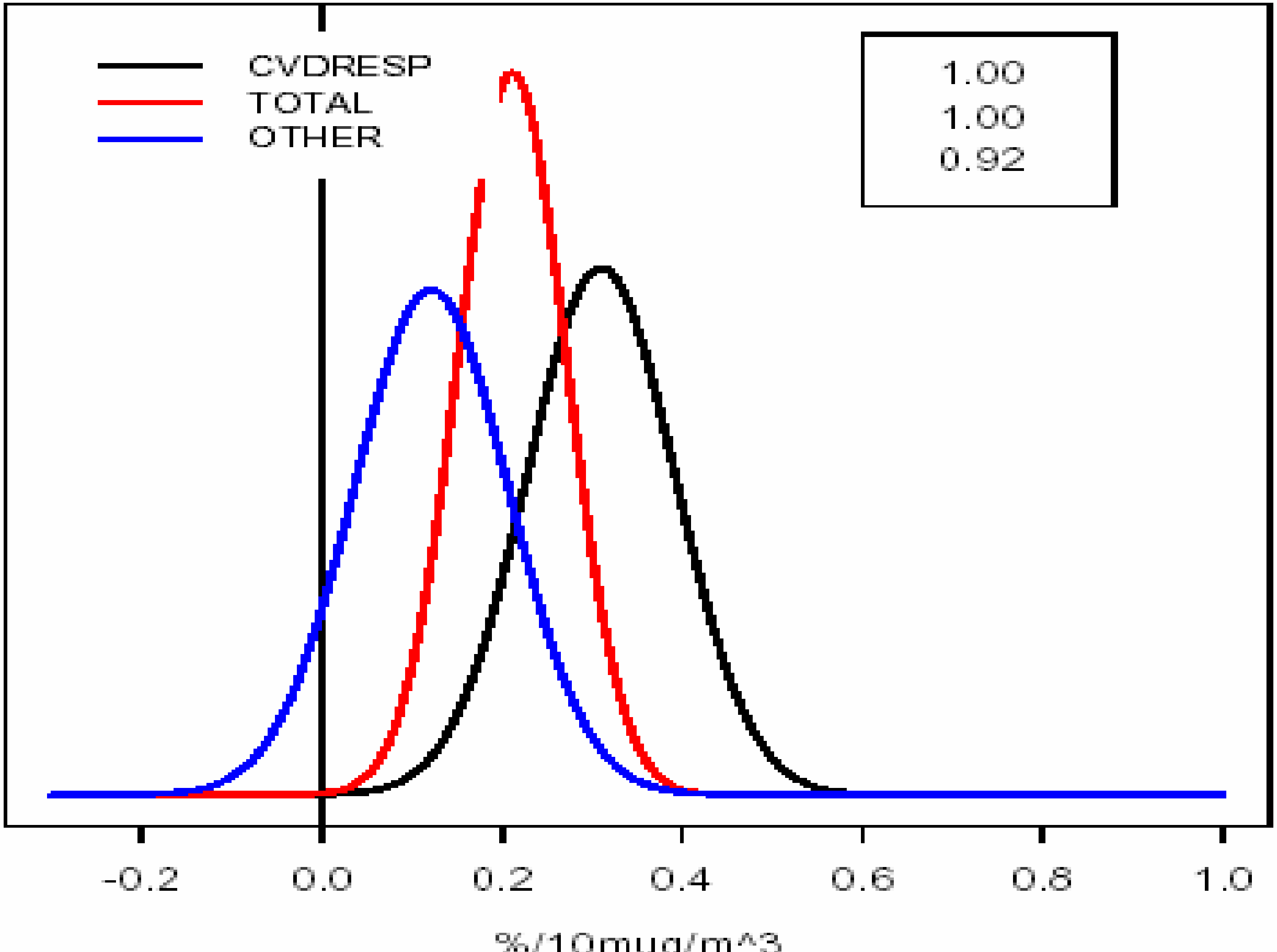


Dominici, McDermott, Zeger, Samet EHP 2003

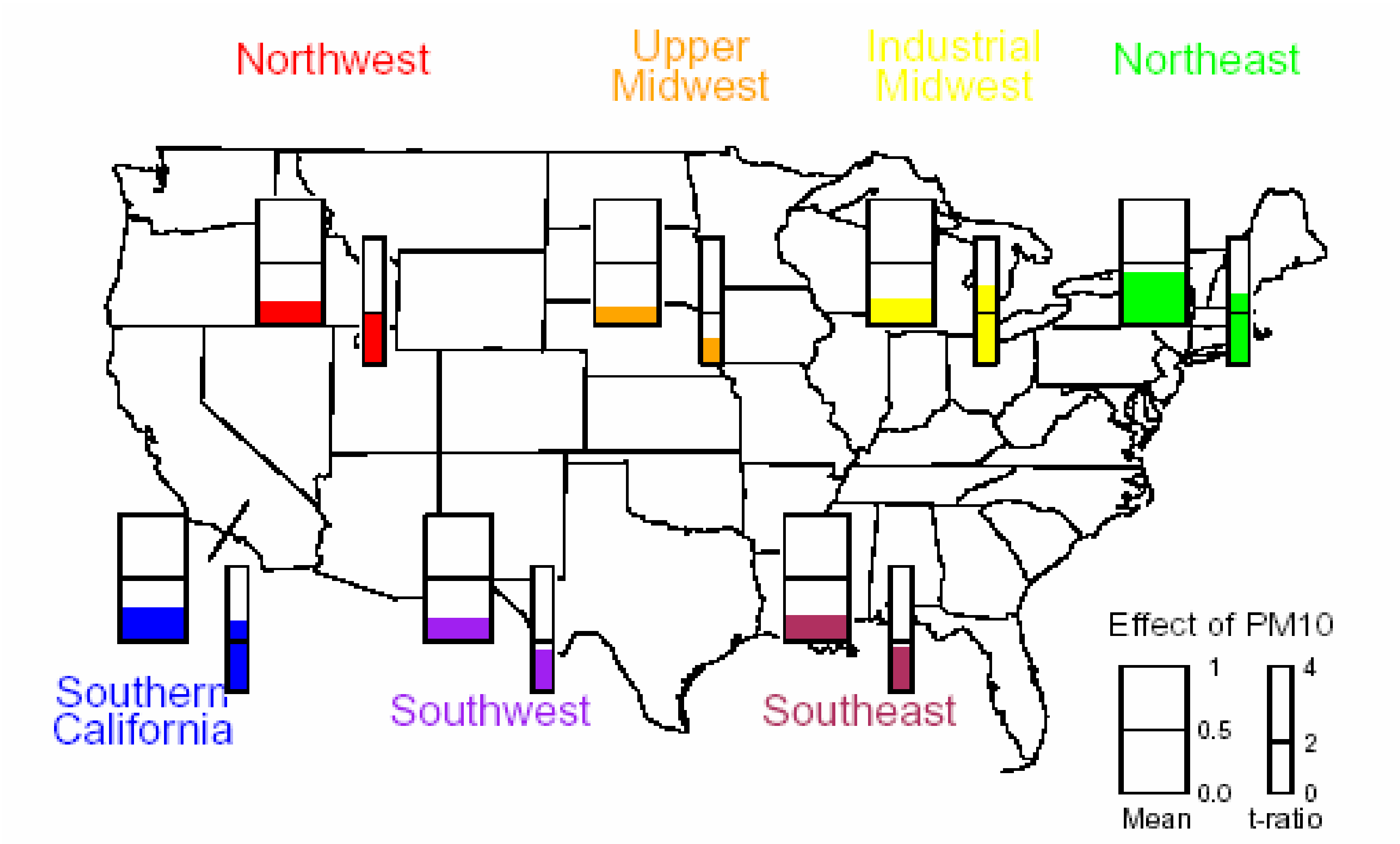
Posterior Distribution of National Average



Results Stratified by Cause of Death



Regional map of air pollution effects



Partition of the United States used in the 1996 Review of the NAAQS

Findings

- NMMAPS has provided at least four important findings about air pollution and mortality

1. There is evidence of an association between acute exposure to particulate air pollution and mortality
2. This association is strongest for cardiovascular and respiratory mortality
3. The association is strongest in the Northeast region of the USA
4. The exposure-response relationship is linear

A Big Challenge

- Doing research in a controversial political context can lead to a process which can be highly non scientific
- Expect to face consultants who use “quasi-scientific” arguments that create confusion about findings

Testimony on the EPA Proposed Decision on Particulate Matter

Suresh H. Moolgavkar, M.D., Ph.D.

Member, Fred Hutchinson Cancer Research Center

Professor of Epidemiology and Biostatistics, University of Washington

"The proposed new regulations for particulate matter are based on the assumption that the magnitude of the associations between these pollutants and adverse human health effects reported in some epidemiologic studies is predictive of the gains in human health that would accrue by lowering ambient concentrations. **The evidence simply does not support this assumption**. Briefly, the dearth of toxicological information, the absence of biological understanding of underlying mechanism, and the potential for uncontrolled confounding by copollutants currently preclude the conclusion that the particulate component of air pollution is causally associated with adverse effects on human health."

Criticisms

- **Heterogeneity**: in presence of heterogeneity of air pollution effects across the country, the national-average estimate is un-meaningful
- **Adjustment for confounders**: the associations are spurious and are the results of inadequate adjustments for confounders
- **Other Pollutants**: associations are not due to PM but to other pollutants and extreme weather

Heterogeneity

Is it appropriate to pool?

What are the data saying about heterogeneity?

- Chi-squared tests of homogeneity are always accepted (need to have 30% smaller standard errors to reject the null)
- Profile likelihood has a peak at zero
- Bayesian approach: marginal posterior distribution of the between-city standard deviation indicates that heterogeneity is very small

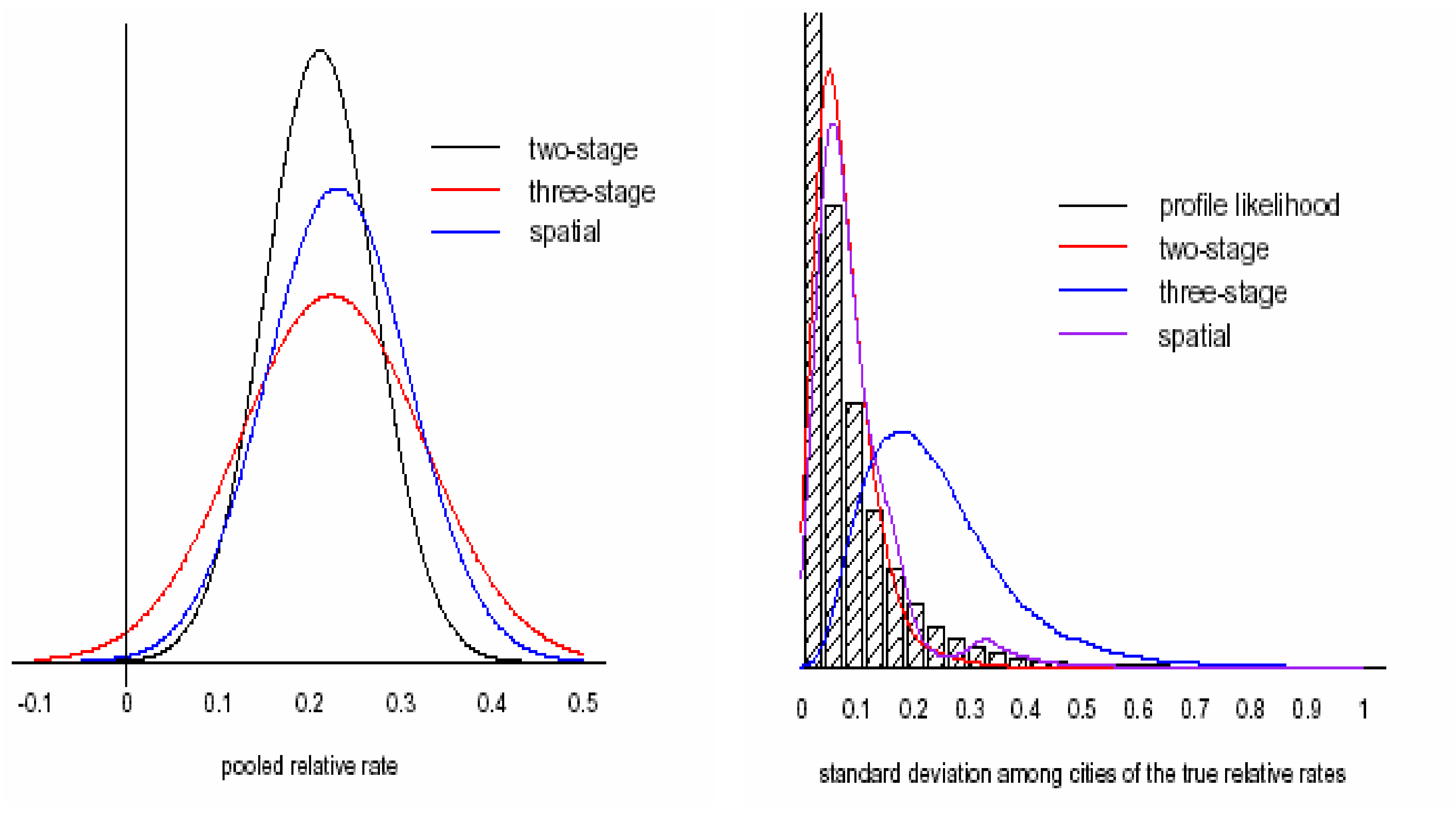
Why do a joint analysis of all the cities?

- Individual cities can be selected to show one point or another
- Results from individual cities are swamped by statistical error
- There is no reason to expect that two neighboring cities with similar sources of particles would have qualitative different relative risks

What are the public policy implications?

- A national estimate of the air pollution effect provides evidence on the amount of hazard from exposure to air pollution
- EPA needs a single number for the entire country

Comparison between heterogeneity models



Other Pollutants

Have PM studies adequately separated the effects of PM, weather, and co-pollutant?

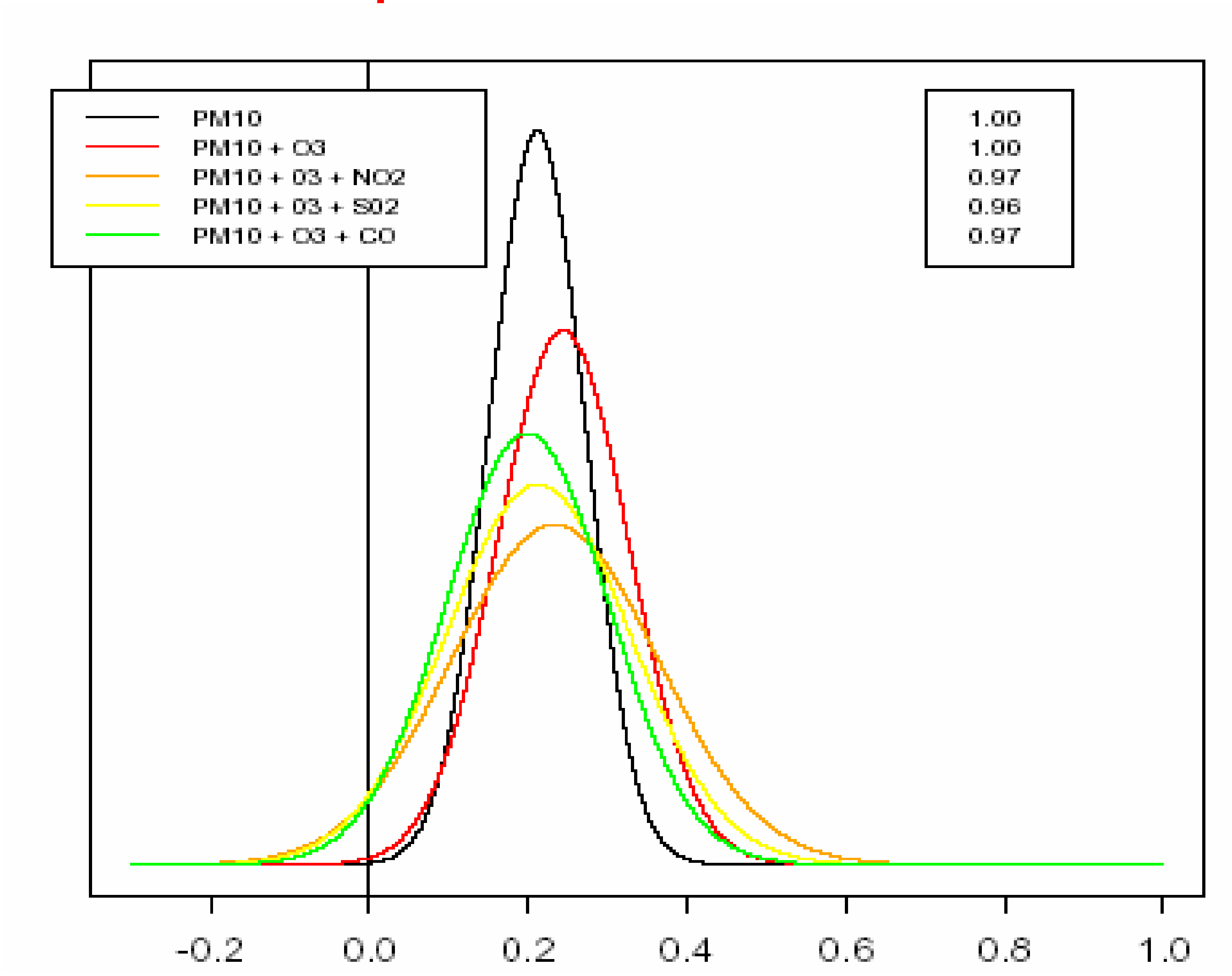
Other pollutants

- This is a complicate matter since many of the same mechanisms are postulated to underlie the effects of different pollutants

A simpler question is:

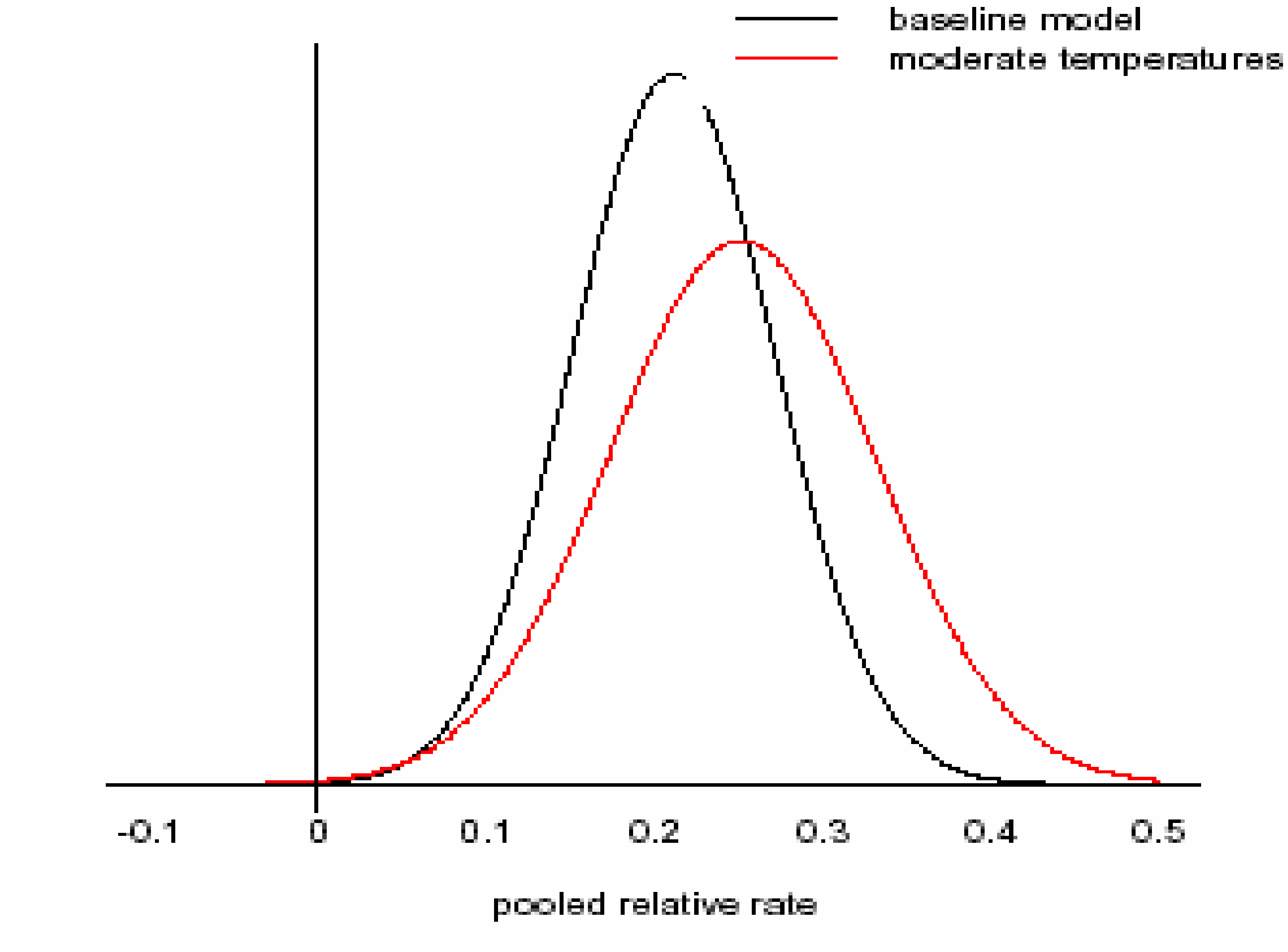
- Does the effect of PM on mortality sensitive to the adjustment for weather, seasonality and other pollutants?

Sensitivity of the pooled effect to the inclusion of other pollutants in the model



Posterior distributions of the pooled PM effects under 5 multi pollutant models

Sensitivity of the pooled effect to adjustment for weather



Findings

- Pooled estimates of the PM effects on mortality are robust to:
- Adjustment for confounding factors
- Inclusion of other pollutant in the models
- Exclusion of days with more extreme temperatures

Discussion

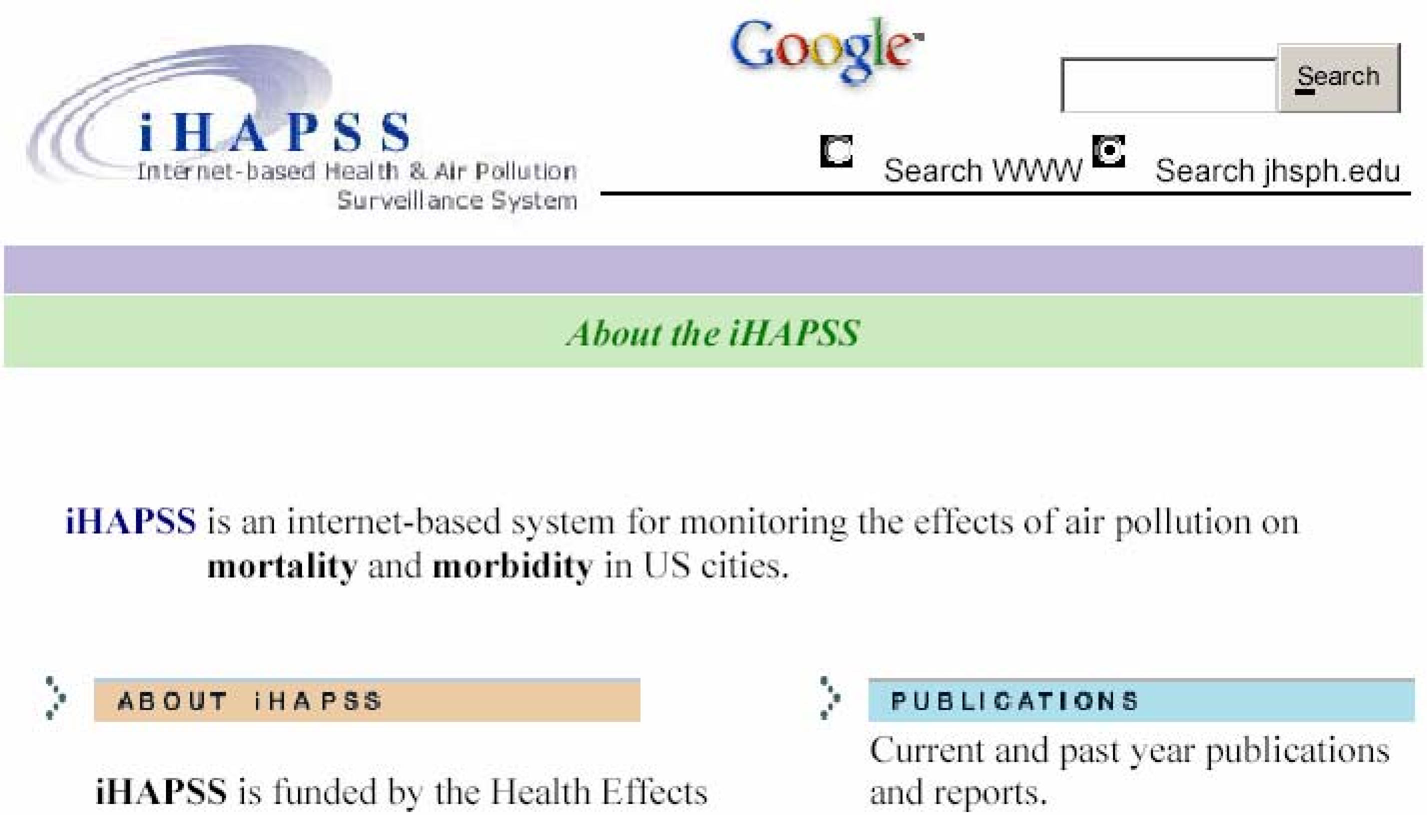
- To disentangle the effects of particulate matter from the effects of the other pollutants is difficult
- Very limited data is currently available on PM composition to better characterize the risk
- Multi site analyses provide a robust approach for exploring confounding and effect modifications to other pollutant and weather

A)lthough many questions remain about how fine particles kill people, the NMMAPS study shows there's no mistaking that PM is the culprit…?

Health Effects of PM, Model Uncertainty, & Environmental Policy

- **The use of epidemiological evidence for policy places heavy weight on statistical methods**
- **To enhance the utilization of time series studies for regulatory policy, a fully assessment of all sources of uncertainty is necessary**
- **Our methods improve estimation of uncertainty in the estimated risks and introduce diagnostic tools for exploring sensitivity of risk estimates to model choices**

internet Health and Air Pollution Surveillance System (iHAPSS)



A Multi-Site Time Series Study of Hospital Admissions and Fine Particles: A Case-Study for National Public Health Surveillance

A NATIONAL ANALYTIC SYSTEM FOR TRACKING POPULATION HEALTH

- **Multiple government databases contain massive amounts of information on the environmental, social, and economic factors that determine health**
- **Research on population health could be rapidly advanced by integrating these existing databases and bringing to bear new statistical models that would describe major threats and their causes**
- **These integrated databases and new analysis tools would create a national analytic system for population health research**

Hospital Admissions and Fine Particles: Objectives

1. assemble a national database of time series data for the period 1999-2002 on hospital admissions rates for cardiovascular and respiratory diseases, fine particulates, and weather for approximately 400 US counties
2. develop state-of-the-art statistical methods
3. estimate maps of relative risks of hospital admissions associated with short-term changes in fine particles
4. illustrate how integration and analysis of national databases can lead to a national health monitoring system

News Focus

Particle air pollution clearly causes substantial deaths and illness, but what makes fine particles so toxic—the size, the chemical compound, or both?

Mounting Evidence Indicts Fine-Particle Pollution

Now the issue is getting another look as EPA faces a December 2005 deadline for revisiting its $PM_{2.5}$ standard. EPA scientists, after reviewing piles of new data implicating $PM_{2.5}$ in health effects, have proposed tightening the 1997 standard to further reduce ambient concentrations of fine particles. Some scientists and industry groups remain skeptical, noting that researchers still haven't pinned down what makes particles dangerous—whether it's mainly size, and that the tiniest particles are most potent; or chemistry, such as metal content; or some combination of the two. Despite 8 years and some $400 million in research, finding out exactly how fine particles do their dirty work has proved frustratingly elusive, researchers say. "We've gotten glimpses, but we don't yet have enough systematic coverage of the problem," says epidemiologist Jon Samet of Johns Hopkins University in Baltimore, Maryland.

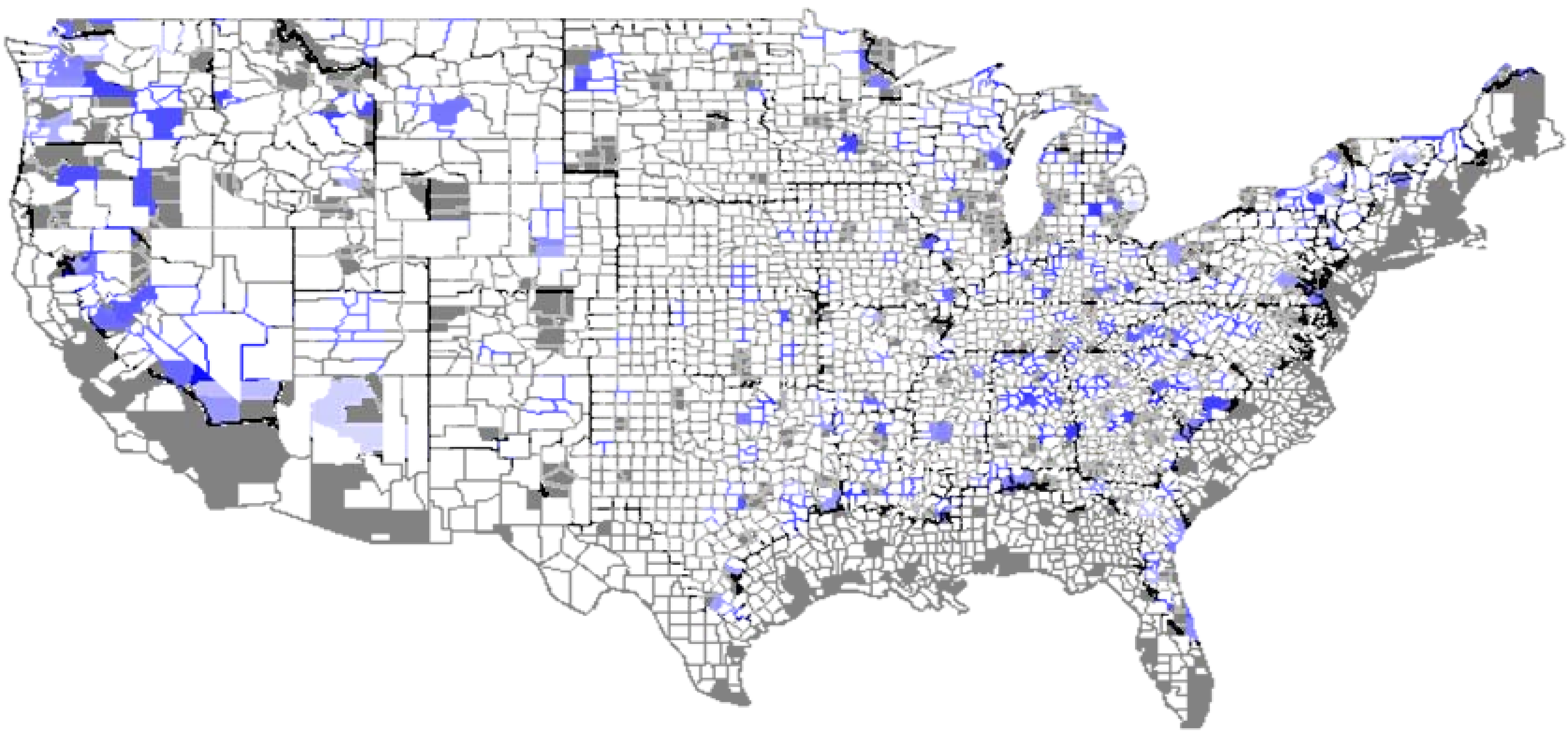
Urgent need to estimate:

- **short-term effects of PM2.5 on mortality and morbidity on average for the entire country**

National Medicare Cohort (1999—2002)

- **National study of fine particles ($PM_{2.5}$) and hospital admissions in Medicare**
- **Data include:**
  - **Billing claims (NCHF) for everyone over 65 enrolled in Medicare (~48 million people),**
    - **date of service**
    - **treatment, disease (ICD 9), costs**
    - **age, gender, and race**
    - **place of residence (ZIP code/county)**
  - **Approximately 400 counties linked to the air pollution monitoring**

Medicare counties linked to air pollution monitoring



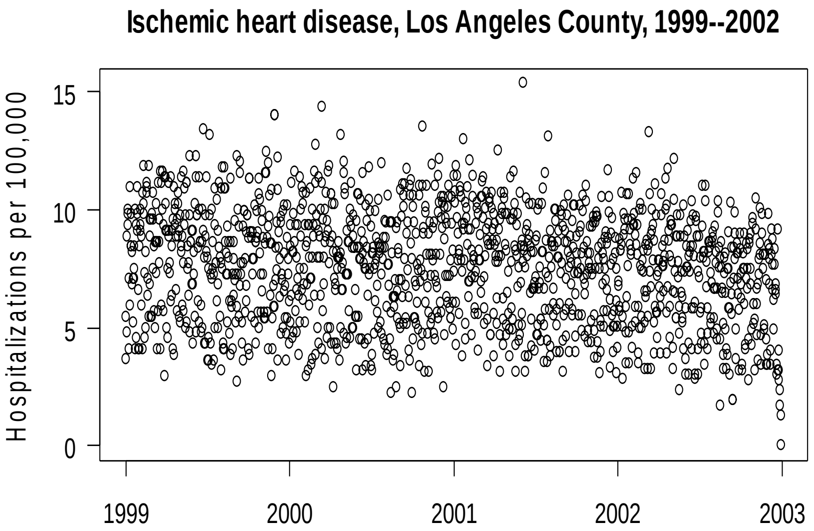
Linking National Databases for Tracking Population Health

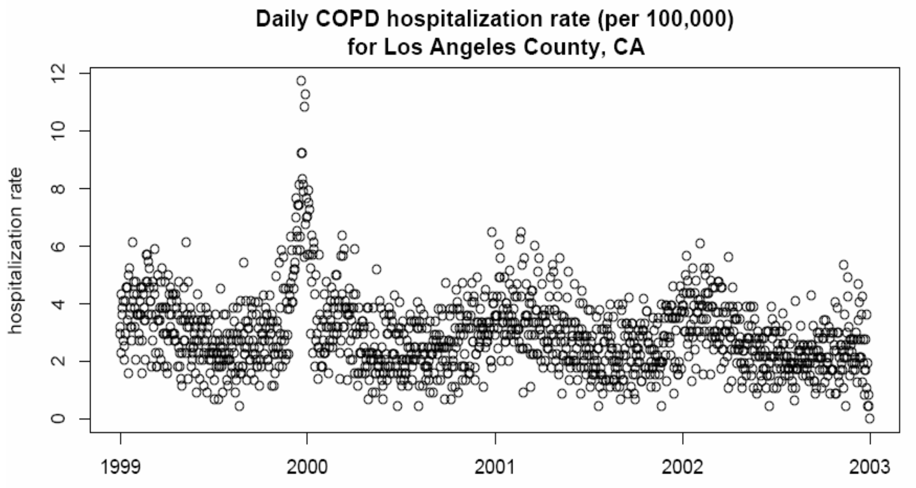
- **We have identified the largest 400 counties in the USA with PM2.5 daily data available for the period 1999-2002**
- **For each of these counties, we have constructed daily time series of hospitalization rates for the following diseases:**
  - **COPD *(239 counties)***
  - **Respiratory Infections *(239)***
  - **Ischemic heart diseases *(251)***
  - **Heart failure *(247)***
  - **Cerebrovascular diseases (250)**
  - **Heart rhythm (241)**
  - **Accidents (265)**

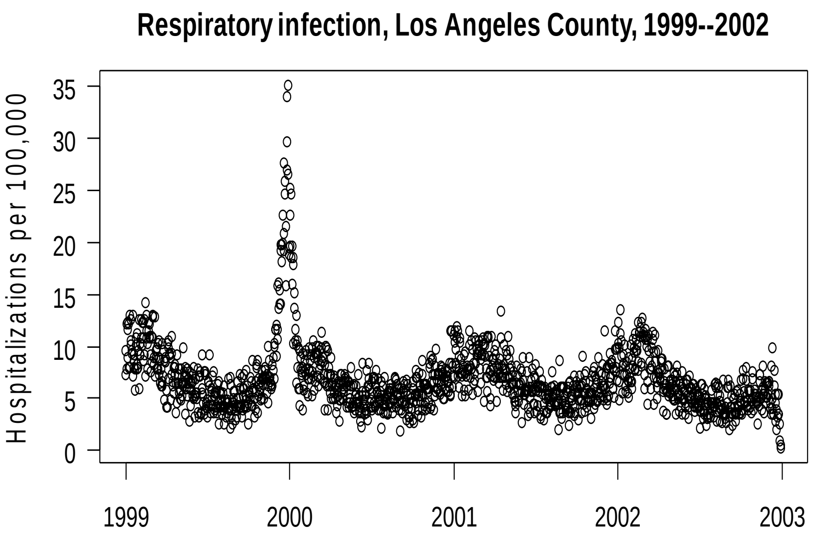
Multi-site time series studies

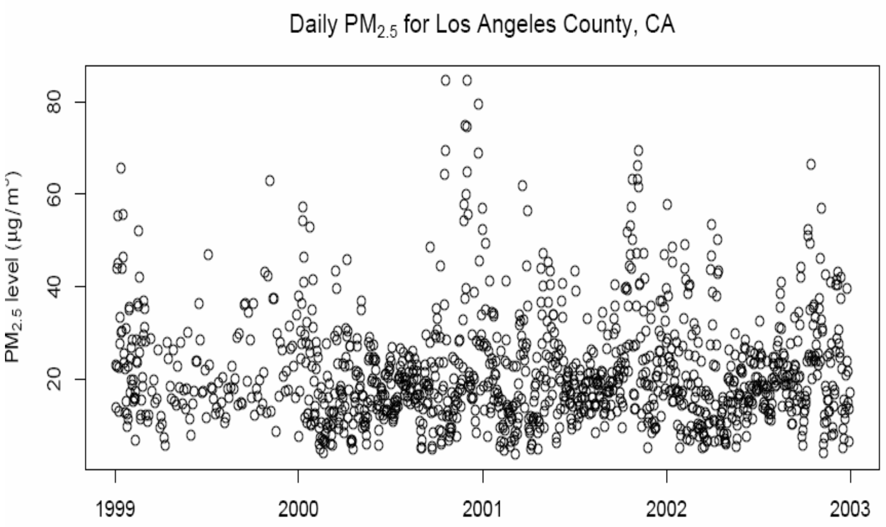
- **Compare day-to-day variations in hospital admission rates with day-to-day variations in pollution levels within the same community**
- **Avoid problem of unmeasured differences among populations**
- **Key confounders**
  - **Seasonal effects of infectious diseases and weather**

Daily time series of hospitalization rates and PM2.5 levels in Los Angeles county (1999-2002)









Why do a joint analysis of all the counties?

- **Individual counties can be selected to show one point or another**
- **Results from individual counties are much more sensitive to model assumptions and are swamped by statistical error**
- **There is not reason to expect that two neighboring counties with similar sources of particles would have qualitative different relative risks**

Results

- **We have illustrated a case study of tracking health risks associated to a short-term exposure to fine particles on a national scale**
- **We have linked by county of residence Medicare hospital rates for different diseases to daily ambient concentration of pollution and weather variables**
- **National analyses indicate that short-term exposure to PM2.5 is significantly associated with an increase of hospital admission rates for respiratory and cardiovascular outcomes**

Air pollution and health: Questions and (some) answers

- **Is there a risk?**
  - **Multi-site time series studies such as NMMAPS (1987—2000) provide strong evidence of short-term association between air pollution and mortality**
  - **Preliminary results from Medicare data (1999—2002) indicate that current air pollution levels still affect health**
- **How can we estimate it?**
  - **National datasets are powerful resources for assessing the health effects of air pollution**
  - **Statistical models that can integrate information across space and time**
  - **National average estimates for the effect of PM are robust to various model formulations and statistical methods**

Discussion

- Linking national databases and developing statistical methods that can properly analyze these them, are essential steps for a successful national public health tracking system
- Because of the small risks to be detected and the large number of potential confounders, single-site studies are generally swamped by statistical error
- A national analytic system, that routinely analyze data from multiple locations in a systematic fashion, is a very promising approach for tracking population health

Acknowledgments

- **Our team:**
  - R. Peng
  - S. Zeger
  - J. Samet
  - A. McDermott
  - Mary Fox
- **Our sponsors:**
  - EPA
  - JHU CDC Center of Excellence