Big Data in Environmental Health

Joel Schwartz

Harvard TH Chan School of Public Health

Big Data plays several key roles in Environmental Health

- Understanding Mechanisms
  - Epigenomics, metabalomics, p > n problems
- Exposomics
  - Analyzing many exposures
  - Modeling exposure with Big Data
- Big n to allow for better exploration of
  - Susceptibility
  - Dose Response

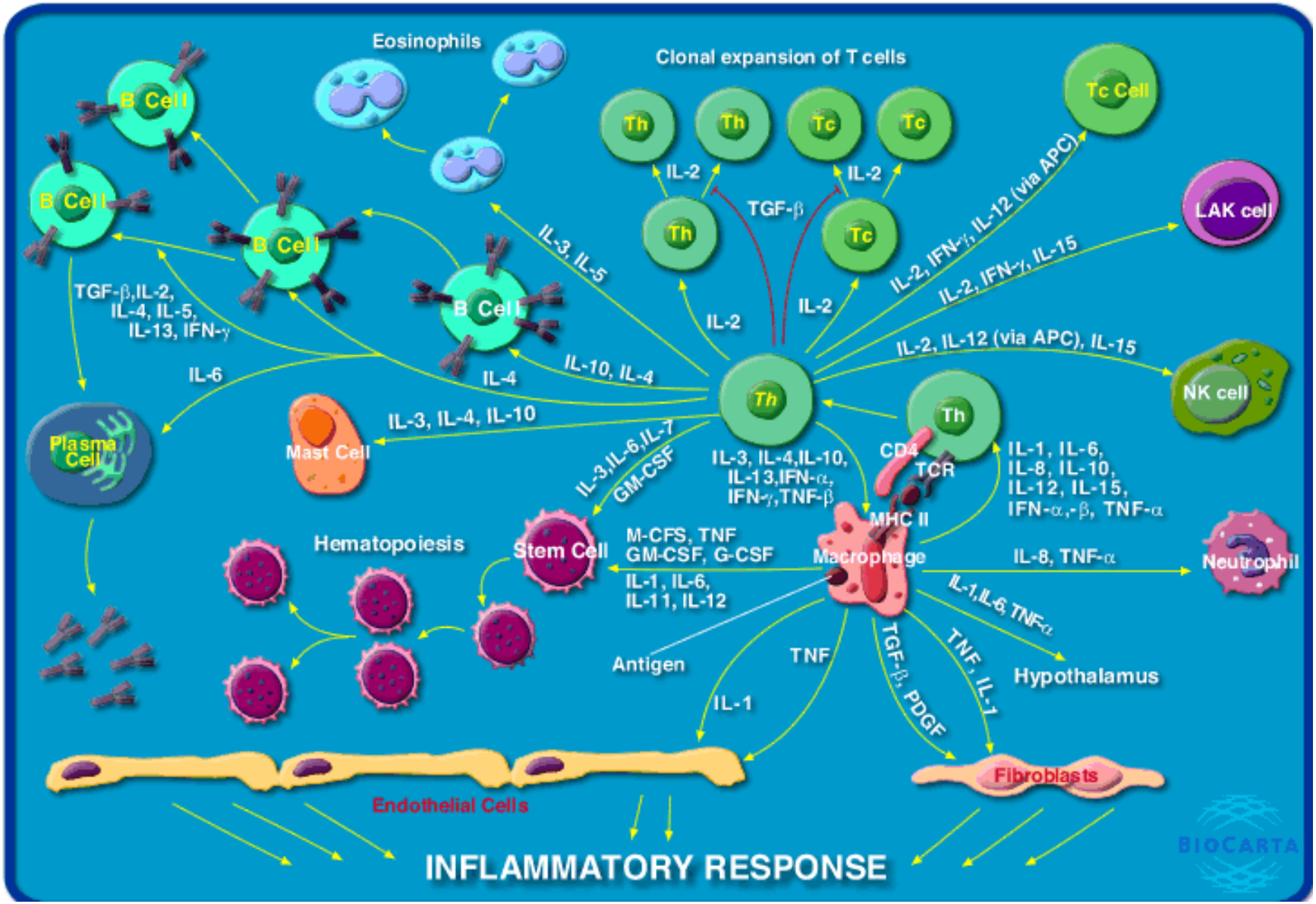
Epigene and Gene

- Share many letters
- But this is misleading
- Methylation is continuous
- Its measurement has more error, and more issues relative to chip position, type of probe, etc.
- Genes can modify exposure effects, epigenetic marks can also mediate exposure effects

Epigenetic marks typically control gene expression

- This is the source of another key difference
- Protein expression occurs in complex networks in the body

Consider this Schematic of Inflammation



Epigenetic Marks must be Correlated

- This is fundamentally different from SNPs
- The pattern of correlation among inflammatory cytokines will be different for different disease states and for different triggers of inflammation
- Seeing which pattern is associated with a pollutant can tell us a lot about how it is acting

Data Reduction

- Use approaches to reduce the number of variables considered
- Summary Measures
- Hypothesized Pathways
- Machine Learning

For Example

- DNA methylation age:
  - Horvath used an Elastic Net to reduce methylation sites to the sum of contributions of 353 CpG sites that is well correlated with age, BUT predicts health events INDEPENDENT of chronological age, such as age at death (Chen BH et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging, 2016. Vol8, No 9.)
- How is this related to air pollution?

Normative Aging Study

- Difference in methylation age per 1 μg/m^3 in PM2.5:
  0.54 years (0.24, 0.84)

Nwanaji-Enwerem, et al, Environmental Epigenomics 2016

Candidate Pathway Analysis: Air Pollution and DNA Methylation

- Normative Aging Study
- BC and SO4 as exposures
- Asthma Pathway as identified via KEGG
- MAPK Pathway as Identified via KEGG
- p >> n

Sparse Canonical Correlation Analysis

- Take linear combination of e.g. methylation in a pathway, with weights
- Choose the weights to maximize the correlation with exposure or sets of exposures
- We then quantified the correlation between them by calculating a single Canonical Correlation for the pathway

Sofer et al Stat Biosci. 2012 November 1; 4(2): 319–338.
doi:10.1007/s12561-012-9072-7.

Methods

- Subject to some penalty for too many high weights
- Stepwise selection to drop unimportant methylation sites correlation into account
- We used a permutation test to correct for the selection process while accounting for correlation in the data

Sofer et al Stat Biosci. 2012 November 1; 4(2): 319–338. doi:10.1007/s12561-012-9072-7.

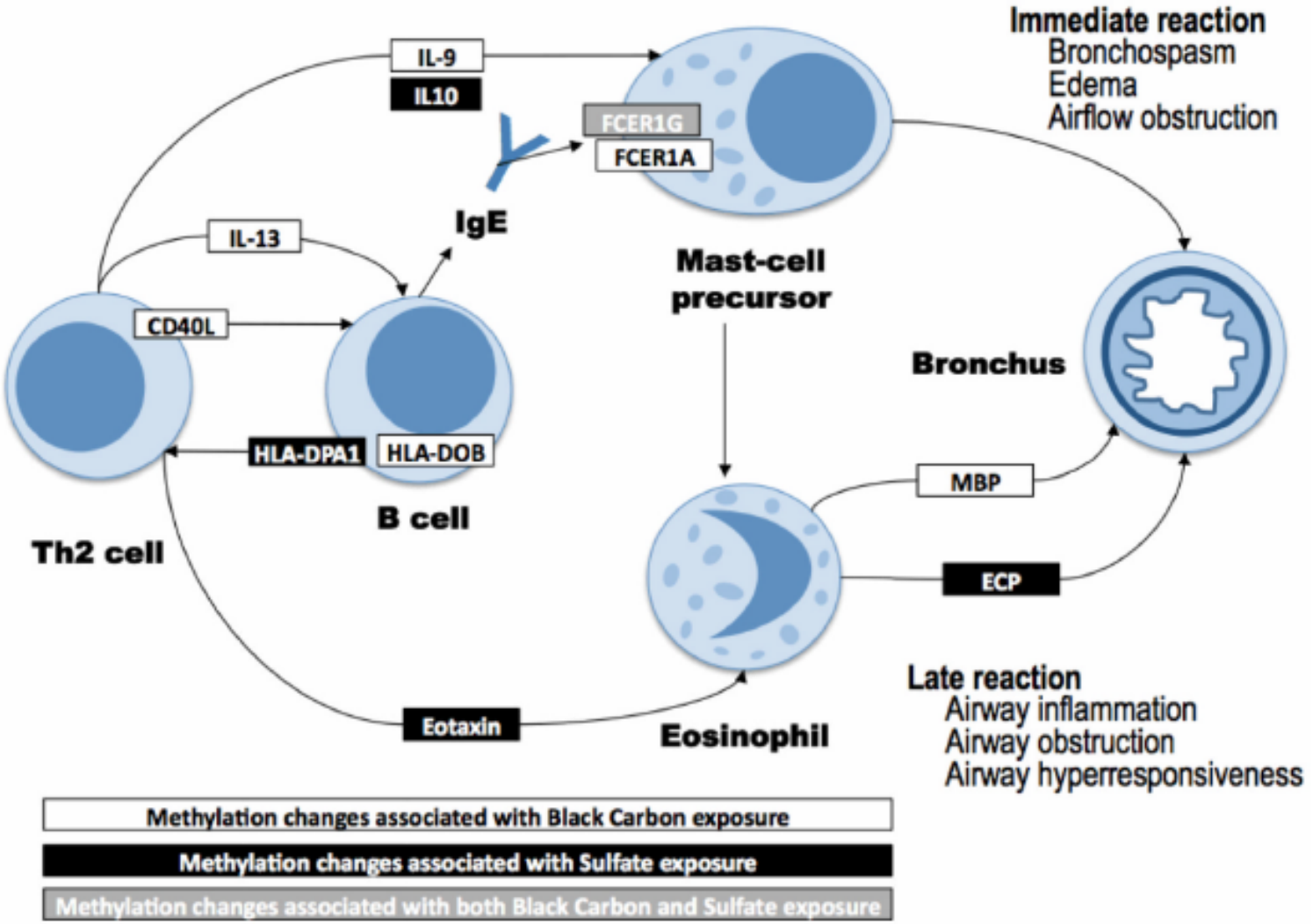
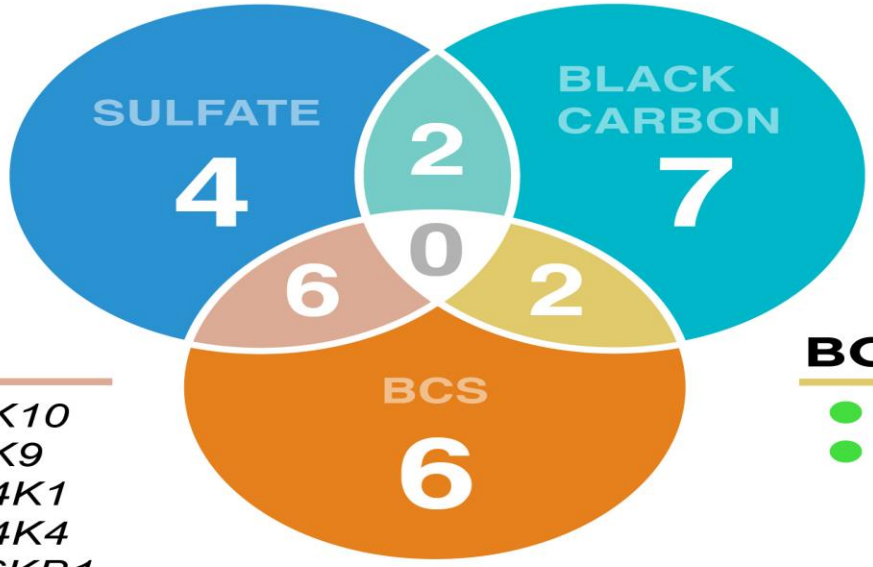


Figure 1.

The proteins coded by the genes identified in the analysis as associated with sulfate and black carbon exposures, and their roles within the KEGG asthma pathway.
Pathway information adapted from the KEGG Asthma pathway, Hsa-05310

LEGEND:

- Hypomethylation
- Hypermethylation



S

- *MAP3K11*
- *MAP3K7*
- *RPS6KA3*
- *TGFB1*

S BC

- *MAPK13*
- *MAPKAPK2*

BC

- *MKNK2*
- *MAPK6*
- *MAP3K2*
- *MAP3K6*
- *MAP4K3*
- *RPS6KB2*
- *SHC1*

S BCS

- *MAPK10*
- *MAPK9*
- *MAP4K1*
- *MAP4K4*
- *RPS6KB1*
- *STAT1*

BC BCS

- *MEF2A*
- *RELA*

BCS

- *MAP2K1*
- *MAP2K5*
- *MAP2K6*
- *MAP3K14*
- *PAK1*
- *MYC*

Pros and Cons

- Can begin to capture patterns of association of multiple exposures with multiple intermediaries
- Linear and no interactions (although they can be put in by hand)

Agnostic Pathway Analysis

- What do I mean by this?
- We hypothesize that there is a network of connected probes whose methylation is, collectively, associated with exposure (or outcome)
- Use the data to identify the network

Supervised PCA

- Do the standard EWAS analysis (500,000 univariate associations with phenotype)
- Take the e.g. 2000 with the highest correlations with phenotype
- Do a Principal Component Analysis on them
- Regress phenotype against 1st PCA
- This identifies a network of correlated probes that are also associated with our Exposure/Outcome
- We see this for PM2.5, lead, etc.

- We can also get importance scores, identify the top probes, put those genes into DAVID or IPA
- This can address the question: What does the pathway do for a living?
- Maybe nothing, but it may be a biomarker of long term exposure
- Which can then be used in other cohorts with blood, but not exposure history

The Lasso and Elastic Net

- The lasso shrinks coefficients of predictors toward zero based on a penalty proportional to the sum of the absolute values of the coefficients
- The adaptive lasso weights those penalties down for variables whose standardized coefficients are larger in an initial analysis
- The Elastic Net adds another weight proportional to the square of the coefficients

The Lasso is Consistent

- If the correlation among predictors is not large
- This is a problematic assumption for epigenetics
- The adaptive lasso and elastic net are consistent even with high correlation

Exposure and Exposomics

- Exposomics adds large numbers of Exposures to the Mix
- Exposure modeling adds large amounts of data to predictive models

Clustering with prior categories

- Kioumourtzoglou et al looked at daily air concentrations of 58 organic compounds and classified them by their chemical properties into 5 groups: n-alkanes, hopanes, cyclohexanes, PAHs and isoalkanes, in Atlanta, Dallas, and Birmingham.
- The first stage analysis included all compounds simultaneously in the model for health phenotype
- A second stage meta-regression regressed the 58 β's against indicators for group, with inverse covariance weights. Cyclohexanes were associated with hospital admissions for CVD.

Kioumourtzoglou MA, Zanobetti A, Schwartz J, Coull BA, Dominici F and Suh HH. The Effect of Primary Organic Particles on Emergency Hospital Admissions among the Elderly in 3 US Cities, Environmental Health, 2013; 12: 68

Least Squares Kernel Machine

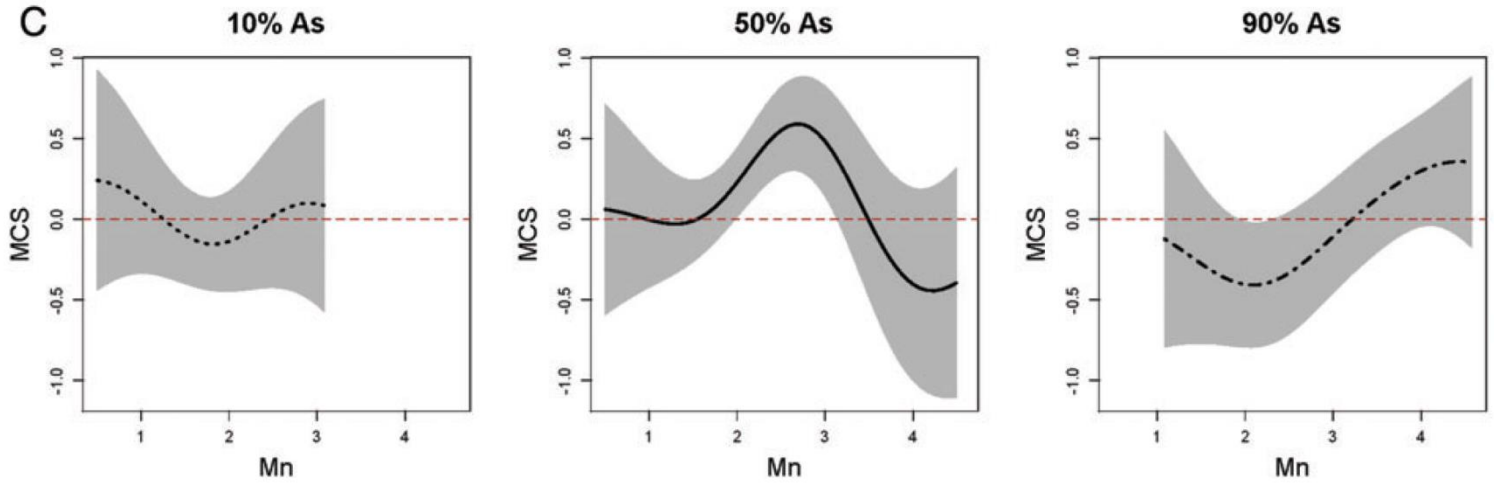
- Our basic problem is:

$$Y_i = h(\mathbf{z}_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$$

- Where **z** is a set of multiple exposures, X are the covariates, and h(**z**) is an unknown function possibly containing nonlinearities, high order interactions, lags,etc
- Nevertheless, we want to know
  - Which Z's are important
  - What is the overall impact

Prenatal Exposure to Metals and Bayley scales of Psychomotor Development

Nonlinear and interactive effect of Mn at different levels of As, and fixed Level of Pb



Bobb J, Valeri L et al. Biostatistics (2015), **16**, 3, *pp.* 493–508.
Maity, A. AND Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics* **67**(4), 1271–1284.

Estimating PM2.5 using Fusion of Satellite Remote Sensing, GEOS-Chem, and other Parameters

Goal: Estimate Daily PM2.5 concentrations for every 1km^2 in North America, Europe, and Asia

The U.S. is 11 million km^2, so this is big data

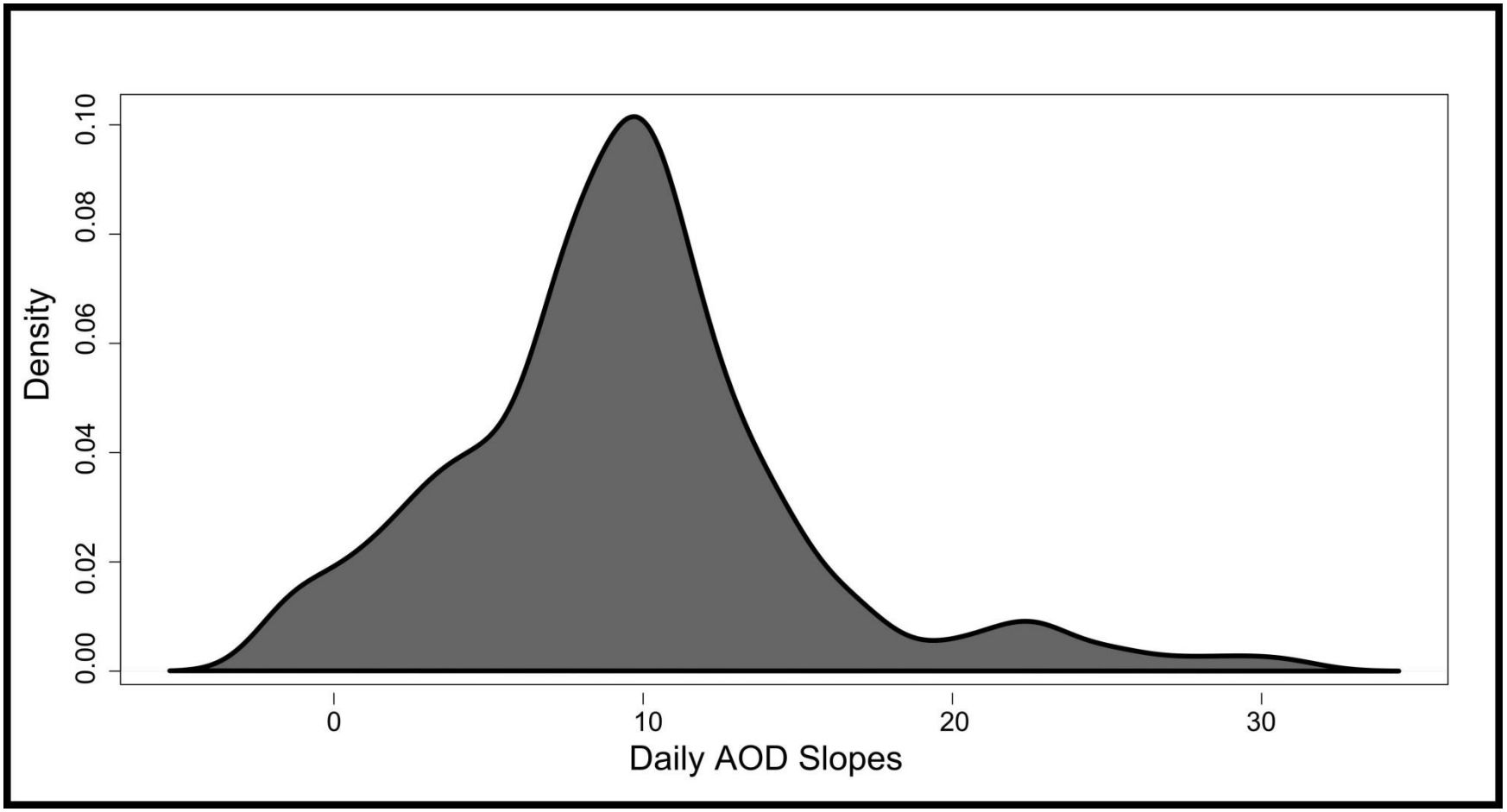
Inputs: Daily measurements everywhere using satellite remote sensing, meteorology, GIS data, Chemical Transport Models, etc.

Critical Issues in Modeling/Fusion

- Missing Data
- Surrogates for emissions (e.g. traffic density) have time varying impacts
- Aerosol Optical Depth has a time varying relationship to ground level particle concentrations
- Nonlinearities
- Many factors have multiplicative impacts on particle concentrations: High order interactions

Aerosol Optical Depth

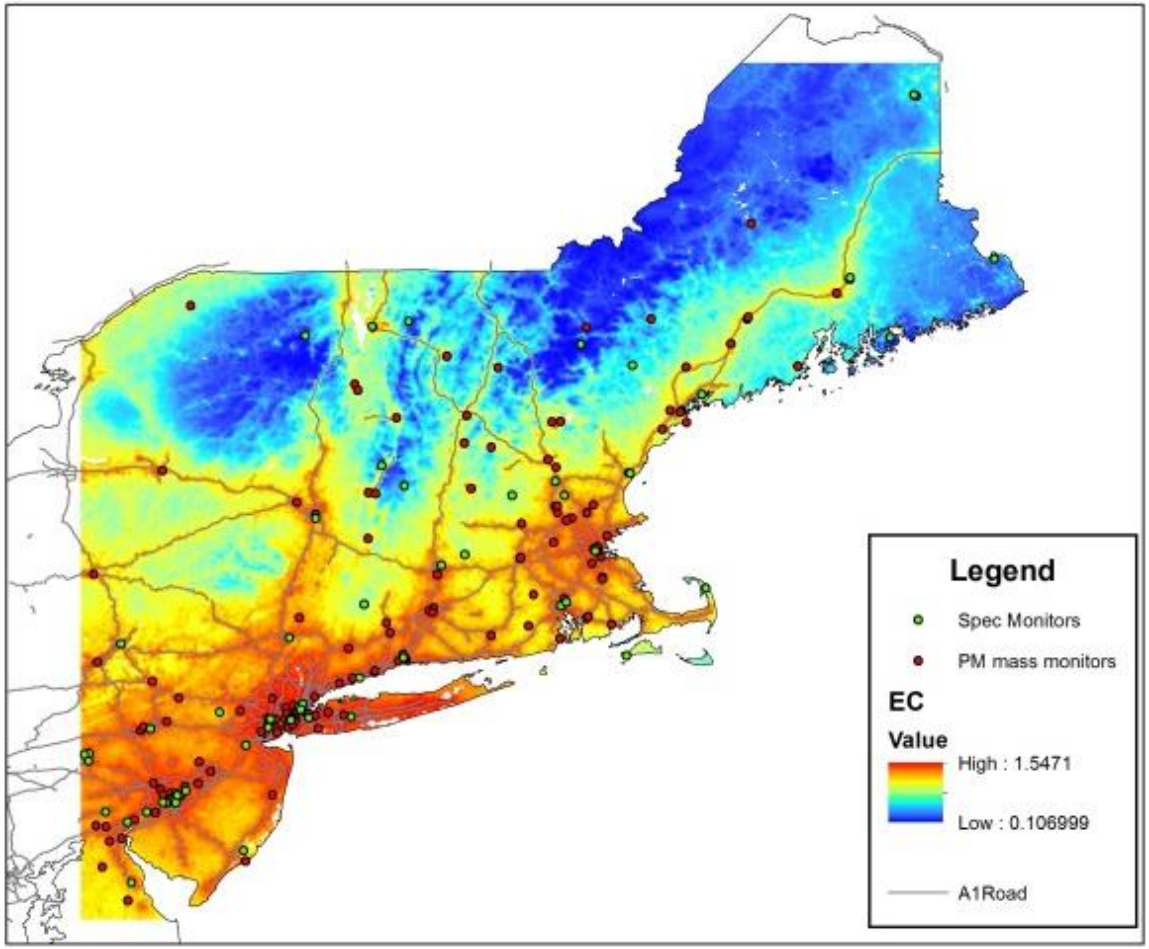
- Is based on direct physical measures at every square kilometer of earth twice per day (and in several different wavelengths)
- This provides high spatial and temporal resolution, BUT
- It is based on scattering and absorption of light in the entire column of air, not just ground level
- The scattering and absorption varies with particle color and size



A density plot exhibiting the daily variation of AOD slopes between 2000-2008 during the stage 1 calibrations

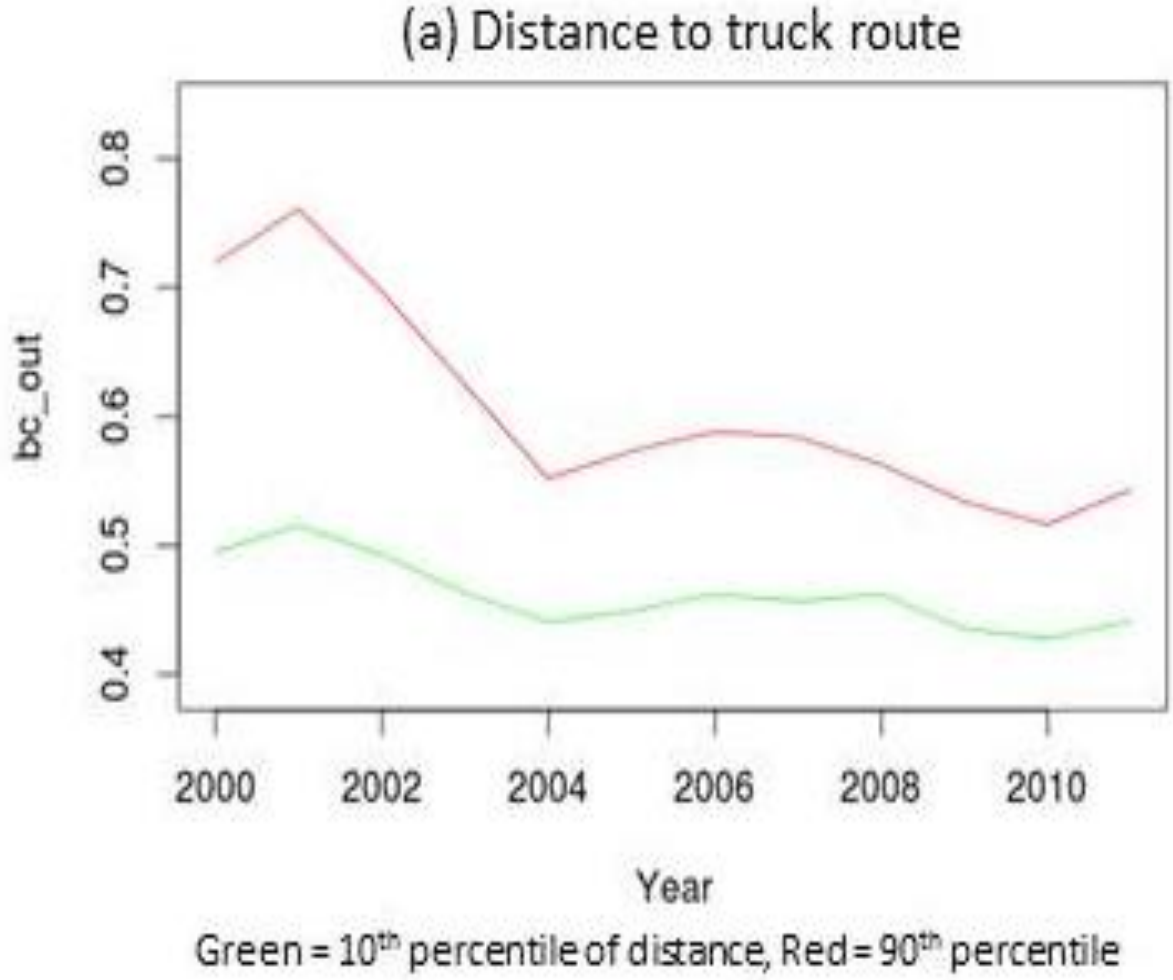
Chemical Transport Models

- Also provide daily (or better) time resolution
- Hybrid Models (CTM + Land Use + Weather)



Land Use Variables

- Offer the possibility of highly geographically resolved estimates of exposures (e.g. address)
- But the impact of those terms changes over time


(a) Distance to truck route
bc_out
0.4
0.5
0.6
0.7
0.8
2000
2002
2004
2006
2008
2010
Year
Green = 10th percentile of distance, Red = 90th percentile

Put it All Together

- MAIAC AOD from Aqua and Terra
- AAI, O3 and NO2 from OMI
- GEOS-Chem output
- Land use and Meteorology
- Monitoring Data
- Neural Network Algorithm
- Entire US, Daily 2000-2012
- Italy, Israel, Mexico City

Out of Sample R2

- 0.85 for PM2.5
- 0.76 for Ozone
- Daily predictions for each of 11 million 1km cells in the Continental US for each day Jan 1 2000-Dec 31 2012.

Large Administrative Datasets

- For Example, Medicare provides medical insurance for everyone in the US over age 65 who is not still employed
- We followed 61 million Medicare Enrollees, with 460 million person years of followup
- We has postal codes, and merged each year in each address with the average PM2.5 and Ozone from our National Model

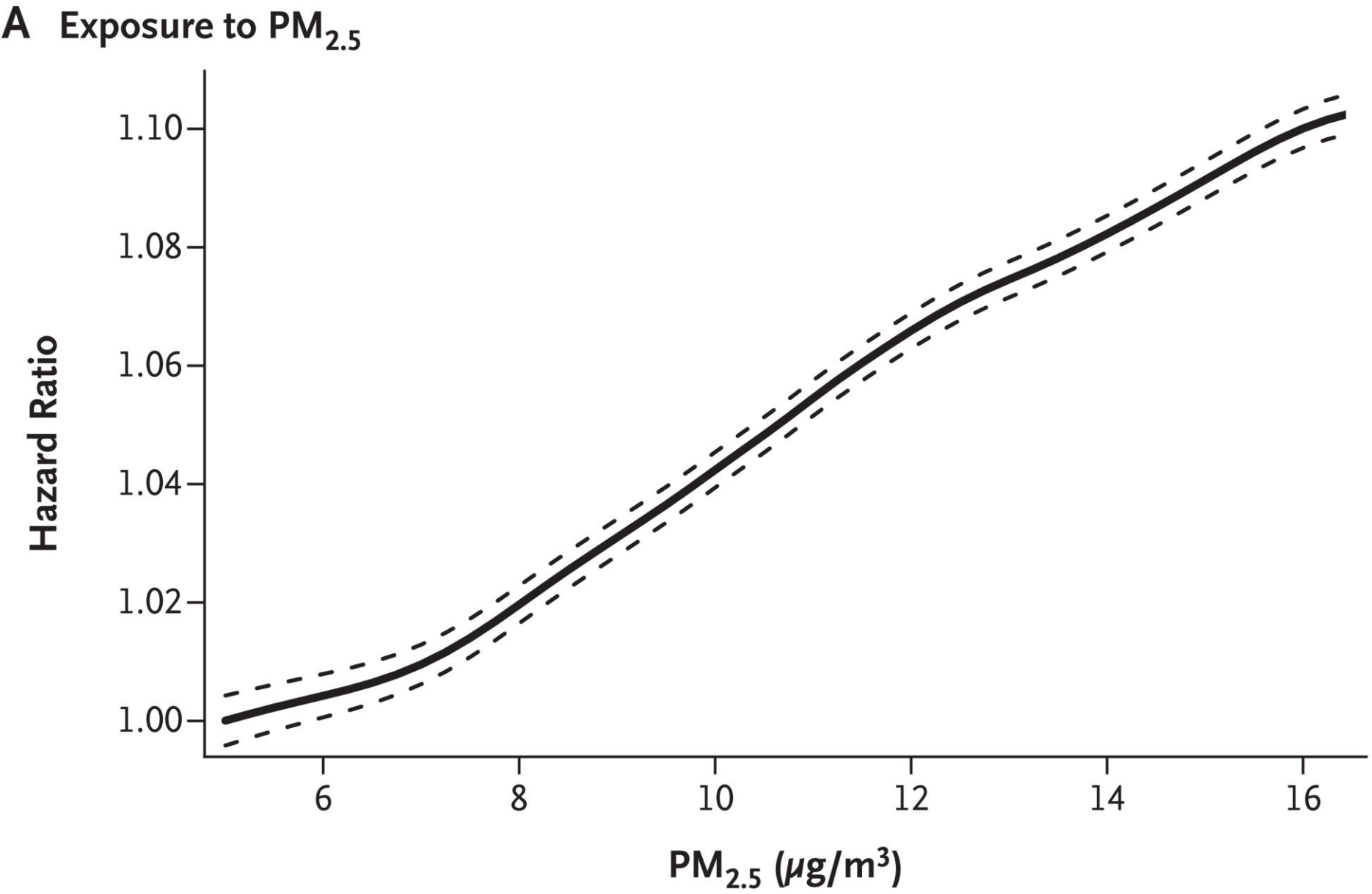
We had enough power to

- Look at people living in smaller cities and towns, and rural areas
- Look at groups separately by Race, Sex, and poverty
- Look at low exposure dose response (over 200 million person years at concentrations less than 12 $\mu g/m^3$)

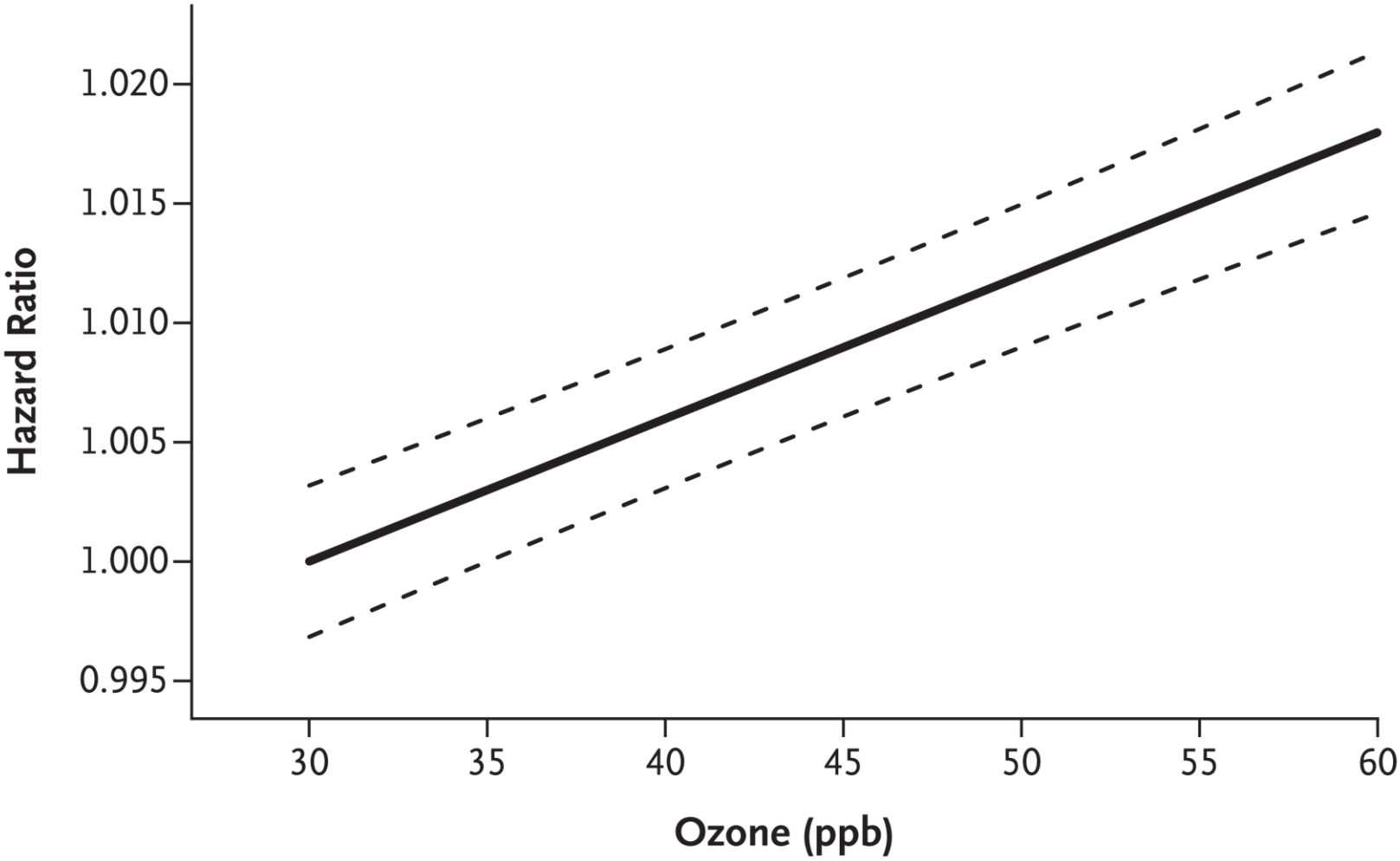
In Fact

- We couldn’t get a Cox Proportionate Hazard Model to run on the full dataset, so we used Divide and Conquer
- Randomly split into 50 datasets, and meta-analyzed the coefficients

Dose-Response Curve



B **Exposure to Ozone**



Analysis

- Candidate Gene Methylation
  - Is it a Modifier?
  - Is it a Mediator?
  - Not really Big Data
- Candidate Pathway Analysis
  - Is the pathway a Modifier/Mediator?
- Agnostic Pathway Analysis
- Agnostic Probe level Analysis

Difference in DNAm-age for IQR Increase in Particle Components

PM2.5 + Individual Components	Difference in Horvath DNAm-age for IQR (95% CI)	*P*	Difference in Hannum DNAm-age for IQR (95% CI)	*P*	N
EC	0.27 (-0.25, 0.80)	0.30	-0.09 (-0.48, 0.29)	0.64	940
OC	0.93 (0.37, 1.50)	0.001	0.35 (-0.05, 0.77)	0.09	940
Sulfate	0.59 (0.37, 0.81)	<0.0001	0.08 (-0.09, 0.25)	0.36	940
Nitrate	0.58 (0.11, 1.04)	0.01	0.30 (-0.04, 0.65)	0.08	940
Ammonium	0.59 (0.26, 0.92)	0.0004	0.06 (-0.18, 0.30)	0.63	940
PM2.5 plus Adaptive Lasso					
$PM_{2.5}$	0.18 (-0.30, 0.66)	0.45	-	-	940
Sulfate	0.51 (0.28, 0.74)	<0.0001	-	-	940
Ammonium	0.36 (0.02, 0.70)	0.04	-	-	940

Does the Pattern of Methylation Change with Exposure?

- Take empirical CDF of 500k methylation scores
- Fit a spline to the CDF
- Look at association of coefficients of the spline with exposure
- This can identify distortions of the curve associated with exposure
- Good if there are general distortions as opposed to pathway specific or location specific

Elastic Net Regression

Want to minimize the sum of squared errors:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - x_i' \hat{\beta})^2$$

Now we minimize:

$$\sum_{i=1}^{n} (y_i - x_i^T b)^2 + \lambda P_\alpha(b_1, \ldots, b_p)$$

Ridge regression, LASSO, and elastic net are part of the same family with penalty term:

$$P_\alpha = \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha) b_j^2 + \alpha |b_j| \right]$$

$\alpha = 0 \rightarrow$ ridge regression

$\alpha = 1 \rightarrow$ LASSO

$0 < \alpha < 1 \rightarrow$ elastic net!

This allows identification of multiple CpG sites associated with phenotype where a large number of small-effect CpG sites contribute to risk of disease or to the exposure

We can look at disease and exposure and capture mediation effects

Possibly interacting CpG sites