Supplementary Digital Content for “Pollutant composition modification of the effect of air pollution on progression of coronary artery calcium: the Multi-Ethnic Study of Atherosclerosis”

eAppendix 1: Predictive *k*-means Method

Here we provide a brief technical overview of the predictive k-means method. For complete details, see Keller *et al.* (2017)[1]. This method is implemented in the R package ‘predkmeans’, available at https://cran.r-project.org/package=predkmeans.

Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$ be a vector of measured values for p pollutants at monitor i. The predictive k-means computes cluster centers by assuming a mixture of normal distributions using a latent (unobserved) variable z_i that represents cluster membership. Latent cluster membership, i.e. $P(z_i = k)$ for $k = 1, \dots, K$, is modelled using a multinomial logit model that depends upon prediction variables. Conditional on $z_i = k$, the observations are modelled as multivariate normal: $(\boldsymbol{x}_i | z_i = k) \sim N(\boldsymbol{m}_k, \sigma^2 \boldsymbol{I})$. These criteria are optimized jointly to obtain the cluster centers $\boldsymbol{m}_k$. Once the cluster centers are identified, monitors are assigned to their closest center in the manner of traditional *k*-means. That is, cluster membership is assigned by minimizing the squared Euclidean distance between the multipollutant observation ($\boldsymbol{x}_i$) and the cluster center ($\boldsymbol{m}_k$).

eTable 1 provides a list of the geographic variables used to identify cluster centers and predict cluster membership. The variables are ‘point’ values at the monitor and residential location, ‘buffer’ values that are summed values within circular regions of different radius, or ‘distance’ values, which are linear distances from the location to a feature. Additional detail for each individual variable is provided in Keller *et al.* (2015)[2]. As described in the main text, PCA scores derived from these geographic covariates are used as the prediction variables in predictive k-means.

eTable 1: Geographic Variables

Variables	**Type**
Latitude and Longitude	Point
Population	Buffer
Major Road Length; Truck Route Length; Major Intersections	Buffer
Land Use (Multiple categories); Vegetation Index; Impervious Surface	Buffer
Elevation	Point
Major Road; Major Intersection; Port; Railyard; Truck Route; Coast	Distance
Emissions Inventory	Buffer

Once cluster centers were identified and cluster assignments made at monitor locations, cluster assignment at subject locations was predicted using multinomial logistic regression. The probability of subject location *j* belonging to cluster *k* was modelled as

$P(z_j = k) = \frac{\exp(\boldsymbol{r}_j^T \boldsymbol{\gamma}_k)}{\sum_{k'=1}^{K} \exp(\boldsymbol{r}_j^T \boldsymbol{\gamma}_{k'})}$. Here, the vectors $\boldsymbol{r}_j$ represent the prediction variables (in this case, PCA scores derived from geographic variables) at each location and $\boldsymbol{\gamma}_k$ are model coefficients. The coefficients are estimated based upon the cluster assignments at monitor locations. Participant locations are assigned to the cluster for which they have the largest probability of membership.

eAppendix 2: Cross-validation Details

In the leave-one-out cross-validation procedure, a single monitoring location was removed from the dataset, the remaining data were clustered and a prediction model for cluster was fit. Cluster membership at the left-out cluster was then predicted.

The primary metric for cross-validation performance of the predictive k-means clustering was mean-squared prediction error (MSPE) [1]. Let $k(i)$ denote the cluster to which monitor i is predicted to belong and $\boldsymbol{m}_k = (m_{k1}, \dots, m_{kp})$ be a vector containing the center of cluster k. We define MSPE as

$$MSPE = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - m_{k(i)j}\right)^2.$$

This is the sum of the squared differences between the monitor observations and their predicted cluster centers. CV results were also quantified using: mean-squared misclassification error (MSME), which assesses accuracy by summing the squared difference between the predicted cluster center and the closest cluster center; within-cluster sum-of-squares (wSS), which assess cluster centers by summing the squared differences between monitor observations and their closest cluster center; and classification accuracy (Acc), which is the proportion of clusters correctly labelled. MSPE incorporates information that is summarized separately by MSME and wSS, and all three quantities are on the same scale.

eTable 1. Cross-validation results for the best-performing models in each season.

Analysis	**Season**	**K**	**# PCA Scores**	**MSPE**	**MSME**	**wSS**	**Acc**
Primary	Cold	3	1	16.94[a]	4.15	15.59	0.85
		3	2	16.82	3.72	15.47	0.88
	Warm	3	2	15.32	4.22	13.44	0.75
Sensitivity	Cold	3	2	18.27	14.74	3.91	0.70
	Warm	2	4	17.19	13.84	3.96	0.75

[a] This model chosen because model with 2 PCA scores was overdetermined for the cluster with only two members.

eAppendix 3: Results from Sensitivity Analysis

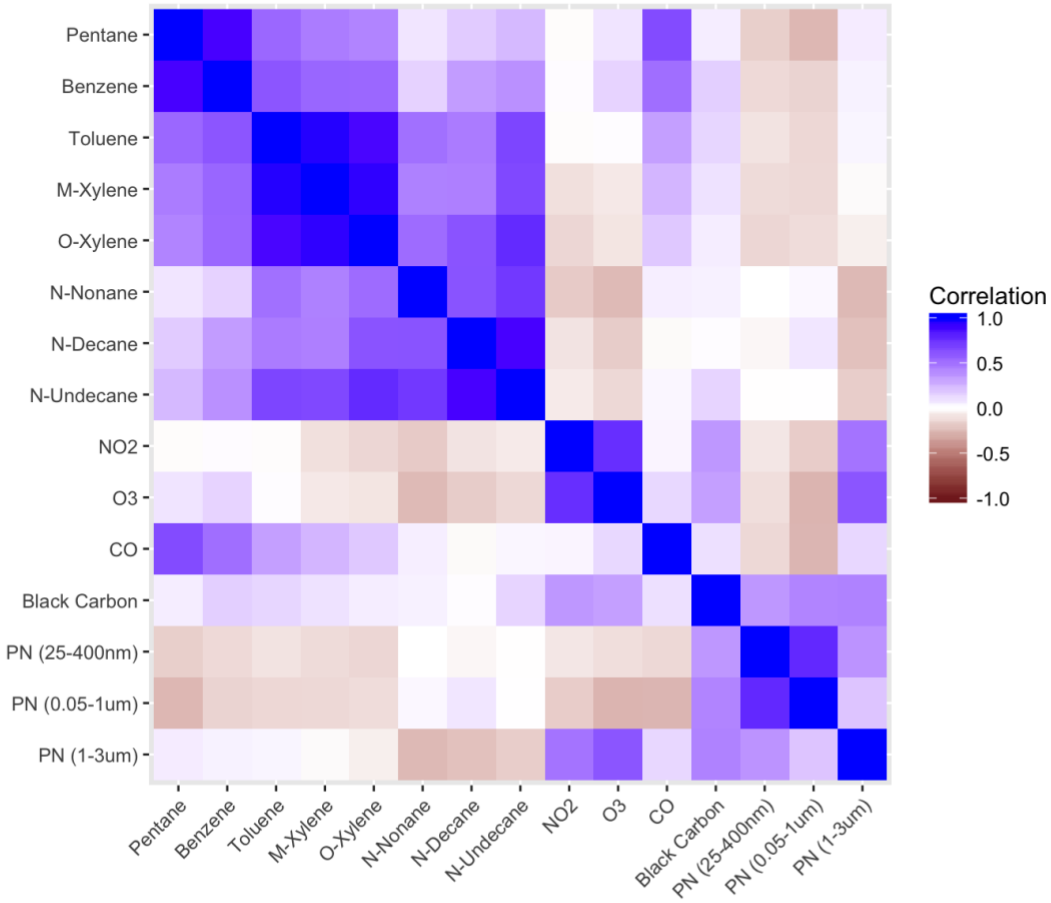
eTable 2. Estimates of the association between CAC progression, in Agatston units per year, and differences of 5 μg/m^3 $PM_{2.5}$ or 40 ppb NO_X, among clusters derived using covariates orthogonalized against baseline NO_X.

Exposure	**Source of Clusters**	**Cluster Name**	**Estimate (95% Confidence Interval)**	**Effect Modification *p*-value**[a]
$PM_{2.5}$	Cold Season[b]	Cluster 1	28.1 (18.2, 37.9)	0.023
		Cluster 2	8.8 (-6.0, 23.6)	
	Warm Season	Cluster 1	23.4 (13.7, 33.2)	0.83
		Cluster 2	21.6 (6.3, 36.9)	
NO_X	Cold Season[b]	Cluster 1	34.2 (22.0, 46.4)	0.017
		Cluster 2	9.8 (-8.9, 28.5)	
	Warm Season	Cluster 1	28.4 (16.2, 40.7)	0.84
		Cluster 2	26.8 (10.4, 43.1)	

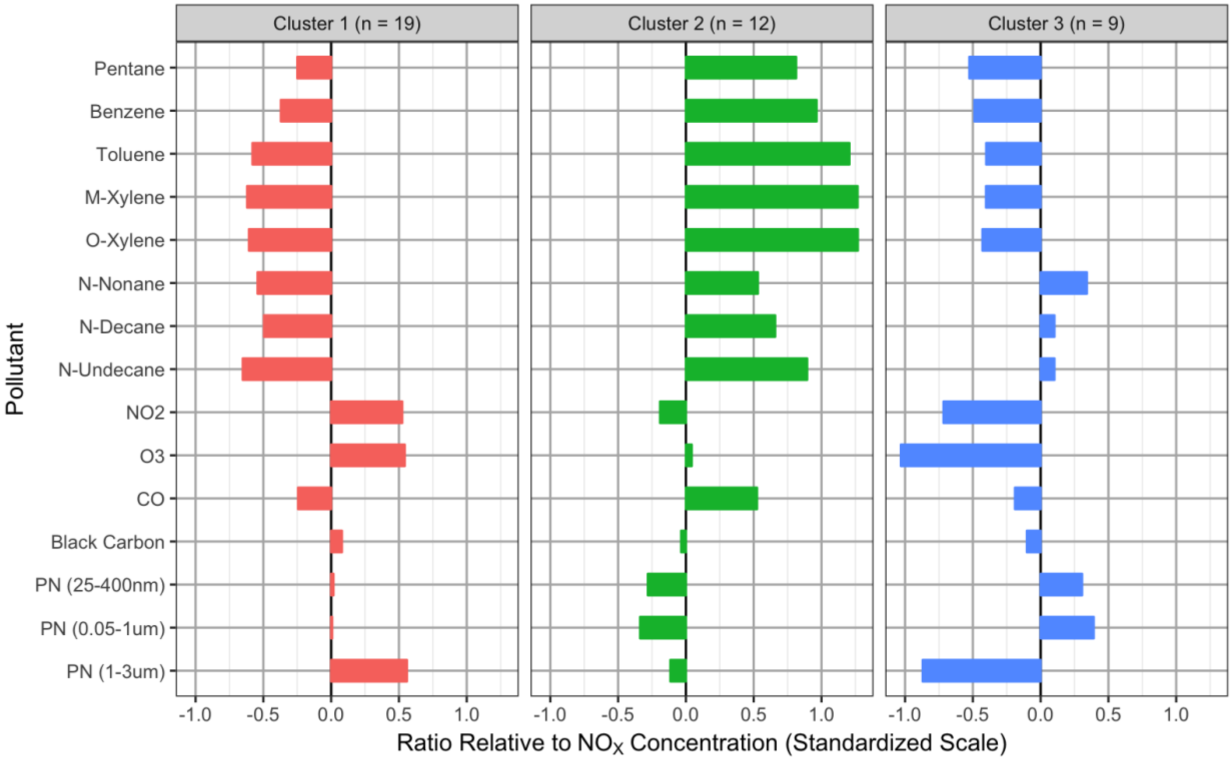
[a] *p*-values are from a likelihood ratio test comparing against the model without cluster-specific progression estimates.

[b] No subjects were predicted to belong to Cluster 3 in the Cold Season.

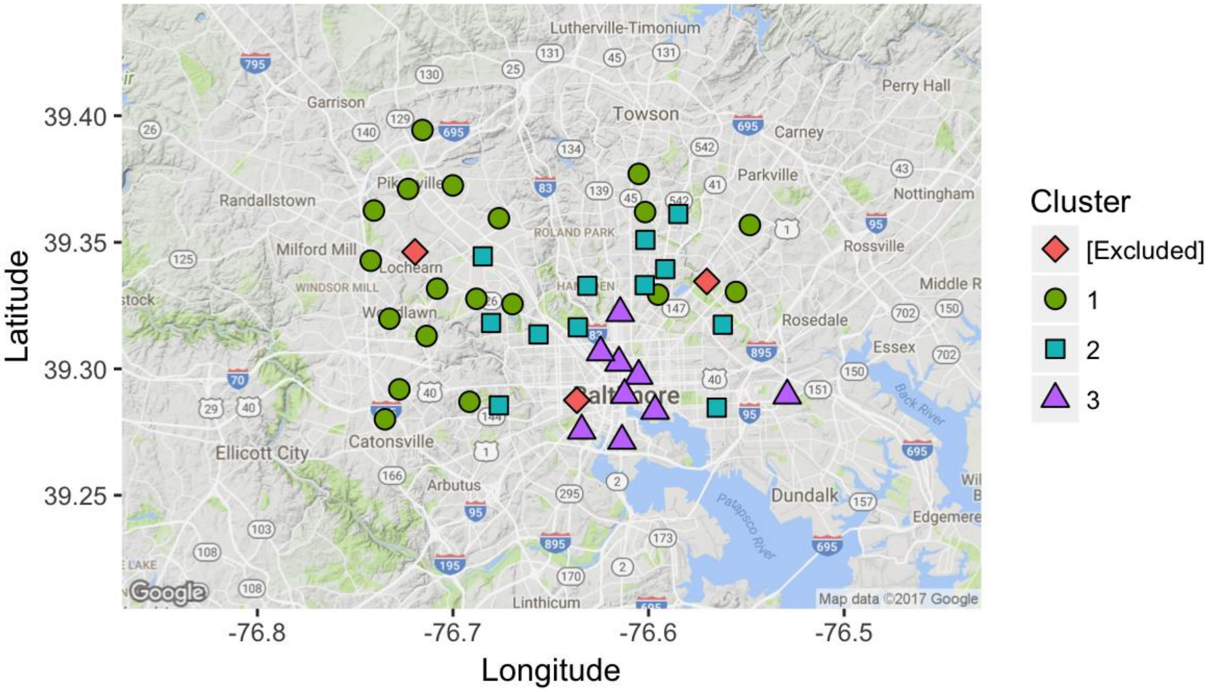
eAppendix 4: Figures for Warm Season



eFigure 1. Heatmap of the correlation between measurements in the warm season.

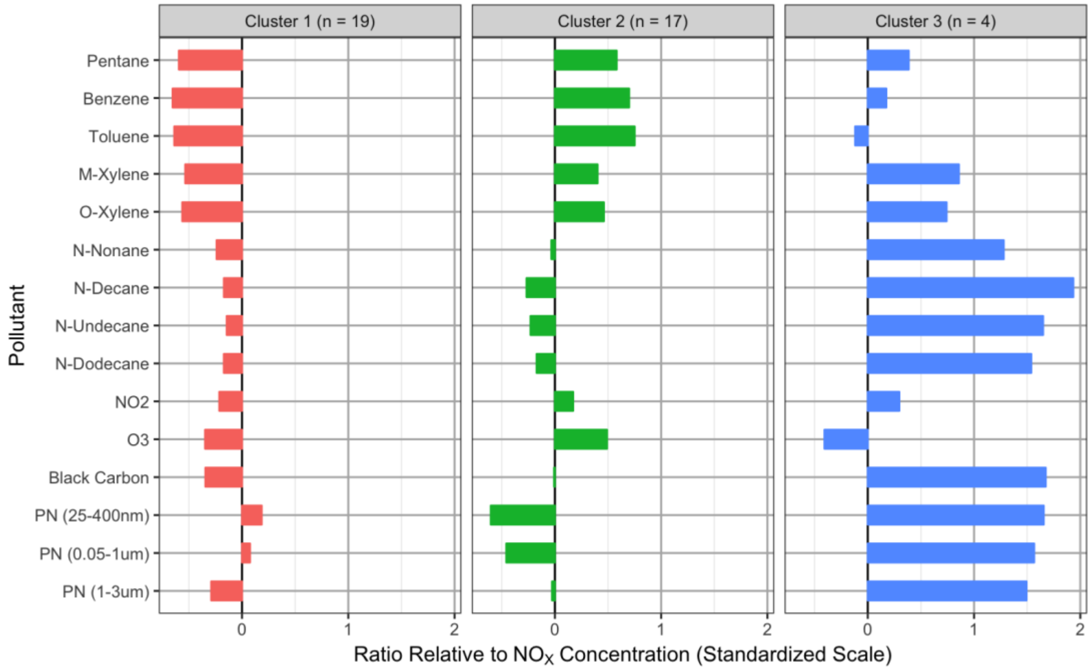


eFigure 2. Warm season cluster centers

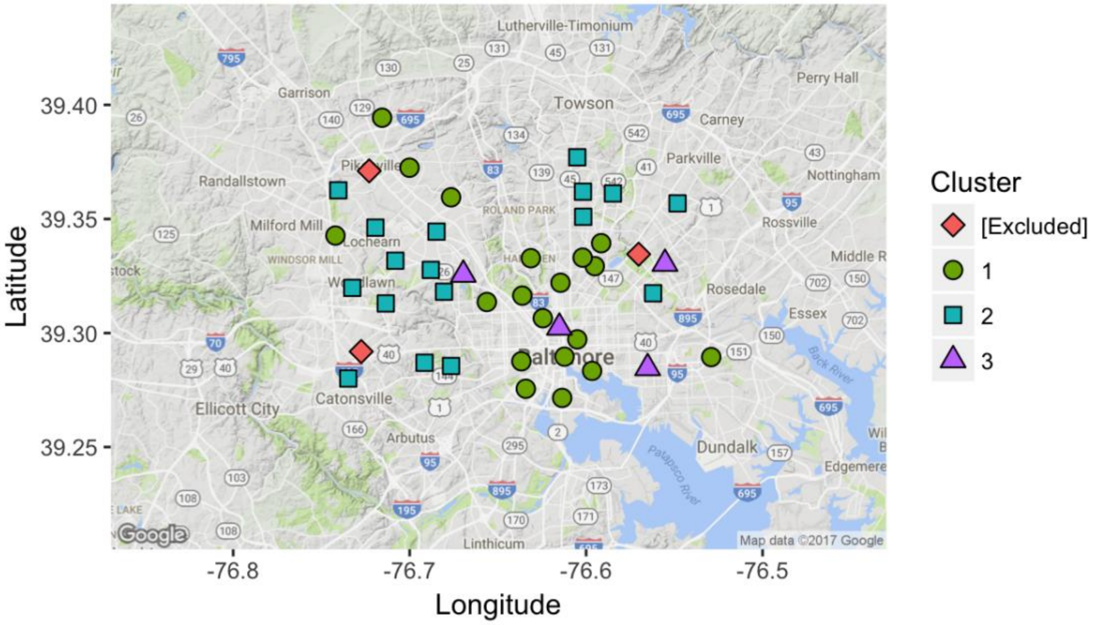


eFigure 3. Monitoring locations, colored by membership in warm season cluster.

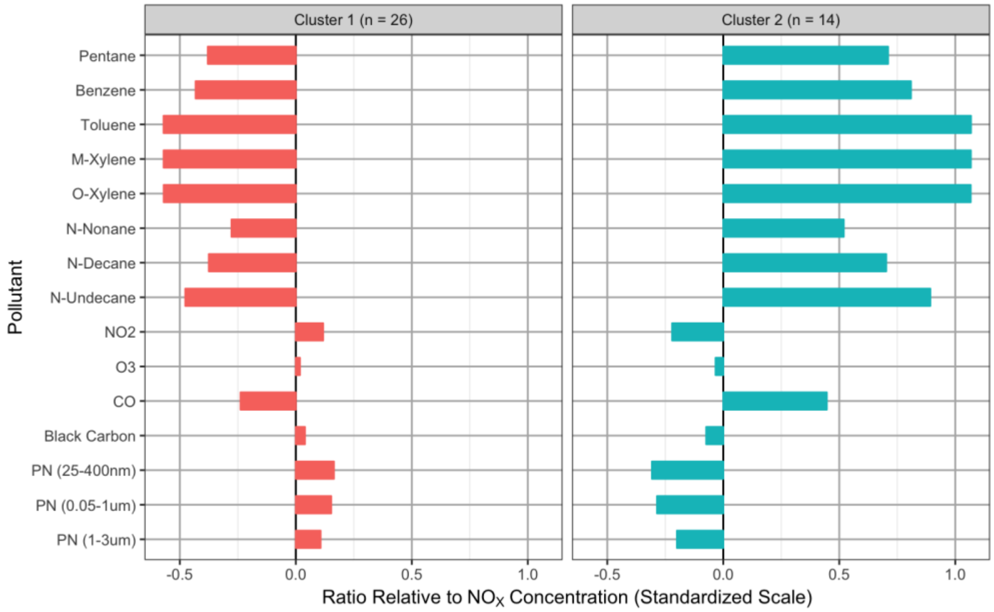
eAppendix 5: Figures for Sensitivity Analysis



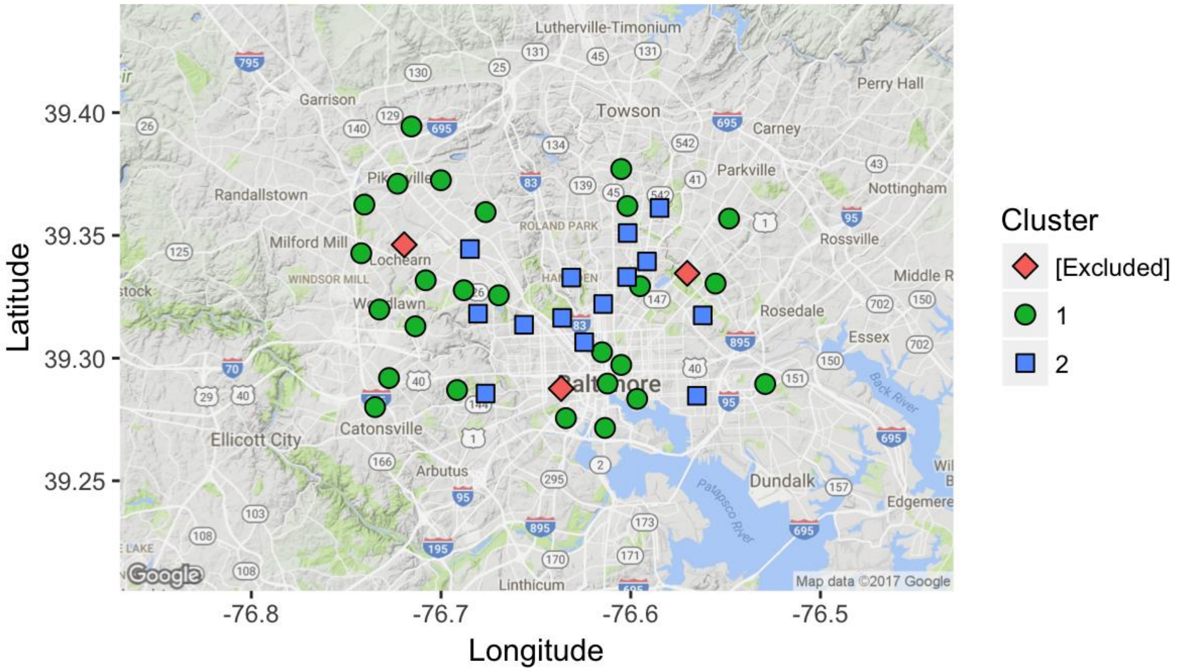
eFigure 4. Cold season cluster centers, from sensitivity analysis that orthogonalized covariates against baseline NO$_X$ concentrations.



eFigure 5. Monitoring locations, colored by membership in cold season cluster, from sensitivity analysis that orthogonalized covariates against baseline NO$_X$ concentrations.



eFigure 6. Warm season cluster centers, from sensitivity analysis that orthogonalized covariates against baseline NO_X concentrations.



eFigure 7. Monitoring locations, colored by membership in warm season cluster, from sensitivity analysis that orthogonalized covariates against baseline NO_X concentrations.

References

1. Keller JP, Drton M, Larson T, Kaufman JD, Sandler DP, Szpiro AA. Covariate-adaptive clustering of exposures for air pollution epidemiology cohorts. *Ann Appl Stat* 2017;11(1):93-113. doi:10.1214/16-AOAS992.
2. Keller JP, Olives C, Kim S-Y, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environ Health Perspect* 2015;123(4):301-309. doi:10.1289/ehp.1408145.