Received: 12 February 2021 | Revised: 29 December 2021 | Accepted: 6 January 2022

DOI: 10.1002/sim.9330

RESEARCH ARTICLE

Statistics in Medicine WILEY

Estimating the health effects of environmental mixtures using principal stratification

Roger D. Peng[1] | **Jia C. Liu**[1] | **Meredith C. McCormack**[2] | **Loretta J. Mickley**[3] | **Michelle L. Bell**[4]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland,

[2]Division of Pulmonary and Critical Care Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland,

[3]School of Engineering and Applied Sciences, Harvard University, Boston, Massachusetts,

[4]School of the Environment, Yale University, New Haven, Connecticut,

Correspondence
Roger D. Peng, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street E3527, Baltimore, MD 21205, USA.
Email: rdpeng@jhu.edu

Funding information
National Institutes of Health, Grant/Award Number: ES021427; US Environmental Protection Agency, Grant/Award Numbers: RD835871, RD83587201

The control of ambient air quality in the United States has been a major public health success since the passing of the Clean Air Act, with particulate matter (PM) reductions resulting in an estimated 160 000 premature deaths prevented in 2010 alone. Currently, public policy is oriented around lowering the levels of individual pollutants and this focus has driven the nature of much epidemiological research. Recently, attention has been given to viewing air pollution as a complex mixture and to developing a multi-pollutant approach to controlling ambient concentrations. We present a statistical approach for estimating the health impacts of complex environmental mixtures using a mixture-altering contrast, which is any comparison, intervention, policy, or natural experiment that changes a mixture's composition. We combine the notion of mixture-altering contrasts with sliced inverse regression, propensity score matching, and principal stratification to assess the health effects of different air pollution chemical mixtures. We demonstrate the application of this approach in an analysis of the health effects of wildfire PM air pollution in the Western US.

KEYWORDS
dimension reduction, mixtures, particulate matter

1 | INTRODUCTION

The control of ambient air quality in the United States has been a major public health success since the passing of the Clean Air Act, with particulate matter (PM) reductions resulting in an estimated 160 000 premature deaths prevented in 2010 alone.[1] Currently, public policy is oriented around lowering the levels of *individual pollutants* and this focus has driven the nature of much epidemiological research. The individual pollutant framework made sense at a time when the primary focus was lowering extreme levels seen in catastrophic episodes such as the London Fog[2] or Donora Pennsylvania.[3] Recently, greater attention has been brought to viewing air pollution as a complex mixture and to developing a *multi-pollutant* approach to controlling ambient concentrations.[4]

While advances in data collection have produced more detailed measurements of the complex air pollution mixture, the statistical approaches used in traditional air pollution studies are not well-suited to studying the health effects of these

mixtures. The traditional statistical approach to studying complex air pollution mixtures has been to select individual components and adjust for the presence of other components. This approach greatly simplified modeling strategies and allowed for the straightforward translation of research on health effects to interventions or policies: If pollutant X is harmful, then we should *reduce* exposure to pollutant X. This inherent directionality induced by the one-dimensional nature of the exposure ("lower" vs "higher") leads to a clear next step.

Considering sets of pollutants as exposures introduces multi-dimensional measurements whose values do not lie on an ordered line, but rather in an unordered high-dimensional space. The geometry of the multi-pollutant exposure space removes the natural directionality of the single pollutant approach, breaking the simplicity that had previously connected health effects research and potential interventions. A common way to circumvent this problem is to focus on one or a few key pollutants and interpret health effects in the context of holding the concentrations of other pollutants constant. However, such an approach is impractical because it may be impossible to design an intervention that modifies the mixture in that exact manner.

One natural phenomenon that can significantly alter the composition of ambient air pollution is a wildfire.[5] Previous work on wildfires in the Western US has shown that wildfires generally increase the level of fine particulate matter air pollution ($PM_{2.5}$), defined as particles less than 2.5 μm in aerodynamic diameter, and are harmful to population health.[6,7] In particular, Liu et al[6] found that days with high intensity wildfire $PM_{2.5}$, which they label as "smoke waves," are associated with a 7.2% (95% CI: 0.25%, 15%) increase in hospitalizations for respiratory diseases amongst elderly people enrolled in the Medicare insurance system. Furthermore, it has been shown that the chemical composition of wildfire $PM_{2.5}$ differs substantially from the typical urban $PM_{2.5}$ mixture and that the composition of wildfire $PM_{2.5}$ depends in large part on the vegetation of the local ecoregion.[5]

Given that previous work has identified wildfires as a phenomenon that can cause the composition of $PM_{2.5}$ to change, we saw this as an opportunity to directly study the health impacts of $PM_{2.5}$ chemical composition in a real-world setting. The goal of this article is to introduce and demonstrate a statistical method that addresses the challenges posed by studying air pollution mixtures and overcomes the limitations of current approaches. In our method, we leverage natural changes to mixtures that are induced by outside forces, such as wildfires, and connect the subsequent changes to the pollution mixtures to population health outcomes. We then apply this approach a dataset of Medicare hospitalizations and wildfire PM concentrations in the Western US.

2 | METHODS

Estimating the health impacts of different mixtures of air pollutants is a challenging task given the number of competing aspects that must be balanced. One must first be able to characterize the change in a mixture, which may be high-dimensional, and then relate that change to a health outcome while controlling for potential confounders and changes in the overall concentration of the pollutant. Finally, one must provide a reasonable interpretation for why the mixture has changed in a given way and suggest next steps for mitigating any potential harm from particularly toxic mixtures. The approach we describe in this section attempts to address all of these challenges in order to provide useful insights into the health effects of PM air pollution mixtures.

Our overall goal is to develop evidence relating to the toxicity of certain air pollution chemical mixtures. In our application (Section 4), we want to compare respiratory hospitalization rates on days with a high proportion of wildfire smoke $PM_{2.5}$ ("smoke wave" days) to hospitalization rates on days with more typical $PM_{2.5}$. The approach that we have developed for exploring the short-term health impacts of air pollution mixtures combines techniques from dimension reduction and causal inference methodology:

1. We first apply propensity score matching to create a dataset where "smoke wave" days, that is, days with high levels of wildfire $PM_{2.5}$, are matched to non-smoke wave days.
2. We then apply sliced inverse regression (SIR) to identify a projection of the $PM_{2.5}$ chemical constituents that best characterizes the variation in $PM_{2.5}$ chemical composition between smoke wave and non-smoke wave days.[8,9]
3. Finally, we make use of principal stratification[10] to estimate the effect of changing the $PM_{2.5}$ composition (as summarized by the SIR projection) on respiratory hospitalizations.

Our methodology leverages the benefits obtained from dimension reduction methods but also produces interpretable estimates of risk because of the presence of an external mixture-altering contrast.

2.1 | Matching

Let Z_t be the binary indicator of a smoke wave day, where $Z_t = 1$ indicates that day t is a smoke wave day and $Z_t = 0$ indicates an otherwise normal day. For the propensity score matching, we model $\mathbb{P}(Z_t = 1|\mathbf{w}_t)$ using a logistic regression where $\mathbf{w}_t$ is a vector of potential confounding covariates such as temperature, season, or humidity. To execute the matching algorithm and to create the matched dataset, we used the MatchIt package of Ho et al[11] with the default nearest neighbor matching option selecting four controls for each observation in the treatment group with replacement. Sensitivity analysis did not indicate that the results would be substantially altered by varying the ratio of controls to treatment from 1 to 4.

A key challenge to studying the effects of chemical composition is controlling for the overall level of the pollutant. Because it is already known that higher levels of $PM_{2.5}$ are associated with worse respiratory outcomes, we would naturally expect higher respiratory hospitalizations associated with smoke waves due to the increase in $PM_{2.5}$ levels.[6] However, we can use the matched dataset to identify days where the smoke wave is not predicted to change the overall level and make our comparisons weighting more heavily those days.

If $PM_t(1)$ and $PM_t(0)$ are the potential outcomes for the overall level of $PM_{2.5}$ on day t when there is and is not a smoke wave, respectively, then we want to identify days where $\Delta_t = PM_t(1) - PM_t(0) < \varepsilon$ for some chosen small value of ε. The challenge of course is that for any given day t we only ever observe one of $PM_t(1)$ or $PM_t(0)$. Our approach models the joint distribution of the $PM_{2.5}$ potential outcomes as a bivariate Normal on the matched dataset with baseline covariates $\mathbf{v}_t$ and correlation coefficient ω. With this formulation, we can predict the conditional mean of each potential outcome as

$$\begin{aligned}\mathbb{E}\left[PM_t(1)|PM_t(0), \mathbf{v}_t\right] &= \mathbf{v}_t'\xi + \tau + \omega\frac{\eta_1}{\eta_0}(PM_t(0) - \mathbf{v}_t'\xi),\\ \mathbb{E}\left[PM_t(0)|PM_t(1), \mathbf{v}_t\right] &= \mathbf{v}_t'\xi + \omega\frac{\eta_0}{\eta_1}(PM_t(1) - \mathbf{v}_t'\xi - \tau),\end{aligned}$$

where $\tau = \mathbb{E}\left[PM_t(1) - PM_t(0)\right]$, η_1 and η_0 are the standard deviations of $PM_t(1)$ and $PM_t(0)$, respectively, and ξ is a vector of regression coefficients. The correlation parameter ω is not estimable from the data and so we choose a range of values between 0 and 1 and examine the sensitivity of our results to the different values. The remaining unknown parameters can be estimated from the observed sample from the marginal distributions of the potential outcomes.

2.1.1 | Weighting

For each t now we have a pair values $PM_t(1)$ and $PM_t(0)$, one of which is observed and one of which is predicted. We then develop a set of weights $u_t \propto \exp\left(-\frac{1}{2}(PM_t(1) - PM_t(0))^2/b\right)$, where b is chosen to define when $PM_t(1)$ and $PM_t(0)$ are thought to be "close." The weights u_t then allow us to downweight days where this a large change in $PM_{2.5}$ level when estimating the change in pollution mixture. Smaller values of b restrict the allowable difference between the $PM_{2.5}$ potential outcomes but may not allow for enough information to estimate the mixture change. Larger values of b allow for the inclusion of more observations but have greater potential to confound the change in mixture with the change in overall level.

2.2 | Principal mixture direction

We use weighted SIR for binary outcomes[8,9] to estimate the change in $PM_{2.5}$ mixture between smoke wave and non-smoke wave days. On each day t we have the concentrations of 28 chemical constituents of $PM_{2.5}$. SIR gives us the projection of the chemical constituent data that best explains the difference between the smoke wave and non-smoke wave days. We call this projection the *principal mixture direction* as it shows how the $PM_{2.5}$ mixture can change with the presence of wildfires. Because SIR only uses information about the mean of the data, there is only one such projection (equivalent to linear discriminant analysis in this case). However, the approach of framing the estimation of the mixture direction as an SIR problem allows for straightforward extension to potential drivers of mixture change that have more than two levels. We address this possible extension in Section 6.

Let M be an $n \times p$ matrix where n represents the total number of days of observations in the matched dataset and p is the number of chemical constituents of $PM_{2.5}$ that we measure. Let $\mathbf{z} = (z_1, \ldots, z_n)$ be an indicator vector where $z_t = 1$ indicates a smoke wave day and $z_t = 0$ indicates a non-smoke wave day. Finally, let $U = \text{diag}(u_1, u_2, \ldots, u_n)$ be an $n \times n$ diagonal matrix of weights where $u_1, \ldots, u_n$ are defined in Section 2.1.

First, we create the weighted chemical constituent matrix $\tilde{M} = U^{1/2}M$ and let Σ_M be the covariance matrix of $\tilde{M}$. Then the standardized constituent matrix is $R = \Sigma_M^{-1/2}\left(\tilde{M} - \mathbb{E}\left[\tilde{M}\right]\right)$. Then we can estimate the first SIR direction as

$$\boldsymbol{\gamma} = \left(\frac{R'\mathbf{z}}{\mathbf{z}'\mathbf{z}} - \frac{R'(1-\mathbf{z})}{(1-\mathbf{z})'(1-\mathbf{z})}\right)\Sigma_M^{-1/2},$$

that is, the group mean difference of the rows of R between smoke wave and non-smoke wave days. The vector γ is the principal mixture direction and can be used to score each day's $PM_{2.5}$ mixture as being more "smoke-wave-like" or more "non-smoke wave-like." For an observation of a set of chemical constituents $\mathbf{m}_t$, this score can be computed for each day t as $x_t = (\mathbf{m}_t - \overline{\mathbf{m}})'\boldsymbol{\gamma}$, where $\overline{\mathbf{m}}$ is the vector of means for each chemical constituent. We then refer to x_t as the *principal mixture score*. If x_t is strongly positive, then the $PM_{2.5}$ mixture on day t is more like a smoke wave. If it is strongly negative, the $PM_{2.5}$ mixture is more like a typical day (non-smoke wave).

2.3 | Principal stratification

Let $Y_t(1)$ and $Y_t(0)$ represent the potential outcomes of a population health response on day t. In order to assess the health impact of changes to an air pollution mixtures, we want to examine the outcomes $Y_t(1)$ and $Y_t(0)$ and their relationship with the potential outcomes of the principal mixtures score, which we will denote as $x_t(1)$ and $x_t(0)$. For our application, $Y_t(1)$ and $Y_t(0)$ will represent the rate of respiratory hospitalizations on days with or without a smoke wave, respectively, and the principal strata will be defined by the difference $x_t(1) - x_t(0)$. We are interested in the log of the ratio $\mathbb{E}[Y_t(1)]/\mathbb{E}[Y_t(0)]$ (ie, the log-relative risk) within those principal strata.

Note that in Section 2.2, we showed how to estimate x_t, which represents the *observed* principal mixture score for a given day t. If day t was a smoke wave day, then we have $z_t = 1$ and we observe the potential outcome $x_t = x_t(1)$. If day t was a non-smoke wave day, then we have $z_t = 0$ and we observe the potential outcome $x_t = x_t(0)$. In either case, we can only observe one of the potential outcomes on any given day t and the task in this section is to build a model for estimating the unobserved potential outcome on day t. Briefly, our approach is to consider the unobserved potential outcome as missing data, build a model for the joint distribution of both potential outcomes (the "complete data"), and then infer the unobserved potential outcome given the observed outcome.

The health outcome is modeled with a Poisson distribution, with,

$$\begin{aligned} Y_t(z_t)|x_t(1), x_t(0) &\sim \text{Poisson}(\mu_t(z_t)), \\ \log \mu_t(z_t) &= \mathbf{r}_t'\alpha + \delta z_t + z_t f(x_t(1) - x_t(0)), \end{aligned} \tag{1}$$

where r is a vector of baseline covariates, z_t is the smoke wave indicator for day t, α is a vector of regression coefficients, and f is a smooth function of the difference in principal mixtures score, $x_t(1) - x_t(0)$. Our focus is on estimating the function f to see what we can learn about the relationship between a change in principal mixture score and risk of hospitalization.

The potential outcomes for the principal mixture scores $x_t(1)$ and $x_t(0)$ are modeled as bivariate Normal,

$$\begin{pmatrix} x_t(1) \\ x_t(0) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{h}_t'\theta + \zeta \\ \mathbf{h}_t'\theta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix}\right), \tag{2}$$

where $\mathbf{h}$ is a vector of covariates for predicting $x_t(1), x_t(0)$, θ is a vector of coefficients, and ζ is the additive effect of smoke waves on the mixture score. The covariates in $\mathbf{h}$ are not $PM_{2.5}$ chemical constituents, but rather are other factors that may be useful for predicting the mixture score (eg, weather or seasonal indicators). Note that on every day t we will observe one of $x_t(1)$ or $x_t(0)$ and so our primary task is to infer the value of the unobserved potential outcome. Given a value for ρ and a vector of covariates $\mathbf{h}$, along with estimates of θ and ζ, we can infer the value of either $x_t(1)$ or $x_t(0)$ using the full conditional distributions.

Our goal is to estimate the quantity

$$\log \frac{\mathbb{E}\left[Y_t(1)|x_t(1), x_t(0)\right]}{\mathbb{E}\left[Y_t(0)|x_t(1), x_t(0)\right]} = \delta + f(x_t(1) - x_t(0)). \tag{3}$$

Here, δ is the change in risk when the change in the mixture score is zero and could potentially be interpreted as a direct effect of the smoke wave. The function f tells us the log change in hospitalization risk associated with a change in the difference of potential mixture scores. While the shape of this function is potentially of interest, it does not specifically have a causal interpretation. Rather, we may be more interested in specific values of the log-relative risk within strata defined by the potential outcome difference $x_t(1) - x_t(0)$. In order to assess the uncertainty of our estimate of this quantity, we use the bootstrap procedure. Within the matched dataset, observations in the smoke wave and non-smoke wave groups are resampled separately with replacement and we reimplement the principal stratification estimation procedure after each resampling. We use 5000 bootstrap replications and compute 95% confidence intervals for our log-relative risk estimates using the percentile method.

3 | SIMULATION STUDY

We designed and conducted a simulation study to assess the ability of the approach described above to detect any relationship between a change in chemical composition of a mixture (as measured by the mixture score) and a health outcome. In particular, we focused on the method's ability to estimate the function f in (3), which characterizes the relationship between x_t and Y_t, under some model misspecification. For the purposes of the simulation study, we specified f as a linear function so that the relative risk of hospitalization is $\log \frac{\mathbb{E}\left[Y_t(1)|x_t(1), x_t(0)\right]}{\mathbb{E}\left[Y_t(0)|x_t(1), x_t(0)\right]} = \delta + \beta(x_t(1) - x_t(0))$ and the primary target of inference is β. We will examine the bias, variance, and overall root mean squared error of our estimate of β.

We first simulated daily observations of the 28 $PM_{2.5}$ chemical constituents for both smoke wave and non-smoke wave days using a multivariate Normal distribution with covariance matrix equal to the empirical covariance matrix estimated from the observed data. We specified the difference between the smoke wave and non-smoke wave days as a random shift in the means of the 28 constituents. For the purpose of the simulation, we assumed the observations were already "matched" and did not simulate values for any potential confounding variables. Because $PM_{2.5}$ chemical constituent data were simulated for both smoke wave and non-smoke wave potential outcomes, we subsequently simulated the principal mixture scores $x_t(1)$ and $x_t(0)$ for both smoke wave and non-smoke wave days, respectively using a multivariate Normal distribution, as in (2). Marginal variances of $x_t(1)$ and $x_t(0)$ were chosen to match what we observe in the dataset and we specified the correlation $\rho^\star$ to vary from 0 to 0.8 across a series of simulation runs. With both potential outcomes $x_t(1)$ and $x_t(0)$ simulated we could simulate the hospitalization data using the model in (1).

With the simulated potential outcomes data, we derived an "observed" dataset and applied our procedure. We first estimated the principal mixture direction using SIR and then calculated principal mixture scores for each observation. From there we applied principal stratification to estimate the linear relationship between the mixtures score and hospitalizations. For that aspect of the approach we specified a value for ρ to produce an estimate $\hat{\beta}$. We then compared $\hat{\beta}$ to the true β used to simulate the hospitalization data and computed the absolute bias, standard deviation (SD), and root mean squared error (RMSE).

The results of the simulation study are summarized in Figure 1A-C, which show the absolute bias, standard deviation, and root mean squared error of $\hat{\beta}$, respectively. On the x-axis of each figure is the true value of $\rho^\star$ used to simulate the potential outcome data. Each of the different colors represents the model being applied using a specific value of ρ ranging from 0 to 0.8 and each circle represents the average over 1000 simulations.

From Figure 1, it is clear that specifying a value of ρ equal to that used to simulate the data results in the smallest bias, in general. However, certain values of ρ can result in overall less bias, regardless of the true value of $\rho^\star$. For example, specifying value of ρ of 0.7 is almost uniformly better with respect to bias than specifying $\rho = 0.8$. With the SD and RMSE, there is very little variation across true values of $\rho^\star$. It seems clear from Figure 1B,C that specifying $\rho = 0$ results in the smallest SD and overall RMSE. Although the bias in $\hat{\beta}$ seems to vary somewhat with the true value of $\rho^\star$, the overall RMSE is dominated by the variance of the estimator, which does not vary with $\rho^\star$. Hence, the choice of ρ that minimizes the overall RMSE is $\rho = 0$. However, if one is more interested in minimizing the overall absolute bias (across all values of $\rho^\star$) while sacrificing some variance, the optimal ρ from the simulation is 0.3.

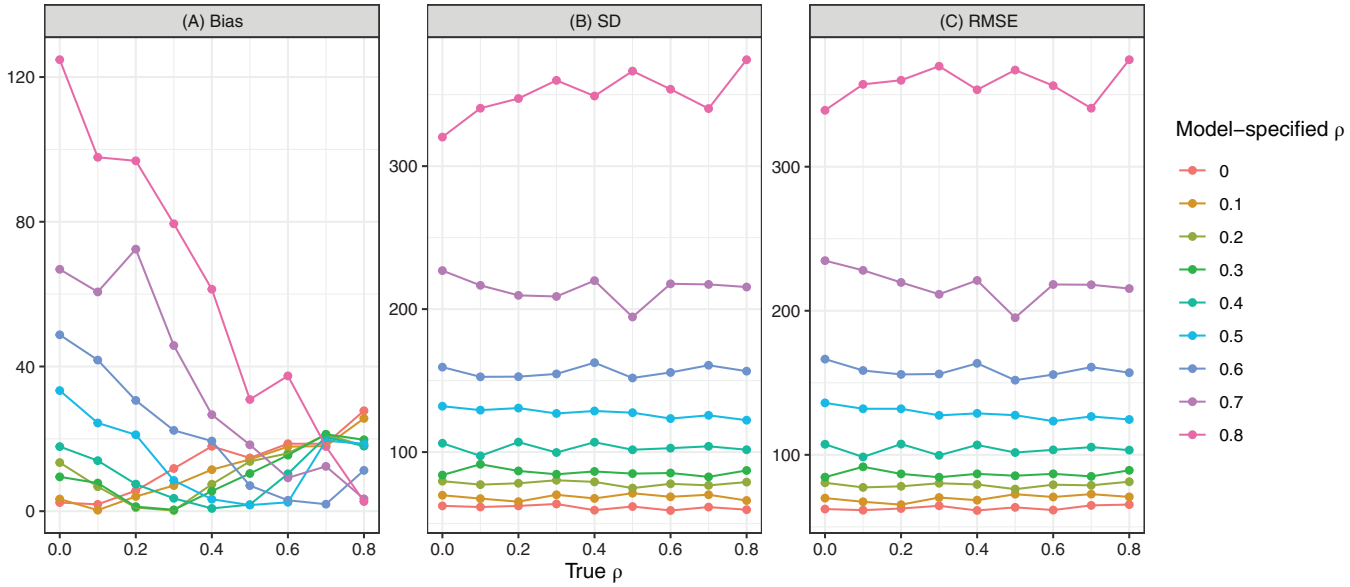

FIGURE 1 Results from the simulation study on bias, variance (SD), and root mean squared error of the estimate of β

4 | APPLICATION: WILDFIRES, AIR POLLUTION, AND HEALTH

Air pollution from wildfires represents a topic of important public health interest, but we take a slightly different interest in wildfires for the purposes of this article. Because wildfires introduce a rapid shift in the chemical composition of ambient air pollution in their surrounding regions, they provide an opportunity to study the short-term health impacts of changing the composition of $PM_{2.5}$. Although wildfires are typically started by humans and are predictable in certain ways (ie, there is a well-defined fire season in most areas), from the perspective of people living in major population centers, the occurrence of wildfires is reasonably modeled as random.[12] In addition, while wildfires can certainly pose a direct threat to human populations, most wildfires originate in largely uninhabited areas and so primarily affect people's health via the pollution that they generate and transport. As such, wildfires present an interesting opportunity to study the effect of air pollution composition changes if the appropriate methodology could be developed and applied.

Recently, there have been calls in the air pollution community to leverage novel study designs in order to develop stronger forms of evidence.[13] Our aim here is to take this opportunity as a case study on which to build the methodology for studying air pollution mixtures that could be generalized to a variety of other scenarios.

4.1 | Data

The analysis that is the focus of this article incorporates data on (1) respiratory hospitalizations amongst enrollees in the US Medicare insurance system; (2) the chemical composition of ambient $PM_{2.5}$; and (3) wildfires in the Western US for the years 2004 to 2009. This section provides some details on these datasets and how they were obtained.

Information about hospitalizations for respiratory diseases was obtained from the US Medicare system, which is a national health insurance system for people aged 65 years and older. The Center for Medicare and Medicaid Studies (CMS) provides billing claims for hospitalizations for Medicare enrollees. We obtained records for the entire US for the years 2004 to 2009 and included any Medicare Part A (hospital inpatient) claim with a primary International Classification of Disease (version 9) code in the range of 490 to 492 (chronic obstructive pulmonary disease), 464 to 466, or 480 to 487 (respiratory tract infection). This dataset was subsequently subsetted to include counties in the Western US for which wildfire information was available.

The chemical composition of ambient $PM_{2.5}$ can be obtained from the US Environmental Protection Agency's Chemical Speciation Network, as well as numerous state and local air monitors. Together, they provide 24-hour average concentrations of over 50 chemical constituents of $PM_{2.5}$ on a 1-in-3 day or 1-in-6 day basis. We obtained $PM_{2.5}$ chemical

speciation data from the EPA Air Quality System (which assembles all these data) for the entire US for the years 2004 to 2009. There are approximately 300 monitoring locations across the United States that measure $PM_{2.5}$ chemical speciation. The analysis presented in Section 5 focuses on 28 constituents that make up most of the mass of ambient $PM_{2.5}$ and are commonly measured across all sites.

The locations used for this article were at the intersection of the hospitalization and chemical speciation datasets. Because the $PM_{2.5}$ chemical speciation data were primarily collected near urban locations, the included locations primarily consist of urban communities. In total, there were 48 counties included in the analysis and they are shown in Figure 2. Summary information about the counties, including average chemical constituent levels and county demographics is presented in Supplementary Figure 1.

4.2 | Wildfires and smoke waves

We employed wildfire outputs from the GEOS-Chem chemical transport model (v9-01-03) to generate daily wildfire-specific $PM_{2.5}$ levels for 6 years (2004-2009). GEOS-Chem is a global 3D atmospheric chemistry model driven by meteorology.[14] The modeling integrates meteorological data from Goddard Earth Observing System (GEOS-5) of the NASA Modeling and Assimilation Office and observed wildfire area burned based on the Global Fire Emissions Database (GFED3). GFED3 combines satellite observations of fire counts, area burned, and fuel load to produce gridded, daily maps of wildfire emissions.[15] The GEOS-Chem model outputs used in this study are daily (24-hour-average), gridded surface $PM_{2.5}$ concentrations for fire seasons (May 1-October 31) 2004 to 2009. The grid size is 0.5×0.67 degrees (approximately 50×75 km) latitude-by-longitude.

The GEOS-Chem model provides an estimate of the proportion of ambient $PM_{2.5}$ at a given location and time that originates from a wildfire, regardless of the location of the wildfire because wildfire smoke can travel large distances. We defined population exposure to wildfire $PM_{2.5}$ based on daily wildfire-specific $PM_{2.5}$ estimates

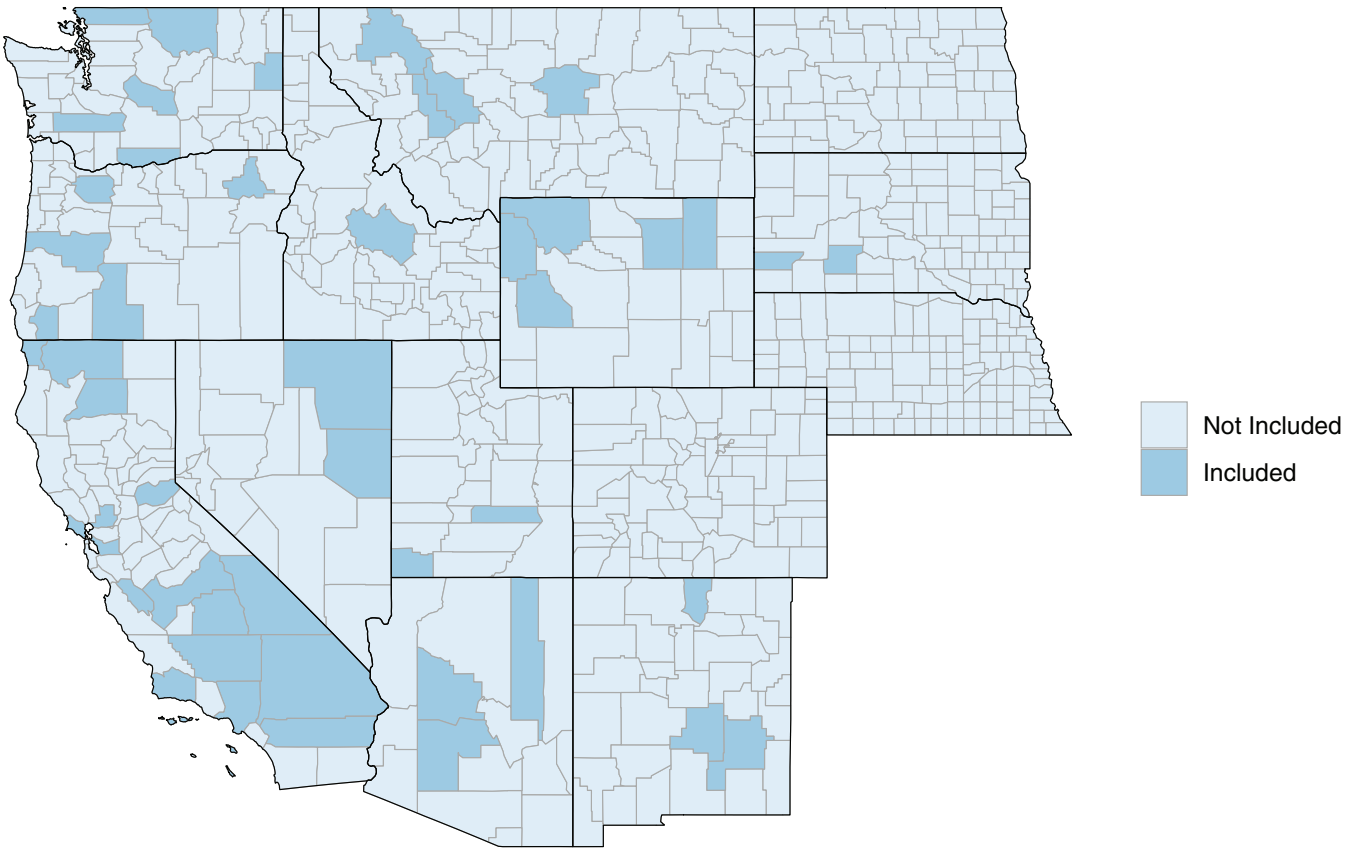

FIGURE 2 Western US counties included in data analysis, 2004 to 2009

from the GEOS-Chem model. The GEOS-Chem model was able to provide wildfire smoke exposure estimates for all subjects in the domain of the study. Further details of the validation of the model can be found in Liu et al.[6]

In our previous work, in order to estimate health effects associated with rare but extreme episodes of wildfire-specific $PM_{2.5}$ we introduced the concept of a "smoke wave" which allows us to capture periods with high concentration, sporadic, and short-lived characteristics of wildfire $PM_{2.5}$. We define a smoke wave as at least 2 consecutive days with daily calibrated wildfire-specific $PM_{2.5} > 20\ \mu g/m^3$ (near the 98th percentile of all county-days across all counties). This definition is based on daily wildfire-specific $PM_{2.5}$ levels above a designated threshold and the daily levels in all days in a smoke wave must exceed the threshold. Unlike the EPA monitoring data, the GEOS-Chem model provides daily data, including daily estimates of wildfire-specific $PM_{2.5}$. Therefore, we were able to identify smoke wave days throughout the study period regardless of the availability of data from the EPA monitoring network. While the GEOS-Chem model, as used here, gives us the proportion of wildfire-specific $PM_{2.5}$, it does not give us the specific chemical composition of $PM_{2.5}$ on each day. That is why we must use the SIR method to estimate the change in mixture from non-smoke wave to smoke wave days.

5 | RESULTS

In the first step, we conduct propensity score matching to on the non-smoke wave and smoke wave days to balance key covariates between the two types of days. We include daily temperature, dew point temperature, month, and year (to capture time trends), and an indicator of the county location (to capture approximate spatial variation). The data on wildfires cover months in the wildfire season of May to October, so the winter and spring seasons are already excluded. Supplementary Figure 2 shows the standardized difference in year, temperature, month, and dew point temperature between smoke wave and non-smoke wave days in the matched and unmatched datasets. In the unmatched dataset, it is clear that the smoke wave days generally have higher temperature (and dew point temperature) and tend to fall in later years and later months within the year. In the matched dataset, these differences are narrowed substantially.

One important feature of the resulting matched dataset that is the representation of the counties themselves is similar across smoke wave and non-smoke wave days. We would not want our comparison of the $PM_{2.5}$ mixture across smoke wave and non-smoke wave days to be confounded with a large-scale spatial contrast across, for example, northern and southern counties. Such a comparison would likely be confounded by other unobserved factors that vary with spatial location. The matching process ensures that we can preserve a similar configuration of counties between the smoke wave and non-smoke wave observations. Supplementary Figure 3 presents for each county the difference in the proportion of observations coming from that county between smoke wave and non-smoke wave days. In the unmatched dataset, we see a greater spread around the zero-line with some counties more under-represented and others over-represented. In the matched dataset, this spread narrows somewhat with fewer extremes of inbalance in the representation of counties.

The two parameters ω and ρ that characterize the correlation between the potential outcomes for $PM_{2.5}$ and the principal mixture score, respectively, cannot be estimated from the data. For choosing ρ we used $\rho = 0$ based on the results of our simulation study in Section 3. We used a value of $\omega = 0.2$ after extensive sensitivity analyses did not show much change in the estimates of risk with different combinations of values for this parameter (Supplementary Figures 4-6). The parameter b described in Section 2.1.1 controls how different the $PM_{2.5}$ potential outcomes can be in order to play a role in estimating the principal mixtures score. We used a value of $b = 0.75$ in the primary analysis and also tried values of $b = 0.4$ and $b = 1.5$. The smaller value of b increased the uncertainty around the health risk estimation relative to the larger value, but overall shape of the risk curve was the same for all values of b (see Supplementary Figures 7 and 8). Hence, it appeared that the choice of b largely affected the variance of the resulting estimate rather than the bias given that b effectively controls how many observations are used to estimate the principal mixture direction. We settled on using $b = 0.75$ to be somewhat more conservative with respect to the estimates of health risk.

Smoke waves significantly increase the overall levels of $PM_{2.5}$ in the air, as we might expect. Figure 3 shows boxplots of the log total $PM_{2.5}$ daily mass concentrations by non-smoke wave and smoke wave days in the matched dataset. From the figure, we can see that smoke wave days tend to experience about twice the levels of $PM_{2.5}$ as non-smoke wave days, although with large variation in the levels on both categories of days.

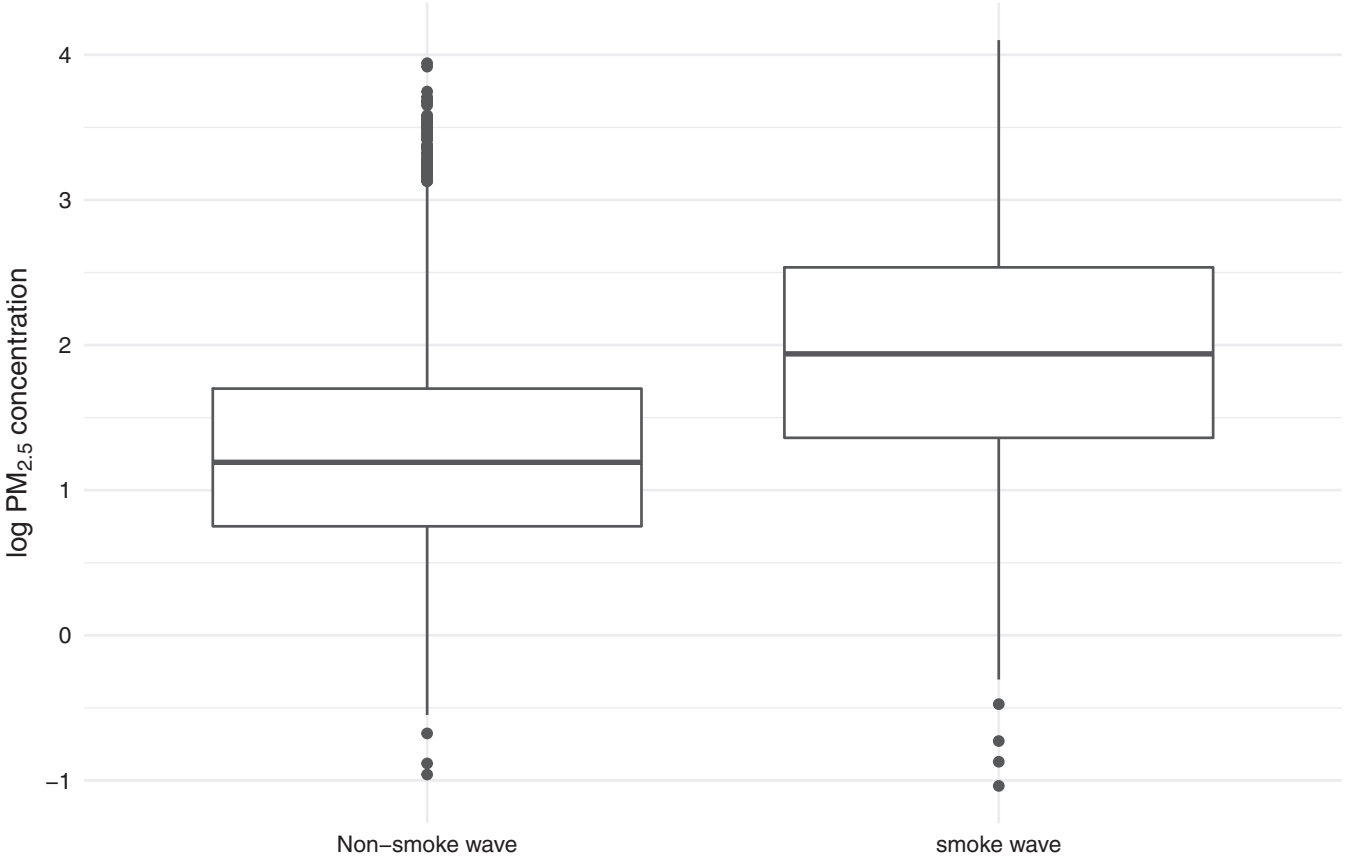

FIGURE 3 Log total $PM_{2.5}$ by non-smoke wave and smoke wave day

The predicted and observed potential outcomes for log total mass $PM_{2.5}$ on each day are shown in Figure 4. The tick marks on the x- and y-axes show the observed values of $PM_{2.5}$ on non-smoke wave and smoke wave days, respectively. The interior of the scatterplot shows the predicted value assigned to each observed value. The solid line is the $y = x$ line and the dashed lines represent $y = x \pm 0.75$, indicating the region where observations will carry the most weight in the weighted SIR described in Section 2.2. In this case, we have specified that $|PM_t(1) - PM_t(0)| < 0.75$ is considered a relatively small change.

Given the predicted potential outcomes for $PM_{2.5}$ on each day t, we can identify days where the predicted change in $PM_{2.5}$ concentration is small and weight those days more heavily when estimating the principal mixture direction with SIR. Similarly, we will downweight observations that are predicted to have a large change in $PM_{2.5}$ concentration.

5.1 | Principal mixture score estimation

The principal mixture direction found by running the weighted SIR procedure described in Section 2.2 is shown in Figure 5. Each vertical bar indicates the SIR coefficient corresponding to each of the 28 $PM_{2.5}$ chemical constituents, with a downward bar indicating a negative coefficient and an upward bar indicating a positive coefficient. We can see from Figure 5 that smoke waves appear to have relatively more selenium, vanadium, bromine, cand chromium, while having relatively less arsenic, copper, lead, rubidium, strontium, and zirconium.

The units of the mixture score do not take any special meaning except that positive values indicate a composition that is like a smoke wave and negative values indicate a composition that is less like a smoke wave. Nevertheless, we need to identify a range of variation that is perhaps typical in order to calibrate what is meant by a "change in mixtures score," that is, $x_t(1) - x_t(0)$, on a given day. Using the observed data, the average difference in mixture score between smoke wave and non-smoke wave days is 0.0027.

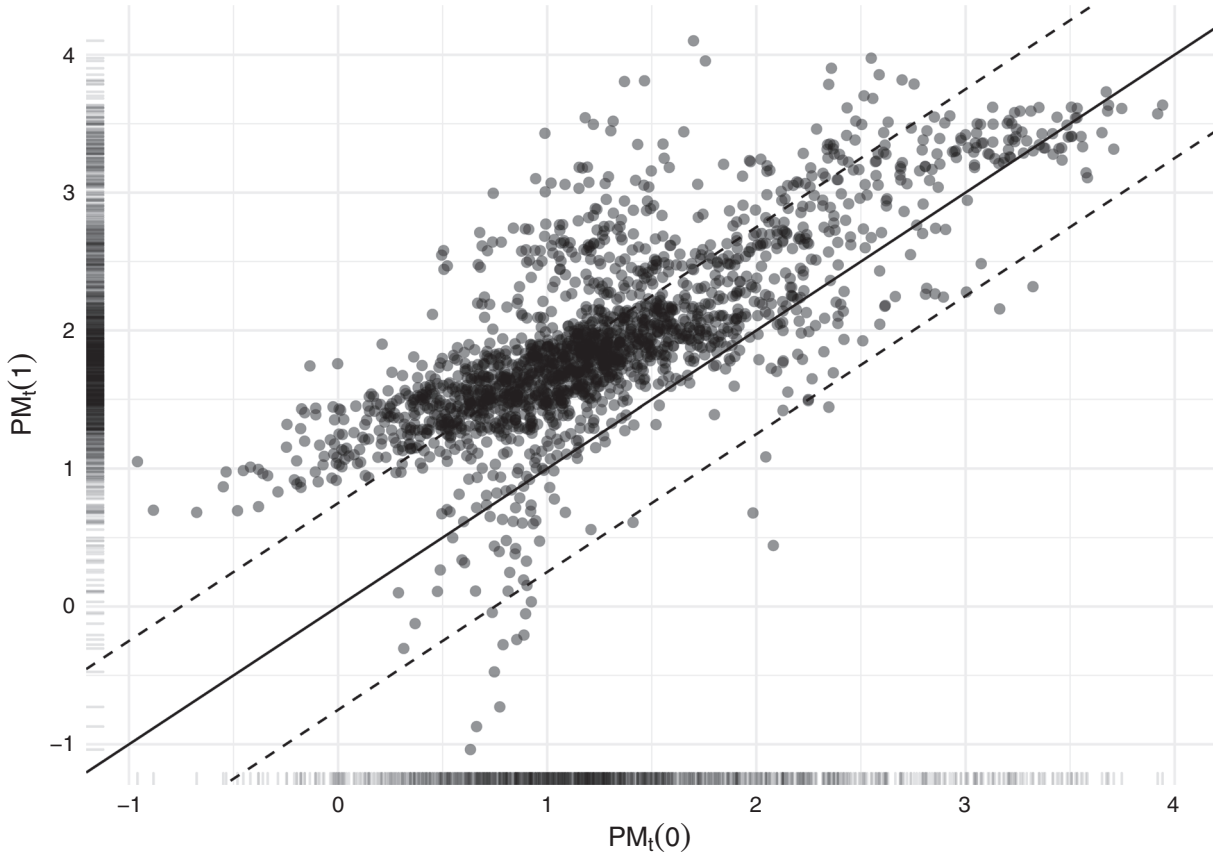

FIGURE 4 Predicted and observed potential outcomes for log $PM_{2.5}$ with ±0.75 bands

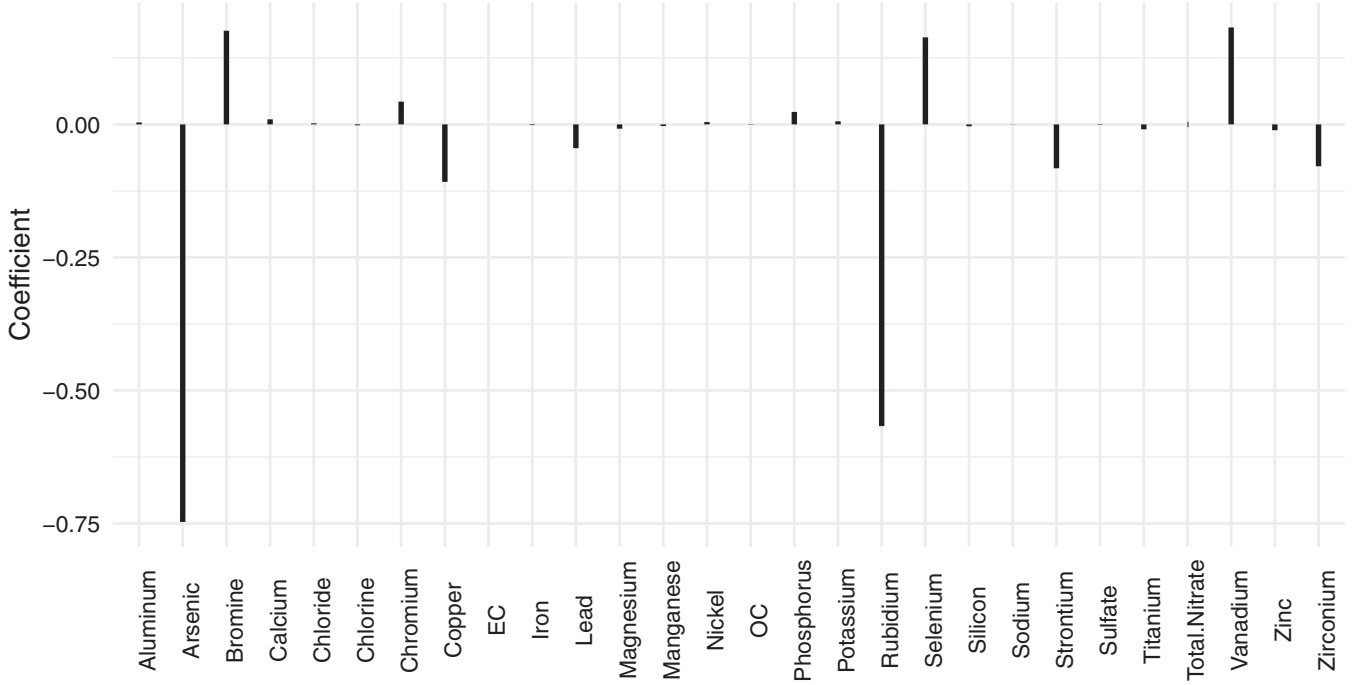

FIGURE 5 Principal mixture direction coefficients

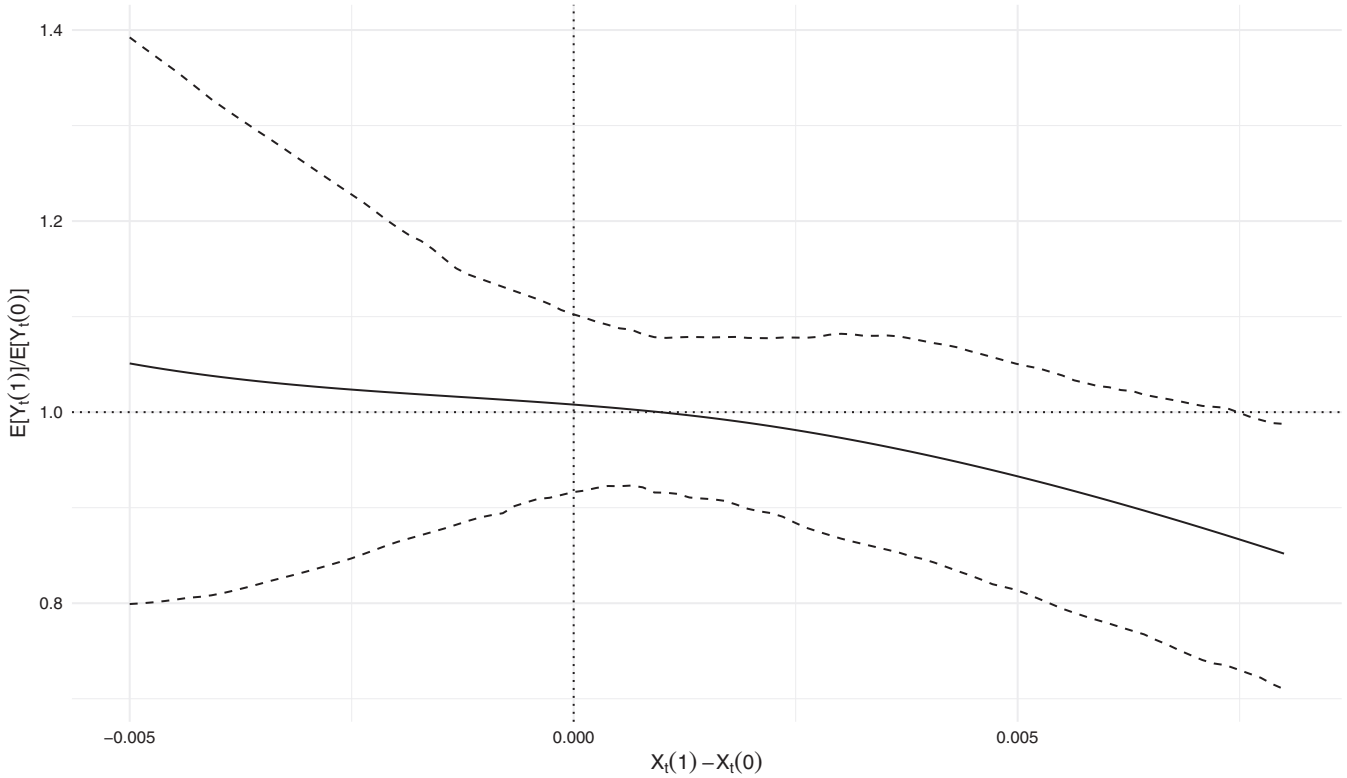

FIGURE 6 Relative risk of respiratory hospitalization by change in principal mixture score

5.2 | Health risk estimation

The estimated relative risk as a function of the difference in predicted mixture score $x_t(1) - x_t(0)$ is shown in Figure 6. In this figure, we used a 2-degree of freedom natural spline to model the relationship between the mixture score difference and the hospitalization relative risk to allow for flexibility in the relationship beyond linearity. It seems clear from Figure 6 that there is a decreasing trend in risk as the difference in mixture score goes from negative to positive. This trend suggests that as the chemical composition of $PM_{2.5}$ concentrations are changed to be more like a smoke wave, we observe a decrease in respiratory hospitalizations from what we might otherwise expect. On the other hand, as $PM_{2.5}$ composition becomes more like the typical mixture (non-smoke wave like), we see an increase in respiratory hospitalizations.

All things being equal (including, critically, the overall $PM_{2.5}$ level), Figure 6 suggests that the chemical composition of wildfire $PM_{2.5}$ is somewhat less harmful than $PM_{2.5}$ from non-wildfire sources in these communities. At a difference in mixture score of 0.0027, which was the average difference in mixture score between smoke wave and non-smoke wave days, we estimate a −2.2% (95% CI: −12.5%, 7.8%) change in respiratory hospitalizations. At a difference in mixtures score of 0.0076 we estimate a −13.6% (95% CI: −27.3%, −0.45%) change in respiratory hospitalizations. If we look at negative differences in the mixture score, meaning that the $PM_{2.5}$ mixture changes to be less like a smoke wave, we see that for a mixture difference of −0.005 we estimate a 5.1% (95% CI: −20%, 40%) change in respiratory hospitalizations. Across the range of the x-axis in Figure 6, we see that there is substantial uncertainty in the estimate of the relative risk associated with a change in composition. Nevertheless, the data suggest that while small changes in chemical composition of $PM_{2.5}$ are not strongly associated with changes in risk, larger changes in composition, particular toward a more "wildfire-like" composition, may substantially change the risk of hospitalization.

5.3 | Sensitivity analysis

We conducted a sensitivity analysis of the main results with respect to the specified parameters ρ, ω, and b and the resulting estimated curves are shown in the Supplementary Material. In general, we found that varying the values of ρ and ω

between 0.2 and 0.8 (Supplementary Figures 4-6) did not significantly alter the estimated mixture risk curve and did not lead to any different conclusions. Similarly, altering the value of b did not change the shape of the curve (Supplementary Figures 7 and 8) but a larger value of b appeared to decrease the uncertainty around the curve. This is perhaps expected as increasing the value of b widens the definition of similarity between potential outcomes $PM_t(1)$ and $PM_t(0)$ and therefore allows more observations to be used to estimate the principal mixture direction.

6 | DISCUSSION

The primary contribution of our work is the linkage of the principal mixture score, which characterizes the pollution mixture in a low-dimensional manner, with both the occurrence of smoke waves and respiratory hospitalizations. While the mixture score we propose shares properties with traditional dimension reduction approaches, the improved interpretability of the score comes from the direct linkage of the mixture score to the occurrence of a smoke wave. Therefore, our approach would suggest that if one wanted to modify the mixture score, one could, for example, dedicate more resources to wildfire suppression or prevention. Traditional dimension reduction approaches do not by default make such a direct linkage, so even if a change in the score were shown to be harmful or beneficial, there would be no obvious intervention designed to modify the score.

We found that wildfires in the Western US from 2004 to 2009 modified both the average level and chemical composition of $PM_{2.5}$. Figures 4 and 5 indicate the increase in $PM_{2.5}$ concentrations on smoke wave days and how the chemical composition of $PM_{2.5}$ changes between smoke wave and non-smoke wave days. The propensity score matching described in Section 2.1 was employed to account for potential confounding effects of temperature, dew point temperature, season, and spatial and temporal trends. In addition, the propensity score matching may provide some robustness to possible model misspecification.[16]

This analysis only examined a single lag of exposure, so that the exposure to a smoke wave and hospitalization were assumed to occur on the same day (lag 0). While previous work in air pollution and health has suggested that the effect of changes in air pollution could be spread out over multiple days,[17] especially for mortality outcomes, work focusing on hospitalization outcomes has largely shown that the strongest effects occur at lag 0.[18] Furthermore, limitations of the $PM_{2.5}$ monitoring network in the United States make conducting more sophisticated distributed lag analyses difficult without significant imputation of missing data.[19]

One potential extension of this methodology is to examine the change in $PM_{2.5}$ mixture in different regions of the Western US. We previously found that the composition of wildfire $PM_{2.5}$ can change between eco-regions[5] and the locations included in this study (Figure 2) span multiple eco-regions. Supplementary Figure 9 shows the average concentrations of the 28 $PM_{2.5}$ constituents on smoke wave and non-smoke wave days. We separate out California here as it represents a large eco-region within the study area. In general, there could be residual confounding due to the differences in $PM_{2.5}$ composition due to eco-region. While the propensity score matching might have mitigated this problem somewhat, it would be worth conducting an analysis on a larger dataset that could stratify by eco-region.

We found that the change in $PM_{2.5}$ mixture induced by wildfire smoke wave days appeared to be associated with fewer respiratory hospitalizations relative to the $PM_{2.5}$ mixture observed on non-smoke wave days. As the principal mixture score that we estimated increased to be more "smoke wave like" we found a decreasing trend in the relative rate of respiratory hospitalizations (Figure 6). There was, however, substantial uncertainty around the estimation of the risk function, suggesting caution should be used when interpreting these results. Nevertheless, the notion that $PM_{2.5}$ from biogenic sources is less harmful than $PM_{2.5}$ from other combustion sources has some precedence in the literature.[20]

It is worth reiterating here that because wildfires increase the overall concentration of $PM_{2.5}$ in the air, there is a positive association between total wildfire $PM_{2.5}$ and respiratory hospitalizations.[6] Our approach separates the issues of examining the increase in the total $PM_{2.5}$ mixture and the change in the composition. An alternative approach could be to look at the change in the percentage of each $PM_{2.5}$ constituent in the total mixture.[21] However, such a compositional approach combines the issue of change in level and change in mixture, which may be appropriate in some settings. In our setting, it is useful to separate these issues out given how dramatically ambient $PM_{2.5}$ can change with the occurrence of a wildfire. With our methodology, if a health effect is detected with a change in mixture score, we can have some confidence that it is not explained by a change in the overall level of pollution occurring in the background.

It is clear from the simulation study and the data analysis that there is significant uncertainty associated with the estimate of the effect of the mixture score on hospitalizations. From the simulation study, the variance appeared to dominate

the overall mean squared error. This is perhaps not surprising given the large amount of "missing data" (ie, potential outcomes) that must be imputed in order to estimate the difference $x_t(1) - x_t(0)$. Given that the imputation of these potential outcomes will always be necessary for this approach, it will likely remain a limitation of the approach.

Considering mixtures in the context of mediation analysis has also been discussed in the literature in conjunction with dimension reduction approaches like PCA or mixture models. Bellavia et al[22] review a variety of approaches in which components of a complex mixture might be incorporated into mediation models in order to estimate direct and indirect effects of environmental exposures. The approach presented here is perhaps closest their "two-stage" approach except in the first stage, we model the mixture based on the intervention rather than the outcome. Our work significantly builds on the ideas presented in Bellavia et al[22] by demonstrating that such approaches can be applied to large-scale population-level data to produce meaningful and interpretable results.

One limitation of the current approach stems from our dichotomization into smoke wave and non-smoke wave days, which constrains the SIR procedure to produce a one-dimensional mixture score. As a result, we could be limited in our ability to distinguish between days with different $PM_{2.5}$ mixtures but similar score values. One extension of our approach would be to treat the proportion of wildfire-specific $PM_{2.5}$ as a continuous measure, which would allow for the exploration of higher dimensions with the SIR method. Another limitation related to the use of SIR is the possibility that the change in $PM_{2.5}$ composition due to wildfires is better captured using higher order information, which the SIR method ignores.[8] It is possible that approaches such as principal Hessian directions or sliced average variance estimation, which use the covariance structure of the data, could better capture those changes.[9,23]

The statistical approach described here provides a way to explore the health effects of air pollution mixtures using epidemiological data. The advantage of such an approach is that it addresses the question of air pollution toxicity using data directly observed on the population of interest. The analysis described here looked at changes in air pollution composition over short periods of time. However, it may be possible to extend this approach to examine spatial contrasts, comparing different locations that may be exposed to different mixtures of pollutants. Such an extension may be explored in future work.

ACKNOWLEDGEMENTS

Drs. Peng, Liu, and Bell, were supported in part by the US Environmental Protection Agency (EPA) through award RD835871. This work has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the agency. EPA does not endorse any products or commercial services mentioned in this publication. Dr. Bell was supported by award ES021427 from the National Institutes of Health. Dr. Mickley was partially supported by award RD83587201 from the US EPA. This work was conducted while Dr. Liu was a postdoctoral fellow in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.

DATA AVAILABILITY STATEMENT

The air pollution, wildfire composition data, and analysis code are available on GitHub at https://github.com/rdpeng/ps_mixtures. The Medicare data used for the health analysis were obtained from the Centers for Medicare and Medicaid Services under a Data Use Agreement.

ORCID

Roger D. Peng https://orcid.org/0000-0003-0572-9055

REFERENCES

1. EPA. The benefits and costs of the clean air act from 1990 to 2020. U. S. Environmental Protection Agency Office of Air and Radiation; 2011.
2. Bell ML, Davis DL. Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environ Health Perspect*. 2001;109:389-394.
3. Ciocco A, Thompson DJ. A follow-up of Donora ten years after: methodology and findings. *Am J Public Health*. 1961;51:155-164.
4. National Research Council. *Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress*. Washington, D.C.: National Research Council of the National Academies; 2004.
5. Liu JC, Peng RD. The impact of wildfire smoke on compositions of fine particulate matter by ecoregion in the Western US. *J Expo Sci Environ Epidemiol*. 2019;29:765-776.
6. Liu JC, Wilson A, Mickley LJ, et al. Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. *Epidimiology*. 2017;28:77.

7. Liu JC, Wilson A, Mickley LJ, et al. Who among the elderly is most vulnerable to exposure to and health risks of fine particulate matter from wildfire smoke? *Am J Epidemiol*. 2017;186:730-735.
8. Li K-C. Sliced inverse regression for dimension reduction. *J Am Stat Assoc*. 1991;86:316-327.
9. Cook RD, Lee H. Dimension reduction in binary response regression. *J Am Stat Assoc*. 1999;94:1187-1200.
10. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21-29.
11. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42:1-28.
12. Peng RD, Schoenberg FP, Woods JA. A space-time conditional intensity model for evaluating a wildfire hazard index. *J Am Stat Assoc*. 2003;100:26-35.
13. Dominici F, Greenstone M, Sunstein CR. Particulate matter matters. *Science*. 2014;344:257-259.
14. Bey I, Jacob DJ, Yantosca RM, et al. Global modeling of tropospheric chemistry with assimilated meteorology: model description and evaluation. *J Geophys Res Atmos*. 2001;106:23073-23095.
15. Mu M, Randerson JT, Werf GR, et al. Daily and 3-hourly variability in global fire emissions and consequences for atmospheric model predictions of carbon monoxide. *J Geophys Res Atmos*. 2011;116:D24303.
16. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15:199-236.
17. Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology*. 2000;11:320-326.
18. Peng RD, Dominici F, Welty LJ. A Bayesian hierarchical distributed lag model for estimating the time course of hospitalization risk associated with particulate matter air pollution. *J R Stat Soc Ser C*. 2009;58:3-24.
19. Caffo B, Peng R, Dominici F, Louis TA, Zeger S. Parallel Bayesian MCMC imputation for multiple distributed lag models: a case study in environmental epidemiology. In: Brooks S, Gelman A, Jones GL, Meng X-L, eds. *Handbook of Markov Chain Monte Carlo*. New York, NY: Chapman and Hall/CRC; 2011:493-512.
20. Krall JR, Hackstadt AJ, Peng RD. A hierarchical modeling approach to estimate regional acute health effects of particulate matter sources. *Stat Med*. 2017;36:1461-1475.
21. Aitchison J. *The Statistical Analysis of Compositional Data*. London, UK: Chapman & Hall; 1986.
22. Bellavia A, James-Todd T, Williams PL. Approaches for incorporating environmental mixtures as mediators in mediation analysis. *Environ Int*. 2019;123:368-374.
23. Li K-C. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J Am Stat Assoc*. 1992;87:1025-1039.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Peng RD, Liu JC, McCormack MC, Mickley LJ, Bell ML. Estimating the health effects of environmental mixtures using principal stratification. *Statistics in Medicine*. 2022;41(10):1815-1828. doi: 10.1002/sim.9330