American Journal of Epidemiology, 2026, 195(8), 2253–2262
https://doi.org/10.1093/aje/kwag069
Published: March 26, 2026
Practice of Epidemiology

Two-way fixed effects models in air pollution epidemiology: a proposed framework for model specifications and recommendations for future studies

Yiqun Ma*[,1], Jordan Kaplan[2,3], David H. Rehkopf[3,4,5], Benjamin Q. Huynh[6], Mathew V. Kiang†[,3], Tarik Benmarhnia†[,1,7]

[1]Scripps Institution of Oceanography, University of California, San Diego , La Jolla, CA, United States
[2]Johns Hopkins University School of Medicine, Baltimore, MD, United States
[3]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, United States
[4]Department of Health Policy, Stanford University School of Medicine, Stanford, CA, United States
[5]Department of Medicine (Primary Care and Population Health), Stanford University School of Medicine, Stanford, CA, United States
[6]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[7]Irset Institut de Recherche en Santé, Environnement et Travail, UMR-S 1085, Inserm, University of Rennes, EHESP, Rennes, France

*Corresponding author: Yiqun Ma, Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (yim022@ucsd.edu)
†Mathew V. Kiang and Tarik Benmarhnia are co-senior authors.

Abstract

Two-way fixed effects (TWFE) models are increasingly applied in air pollution epidemiology. However, it remains unclear how sensitive TWFE models are to different modeling choices and how to transparently select a final model from a wide range of options. This study aims to systematically assess the sensitivity of TWFE models to different fixed effects (FEs) and propose a comprehensive decision-making framework. As an illustrative case study, we investigated associations between weekly ZIP Code Tabulation Area (ZCTA)-level mean particulate matter ($PM_{2.5}$) concentrations and respiratory hospitalizations in CA, 2011-2019. We proposed 52 TWFE models with all plausible spatiotemporal FE combinations and demonstrated a three-stage decision-making framework for final model selection based on permutation tests, equivalence testing, and model complexity. Substantial variations in estimates were observed among models with different combinations of FEs and their interactions. The final model estimated that a 1 $\mu g/m^3$ increase in weekly mean $PM_{2.5}$ concentration was associated with a 0.06% increase in weekly respiratory hospitalizations (95% CI, 0.03%, 0.09%; clustered SE by ZCTA). This study highlighted that the choice of FEs in TWFE models can have meaningful influences in the inference of interest. Our proposed decision-making offers a practical guide for future air pollution epidemiological studies using TWFE models.

Key words two-way fixed effects model, model specification, fine particulate matter, respiratory hospitalization

Introduction

Two-way fixed effects (TWFE) models, which are widely used in econometrics, have been increasingly applied in air pollution epidemiology in recent years.[1-4] This type of model is typically used to analyze observational time-series cross-sectional data (also known as panel data), which comprises observations of the same units (eg, states or counties) across multiple time points. TWFE models compare each unit to itself over time by including unit fixed effects (FEs), also referred to as spatial FEs, and adjust for secular trends common across units by including time FEs (eg, calendar months or years). This approach theoretically adjusts for both unit-level observable and unobservable time-invariant confounders, as well as spatially invariant time-varying confounders.[5-7]

The increasing popularity of TWFE models in air pollution epidemiology can be attributed to several key advantages. First, this model can be flexibly applied to study the health impacts of air pollution exposure using data with varying temporal resolutions; for daily data with multiple spatial units, however, case-crossover methods and two-stage time-series analyses are most commonly used in environmental epidemiology and perform well.[8-10] Notably, space–time-stratified case crossover methods is also based on FEs to control for spatial and temporal confounders.[8] Second, it offers a feasible alternative when some other study designs are challenging to implement due

Received: September 1, 2025. **Accepted:** March 20, 2026

to the spatial and temporal resolutions of datasets (eg, the monthly county-level mortality dataset from the US National Center for Health Statistics[11]). Although recent methodological advances allow analyses of daily exposures with weekly or monthly outcome data,[12] TWFE remains a practical option when the exposure data is also temporally aggregated and when the available time series is relatively short. Third, when appropriate FEs are included, these models adjust for unmeasured confounders that are fixed over time or space, without the need to collect detailed information on all potential confounders and adjust for each individually.

In previous studies that applied TWFE models to investigate the effects of air pollution exposure on mortality and other health outcomes (Table S1), commonly used spatial units include zip codes, census tracts, municipalities, and counties. Studies that utilized TWFE approaches relied on various choices in the selection and combination of FEs, particularly when the temporal resolution is finer than annual.[4,13] Such decisions can be reasonably made *a priori*, though they involve different assumptions and may lead to nontrivial differences in research findings. Differences in modeling assumptions in air pollution epidemiological studies using TWFE models illustrate what Gelman and Loken describe as the "garden of forking paths," where the flexibility in researcher decision-making can result in differing conclusions, even when reasonable decisions are made at each step.[14]

The term "fixed effects" has been used to mean different things in different contexts. In the context of meta-analyses, FEs assume that the true effect sizes are the same across all studies.[15] In the context of clustered or panel data, FEs (that are typically applied to the intercept) are routinely employed when random effects do not meet the assumption of independence between cluster-specific intercepts and variables of interest. Data can be clustered across both time and spatial units. By fitting a separate intercept for each cluster, FE models adjust for any shared characteristics among units within a cluster. Here, we refer to "fixed effects" in the context of panel data. Many previous epidemiological studies have referred to TWFE models as "a variant of difference-in-differences (DiD)" or used "TWFE" and "DiD" interchangeably. However, TWFE is a particular estimator that can be used in DiD settings as well as other panel-data contexts, while DiD is a study design defined by additional identification assumptions, such as the parallel-trends, and can be implemented with estimators other than TWFE.[16,17] Indeed, the canonical version of DiD is based on 2 FEs (eg, 1 for treated/untreated clusters and another for periods before/after the implementation of a policy or occurrence of a shock of interest) and the interaction term between these two FEs is considered to be the DiD estimator. This approach aims to capitalize on the quasi-random timing of a given intervention (often referred to as a natural experiment) and relies on additional identification assumptions.[18] Recent literature also showed how such canonical DiD approaches lead to biased estimates in multiple settings including staggered interventions, and alternative strategies have been proposed.[19-22] Therefore, using them interchangeably is misleading. To avoid ambiguity, we regard TWFE models as distinct from DiD approaches throughout this study.

The effects of fine particulate matter ($PM_{2.5}$) on respiratory mortality and morbidity are well documented.[23-28] Most prior studies have focused on either short-term $PM_{2.5}$ exposures, using time-series analyses or case-crossover designs with daily data,[23-26] or long-term exposures, using cohort studies with annual average $PM_{2.5}$ concentrations.[27,28] The health effects of $PM_{2.5}$ at a weekly or monthly time scale remain relatively understudied but may provide important insights into the etiology of health outcomes. TWFE constitutes appropriate analytical methods for handling data at this intermediate temporal resolution. In this paper, we utilized TWFE models to conduct a case study of weekly ZIP Code Tabulation Area (ZCTA)-level smoke $PM_{2.5}$ and respiratory hospitalizations in CA from 2011 to 2019. Our goals were to (1) systematically assess the sensitivity of TWFE models to the selection and combination of FEs and their interactions, and (2) propose a comprehensive framework for decision-making when specifying TWFE models in the context of air pollution epidemiology. This framework can be applied to exposures beyond air pollution, such as temperature or precipitation anomalies, as well as time scales other than weeks, such as months, quarters, or years.

Methods

Data sources

This analysis utilized weekly ZCTA-level respiratory hospitalization data from 2011 to 2019 for 1405 ZCTAs in CA with a population of at least 1000 people. The hospitalization data covers all unscheduled hospital and emergency department visits, obtained from the California Department of Health Care Access and Information.[29] Annual total population data for each ZCTA was extracted from American Community Survey 5-Year Data by US Census Bureau.

Daily ZCTA-level $PM_{2.5}$ concentrations in CA was estimated using an ensemble-based statistical approach.[30] The ensemble model showed good predictive performance, with a prediction R^2 of 0.83 for all sites. We calculated weekly mean $PM_{2.5}$ concentrations for each ZCTA to match the weekly hospitalization data.

Daily mean air temperature and total precipitation data were obtained from the Gridded Meteorological reanalysis product (gridMET), which is a high-resolution (4×4 km) gridded dataset of surface meteorological variables for the contiguous United States.[31] We extracted daily ZCTA-level meteorological data using the population-weighted centroid of each ZCTA, and then calculated weekly mean values.

TWFE models with different FE combinations

We applied TWFE models with quasi-Poisson regression to estimate the association between weekly mean $PM_{2.5}$ exposure and respiratory hospitalizations. We proposed 52 model specifications that were identical in all aspects except for the FE combinations (Models 1-52). A generic model can be expressed as

$$\begin{aligned}\log\left(E\left[Y_{z,w}\right]\right) = \beta_1 \mathbf{PM}_{z,w} + ns\left(T\mathbf{mean}_{z,w}, df = 5\right) \\ + \mathbf{offset}\left(\log\left[\mathbf{Population}_{z,y}\right]\right) + \mathbf{FE}_{\mathbf{space}} + \mathbf{FE}_{\mathbf{time}} \\ + \mathbf{FE}_{\mathbf{space-time}} \qquad (1)\end{aligned}$$

$Y_{z,w}$ represents the number of respiratory hospitalizations in ZCTA z during week w; $\mathbf{PM}_{z,w}$ is the mean $PM_{2.5}$ concentration in ZCTA z, week w. $T\mathbf{mean}_{z,w}$ is the mean air temperature and it was modeled by a natural cubic spline with 5 degrees of freedom (df). $\log\left(\mathbf{Population}_{z,y}\right)$ is an offset term, which presents the natural log of the population in ZCTA z, year y. $\mathbf{FE}_{\mathbf{space}}$ refers to FEs for space, which adjusts for characteristics that are unique to each spatial unit but do not change over time; $\mathbf{FE}_{\mathbf{time}}$ refers to FEs for time, which accounts for characteristics that are unique to each time unit but do not vary across space; and $\mathbf{FE}_{\mathbf{space-time}}$ refers to FEs for

space–time interactions, which accounts for unmeasured confounding that varies simultaneously across space and time. Note that not all three types of FEs need to be included (see the following paragraphs for details). We report the percent change in weekly respiratory hospitalizations associated with a 1 $\mu g/m^3$ increase in weekly mean $PM_{2.5}$ concentration, calculated by $(\exp(\beta_1) - 1) \times 100\%$.

Models 1-52 are identical in all other aspects, but with different combinations of the following FE terms: ZCTA (α_z), year (γ_y), month (γ_m), week (γ_w), year^month interaction ($\gamma_{m,y}$), ZCTA^year interaction ($\theta_{z,y}$), ZCTA^month interaction ($\theta_{z,m}$), ZCTA^year^month interaction ($\theta_{z,m,y}$), county^year interaction ($\theta_{c,y}$), county^month interaction ($\theta_{c,m}$), county^week interaction ($\theta_{c,w}$), and county^year^month interaction ($\theta_{c,m,y}$). In this study, we used the symbol "^" to denote the interaction between FEs. For example, ZCTA^year interaction accounts for ZCTA-specific year-to-year variations (eg, gradual healthcare improvements in certain ZCTAs).

We selected FE combinations based on the following principles. First, each specification includes both a spatial and a temporal component, which ensures that we adjust for both spatial and temporal confounding. Second, all models contain a ZCTA-level component, either as a main effect or as part of an interaction term, because it captures local spatial confounding at the finest available resolution. For temporal FEs, we avoid redundancy by not including coarser time units if week-level FEs (the finest temporal resolution) are already present. Additionally, in models with interaction FEs, we do not include the same spatial or temporal unit as a separate main effect, since it is already accounted for within the interaction. The proposed 52 models represent all possible FE combinations based on these principles.

We divided these FE combinations into 5 categories: (1) space + time (models 1-5), such as $\alpha_z + \gamma_y$; (2) space–time interaction (models 6-19), such as $\theta_{z,y} + \theta_{z,m}$; (3) space + space–time interaction (models 20-28), such as $\alpha_z + \theta_{c,y}$; (4) time + space–time interaction (models 29-40), such as $\gamma_w + \theta_{z,y}$; and (5) space + time + space–time interaction (models 41-52), such as $\alpha_z + \gamma_y + \theta_{c,m}$. A detailed list of all these model specifications can be found in Methods S1.

Permutation tests

To select a final model from the 52 specifications, as a preliminary step, we conducted spatial and temporal permutation tests for each model to assess the potential presence of residual confounders after adjusting for spatial and temporal confounding, respectively. In the spatial permutation tests, we randomly reassigned the $PM_{2.5}$ observations across ZCTAs 1000 times while keeping their original week allocation. For the temporal permutation test, we randomly shuffled the time order of the $PM_{2.5}$ observations 1000 times within each ZCTA. If the distribution of the 1000 model coefficients obtained using the permuted data is not centered at 0, this model does not pass the test. Whether the permutation distribution is centered at 0 can be assessed visually; we also supplemented a quantitative approach to complement visual inspection (see Methods S2 for details). The key idea behind such permutation test is to deliberately break the exposure–outcome linkage while preserving either the spatial or temporal structure of the exposure; any non–zero association observed after permutation therefore indicates remaining confounding from the preserved dimension. A detailed explanation of the permutation test, along with an example, can be found in Methods S3.

Final model selection and sensitivity analysis

Among the models which passed both the spatial and temporal permutation tests, we selected the final model based on a principle of parsimony. First, we sorted these models by complexity, which was measured by the number of FEs and their interactions in the model. The most complex model was used as the reference model because it accounts for all theoretically plausible confounding structures in the given setting.

Second, we calculated the difference in $PM_{2.5}$ coefficients between each simpler model and the reference model. To construct 95% empirical CIs (eCIs) for these differences, we applied a ZCTA-clustered bootstrap procedure with 1000 iterations. In each iteration, we resampled ZCTAs with replacement and retained all observations within each sampled ZCTA to preserve both temporal and spatial correlation structures. We then re-estimated both models using the bootstrapped data and calculated the difference in the $PM_{2.5}$ coefficients. The 95% eCIs were obtained from the 2.5th and 97.5th percentiles of the 1000 bootstrap differences.

Then, we applied equivalence testing to identify which model estimates were statistically equivalent to that from the reference model. In brief, an equivalence test evaluates whether the difference in estimates falls entirely within a pre-specified margin, indicating that any observed difference is small enough to be considered practically negligible.[32,33] Using the 95% eCIs from the ZCTA-clustered bootstrap, we determined equivalence if the entire interval lay within ±0.1 percentage point margin. Note that the equivalence margin is defined by researchers based on field-specific knowledge or practical considerations. Here, we *a priori* chose ±0.1 percentage point as the equivalence margin because a difference of 0.1 percentage point corresponds to a change in the third decimal place of the risk ratio, which we regard as practically negligible. Compared to traditional null-hypothesis significance tests, which assess whether there is "no evidence of difference" between two estimates, equivalence tests are better aligned with the goal of providing empirical "evidence of equivalence."[33]

Finally, among the models with estimates statistically homogeneous to the reference model, we selected the simplest specification as the final model. This approach balances the need for model accuracy with the desire for simplicity and interpretability. In sensitivity analysis, we checked the robustness of the results of the final model by using (1) alternative number of *df* (3, 4, or 6) for air temperature, (2) adjust for weekly total precipitation using a linear term or a natural cubic spline with 3 *df*, and (3) alternative combinations of FEs (models with equivalent estimates but were not chosen as the final model due to their higher complexity).

Uncertainty estimation

Unless otherwise stated, we reported ZCTA-clustered SEs, which assume that errors are correlated within each ZCTA. To assess how different SE calculation methods may influence uncertainty estimation, we also calculated SEs using 4 alternative methods for the final model: (1) independent and identically distributed (IID) SEs, which assume errors are homoskedastic and not correlated; (2) heteroskedasticity-robust SEs, which assume that errors are heteroskedastic; (3) SEs clustered by county, which assume that errors are correlated within each county; and (4) SEs clustered by both ZCTA and week, which assume that errors are correlated within each county and within each year. We compared the 95% CIs based on these

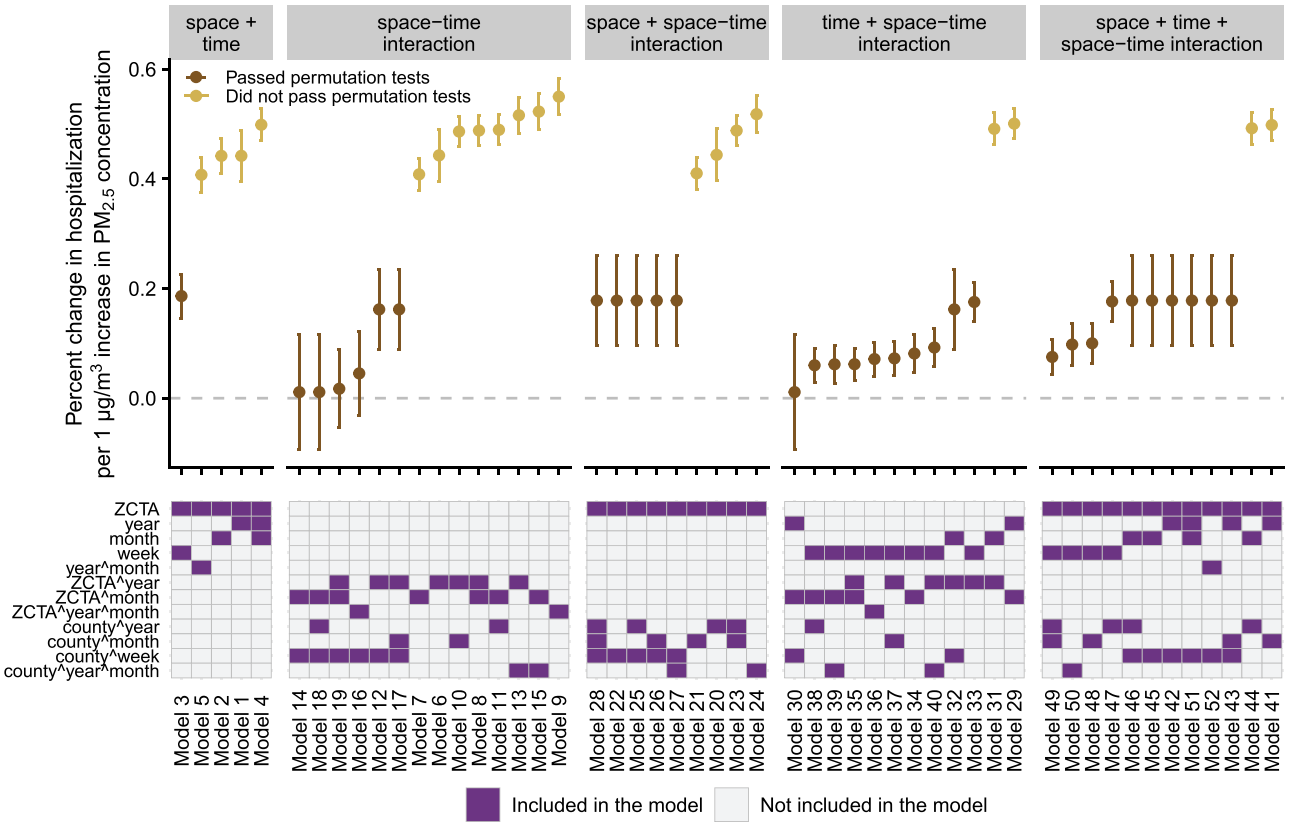

Figure 1 Percent change in weekly respiratory hospitalizations associated with a 1 μg/m^3 increase in weekly mean $PM_{2.5}$ concentration estimated by models with different combinations of fixed effects. The error bars indicate 95% CIs, computed by ZCTA-clustered SEs. Abbreviation: ZCTA, ZIP Code Tabulation Area.

SE calculations with the 95% eCIs from the aforementioned ZCTA-clustered bootstrap approach.

We conclude by proposing a decision-making framework for specifying TWFE models in air pollution epidemiology following the steps described above. This framework can be readily extended to analysis with other time units (see a supplementary case study with year as the time unit in Methods S4). All data analyses were conducted using R software (version 4.4.0; R Development Core Team). The *fixest* package was used for TWFE models. While other statistical packages, such as *gnm*,[34] can also be employed for TWFE models with quasi-Poisson regression, we selected *fixest* for its fast computation and its wide range of options for SE calculation. The final model estimates from these two packages were almost identical when IID SEs were used (Table S2). The R code to replicate these analyses is available at the following link: https://github.com/benmarhnia-lab/TWFE_framework.

Results

Description of main variables

This case study included 12 173 428 respiratory hospitalizations from 2011 to 2019 in CA. The distributions of weekly ZCTA-level total respiratory hospitalizations, mean $PM_{2.5}$ concentrations, mean air temperature, and total precipitation are summarized in Table S3. The weekly ZCTA-level total respiratory hospitalization counts ranged from 0 to 702, with a mean of 18 (SD: 22). The weekly ZCTA-level $PM_{2.5}$ concentrations ranged from 0.27 to 218.75 μg/m^3, with a mean of 9.71 μg/m^3 (SD: 5.56 μg/m^3). Figure S1 shows the spatial distribution of $PM_{2.5}$ concentrations in CA during the study period, with higher concentrations found in the San Joaquin Valley and coastal Southern CA, particularly in the Inland Empire metropolitan area. $PM_{2.5}$ concentrations varied substantially both within years and across years, with a clear seasonal pattern (Figure S2).

Models estimates and permutation test results

Figure 1 presents the estimated percent change in weekly respiratory hospitalizations associated with a 1 μg/m^3 increase in weekly mean $PM_{2.5}$ concentration, estimated by models with different combinations of FEs (Models 1-52). Substantial variation was observed across the estimates, ranging from 0.01% (95% CI, –0.09% to 0.12%) by Model 14 to 0.55% (95% CI, 0.52%-0.58%) by Model 9.

All these models passed the temporal permutation test, with the distribution of $PM_{2.5}$ coefficients from temporally permutated data centered at 0 (Figure S3). However, 20 of them did not pass the spatial permutation test, as the distribution of $PM_{2.5}$ coefficients from spatially permuted data deviated from 0, indicating that the estimates of these models were biased by residual confounding that these models did not account for (Figure S4). A common feature of the models that passed the spatial permutation test is the inclusion of a week-level component, the finest temporal unit available, either as a main effect or as part of an interaction term (Figure 1). Similarly, the fact that all proposed models passed the temporal permutation test is likely due to the inclusion of a ZCTA-level component, the smallest spatial unit in the data, in every specification.

Table 1 Selection of the final model among all models that passed permutation tests.

Model #	Fixed effects	# of FEs[a]	Estimates (95% CI)[b]	Equivalence[c]	Decision
Model 3	ZCTA + week	1874	0.19 (0.15, 0.23)	Not equivalent	Exclude
Model 47	ZCTA + week + county^year	2396	0.18 (0.14, 0.21)	Not equivalent	Exclude
Model 48	ZCTA + week + county^month	2570	0.10 (0.06, 0.14)	Not equivalent	Exclude
Model 49	ZCTA + week + county^year+county^month	3092	0.08 (0.04, 0.11)	Not equivalent	Exclude
Model 50	ZCTA + week + county^year^month	8064	0.10 (0.06, 0.14)	Not equivalent	Exclude
Model 33	Week + ZCTA^year	13 112	0.18 (0.14, 0.21)	Not equivalent	Exclude
Model 37	Week + ZCTA^year + county^month	13 808	0.07 (0.04, 0.10)	Not equivalent	Exclude
Model 34	Week + ZCTA^month	17 318	0.08 (0.05, 0.12)	Not equivalent	Exclude
Model 38	Week + county^year + ZCTA^month	17 840	0.06 (0.03, 0.09)	Statistically equivalent	Final model
Model 40	Week + ZCTA^year + county^year^month	19 302	0.09 (0.06, 0.13)	Not equivalent	Exclude
Model 39	Week + ZCTA^month + county^year^month	23 508	0.06 (0.03, 0.10)	Statistically equivalent	Sensitivity analysis
Model 22	ZCTA + county^week	27 953	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 42	ZCTA + year + county^week	27 962	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 45	ZCTA + month + county^week	27 965	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 51	ZCTA + year + month + county^week	27 974	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 52	ZCTA + year^month + county^week	28 061	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 25	ZCTA + county^year + county^week	28 475	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 46	ZCTA + month + county^year + county^week	28 487	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 26	ZCTA + county^month + county^week	28 649	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 43	ZCTA + year + county^month + county^week	28 658	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 28	ZCTA + county^week + county^year + county^month	29 171	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 35	Week + ZCTA^year + ZCTA^month	29 961	0.06 (0.03, 0.09)	Statistically equivalent	Sensitivity analysis
Model 27	ZCTA + county^week + county^year^month	34 143	0.18 (0.10, 0.26)	Not equivalent	Exclude
Model 12	ZCTA^year + county^week	39 191	0.16 (0.09, 0.23)	Not equivalent	Exclude
Model 32	Month + ZCTA^year + county^week	39 203	0.16 (0.09, 0.24)	Not equivalent	Exclude
Model 17	County^week + ZCTA^year + county^month	39 887	0.16 (0.09, 0.23)	Not equivalent	Exclude
Model 14	ZCTA^month + county^week	43 397	0.01 (–0.09, 0.12)	Not equivalent	Exclude
Model 30	Year + ZCTA^month + county^week	43 406	0.01 (–0.09, 0.12)	Not equivalent	Exclude
Model 18	County^week + county^year + ZCTA^month	43 919	0.01 (–0.09, 0.12)	Not equivalent	Exclude
Model 19	County^week + ZCTA^year + ZCTA^month	56 040	0.02 (–0.05, 0.09)	Statistically equivalent	Sensitivity analysis
Model 36	Week + ZCTA^year^month	149 817	0.07 (0.04, 0.10)	Not equivalent	Exclude
Model 16	ZCTA^year^month + county^week	175 896	0.05 (–0.03, 0.12)	Reference	Sensitivity analysis

Abbreviations: FE, fixed effect; ZCTA, ZIP Code Tabulation Area.
We used the symbol "^" to denote the interaction between fixed effects.
[a]Model complexity was measured by the number of fixed effects and their interactions in the model.
[b]Percent change in weekly respiratory hospitalizations associated with a 1 $\mu g/m^3$ increase in weekly mean $PM_{2.5}$ concentration estimated by each model. The 95% CIs were computed by ZCTA-clustered SEs.
[c]Results of the equivalence test with a margin of ±0.1 percentage point.

Final model selection

All 32 models that passed both temporal and spatial permutation tests are listed in Table 1, sorted by model complexity. Model 16 (FEs: ZCTA^year^month + county^week), which included 175 896 FEs, was the most complex specification and was selected as the reference model. Then, we calculated the difference in $PM_{2.5}$ association estimates between each simpler model and the reference, with 95% eCIs computed from ZCTA-clustered bootstraps (Figure 2). Based on equivalence tests with ±0.1 percentage point margin, we found that estimates from Model 38 (FEs: week + county^year + ZCTA^month), Model 39 (FEs: week + ZCTA^month + county^year^month), Model 35 (FEs: week + ZCTA^year + ZCTA^month), and Model 19 (FEs: county^week + ZCTA^year + ZCTA^month) were statistically equivalent to that from the reference model. Finally, Model 38 was selected as the final model because it is the one with the lowest complexity among models that passed the permutation test and with a statistically equivalent estimate compared to the most complex model (Table 1 and Figure 2). In sensitivity analysis, the results of the final model remain robust when we used alternative number of *df* for air temperature, adjusted for total precipitation, and used alternative combinations of FEs (Figure S5).

The final model (Model 38) estimated that a 1 $\mu g/m^3$ increase in weekly mean $PM_{2.5}$ concentration was associated with a 0.06% increase in weekly respiratory hospitalizations. The 95% CI was the narrowest when the SE was calculated based on the IID assumption (95% CI, 0.04%, 0.08%) and widest when the SE was clustered by both ZCTA and week (95% CI, −0.06%, 0.18%) (Figure 3). The 95% CIs were similar when using the ZCTA-clustered SE, heteroskedasticity-robust SE, and empirical intervals derived from ZCTA-clustered bootstraps (95% CI, 0.03%-0.09%).

Discussion

To the best of our knowledge, this study is the first systematic examination of the sensitivity of TWFE models to different combinations of FE specifications. Using the association between weekly ZCTA-level mean $PM_{2.5}$ concentrations and respiratory hospitalization in CA as a case study, we found that model estimates were highly sensitive to the combination of FEs and their interactions employed. We propose

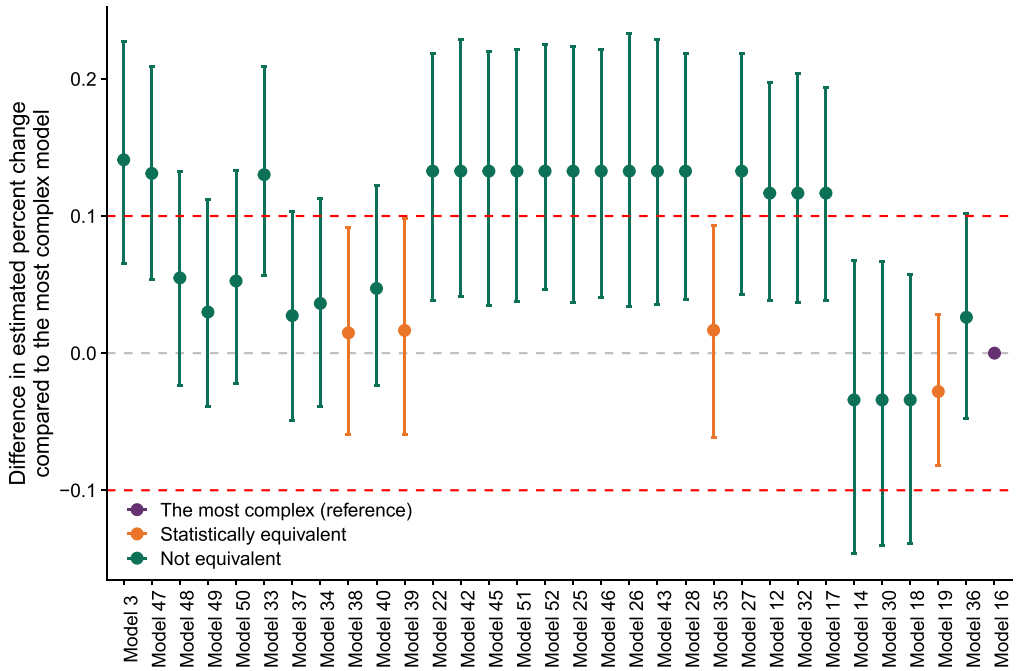

Figure 2 Difference in estimated percent change in weekly respiratory hospitalizations associated with a 1 μg/m^3 increase in weekly mean $PM_{2.5}$ concentration between other simpler models and the reference model. The error bars indicate 95% eCIs of the difference, computed by ZCTA-clustered bootstraps. The dashed lines indicate the equivalence margin used for the equivalence test. Abbreviations: eCIs, empirical CIs; ZCTA, ZIP Code Tabulation Area.

a comprehensive framework to guide decision-making in the specification of TWFE models, which has potential applications in future epidemiological studies across diverse settings.

Figure 4 illustrates our proposed decision-making framework for specifying TWFE models in the context of air pollution epidemiology. In this framework, we recommend a three-stage approach: in the first stage, researchers propose a set of candidate models based on data availability and *a priori* knowledge of the research question; the second stage focuses on selecting a main model from these candidates, guided by permutation tests and model complexity (following a parsimonious principle); and in the third stage, researchers conduct sensitivity analyses for the chosen main model and consider appropriate methods to estimate uncertainties.

In the first stage, it is recommended to propose candidate models by first considering measured time-varying unit-specific confounders and then listing plausible combination of FEs (Steps 1-3). As TWFE models can theoretically adjust for both unit-level time-invariant spatial confounders and spatially invariant time-varying confounders, the focus narrows to time-varying unit-specific confounders. In addition, it is necessary to account for potential non-linear health effects of these covariates.[35] This step heavily relies on researchers' *a priori* knowledge of the research question and should be guided by directed

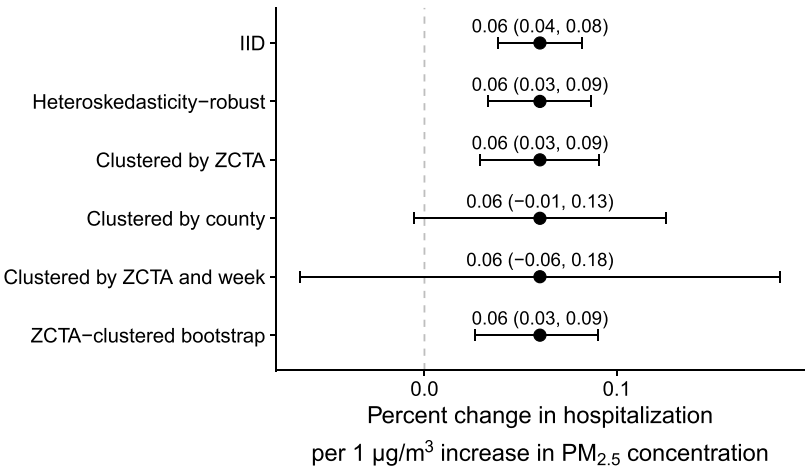

Figure 3 Percent change in weekly respiratory hospitalizations associated with a 1 μg/m^3 increase in weekly mean $PM_{2.5}$ concentration estimated by the final model with different uncertainty estimation approaches. The error bars indicate 95% CI or 95% eCIs. Abbreviation: eCIs, empirical CIs.

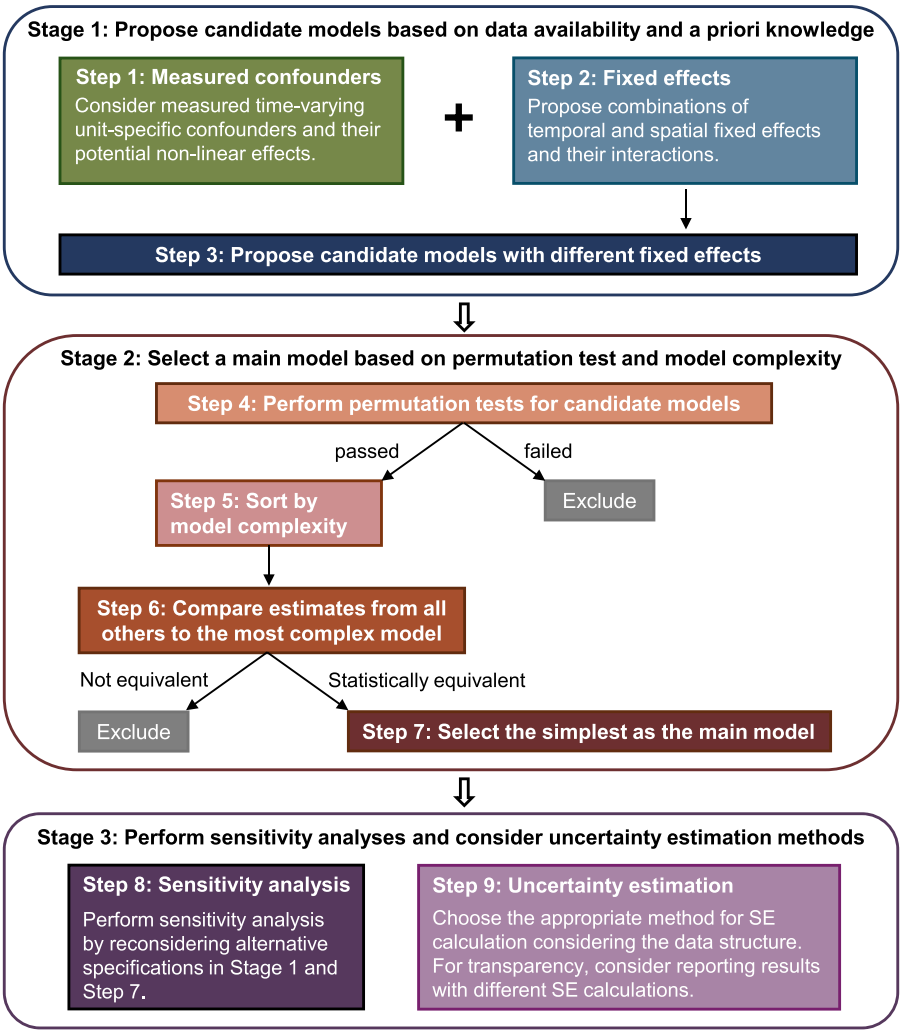

Figure 4 Proposed decision-making framework for specifying two-way fixed effects models in air pollution epidemiology.

acyclic graphs. The second step, selecting the combination of FEs, is critical in developing TWFE models. Based on our findings (Figure 1), it is recommended to include both the finest temporal and spatial unit available, either as a main effect or as part of an interaction term and list all plausible combinations as candidates.

The second stage of our framework provides a structured approach for selecting an optimal model from the candidate models proposed in stage 1 using transparent principles. We recommend first conducting spatial and temporal permutation tests on the candidate models (step 4). Such placebo tests are commonly used to detect spatial and temporal dependence due to model misspecification in panel models.[36,37] Candidate models that fail these tests can be directly excluded from further consideration. For ambiguous cases, we recommend retaining these models at this step because the primary goal here is to exclude clearly invalid models and thereby reduce the burden in subsequent steps, where stricter comparisons will be performed. Among the models that pass the permutation tests, we suggest sorting them based on model complexity (step 5). The estimates from each model are then compared to those from the most complex model (step 6). Finally, among the models with statistically equivalent estimates compared to the most complex model, we recommend selecting the simplest one as the main model based on a principle of parsimony (step 7).

In the final stage, we recommend conducting comprehensive sensitivity analyses to check the robustness of the main model's results. Alternative specifications initially considered in stage 1 and step 7 can be included in this step. In addition, it is important to choose an appropriate method for SE calculation that aligns with the specific data structure and sampling process of the study. Our analysis indicated substantial variations in uncertainty estimation when employing different approaches (Figure 3). As the IID assumption is typically violated in real-world panel data, where there may be serial correlation and heteroskedasticity, we recommend always using at least one type of SEs that are robust to certain violations of the IID assumption.[38] Heteroskedasticity-robust SE allows for heteroskedasticity of unknown form, and cluster-robust SE is a generalization of it for clustered data or repeated outcome measurements in longitudinal data.[38,39] Failure to consider within-cluster error correlation can lead to misleadingly small SEs, and consequent inaccurate narrow CIs.[40] In the context of air pollution epidemiology, such as the example shown in this study, clustering by the smallest geographic unit is usually strongly justified theoretically and empirically. Therefore, we suggest

starting from there and then empirically testing and/or reporting other SE estimates. In addition, bootstrapping serves as a nonparametric approach for uncertainty estimation. Previous papers have provided a guide to select the method for SE in the process of specifying and estimating a model, and a roadmap to check the robustness of the models.[39]

Our selected final model (Model 38) estimated that a 1 $\mu g/m^3$ increase in weekly mean $PM_{2.5}$ concentration was associated with an approximately 0.06% increase in weekly respiratory hospitalizations in CA. This finding generally aligns with previous literature, which consistently reported a positive relationship between short-term $PM_{2.5}$ exposure and respiratory outcomes in CA,[41-43] across the United States,[44-46] and globally.[25] However, the effect estimates are not directly comparable across studies due to the differences in temporal resolution as well as other factors, such as study population, geographic region, time period, exposure measurement, $PM_{2.5}$ composition, and concentration range. In addition, the selected final model is specific to this case study and this selection should not be universally applied. Each individual study needs to determine its own optimal model based on the proposed framework (Figure 4).

Several limitations of this study should be acknowledged. First, our study did not explore potential non-linear relationships between $PM_{2.5}$ and respiratory hospitalizations or consider multi-pollutant models, because the primary focus of this study is to examine the sensitivity of TWFE models to FE combinations and propose a decision-making framework, rather than comprehensively analyzing the air pollution–hospitalization relationship. However, our framework can, in principle, be extended to non-linear and multi-pollutant settings, and future studies are needed to address the complexities introduced by these extensions and to explore optimal procedures. Second, in the selection of the main model, we employed the principle of parsimony in Step 7 of our framework; however, alternative criteria, such as prioritizing models with narrower 95% CIs, are also reasonable and may be applied. We also did not fully discuss the selection of measured confounders or its contribution to uncertainty, as this issue is not unique to TWFE models and is beyond the scope of our study. Alternative approaches, such as "model averaging," which averages results across multiple candidate models,[47] could be considered in future studies to account for multiple plausible adjustment sets and FE combinations. In addition, we did not consider spatial structure and spatial autocorrelation in our data, which could be further explored in future work. Furthermore, in this paper, we aimed at assessing the sensitivity of TWFE models to various combinations of spatial and temporal FEs and their interactions. While this is a commonly used approach to deal with spatial and temporal confounding, we acknowledge that alternative approaches have been recently proposed to deal with spatial confounding.[48] It would be valuable to compare these recent methods to handle spatial confounding to spatial FEs. Finally, our study did not exhaustively test all possible methods for SE calculation. Other approaches, such as Driscoll-Kraay SE,[49] could be explored in future research.

In conclusion, this study found that TWFE models can be sensitive to the combination of FEs employed. To address this challenge, we propose a comprehensive framework, which is designed to guide researchers through the complex process of model specification, selection, and validation. We anticipate that this framework will serve as a practical guide for future studies and enhance the robustness, reliability, and reproducibility of findings from TWFE models across various domains of air pollution epidemiology.

Supplementary material

Supplementary material is available at the *American Journal of Epidemiology* online.

Conflicts of interest

The authors declare they have no conflicts of interest related to this work to disclose.

Funding

This work was supported by the National Institute on Aging (RF1AG080948).

Data availability

The datasets and R scripts used in this study are publicly available at https://github.com/benmarhnia-lab/TWFE_framework.

References

1. Wang Y, Kloog I, Coull BA, et al. Estimating causal effects of long-term $PM_{2.5}$ exposure on mortality in New Jersey. *Environ Health Perspect*. 2016;124(8):1182-1188. https://doi.org/10.1289/ehp.1409671
2. Renzi M, Forastiere F, Schwartz J, et al. Long-term PM_{10} exposure and cause-specific mortality in the Latium region (Italy): a difference-in-differences approach. *Environ Health Perspect*. 2019;127(6):067004. https://doi.org/10.1289/EHP3759
3. Yu P, Xu R, Li S, et al. Exposure to wildfire-related $PM_{2.5}$ and site-specific cancer mortality in Brazil from 2010 to 2016: a retrospective study. *PLoS Med*. 2022;19(9):e1004103. https://doi.org/10.1371/journal.pmed.1004103
4. Heft-Neal S, Gould CF, Childs ML, et al. Emergency department visits respond nonlinearly to wildfire smoke. *Proc Natl Acad Sci U S A*. 2023;120(39):e2302409120. https://doi.org/10.1073/pnas.2302409120
5. Cunningham S. *Causal inference: The mixtape*. Yale University Press; 2021.
6. Liu L, Wang Y, Xu Y. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *Am J Pol Sci*. 2022;68(1):160-176. https://doi.org/10.1111/ajps.12723
7. Strumpf EC, Harper S, Kaufman JS. Fixed effects and difference in differences. In: Oakes JM, Kaufman JS, eds. *Methods in Social Epidemiology*. 2nd ed. San Francisco, CA: Jossey-Bass; 2017:341-368.
8. Tobias A, Kim Y, Madaniyazi L. Time-stratified case-crossover studies for aggregated data in environmental epidemiology: a tutorial. *Int J Epidemiol*. 2024;53(2):dyae020. https://doi.org/10.1093/ije/dyae020
9. Berhane K, Thomas DC. A two-stage model for multiple time series data of counts. *Biostatistics*. 2002;3(1):21-32. https://doi.org/10.1093/biostatistics/3.1.21
10. Sera F, Gasparrini A. Extended two-stage designs for environmental research. *Environ Health*. 2022;21(1):41. https://doi.org/10.1186/s12940-022-00853-z
11. National Center for Health Statistics. *Restricted-Use Vital Statistics Data*, 2022. https://www.cdc.gov/nchs/nvss/nvss-restricted-data.htm
12. Basagaña X, Ballester J. Unbiased temperature-related mortality estimates using weekly and monthly health data: a new method for environmental epidemiology and climate impact studies. *Lancet*

Planet Health. 2024;8(10):e766-e777. https://doi.org/10.1016/S2542-5196(24)00212-2
13. Guo Y, Wu Y, Ye T, et al. A causal modelling framework for short-term effects of $PM_{(2.5)}$ on hospitalisations: a nationwide time series study in Brazil. *Environ Int*. 2023;171:107688. https://doi.org/10.1016/j.envint.2022.107688
14. Gelman A, Loken E. The statistical crisis in science. In: Pitici M, ed. *The Best Writing on Mathematics 2015*. Princeton, NJ: Princeton University Press; 2016:305-318.
15. Borenstein M, Hedges LV, Higgins JP, et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111. https://doi.org/10.1002/jrsm.12
16. Goin DE, Riddell CA. Comparing two-way fixed effects and new estimators for difference-in-differences: a simulation study and empirical example. *Epidemiology*. 2023;34(4):535-543. https://doi.org/10.1097/EDE.0000000000001611
17. Imai K, Kim IS. On the use of two-way fixed effects regression models for causal inference with panel data. *Polit Anal*. 2021;29(3):405-415. https://doi.org/10.1017/pan.2020.33
18. Nianogo RA, Benmarhnia T, O'Neill S. A comparison of quasi-experimental methods with data before and after an intervention: an introduction for epidemiologists and a simulation study. *Int J Epidemiol*. 2023;52(5):1522-1533. https://doi.org/10.1093/ije/dyad032
19. De Chaisemartin C, d'Haultfoeuille X. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *Econom J*. 2023;26(3):C1-C30. https://doi.org/10.1093/ectj/utac017
20. Roth J, Sant'Anna PH, Bilinski A, et al. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J Econom*. 2023;235(2):2218-2244. https://doi.org/10.1016/j.jeconom.2023.03.008
21. Callaway B, Sant'Anna PH. Difference-in-differences with multiple time periods. *J Econom*. 2021;225(2):200-230. https://doi.org/10.1016/j.jeconom.2020.12.001
22. Wang G, Hamad R, White JS. Advances in difference-in-differences methods for policy evaluation research. *Epidemiology*. 2024;35(5):628-637. https://doi.org/10.1097/EDE.0000000000001755
23. Yu W, Huang W, Gasparrini A, et al. Ambient fine particulate matter and daily mortality: a comparative analysis of observed and estimated exposure in 347 cities. *Int J Epidemiol*. 2024;53(3):dyae066. https://doi.org/10.1093/ije/dyae066
24. Liu C, Chen R, Sera F, et al. Ambient particulate air pollution and daily mortality in 652 cities. *N Engl J Med*. 2019;381(8):705-715. https://doi.org/10.1056/NEJMoa1817364
25. Atkinson R, Kang S, Anderson H, et al. Epidemiological time series studies of $PM_{2.5}$ and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*. 2014;69(7):660-665. https://doi.org/10.1136/thoraxjnl-2013-204492
26. Alessandrini ER, Stafoggia M, Faustini A, et al. Association between short-term exposure to $PM_{2.5}$ and PM_{10} and mortality in susceptible subgroups: a multisite case-crossover analysis of individual effect modifiers. *Am J Epidemiol*. 2016;184(10):744-754. https://doi.org/10.1093/aje/kww078
27. Pun VC, Kazemiparkouhi F, Manjourides J, et al. Long-term $PM_{2.5}$ exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *Am J Epidemiol*. 2017;186(8):961-969. https://doi.org/10.1093/aje/kwx166
28. Zang S-T, Wu QJ, Li XY, et al. Long-term $PM_{2.5}$ exposure and various health outcomes: an umbrella review of systematic reviews and meta-analyses of observational studies. *Sci Total Environ*. 2022;812:152381. https://doi.org/10.1016/j.scitotenv.2021.152381
29. State of California. California Department of Health Care Access and Information, 2024. https://hcai.ca.gov/
30. Aguilera R, Luo N, Basu R, et al. A novel ensemble-based statistical approach to estimate daily wildfire-specific $PM_{2.5}$ in California (2006-2020). *Environ Int*. 2023;171:107719. https://doi.org/10.1016/j.envint.2022.107719
31. Abatzoglou JT. Development of gridded surface meteorological data for ecological applications and modelling. *Int J Climatol*. 2013;33(1):121-131. https://doi.org/10.1002/joc.3413
32. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm*. 1987;15(6):657-680. https://doi.org/10.1007/BF01068419
33. Dixon PM, Saint-Maurice PF, Kim Y, et al. A primer on the use of equivalence testing for evaluating measurement agreement. *Med Sci Sports Exerc*. 2018;50(4):837-845. https://doi.org/10.1249/MSS.0000000000001481
34. Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Med Res Methodol*. 2014;14(1):1-6. https://doi.org/10.1186/1471-2288-14-122
35. Gasparrini A, Guo Y, Hashizume M, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *Lancet*. 2015;386(9991):369-375. https://doi.org/10.1016/S0140-6736(14)62114-0
36. Lam PH, Zang E, Chen D, et al. Long-term exposure to fine particulate matter and academic performance among children in North Carolina. *JAMA Netw Open*. 2023;6(10):e2340928. https://doi.org/10.1001/jamanetworkopen.2023.40928
37. Benshaul-Tolonen A. Local industrial shocks and infant mortality. *Econ J*. 2019;129(620):1561-1592. https://doi.org/10.1111/ecoj.12625
38. Mansournia MA, Nazemipour M, Naimi AI, et al. Reflection on modern methods: demystifying robust standard errors for epidemiologists. *Int J Epidemiol*. 2021;50(1):346-351. https://doi.org/10.1093/ije/dyaa260
39. MacKinnon JG, Nielsen MØ, Webb MD. Cluster-robust inference: a guide to empirical practice. *J Econom*. 2023;232(2):272-299. https://doi.org/10.1016/j.jeconom.2022.04.001
40. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *J Hum Resour*. 2015;50(2):317-372. https://doi.org/10.3368/jhr.50.2.317
41. Ostro B, Roth L, Malig B, et al. The effects of fine particle components on respiratory hospital admissions in children. *Environ Health Perspect*. 2009;117(3):475-480. https://doi.org/10.1289/ehp.11848
42. Stowell JD, Wing IS, Romitti Y, et al. Emergency department visits in California associated with wildfire $PM_{2.5}$: differing risk across individuals and communities. *Environ Res Health*. 2024;3(1):015002. https://doi.org/10.1088/2752-5309/ad976d
43. Do V, Chen C, Benmarhnia T, et al. Spatial heterogeneity of the respiratory health impacts of wildfire smoke $PM_{2.5}$ in California. *Geohealth*. 2024;8(4):e2023GH000997. https://doi.org/10.1029/2023GH000997
44. Dominici F, Peng RD, Bell ML, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*. 2006;295(10):1127-1134. https://doi.org/10.1001/jama.295.10.1127
45. Bell ML, Ebisu K, Peng RD, et al. Seasonal and regional short-term effects of fine particles on hospital admissions in 202 US counties, 1999–2005. *Am J Epidemiol*. 2008;168(11):1301-1310. https://doi.org/10.1093/aje/kwn252

46. Sun Y, Milando CW, Spangler KR, et al. Short term exposure to low level ambient fine particulate matter and natural cause, cardiovascular, and respiratory morbidity among US adults with health insurance: case time series study. *BMJ*. 2024;384:e076322. https://doi.org/10.1136/bmj-2023-076322
47. Hamra GB, Kaufman JS, Vahratian A. Model averaging for improving inference from causal diagrams. *Int J Environ Res Public Health*. 2015;12(8):9391-9407. https://doi.org/10.3390/ijerph120809391
48. Akbari K, Winter S, Tomko M. Spatial causality: a systematic review on spatial causal inference. *Geogr Anal*. 2023;55(1):56-89. https://doi.org/10.1111/gean.12312
49. Driscoll JC, Kraay AC. Consistent covariance matrix estimation with spatially dependent panel data. *Rev Econ Stat*. 1998;80(4):549-560. https://doi.org/10.1162/003465398557825

American Journal of Epidemiology, 2026, **195**, 2253–2262
https://doi.org/10.1093/aje/kwag069
Practice of Epidemiology