Supplementary Appendix

for

Short-term exposure to fine particulate matter and hospitalization risks and costs in the Medicare population: time-stratified case-crossover study

Figure S1 (Shiny app available at https://nsaph.shinyapps.io/cause_specific_viz/). Descriptive statistics for 214 disease groups during 2000-2012 among Medicare fee-for-service beneficiaries in the United States. (a) Total number of hospital admissions, decomposed by discharge destinations (deaths at discharge, discharges to skilled nursing facilities, discharges to home healthcare services, and other discharge destinations). (b) Distribution of days of hospitalization across all hospital admission records for each disease group (bars represent means; dots represent medians; and error bars represent 25^{th} and 75^{th} percentiles). (c) Distribution of inpatient cost (in 2017 USD) across all hospital admission records for each disease group (bars represent means; dots represent medians; and error bars represent 25^{th} and 75^{th} percentiles). Disease groups are ranked from highest to lowest total number of hospital admissions.

Figure S2 (Shiny app available at https://nsaph.shinyapps.io/cause_specific_viz/). Absolute and relative increase in risk of hospitalization associated with each 1 $\mu g \cdot m^{-3}$ increase in lag 0–1 $PM_{2.5}$ for each of the 214 disease groups. The absolute increases and relative increases for each of the 214 disease groups are presented with nine model specifications: (1) main model used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% confidence interval (CI); (2) below-guideline model used the same model specification as the main model and was restricted to days with daily $PM_{2.5}$ concentration $\leqslant 25\ \mu g \cdot m^{-3}$ (the current WHO air quality guideline for daily $PM_{2.5}$), with error bar representing Bonferroni-corrected 95% CI; (3) sensitivity model 1 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-4 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (4) sensitivity model 2 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-5 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (5) sensitivity model 3 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-6 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (6) single lag 0 model used lag 0 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (7) single lag 1 model used lag 1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95%

CI; (8) single lag 2 model used lag 2 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; and (9) FDR-corrected model used the same model specification as the main model, with error bar representing false discovery rate-corrected 95% CI.

Figure S3 (Shiny app available at https://nsaph.shinyapps.io/cause_specific_viz/). Absolute and relative increase in risk of hospitalization associated with each 1 $\mu g \cdot m^{-3}$ increase in lag 0–1 $PM_{2.5}$ for each of the 122 broader disease groups. The absolute increases and relative increases for each of the 122 broader disease groups are presented with nine model specifications: (1) main model used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% confidence interval (CI); (2) below-guideline model used the same model specification as the main model and was restricted to days with daily $PM_{2.5}$ concentration ≤ 25 $\mu g \cdot m^{-3}$ (the current WHO air quality guideline for daily $PM_{2.5}$), with error bar representing Bonferroni-corrected 95% CI; (3) sensitivity model 1 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-4 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (4) sensitivity model 2 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-5 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (5) sensitivity model 3 used lag 0-1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-6 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (6) single lag 0 model used lag 0 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (7) single lag 1 model used lag 1 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; (8) single lag 2 model used lag 2 $PM_{2.5}$ as the exposure and adjusted for penalized splines of lag 0-1 air and dew point temperatures, with error bar representing Bonferroni-corrected 95% CI; and (9) FDR-corrected model used the same model specification as the main model, with error bar representing false discovery rate-corrected 95% CI.

Section 1. Predictors for daily $PM_{2.5}$ prediction model.

Multiple sources of predictors were fused by the $PM_{2.5}$ prediction model, including 1 km × 1 km aerosol optical depth (AOD) and normalized vegetation index data (NDVI) measuring vegetation coverage both retrieved from the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua and Terra satellites,[1] predicted $PM_{2.5}$ from a chemical transport (CTM) model (GEOS-Chem), meteorological data [daily air temperature, relative humidity, wind speed, and height of planetary boundary layer from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis Project at approximately 32 km × 32 km spatial resolution[2]], and land use data (e.g., distance to major roads, emission, and land use pattern).

Section 2. Methodology and computation for large conditional logistic regressions with penalized splines.

The conditional logistic regressions used in the present study adjusted for penalized cubic splines for air and dew point temperatures. The reasons for choosing penalized splines over regression splines were: 1) regression splines are known to overfit the data and give wiggly curve estimates;[3] and 2) using regression splines often requires additional sensitivity analyses to explore sensitivity of the coefficient of interest with respect to degrees of freedom of the splines. Penalized splines imposed smoothness penalties to suppress the wiggly behavior of splines, with optimal smoothing parameters and thereby optimal degrees of freedom for the air and dew point temperatures splines selected by maximizing the marginal likelihood.[4] This is a data-driven approach, avoiding manual selection of degrees of freedom for the adjustment.

Due to the large number of inpatient claims, a two-stage fast computation strategy was used to make fitting the large conditional logistic regressions with penalized splines computationally feasible. We denoted the logarithm of the conditional likelihood for conditional logistic regression by $l\big(\mu_1(\boldsymbol{\beta}),\cdots,\mu_n(\boldsymbol{\beta})\big)$, which depends on each observation $i=1,\cdots,n$ via linear predictors $\mu_i(\boldsymbol{\beta})$, where

$$\mu_i(\boldsymbol{\beta}) = \beta_1 X_{i,1} + \beta_T{}' \boldsymbol{X_{i,T}} + \beta_D{}' \boldsymbol{X_{i,D}},$$

$X_{i,1}$is lag 0-1 $PM_{2.5}$, $\boldsymbol{X_{i,T}}$ is the vector of the cubic spline expansion of lag 0–1 air temperature, and $\boldsymbol{X_{i,D}}$ is the vector of the cubic spline expansion of lag 0–1 dew point temperature.[5] Each spline expansion has a maximum of nine degrees of freedom and is centered at zero.

The penalized spline estimator is traditionally defined by

$$\widehat{\boldsymbol{\beta}}_{\text{Trad}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\Big[-2l\big(\mu_1(\boldsymbol{\beta}),\cdots,\mu_n(\boldsymbol{\beta})\big) + \boldsymbol{\beta}'\boldsymbol{S_\lambda}\boldsymbol{\beta}\Big],$$

where $\boldsymbol{S_\lambda}$ is a block diagonal penalization matrix $\text{diag}(0,\lambda_1\boldsymbol{S_T},\lambda_2\boldsymbol{S_D})$ parametrized by smoothing parameters λ_1 and λ_2 and $\boldsymbol{S_T}$ and $\boldsymbol{S_D}$ are known semi-definite matrices, with $\boldsymbol{\beta}'\boldsymbol{S_\lambda}\boldsymbol{\beta}$ measuring total wiggliness of the two splines.

From a Bayesian perspective, the penalized spline estimator $\widehat{\boldsymbol{\beta}}_{\text{Trad}}$ is the mode of the posterior distribution based on a likelihood of $f(\boldsymbol{\beta}|D) \propto \exp\big[l\big(\mu_1(\boldsymbol{\beta}),\cdots,\mu_n(\boldsymbol{\beta})\big)\big]$ and a prior $f_{p,\boldsymbol{\lambda}}(\boldsymbol{\beta}) \sim N(0,\boldsymbol{S_\lambda^-})$, where $\boldsymbol{S_\lambda^-}$ is the pseudo inverse of $\boldsymbol{S_\lambda}$. The optimal smoothing parameter $\boldsymbol{\lambda} = (\lambda_1,\lambda_2)'$ can be selected by maximizing the Laplace approximated marginal likelihood (LAML) as described in Wood et al. (2016),[4] which applies a Laplace approximation to the marginal likelihood,

$$\hat{\boldsymbol{\lambda}}_{\text{Trad}} = \underset{\boldsymbol{\lambda}}{\text{argmax}} \int f_{p,\boldsymbol{\lambda}}(\boldsymbol{\beta}) f(\boldsymbol{\beta}|D)\, d\boldsymbol{\beta}\,.$$

Previous work has demonstrated advantages of maximizing LAML over optimizing generalized cross-validation or Akaike information criterion.[3]

The gam function in the mgcv R package implements an algorithm to estimate $\hat{\boldsymbol{\lambda}}_{\text{Trad}}$ for conditional logistic regressions. However, we found that the fitting procedure was extremely slow for disease groups with a large sample size. To achieve faster computations, we used a novel two-stage approach. When the sample size is large and the dimension of the spline basis is fixed, the unpenalized coefficient $\widetilde{\boldsymbol{\beta}}$ is approximately normal according to the central limit theorem. That is, $\widetilde{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta},\widetilde{\boldsymbol{V}}\big)$, where $\widetilde{\boldsymbol{V}}$ is the variance-covariance estimator for $\widetilde{\boldsymbol{\beta}}$. Locally, $f(\boldsymbol{\beta}|D) \approx \tilde{f}(\boldsymbol{\beta}|D) \propto \exp\Big[-\frac{1}{2}\big(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\big)'\,\widetilde{\boldsymbol{V}}^{-1}\big(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\big)\Big]$. In the two-stage approach, Stage 1 fits an unpenalized conditional logistic regression and Stage 2 fits a penalized Gaussian regression, as if we observed a single sample of $\widetilde{\boldsymbol{\beta}}$ following a Gaussian distribution with its variance-covariance being $\widetilde{\boldsymbol{V}}$, which solves

$$\widehat{\boldsymbol{\beta}}_{\text{two-stage}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})' \, \widetilde{\boldsymbol{V}}^{-1} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \boldsymbol{\beta}' \, \boldsymbol{S}_{\boldsymbol{\lambda}} \boldsymbol{\beta} \right],$$

with the optimal smoothing parameters selected by

$$\widehat{\boldsymbol{\lambda}}_{\text{two-stage}} = \underset{\boldsymbol{\lambda}}{\text{argmax}} \int f_{p,\boldsymbol{\lambda}}(\boldsymbol{\beta}) \tilde{f}(\boldsymbol{\beta}|D) d\boldsymbol{\beta} \, .$$

Stage 1 is an unpenalized conditional logistic regression, which can be fitted efficiently using the coxph function in R. Stage 2 is an optimization based on a sample whose size equals the column dimension of the design matrix, while by comparison the original problem is based on a sample of size *n*, which is substantially larger than the column dimension of the design matrix. Fitting an unpenalized conditional logistic regression in Stage 1 is substantially faster than its penalized counterpart and further Stage 2 can be completed within a trivial amount of time. As a result, we achieved a substantially faster computation for conditional logistic regressions with penalized splines using the two-stage framework.

Section 3. Calculation of absolute increase in risk of hospitalization associated with each 1 $\mu g \cdot m^{-3}$ increase in lag 0–1 $PM_{2.5}$.

To calculate the absolute increase in risk of hospitalization associated with each 1 $\mu g \cdot m^{-3}$ increase in lag 0–1 $PM_{2.5}$, we first calculated the cause-specific baseline rate of hospitalization, denoted by α_c, using the total number of cause-specific admissions during 2000–2012 among Medicare fee-for-service beneficiaries in the United States divided by the total person-days for fee-for-service beneficiaries. The total person-days for fee-for-service beneficiaries was 132 billion, which was the total number of persons at risk on each day during 2000–2012.

Then we calculated the absolute increase in risk of hospitalization associated with each 1 $\mu g \cdot m^{-3}$ increase in lag 0–1 $PM_{2.5}$ using $\alpha_c\left(1 - \exp\left(-\hat{\beta}_c\right)\right)$ with a standard error calculation of $\alpha_c \exp\left(-\hat{\beta}_c\right) \widehat{se}_c$, where α_c is the cause-specific baseline rate of hospital admissions, and $\hat{\beta}_c$ and $\widehat{se}_c$ are the coefficient and standard error, respectively, for lag 0–1 $PM_{2.5}$ extracted from the conditional logistic regression model for disease group *c*. This approach is consistent with Di et al. (2017).[6]

Section 4. Calculation of cause-specific annual increases in number of deaths at discharge, number of discharges to skilled nursing facilities, number of discharges to home healthcare services, and number of other discharge destinations associated with each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$.

For each disease group *c*, we denoted by $\hat{p}_c$ the estimated percent increase in risk of hospitalization associated with each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$. We estimated the annual increase in number of hospitalizations associated with each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$ as $\hat{p}_c N_c$, where N_c is annual averaged number of hospital admissions.

For each type of discharge destinations (denoted by *m*: 1 stands for death at discharge, 2 for discharge to skilled nursing facilities, 3 for discharge to home healthcare services, and 4 for other discharge destinations) of disease group *c*, we created a case-crossover dataset to estimate percent increase in risk of discharge destination *m* conditional on disease group *c* ($\hat{p}_{m|c}$) for each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$.

Then for each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$, we estimated the annual increase of:

1) number of deaths at discharge as $\hat{p}_c \hat{p}_{m=1|c} D_c$ if there was significant association between short-term $PM_{2.5}$ exposure and death at discharge, or as $\hat{p}_c D_c$ if there was no evidence of association between short-term $PM_{2.5}$ exposure and death at discharge, where D_c is annual average number of deaths at discharge;

2) number of discharges to skilled nursing facilities as $\hat{p}_c \hat{p}_{m=2|c} S_c$ if there was significant association between short-term $PM_{2.5}$ exposure and discharge to skilled nursing facilities, or as $\hat{p}_c S_c$ if there was no evidence of association between short-term $PM_{2.5}$ exposure and discharge to skilled nursing facilities, where S_c is annual average number of discharges to skilled nursing homes;

3) number of discharges to home healthcare services as $\hat{p}_c \hat{p}_{m=3|c} H_c$ if there was significant association between short-term $PM_{2.5}$ exposure and discharge to home healthcare services, or as $\hat{p}_c H_c$ if there was no evidence of association between short-term $PM_{2.5}$ exposure and discharge to home healthcare services, where H_c is annual average number of discharges to home healthcare services; and

4) number of other discharge destinations as the difference of the annual increase in number of hospitalizations ($\hat{p}_c N_c$) and the sum of the three quantities above – i.e., number of deaths at discharge, number of discharges to skilled nursing facilities, and number of discharges to home healthcare services.

Section 5. Calculation of cumulative annual increases in number of hospitalizations, days hospitalized, and healthcare costs (inpatient and post-acute) associated with each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$.

Among the disease groups that were found to be statistically significantly associated with short-term $PM_{2.5}$ exposure, increase in number of hospitalizations, days hospitalized, and healthcare costs (inpatient and post-acute) were cumulated for the set of newly identified disease groups and the set of disease groups identified from previous studies.

For each disease group c, we denoted by $\hat{\beta}_c$ the coefficient for short-term $PM_{2.5}$ from the conditional logistic regression model and its corresponding standard error $\widehat{\mathrm{se}}_c$. The cumulative increase for each of the quantities above across the set of newly identified disease groups or the set of disease groups identified from previous studies that were significantly associated with each 1 μg·m^{-3} increase in lag 0–1 $PM_{2.5}$ was

$$\sum_{c \in S} Q_c \left[\exp\left(\hat{\beta}_c\right) - 1\right],$$

with its 95% confidence interval estimate of

$$\left(95\%\ \mathrm{CI},\quad \sum_{c \in S} Q_c \left[\exp\left(\hat{\beta}_c\right) - 1\right] - z_{0.975}\ \widehat{\mathrm{se}}_{\mathrm{cum}} \quad - \quad \sum_{c \in S} Q_c \left[\exp\left(\hat{\beta}_c\right) - 1\right] + z_{0.975} \widehat{\mathrm{se}}_{\mathrm{cum}}\right),$$

where

$$\widehat{\mathrm{se}}_{\mathrm{cum}} = \left(\sum_{c \in S} Q_c^2 \exp\left(2\hat{\beta}_c\right) \widehat{\mathrm{se}}_c^2\right)^{\frac{1}{2}}$$

according to the delta method assuming independence between hospital admissions, S represents the set of newly identified disease groups or the set of disease groups identified from previous studies that were significantly associated with short-term $PM_{2.5}$, $z_{0.975}$ is the 97.5$^{\mathrm{th}}$ percentile of a standard normal distribution, and Q_c corresponds to N_c, $N_c L_c$, and $N_c(K_c + P_c)$, respectively, for the calculation of annual increases in number of hospitalizations, days hospitalized, and healthcare costs (inpatient and post-acute).

Section 6. Classification of principal discharge ICD-9 codes into 122 broader disease groups.

In order to reduce the potential overlap of disease groups and diagnostic misclassifications, and provide more comprehensible results, we used another ICD-9 classification scheme, the 1st and 2nd levels of multi-level CCS, to categorize diagnoses broadly into 122 broader disease groups, excluding conditions in the perinatal period and complications of pregnancy that are biologically implausible for adults ⩾65 years of age. Unlike the single-level CCS that we used in the main analysis, the multi-level CCS is has consecutive hierarchical levels with different numbers of higher levels (up to four).[7] The corresponding ICD-9 diagnosis codes for each broader disease group can be found on Healthcare Cost and Utilization Project website.[8]

References

1. Lyapustin A, Martonchik J, Wang Y, Laszlo I, Korkin S. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *Journal of Geophysical Research*. 2011;116(D3).
2. Kalnay E, Kanamitsu M, Kistler R et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*. 1996;77(3):437–71.
3. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011;73(1):3-36.
4. Wood SN, Pya N, Säfken B. Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*. 2016;111(516);1548-63.
5. Wood SN. Generalized additive models: an introduction with R: Chapman and Hall/CRC. 2006.
6. Di Q, Dai L, Wang Y et al. Association of short-term exposure to air pollution with mortality in older adults. *JAMA*. 2017;318(24):2446–56.
7. HCUP CCS. Healthcare Cost and Utilization Project (HCUP). 2017. Agency for Healthcare Research and Quality, Rockville, MD. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp [Accessed 14 Jun. 2019].
8. HCUP CCS. Healthcare Cost and Utilization Project (HCUP). 2017. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixCMultiDX.txt [Accessed 18 Sep. 2019].