Environmental Sciences Europe

RESEARCH

Open Access

Association between short-term exposure to air pollution and COVID-19 mortality in all German districts: the importance of confounders

Gregor Miller[1*], Annette Menzel[2] and Donna P. Ankerst[1,2]

Abstract

Background: The focus of many studies is to estimate the effect of risk factors on outcomes, yet results may be dependent on the choice of other risk factors or potential confounders to include in a statistical model. For complex and unexplored systems, such as the COVID-19 spreading process, where a priori knowledge of potential confounders is lacking, data-driven empirical variable selection methods may be primarily utilized. Published studies often lack a sensitivity analysis as to how results depend on the choice of confounders in the model. This study showed variability in associations of short-term air pollution with COVID-19 mortality in Germany under multiple approaches accounting for confounders in statistical models.

Methods: Associations between air pollution variables $PM_{2.5}$, PM_{10}, CO, NO, NO_2, and O_3 and cumulative COVID-19 deaths in 400 German districts were assessed via negative binomial models for two time periods, March 2020–February 2021 and March 2021–February 2022. Prevalent methods for adjustment of confounders were identified after a literature search, including change-in-estimate and information criteria approaches. The methods were compared to assess the impact on the association estimates of air pollution and COVID-19 mortality considering 37 potential confounders.

Results: Univariate analyses showed significant negative associations with COVID-19 mortality for CO, NO, and NO_2, and positive associations, at least for the first time period, for O_3 and $PM_{2.5}$. However, these associations became non-significant when other risk factors were accounted for in the model, in particular after adjustment for mobility, political orientation, and age. Model estimates from most selection methods were similar to models including all risk factors.

Conclusion: Results highlight the importance of adequately accounting for high-impact confounders when analyzing associations of air pollution with COVID-19 and show that it can be of help to compare multiple selection approaches. This study showed how model selection processes can be performed using different methods in the context of high-dimensional and correlated covariates, when important confounders are not known a priori. Apparent associations between air pollution and COVID-19 mortality failed to reach significance when leading selection methods were used.

*Correspondence: gregor.miller@tum.de

[1] Department of Mathematics, Technical University of Munich, Boltzmannstrasse 3, Garching, Germany
Full list of author information is available at the end of the article

Keywords: Variable selection, COVID-19, Air quality, Pollution, Change-in-estimate, LASSO, AIC, BIC, Cross-sectional

Background

In light of the worldwide impact of COVID-19 ubiquitously on society sectors, an increasing supply of studies have been conducted to ascertain the risk factors shaping the spread and severity of the disease. Studies based on aggregated and individual-level data have identified a multitude of clinical and demographic risk factors, including age [1–4], gender [2–5], ethnicity [5], income [5, 6], education [5], mobility [7], obesity [4, 8–10], hypertension [3, 4, 10–12], cardiovascular disease [2, 11, 12], respiratory disease [4, 12], pneumonia [1, 2, 4], history of cancer [10, 13], diabetes [3, 4, 10–12], and chronic kidney disease [4, 14, 15]. The high number of potential confounders in COVID-19 studies and large heterogeneity in approaches to adjust for them warrants robustness strategies. Air quality is one of the factors that are hypothesized to play a role, however, the overall results of previous studies are heterogeneous.

To illustrate the issues, this study focuses on the relationship between air quality and COVID-19 as one of the pressing environmental concerns with COVID-19-related morbidity and mortality. Pollution is the largest environmental cause of death being responsible for 16% of all deaths worldwide [16]. Exposure to air pollution increases the risk for hospital admission for cardiovascular and respiratory diseases [17] and enhances general mortality [18]. The predominant effect of COVID-19 on the lower and upper respiratory tract [19] can be anticipated to be compounded by the additional targeted effects of air pollution and individual risk factors, such as smoking and cardiovascular disease history, leading to multiple pathways impacting patient outcomes. For example, pollution could affect susceptibility to COVID-19 via the increase of hypertension and cardiovascular diseases [20] or weaken the host defense system of the respiratory system [21, 22]. Airborne particles might serve as carriers for pathogens, thereby supporting the dominant airborne transmission [23, 24]. Multiple studies have analyzed aspects of the association between air quality and COVID-19 outcomes, including infections and mortality (Table 1). Some studies did not account for any confounders, others only for a small fixed set. Studies adjusting for wider ranges of confounders more often failed to find significant associations.

Targeted association analyses, such as between air pollution and COVID-19 outcomes here, aim to both accurately estimate independent effect sizes as well as determine statistical significance. Determination of independent effects of a risk factor of primary interest requires adjustment for all potential confounders, many of which may be related. Although the liberal use of data-driven variable selection methods to control for confounders has been criticized [25–27], such methods remain in widespread use. Among four major epidemiological journals in 2015, half used prior or causal graphs to select variables, 12% a change in effect estimate approach, 9% stepwise methods, 5% univariate analyses, 2% other methods, and 37% did not report their methods in detail [28]. Within any given study, robustness of primary association analyses to choice of confounders for inclusion is typically omitted, although sensitivity to choice of confounders has been demonstrated even in cases of small numbers of confounders [29]. The objective of this study was to compare the resulting associations of air pollution with COVID-19 mortality in high-dimensional settings when applying leading epidemiological methods for confounder selection.

Methods

The outcomes of interest were cumulative COVID-19 mortality counts from the 400 districts in Germany for two time periods, March 2020–February 2021 and March 2021–February 2022, extracted from the Robert Koch-Institut [30] (dl-de/by-2-0 [31]) (Additional file 1: Figure S1). During the analyzed timeframe, the local health departments of two districts, Eisenach and Wartburgkreis, merged and thus merged their COVID-19 numbers. The two time periods were selected to reflect an initial phase, where lockdowns led to decreased mobility and pollution, and a later re-opening phase with increased levels. The advantage of a single country analysis is that the availability and measuring processes for the observed data are standardized at least to a certain degree, which is especially relevant with respect to the international and temporal differences in reporting COVID-19 statistics [32, 33]. Furthermore, Germany, which holds the largest population in Europe, provides extensive data on potential confounders with high spatial resolution. COVID-19 death counts were used as the outcome as there is considerable, fluctuating under-ascertainment for infection counts. Even though also undercounted and varying with time, mortality data are considered more complete than infection data [32, 33].

Risk factors

For association with cumulative COVID-19 counts, 43 risk factors were assembled for the 400 German districts

Table 1 Overview of selected publications studying associations between air quality and COVID-19 statistics

Study	Approach	Result	Area	Time
Ogen [61]	Categorized NO_2 measurements were compared	The results indicated a strong association between high values of the pollutant and high fatality cases	66 administrative regions in Italy, Spain, France, and Germany	January to February 2020
Bashir et al. [62]	The individual correlation between risk factors and new infections, total infections, and mortality were measured on a daily basis. Kendall and Spearman rank correlation was calculated. It is not clear what measurement was used to determine air quality	Besides temperature, air quality was significantly correlated with the COVID-19 metrics	New York City, USA	March to April 2020
Accarino et al. [63]	The Spearman correlation between $PM_{2.5}$, PM_{10}, NO_2 and COVID-19 incidence rate as well as mortality rate was measured	Significant associations between all of them were found	107 Italian territorial areas	February and March 2020
Zhu et al. [64]	Daily infections, meteorological variables, and air pollution concentrations for $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3 were collected. Generalized additive models were used to estimate the associations between lagged, moving average concentrations of air pollutants and daily infections	Significant positive associations for $PM_{2.5}$, PM_{10}, CO, NO_2, and O_3 and a negative association for SO_2 were shown	120 Chinese cities	January to February 2020
Stieb et al. [41]	A negative binomial model was used to measure the association between $PM_{2.5}$ from 2000 to 2016 and infection count. The Akaike information criterion was used to some extent to select from the socio-demographic, health, time since peak incidence, and temperature variables	The multivariate model did not show a significant association for $PM_{2.5}$	111 Canadian regions	Up to May 13, 2020
Wu et al. [65]	Negative binomial mixed models were used to regress on the mortality rate with $PM_{2.5}$ and 20 other confounders as predictors. The particulate matter between 2000 and 2016 was considered	A notable association was found for $PM_{2.5}$, population density, days since first reported case, household income, percent of owner-occupied housing, high school education, age, and percent of Black residents	3089 US counties	Up to June 18, 2020
Rodriguez-Villamizar et al. [42]	A negative binomial hurdle model was used to analyze the effect of $PM_{2.5}$ measured between 2014 and 2018 on COVID-19 mortality including socio-demographic, socio-economic and health confounders	$PM_{2.5}$ did not show a significant association with mortality	772 Colombian municipalities	Up to July 17, 2020
Adhikari et al. [43]	A negative binomial regression was applied on time-series data. Besides daily $PM_{2.5}$ and ozone, meteorological confounders were included	Ozone was found to be significantly associated with the daily infections but not with deaths	Queens county, New York, USA	March to April 2020
Borro et al. [66]	Simple linear regressions were performed for cumulative COVID-19 incidence, mortality rate, and case-fatality rate with $PM_{2.5}$ as predictor	Significant associations were found for all three metrics	110 Italian provinces	February to March 2020

Table 1 (continued)

Study	Approach	Result	Area	Time
Travaglio et al. [44]	Negative binomial models were used to measure the association between $PM_{2.5}$, PM_{10}, NO, NO_2, O_3 and COVID-19 cases as well as deaths. Population density, average age, and mean earning were included as confounders. Air quality data prior to the pandemic were aggregated over one and five years	Both COVID-19 metrics showed significant associations with the air quality risk factors	England on regional and sub-regional level	February to May 2020
Tieskens et al. [67]	The incidence of five distinct time periods was analyzed via mixed-effect Poisson regression. Besides $PM_{2.5}$, also 19 other socio-demographic, occupational, and mobility variables were incorporated. The variables were selected by excluding covariates with a variance inflation factor higher 2.5 in the regression of the first time period	$PM_{2.5}$ was not selected, yet almost all selected socio-demographic and economic variables indicated strong variance of their association between the time periods	351 cities in Massachusetts, USA	March to October 2020
Liang et al. [45]	Zero-inflated negative binomial models were used to determine the association between NO_2, $PM_{2.5}$, and O_3 and case-fatality and mortality rates. Air quality measurements between 2010 and 2016 were considered. The models also included socio-demographic, socio-economic, health, and mobility variables	For NO_2, a positive association with the COVID-19 metrics was found	3122 US counties	January to July 2020

over the two time periods (Additional file 1: Table S1). Daily CO, NO, NO_2, O_3, PM_{10}, and $PM_{2.5}$ measurements extracted from the ENSEMBLE dataset of the Copernicus Atmosphere Monitoring Service referred to surface estimates at noon with a 0.1° horizontal coverage over all of Germany [34]. The ENSEMBLE dataset extracts the value from nine numerical air quality models and thereby achieves a higher degree of robustness than individual models. Daily district-wide estimates of the air quality values from extracted polygons were aggregated by calculating the weighted mean depending on how much of the respective district area was covered by the corresponding polygon. For each of the two analyzed time periods and each district, the average of the daily values was then calculated for inclusion as risk factors in the models.[1]

Socio-demographic, health infrastructure, political, educational, and socio-economic variables were extracted from the German Federal and State Statistical Offices [35, 36] (license: dl-de/by-2-0 [31], tables: 12411-0015, 11111-0002, 12411-0018, 12521-0040, 12521-0041, 12531-0040, AI014-1, AI014-2, AI003-2, AI005, AI-N-01-2, AI-N-10, AI-S-01, AI007-1). Political variables referred to the federal election in 2017; gross domestic product, disposable income, and employee distribution referred to 2018; education level, socio-demographic, proportion of settlement and traffic area, and health infrastructure, 2019, except for hospital bed density in 2017. Geographic data on district area were acquired from the OpenDataLab [37] (Geodatenzentrum © GeoBasis-DE/BKG 2018 (VG250 31.12., Data changed)).

Daily mobility data were extracted from the Google Community Mobility Report [38] and was only available on a state level for the 16 states in Germany. Mobility data quantified change in number of visits and length of stay for certain places, including groceries, pharmacies, parks, residences, retail and recreational areas, transit stations, and workplaces, with respect to a reference period between January 3 and February 6, 2020. Daily values were averaged over respective time periods. Flu and vaccination data were extracted from the Robert Koch-Institut [39, 40] (dl-de/by-2-0 [31]). Means of the reported yearly flu incidences between 2017 and 2019 were calculated for each district. Daily vaccination rates reported the number of people who had received full vaccination status in the district of vaccination divided by the population of the corresponding district. Vaccination rates at the end of the respective period were used for analysis. Finally, the mean of the reported yearly flu incidences between 2017 and 2019 was calculated for each district.

[1] Neither the European Commission nor the European Centre for Medium-Range Weather Forecasts are responsible for any use that may be made of the information or data this publication contains.

Statistical methods

Two-sample *t*-tests were used to assess differences in risk factors between the two time periods, with two-sided 0.05 levels considered statistically significant. Correlation between risk factors was assessed by the Spearman method and the corresponding *p*-values were approximated by using the *t*-distribution. Negative binomial regression was used for the univariate and multivariate association analyses of risk factors with cumulative COVID-19 mortality counts, with the logarithm of the population size as offset. Negative binomial regression extends the variation of Poisson regression to accommodate overdispersion, and hence is commonly used in COVID-19 studies [41–45]. The exponentiated coefficient estimates of the negative binomial model are called incidence rate ratios (IRR).

Due to the high correlation between some of the air pollution variables, CO, NO, NO_2, O_3, PM_{10}, and $PM_{2.5}$, each was analyzed separately. A literature search identified leading methods for variable selection, which were investigated in the study as part of a sensitivity analysis [28, 46, 47]. Additionally, basic and full models were analyzed, either including only the considered air pollution variables, or all other risk factors as well. Selection methods were applied such that the respective air pollution variable, the target parameter, was always included in the model. This separates the approaches of this paper from other applications concerned only with prediction or interest in all risk factor effects equally.

Variable selection and model fitting, including basic and full models, were performed utilizing bootstrap sampling with 100 samples to obtain confidence intervals of coefficient estimates and included covariate numbers using quantiles [48]. All calculations were implemented using R version 4.1.2 [49] including the packages MASS [50], furrr [51], mpath [52], Hmisc [53], and lmtest [54].

Selection methods

The traditional stepwise selection method based on significance uses *p*-values to determine if the corresponding covariate should be included in the model. In this study, the selection criterion $p < 0.05$ was used. For the forward approach, the starting model is the basic model only including the current air pollution covariate. Iteratively a single new variable at a time is included in the current model. Each of the new models is compared to the current model via the likelihood ratio test, selecting the model with the smallest *p*-value. The process is repeated until all of the new potential models have $p \geq 0.05$ or all of the potential covariates

are included. In the backwards variant, the full model is the starting model and variables are discarded when their exclusion leads to the largest p-value. The process is stopped if all new potential models have $p < 0.05$ or only the air pollution covariate remains. The problem of the significance approach is that it can only determine if a risk factor is relevant given the other risk factors incorporated in the model.

Again starting with the basic or full model according to the forwards or backwards specification, also the Akaike (AIC) and Bayesian Information Criterion (BIC) were used. In this case, the models were selected with the smallest AIC or BIC value, respectively. With these criteria, it was possible to consider not only either inclusion or exclusion of covariates when comparing models, but both, regardless of the initial model. In general, the BIC penalizes larger numbers of covariates more severely and therefore favors smaller models. Information criteria allow the user to sort through huge numbers of models, while being computationally very efficient. However, as any of the stepwise approaches, they do not guarantee stable results such that small changes in the data may lead to very different selections.

In the change-in-estimate method (CIE), the selection criterion is based on the change of the coefficient estimate of the target parameter, in our case the air pollution variable. The implementation of the method occurs in many different flavors. In the predominant variant [55], a full model is fitted including the target parameter and all possible confounders. Confounders are then removed from the model one at a time until it becomes impossible to remove a confounder without altering the target parameter effect estimate too much compared to the estimate produced by the initial model. The change-in-estimate is defined as:

$$\Delta \mathrm{CE} = \frac{|\mathrm{CE}_i - \mathrm{CE}_0|}{\mathrm{CE}_0},$$

where CE_i is the target parameter effect estimate of the considered model with one of the confounders removed and CE_0 is the estimate of the initial model. In this backward variant, the confounder leading to the smallest change is selected as long as it is smaller than ten percent. A different option is the forward approach, where the initial model is the basic model including no confounders and confounders leading to the largest change are added as long as the change-in-estimate is larger than ten percent. The variant, where the change-in-estimate is not calculated with respect to the estimate of the initial model but with respect to the estimate of the model of the previous step, was also considered. The CIE approach offers an intuitive way to exclude and include risk factors in a model based directly on the changes in the coefficient estimates; however, setting the threshold of decision may even be more arbitrary than in other methods.

Finally, a variable selection, which is usually not presented as part of the traditional selection methods but has found its use in various studies, was also implemented [56, 57]. In this approach, the least absolute shrinkage and selection operator (LASSO) is used to select the relevant variables. LASSO is a shrinkage estimator penalizing the likelihood, thereby shrinking some coefficient estimates to zero. As all coefficient estimates are biased, the non-zero coefficients are then selected and used to refit the model to receive interpretable coefficient estimates. Cross-validation was used to set the hyperparameter of the procedure and no penalty factor for the air pollution variable was set to guarantee that it stayed in the model. The shrinkage approach of LASSO allows a more robust selection of risk factors than the other methods; however, it prohibits the direct interpretation of coefficient estimates.

Results

Before comparing the model selection approaches to evaluate the impact of the air pollution effects on COVID-19 mortality, the data are first visually and quantitatively explored. Comparisons between the first and second year after the start of the COVID-19 pandemic (Additional file 1: Table S1) showed that the air pollution variables all increased significantly except for ozone, which showed a significant decrease (all $p < 0.001$). NO, NO_2, $PM_{2.5}$, and PM_{10} more than doubled in the second period. Visits to grocery stores and pharmacies increased in comparison to the reference in the first year (median of district values: 16.5%), this dropped down again in the second year (7.0%, $p < 0.001$). Activity in parks remained on an increased level (57.0% and 58.5%), while activity in transit stations and at workplaces decreased in the first period (-13.5% and -2.0%) and then dropped even further (-31.0% and -26.0%, both $p < 0.001$). While the number of infections increased from the first to second period (27.9 to 151.1 infections per 1000 inhabitants, $p < 0.001$), the number of deaths decreased (86.6 to 51.5 deaths per 100 000 inhabitants, $p < 0.001$).

High correlations between some of the air pollution variables indicated the necessity to estimate their association to mortality separately (Additional file 1: Figure S2). NO_2 and NO (Spearman rank correlation coefficient: 0.92) as well as $PM_{2.5}$ and PM_{10} (0.91) were highly correlated. Other covariates also showed high correlations that could lead to multicollinearity. For example, transit station mobility was highly correlated with activity in retail and recreation (0.90) and workplaces (0.89), while residential and workplace mobility were

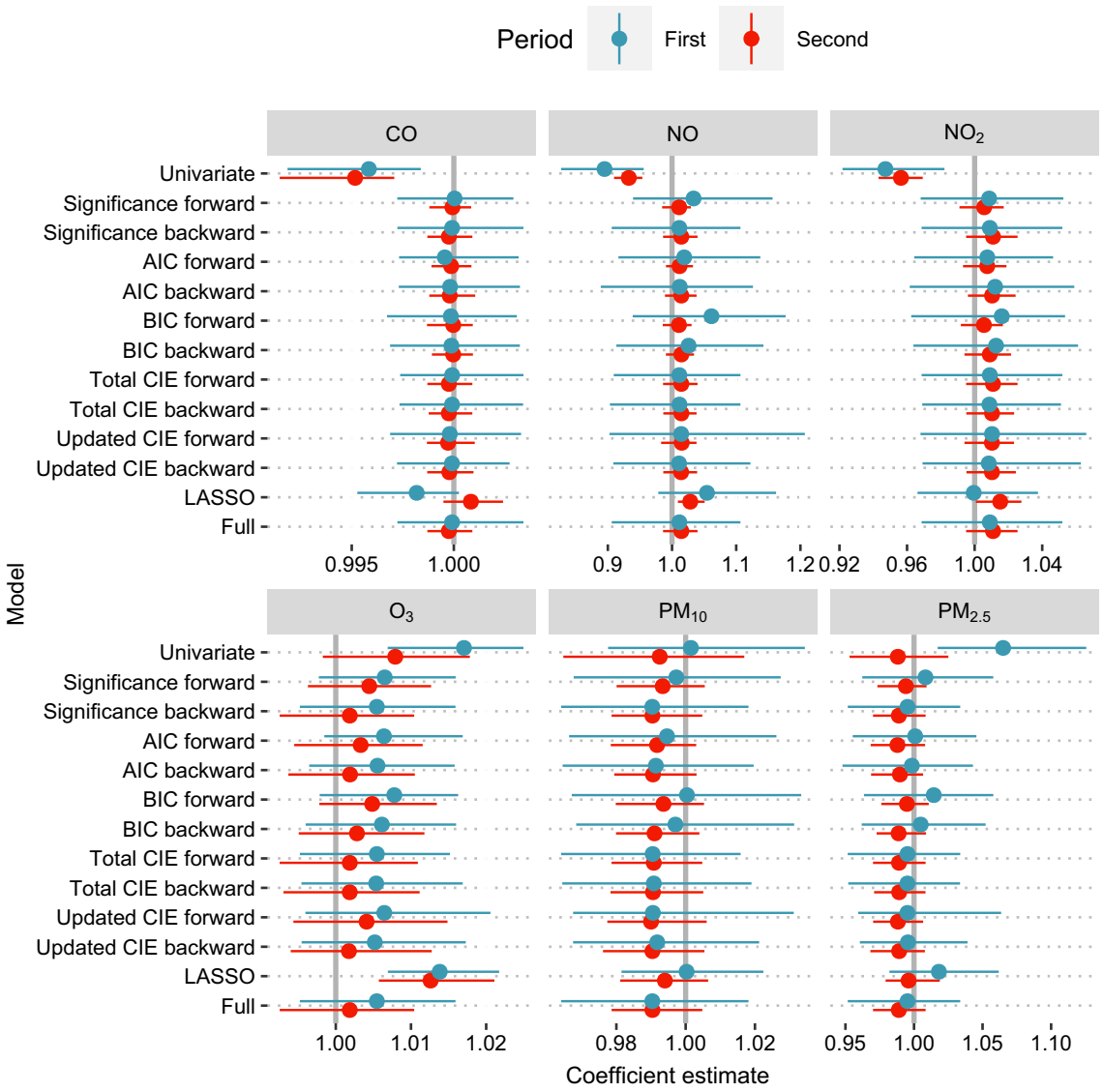

Fig. 1 Coefficient estimates of bootstrapped variable selection processes for air pollution covariates with 95% quantiles from bootstrap samples. Generally, higher mortality rates and larger dispersion in the first period lead to increased quantiles in comparison to the second period

negatively correlated (−0.94). Other examples of significant correlations were between population density and proportion of urban area in a district (0.95), proportion of males and females at least 75 years old (0.91), as well as proportion of people working in the service and people working in manufacturing (−0.99). All of these examples had *p*-values smaller than 0.0001.

The univariate analyses showed that O_3 had a positive association with COVID-19 mortality for both considered time periods (IRR of first period: 1.02, *p*-value < 0.001; IRR second period: 1.01, *p*: 0.031) (Additional file 1: Table S2). Another significant positive association was shown for $PM_{2.5}$ in the first period (IRR: 1.07, *p*: 0.009), this however lost significance in the second period (*p*: 0.4). Significant negative associations were estimated for NO, NO_2, and CO in the first period (IRR: 0.90, 0.95, 1.00; *p*: 0.013, 0.002, 0.022). This remained stable for the second time period.

Many of the other covariates also showed significant associations with mortality (Additional file 1: Table S2). Generally, indicators positively associated with increased mortality included a higher proportion of older people, less foreigners, less education, more mobility in workplaces, transit stations, retail and recreation instead of residences, more votes for political parties at the outer spectrum, higher proportion of people working

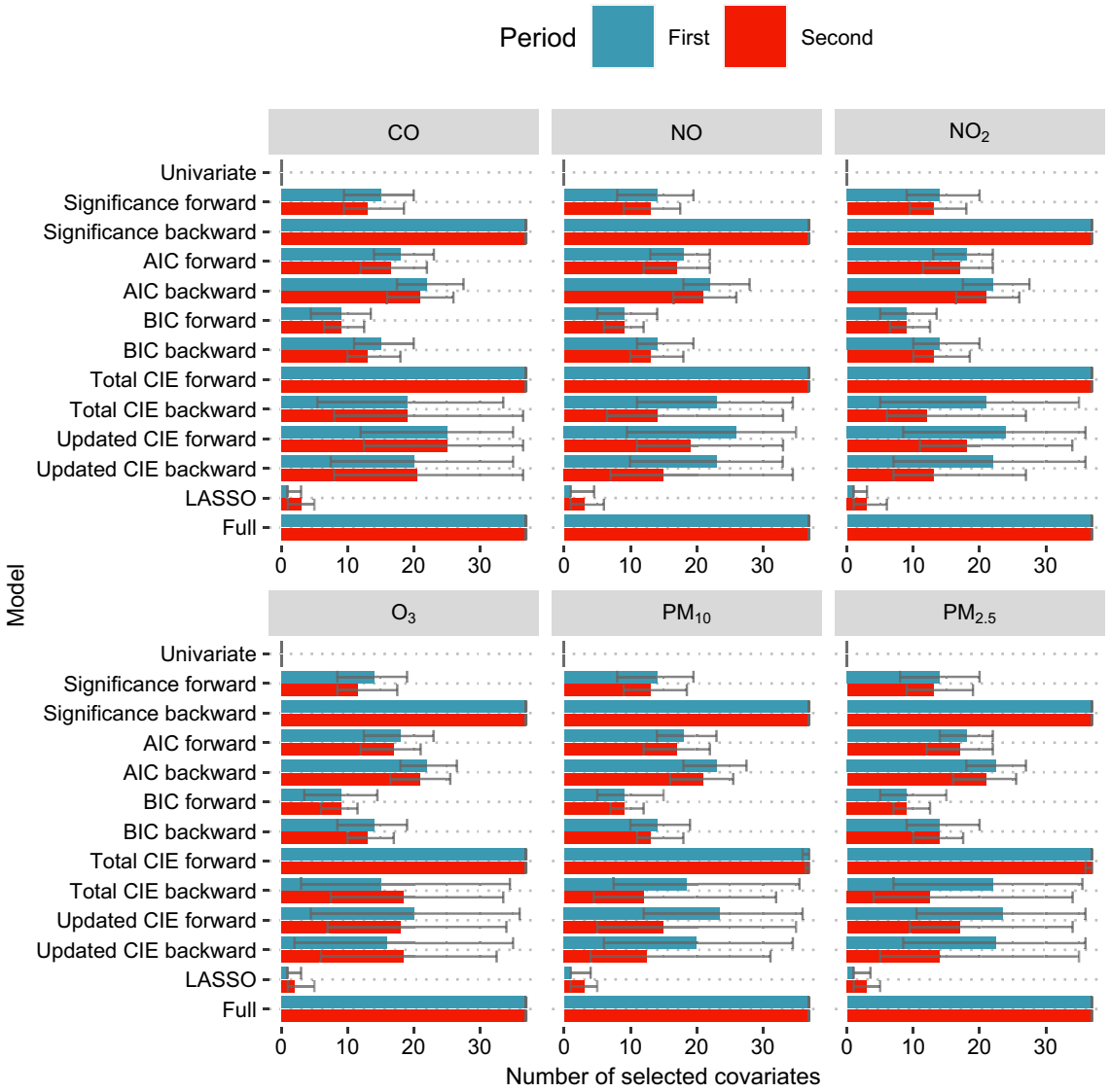

Fig. 2 Median number of selected confounders after variable selection process with 95% quantiles from bootstrap samples

in manufacturing and construction, and less health personnel per persons needing inpatient care. Many of the associations remained similar between the two time periods, however, some showed changes such as the vaccination rate which was first positive (IRR: 1.23, *p*: 0.8), when barely any full vaccinations were performed, then negative one year later, although still not quite significant (IRR: 0.83, *p*: 0.07).

Comparison of the multivariate model selection algorithms

Coefficient estimates for all selection methods can be found in Fig. 1. The often significant association of air pollution with mortality was diminished if further variables were included. This was generally independent of the selection method. For example, NO_2 is one of the clearest cases, where significant estimates in the univariates case were not visible anymore in the multivariate case. The coefficient estimates from first to second time period were somewhat decreased for O_3 and $PM_{2.5}$, otherwise, the estimates were very close between the time periods for the pollution variables, even though the variable selection methods ran independently and there were various changes for the effects of the other covariates. Another important result was that, for most selection methods, the coefficient estimates were equivalent to the full model. The LASSO selection as the only non-standard method led to larger deviations and sometimes did not converge properly. The otherwise

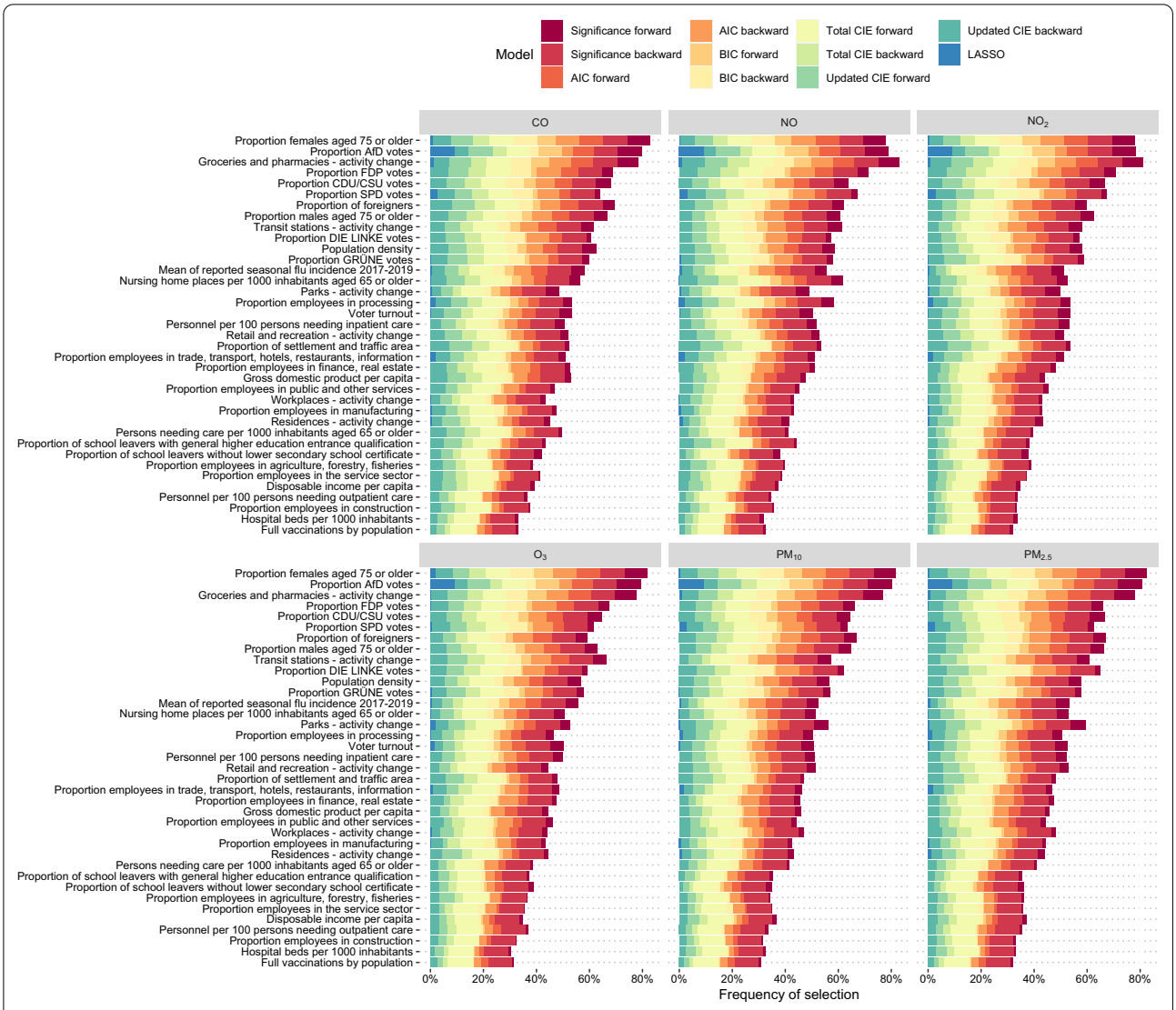

Fig. 3 Selection frequency of confounders depending on variable selection method aggregated for both analyzed time periods excluding the univariate and full model. For example, for CO, the proportion of females aged 75 or older was selected in 83% of the models with 8.7% being from the significance forward selection models

homogeneity between the selection methods, however, did not translate to the number of selected covariates. In addition, multivariate analyses were performed for all air quality metrics and selection methods except LASSO with two additional risk factors, temperature and precipitation, which yielded similar results and were therefore not considered further.

The number of selected covariates can be seen in Fig. 2. The BIC forward and LASSO selection methods led to the smallest number of covariates, but also showed larger differences to the full model. Almost all CIE methods had very large variances in the number of covariates, with the total CIE forward and significance backward consistently picking all covariates. Generally, the number of selected covariates was very consistent between the pollution variables. The most consistently selected covariates were the population proportion of females at least 75 years of age, the proportion of votes for the right-wing party AfD, and the activity in groceries and pharmacies, independent of the considered air pollution variable (Fig. 3).

As an example, the confidence intervals of the NO_2 coefficient extracted as quantiles from the bootstrap estimates were also compared with those calculated

analytically in a single selection run for the entire data set (Additional file 1: Table S3). The confidence intervals were extremely similar. The number of selected covariates in the single run was also very close to the median of the bootstrap results.

Conclusion

While previous studies have investigated the impact of air pollution on COVID-19 mortality on a very short time frame with often limited confounders, leading to different conclusions, this is the first study to consider the association over two years while incorporating high dimensional confounders, as well as propose a sensitivity analysis comparing the effect of commonly proposed variable selection methods. Univariate analyses of one air pollution risk factor at a time yielded many significant results, with some pollution variables even showing negative associations with COVID-19 mortality, which failed to reach significance after adjustment for confounders by nearly all methods. One reason could be that other risk factors, such as mobility, also drive air pollution, leading to surrogacy effects. The traditional variable selection methods provided similar results and bootstrap confidence intervals were close to those of a single iteration. If there are considerable correlations of the main exposure to other risk factors, the multicollinearity effect needs to be considered and quantified. If possible, separate analyses should be considered such as in our case where separate models were created for each of the pollution variables. The analyses here demonstrate the importance of performing sensitivity analyses of targeted risk factor outcome results to multiple methods for confounder adjustment.

There are a number of limitations with respect to previous cross-sectional studies on air pollution and COVID-19, such as ignored time differences in the introduction of the virus, confounding due to aggregation of the data on a crude level [58], and omitted confounders. These vulnerabilities were avoided or at least mitigated in this study by using the highest available spatial resolution of the data and by selection of likely confounders. In this study, a single country was analyzed over a long time span starting after introduction of the virus, while many early studies considered only the first two or three months. Use of aggregated data rather than individual-level data lead to loss of specificity in risk factor outcome association precision. However, area-specific analyses are crucial to highlight the necessity of policy decisions and more feasible in the presence of large numbers of confounders, all of which could not be easily obtained for large numbers of individuals. Another limitation is that the considered air pollution metrics may be too low to measure a significant effect on the severity of COVID-19 in comparison, for example to the highly industrialized regions, Lombardy, Veneto, and Emilia-Romagna, where the initial surge of infections and deaths in Italy appeared most severely [59].

Further studies are required to determine and gauge associations of risk factors with the spread of COVID-19. Moreover, necessary data need to be available and be standardized between countries. For example, it would be necessary to know the place of residence of vaccinated people not only the place of their vaccinations, a standardized and reliable database of interventions with a high spatial resolution is necessary, and higher reliability of COVID-19 numbers is crucial. This study has focused on mortality, but when available, excess mortality with appropriate resolution should be considered as a potentially more reliable mortality measure to compare with reported deaths [60]. Comparable sensitivity analyses as performed here should be performed in other COVID-19 association studies to assess the robustness of targeted risk factor effects on outcomes, thus avoiding unnecessary or false public health actions based on spurious results.

Abbreviations

AIC: Akaike information criterion; BIC: Bayesian information criterion; CI: Confidence interval; CIE: Change-in-estimate; IQR: Inter-quartile range; IRR: Incidence rate ratios; LASSO: Least absolute shrinkage and selection operator.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12302-022-00657-5.

Additional file 1: Figure S1. Cumulative mortality rate and average NO_2 in µg m^{-3} for the full considered time frame between March 2020 and February 2022 of the 400 German districts. **Figure S2.** Correlation plot of risk factors between German districts aggregated over the time frame between March 2020 and February 2022. Black borders indicate p<0.0005. **Table S1.** Risk factors and outcomes for first time period March 2020 – February 2021 and second time period March 2021 – February 2022. **Table S2.** Univariate association of variables with COVID-19 mortality for first time period March 2020 – February 2021 and second period March 2021 – February 2022. **Table S3.** Comparison of bootstrapped selection process with confidence intervals derived from the bootstrap quantiles and a single selection execution on the full dataset for NO2.

Acknowledgements

Not applicable.

Author contributions

GM, DPA, and AM contributed to the conception of the work and revised the manuscript. GM and DPA analyzed and interpreted the data. GM drafted the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

All data were acquired from publicly available databases.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematics, Technical University of Munich, Boltzmannstrasse 3, Garching, Germany. [2]Department of Life Science Systems, Technical University of Munich, Freising, Germany.

Received: 5 May 2022 Accepted: 7 August 2022
Published online: 27 August 2022

References

1. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN (2021) Predicting COVID-19 mortality with electronic medical records. Digit Med. 4:1–10
2. Redondo-Bravo L, Sierra Moros MJ, Martínez Sánchez EV, Lorusso N, Carmona Ubago A, Gallardo García V et al (2020) The first wave of the COVID-19 pandemic in Spain: characterisation of cases and risk factors for severe outcomes, as at 27 April 2020. Euro Surveill 25:2001431
3. Li J, Huang DQ, Zou B, Yang H, Hui WZ, Rui F et al (2021) Epidemiology of COVID-19: A systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes. J Med Virol 93:1449–1458
4. Parra-Bracamonte GM, Lopez-Villalobos N, Parra-Bracamonte FE (2020) Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico. Ann Epidemiol 52:93-98.e2
5. Drefahl S, Wallace M, Mussino E, Aradhya S, Kolk M, Brandén M et al (2020) A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. Nat Commun 11:5097
6. Baena-Díez JM, Barroso M, Cordeiro-Coelho SI, Díaz JL, Grau M (2020) Impact of COVID-19 outbreak by income: hitting hardest the most deprived. J Public Health (Oxf) 9:136
7. Kephart JL, Delclòs-Alió X, Rodríguez DA, Sarmiento OL, Barrientos-Gutiérrez T, Ramirez-Zea M et al (2021) The effect of population mobility on COVID-19 incidence in 314 Latin American cities: a longitudinal ecological study with mobile phone location data. Lancet Digital Health 3:e716–e722
8. Kwok S, Adam S, Ho JH, Iqbal Z, Turkington P, Razvi S et al (2020) Obesity: A critical risk factor in the COVID-19 pandemic. Clinical Obesity 10:e12403
9. Malik VS, Ravindra K, Attri SV, Bhadada SK, Singh M (2020) Higher body mass index is an important risk factor in COVID-19 patients: a systematic review and meta-analysis. Environ Sci Pollut Res 27:42115–42123
10. Kirillov Y, Timofeev S, Avdalyan A, Nikolenko VN, Gridin L, Sinelnikov MY (2021) Analysis of Risk Factors in COVID-19 Adult Mortality in Russia. J Prim Care Community Health 12:21501327211008050
11. Bae S, Kim SR, Kim M-N, Shim WJ, Park S-M (2021) Impact of cardiovascular disease and risk factors on fatal outcomes in patients with COVID-19 according to age: a systematic review and meta-analysis. Heart 107:373–380
12. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J et al (2020) Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis. J Infect 81:e16-25
13. Meng Y, Lu W, Guo E, Liu J, Yang B, Wu P et al (2020) Cancer history is an independent risk factor for mortality in hospitalized COVID-19 patients: a propensity score-matched analysis. J Hematol Oncol 13:75
14. Ozturk S, Turgutalp K, Arici M, Odabas AR, Altiparmak MR, Aydin Z et al (2020) Mortality analysis of COVID-19 infection in chronic kidney disease, haemodialysis and renal transplant patients compared with patients without kidney disease: a nationwide analysis from Turkey. Nephrol Dial Transplant 35:2083–2095
15. Cai R, Zhang J, Zhu Y, Liu L, Liu Y, He Q (2021) Mortality in chronic kidney disease patients with COVID-19: a systematic review and meta-analysis. Int Urol Nephrol 53:1623–1629
16. Landrigan PJ, Fuller R, Acosta NJR, Adeyi O, Arnold R, Basu N et al (2018) The Lancet Commission on pollution and health. Lancet 391:462–512
17. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL et al (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295:1127–1134
18. Faustini A, Rapp R, Forastiere F (2014) Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. Eur Respir J 44:744–753
19. Harrison AG, Lin T, Wang P (2020) Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends Immunol 41:1100–1115
20. Meo SA, Suraya F (2015) Effect of environmental air pollution on cardiovascular diseases. Eur Rev Med Pharmacol Sci 19:4890–4897
21. Yang L, Li C, Tang X (2020) The Impact of PM2.5 on the Host Defense of Respiratory System. Front Cell Develop Biol. 8:89
22. Cao Y, Chen M, Dong D, Xie S, Liu M (2020) Environmental pollutants damage airway epithelial cell cilia: Implications for the prevention of obstructive lung diseases. Thorac Cancer 11:505–510
23. Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ (2020) Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc Natl Acad Sci U S A 117:14857–14863
24. Setti L, Passarini F, De Gennaro G, Barbieri P, Perrone MG, Borelli M et al (2020) SARS-Cov-2RNA found on particulate matter of Bergamo in Northern Italy: First evidence. Environ Res 188:109754
25. Harrell FE. Multivariable Modeling Strategies. In: Harrell J Frank E, editor. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer International Publishing; 2015. p. 63–102.
26. Steyerberg EW (2009) Selection of main effects. In: Steyerberg EW (ed) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer, New York, pp 191–211
27. Chatfield C (1995) Model Uncertainty, Data Mining and Statistical Inference. J R Stat Soc A Stat Soc 158:419–444
28. Talbot D, Massamba VK (2019) A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol 34:725–730
29. Dominici F, Greenstone M, Sunstein CR (2014) Particulate Matter Matters. Science 344:257–259
30. Robert Koch-Institut. RKI_COVID19 - Übersicht. https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74. Accessed 7 Mar 2022.
31. GovData. DL-DE->BY-2.0. DL-DE->BY-2.0. https://www.govdata.de/dl-de/by-2-0. Accessed 23 Mar 2022.
32. Russell TW, Golding N, Hellewell J, Abbott S, Wright L, Pearson CAB et al (2020) Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC Med 18:332
33. Whittaker C, Walker PGT, Alhaffar M, Hamlet A, Djaafara BA, Ghani A et al (2021) Under-reporting of deaths limits our understanding of true burden of covid-19. BMJ 375:n2239
34. METEO FRANCE, Institut national de l'environnement industriel et des risques (Ineris), Aarhus University, Norwegian Meteorological Institute (MET Norway), Jülich Institut für Energie- und Klimaforschung (IEK), Institute of Environmental Protection – National Research Institute (IEP-NRI), Koninklijk Nederlands Meteorologisch Instituut (KNMI), Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO), Swedish Meteorological and Hydrological Institute (SMHI) and Finnish Meteorological Institute (FMI). CAMS European air quality forecasts, ENSEMBLE data. 2020. https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-forecasts?tab=overview. Accessed 7 Mar 2022.
35. Statistisches Bundesamt Deutschland. GENESIS-Online. 2022. https://www-genesis.destatis.de/genesis/online. Accessed 29 Apr 2022.
36. Statistische Ämter des Bundes und der Länder. Regionaldatenbank Deutschland. 2022. https://www.regionalstatistik.de/genesis/online/. Accessed 29 Apr 2022.
37. GeoJSON Utilities. http://opendatalab.de/projects/geojson-utilities/. Accessed 2 Jun 2020.
38. Google LLC. COVID-19 Community Mobility Report. COVID-19 Community Mobility Report. https://www.google.com/covid19/mobility?hl=de. Accessed 7 Mar 2022.

39. Robert Koch-Institut. SurvStat@RKI 2.0. 2021. https://survstat.rki.de/. Accessed 10 Dec 2021.
40. Robert Koch-Institut F 33. COVID-19-Impfungen in Deutschland. 2021.
41. Stieb DM, Evans GJ, To TM, Brook JR, Burnett RT (2020) An ecological analysis of long-term exposure to PM2.5 and incidence of COVID-19 in Canadian health regions. Environ Res 191:110052
42. Rodriguez-Villamizar LA, Belalcázar-Ceron LC, Fernández-Niño JA, Marín-Pineda DM, Rojas-Sánchez OA, Acuña-Merchán LA et al (2021) Air pollution, sociodemographic and health conditions effects on COVID-19 mortality in Colombia: An ecological study. Sci Total Environ 756:144020
43. Adhikari A, Yin J (2020) Short-Term Effects of Ambient Ozone, PM2.5, and Meteorological Factors on COVID-19 Confirmed Cases and Deaths in Queens, New York. Int J Environ Res Public Health 17:4047
44. Travaglio M, Yu Y, Popovic R, Selley L, Leal NS, Martins LM (2021) Links between air pollution and COVID-19 in England. Environ Pollut 268:115859
45. Liang D, Shi L, Zhao J, Liu P, Sarnat JA, Gao S et al (2020) Urban Air Pollution May Enhance COVID-19 Case-Fatality and Mortality Rates in the United States. Innovation (N Y) 1:100047
46. Heinze G, Wallisch C, Dunkler D (2018) Variable selection – A review and recommendations for the practicing statistician. Biom J 60:431–449
47. Greenland S, Daniel R, Pearce N (2016) Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int J Epidemiol 45:565–575
48. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM (2003) Internal and external validation of predictive models: A simulation study of bias and precision in small samples. J Clin Epidemiol 56:441–447
49. R Core Team. R: A language and environment for statistical computing. 2021.
50. Venables WN, Ripley BD, Venables WN (2002) Modern applied statistics with S, 4th edn. Springer, New York
51. Vaughan D, Dancho M. furrr: Apply Mapping Functions in Parallel using Futures. 2021.
52. Wang Z. mpath: Regularized Linear Models. 2021.
53. Harrell F. Hmisc: Harrell Miscellaneous. 2021.
54. Zeileis A, Hothorn T (2002) Diagnostic Checking in Regression Relationships. R News 2:7–10
55. Weng H-Y, Hsueh Y-H, Messam LLM, Hertz-Picciotto I (2009) Methods of Covariate Selection: Directed Acyclic Graphs and the Change-in-Estimate Procedure. Am J Epidemiol 169:1182–1190
56. Meintrup D, Borgmann S, Seidl K, Stecher M, Jakob CEM, Pilgram L et al (2021) Specific Risk Factors for Fatal Outcome in Critically Ill COVID-19 Patients: Results from a European Multicenter Study. J Clin Med 10:3855
57. Nomura S, Eguchi A, Yoneoka D, Kawashima T, Tanoue Y, Murakami M et al (2021) Reasons for being unsure or unwilling regarding intention to take COVID-19 vaccine among Japanese people: A large cross-sectional national survey. Lancet Reg Health West Pac 14:100223
58. Heederik DJJ, Smit LAM, Vermeulen RCH (2020) Go slow to go fast: a plea for sustained scientific rigour in air pollution research during the COVID-19 pandemic. Eur Respir J 56:2001361
59. Filippini T, Rothman KJ, Goffi A, Ferrari F, Maffeis G, Orsini N et al (2020) Satellite-detected tropospheric nitrogen dioxide and spread of SARS-CoV-2 infection in Northern Italy. Sci Total Environ 739:140278
60. Karlinsky A, Kobak D. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. eLife. 10:e69336.
61. Ogen Y (2020) Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality. Sci Total Environ 726:138605
62. Bashir MF, Ma B, Bilal, Komal B, Bashir MA, Tan D, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci Total Environ. 2020;728:138835.
63. Accarino G, Lorenzetti S, Aloisio G (2021) Assessing correlations between short-term exposure to atmospheric pollutants and COVID-19 spread in all Italian territorial areas. Environ Pollut 268:115714
64. Zhu Y, Xie J, Huang F, Cao L (2020) Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. Sci Total Environ 727:138704
65. Wu X, Nethery RC, Sabath MB, Braun D, Dominici F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. Science Advances. 6:eabd4049.
66. Borro M, Di Girolamo P, Gentile G, De Luca O, Preissner R, Marcolongo A, et al. Evidence-Based Considerations Exploring Relations between SARS-CoV-2 Pandemic and Air Pollution: Involvement of PM2.5-Mediated Up-Regulation of the Viral Receptor ACE-2. International Journal of Environmental Research and Public Health. 2020;17:5573.
67. Tieskens KF, Patil P, Levy JI, Brochu P, Lane KJ, Fabian MP et al (2021) Time-varying associations between COVID-19 case incidence and community-level sociodemographic, occupational, environmental, and mobility risk factors in Massachusetts. BMC Infect Dis 21:686

Publisher's Note