Agreement No. 19RD004

Sources of On-Road Vehicle Emissions and their Impacts on Respiratory Disease Symptoms in California

Guangquan (Jason) Su, PhD, Principal Investigator
Shadi Aslebagh, PhD
Stacey Jung
Emma Yakutis
University of California, Berkeley

Meredith Barrett, PhD, Subcontract PI
Vy Vuong,
ResMed/Propeller Health

March 2022

Prepared for the California Air Resources Board

Disclaimer

The statements and conclusions in this Report are those of the contractor and not necessarily those of the California Air Resources Board. The mention of commercial products, their source, or their use in connection with material reported herein is not to be construed as actual or implied endorsement of such products.

Acknowledgement

This Report was submitted in fulfillment of CARB 19RD004 agreement: “Sources of On-Road Vehicle Emissions and their Impacts on Respiratory Disease Symptoms in California” by the University of California, Berkeley, and its subcontractor ResMed/Propeller Health under the sponsorship of the California Air Resources Board. Work was completed as of March 2022.

This project was funded under the CARB’s Dr. William F. Friedman Health Research Program. During Dr. Friedman’s tenure on the Board, he played a major role in guiding CARB’s health research program. His commitment to the citizens of California was evident through his personal and professional interest in the Board’s health research, especially in studies related to children’s health. The Board is sincerely grateful for all of Dr. Friedman’s personal and professional contributions to the State of California.

Abbreviations

ACT: Asthma Control Test
AOD: Aerosol Optical Depth
Ba: Barium
BC: Black Carbon
CARB: California Air Resources Board
CalTrans: California Department of Transportation
CAT: COPD Assessment Test
CDC: Centers for Disease Control and Prevention
CH_4: Methane
CI: Confidence Interval
CO_2: Carbon Dioxide
COPD: Chronic Obstructive Pulmonary Disease
Cr: Chromium
CTRL: Control Areas
Cu: Copper
DEM: Digital Elevation Modeling
DOAS: Differential Absorption Spectroscopy
D/S/A: Deletion/Substitution/Addition
ED: Emergency Department
EPA: Environmental Protection Agency
ESRI: Environmental Systems Research Institute
EXP: Exponential function
FCC: Feature Class Classification
GB: Gigabyte
GIS: Geographic Information System
glmerMod: Generalized Linear Mixed Model Fit by Maximum Likelihood
glmmTMB: Generalized Linear Mixed Models using Template Model Builder
GMC: Goods Movement Corridors
GPS: Global Positioning System
Ha: Hectare
HEI: Health Effects Institute
HIPAA: Health Insurance Portability and Accountability Act of 1996

HPMS: Highway Performance Monitoring System
Humid: Humidity
Hz: Hertz
IDW: Inverse Distance Weighting
IRI: International Roughness Index
K: Kelvin
km: Kilometer
kPa: Kilopascal
LA: Los Angeles
LB: Long Beach
LUR: Land Use Regression
m: Meter
Max.: Maximum
Min.: Minimum
mm: Millimeter
m/s: Meters Per Second
MB: Megabyte
Mn: Manganese
µg: Microgram
NAAQS: National Ambient Air Quality Standard
NDVI: Normalized Difference Vegetation Index
NGMC: Non-Goods Movement Corridors
Ni: Nickel
NLCD: National Land Cover Database
NO: Nitrogen Monoxide
NO$_2$: Nitrogen Dioxide
OMDOAO3e: OMI/Aura Ozone (O3) DOAS Total Column 1-Orbit L2 Swath 13x24 km V003
OMI: Ozone Monitoring Instrument
O$_3$: Ozone
OR: Odds Ratio
Pb: Lead
Pctl: Percentile
PEF: Peak Expiratory Flow
PeMS: Performance Measuring System
PM$_{2.5}$ and PM$_{10}$: particulate matter with diameter less than 2.5 and 10 microns
ppb: Parts Per Billion
RR: Rate Ratio
S: Sulfur
SABA: Short-Acting Beta Agonist
Sb: Antimony
Se: Selenium
SO$_2$: Sulfur Dioxide
std. dev.: Standard Deviation
TB: Terabyte
Temp.: Temperature

TEU: Twenty-foot Equivalent Units
TRAP: Traffic-Related Air Pollution
TBW: Tire- and Brake-Wear
UA: Urban Areas with a Population Over 50,000
UC: Urban Areas with a Population Between 25,000 and 50,000
UCB: University of California, Berkeley
UCLA: University of California, Los Angeles
UFP: Ultrafine Particles
UV: Ultraviolet
USGS: United States Geological Survey
UV: Ultraviolet
VIF: Variance Inflation Factors
VKT: Vehicle Kilometer Traveled
WHO: World Health Organization
Zn: Zinc

Table of Contents

List of Tables

List of Figures

Project Summary/Abstract

Regulations and technological upgrades have resulted in a steady decline in vehicle tailpipe emissions in California. However, these positive developments have not been able to fully compensate for the rapid growth of the motor vehicle fleet, the accompanying peak traffic and vehicular congestion on California roadways, and the increased area-based contributions from on-road vehicles while off-road (such as shopping centers, parking lots, and freight distribution centers). Some communities in proximity to heavy traffic emissions sources continue to be disproportionately exposed. Another important on-road vehicle pollution source is non-exhaust emissions from tire and brake wear, which will become increasingly important as the benefits of implementation of tailpipe emission regulations become more widespread. Previous studies on the effects of air pollution on respiratory disease symptoms, such as asthma and chronic obstructive pulmonary disease (COPD) often used an individuals' residential address in defining the location of air pollution exposure. Air pollution exposure can occur in the community, at work, at home, at school, and elsewhere; therefore, a residential address does not capture the full signature of exposure for an individual.

Digital sensors fitted onto inhalers can capture the date, time, and location of rescue inhaler medication use (i.e., use of Short-Acting Beta Agonist for the acute relief of respiratory disease symptoms, referred to as SABA use here after) and identify activity space through sensor "heartbeats" — sensor checking into battery life every 3-4 hours with location information; thereby, offering an objective signal of respiratory disease symptoms and exposure space in real-time. The spatiotemporally rich data in SABA use, locations of activity space, and extensive information on environmental exposures; however, raise methodological challenges in modeling the impacts of environmental exposures on respiratory disease symptoms. Recent advances in artificial intelligence and machine learning techniques like random forest, have boosted the potential of post-analyzing the high dimension patient data. The random forest modeling technique has the advantage over traditional linear mixed models in being free of assumption to the underlying data structure, in resistance to model overfit and multicollinearity, in dealing with interactions between simultaneous multiple pollutant exposures, and in increased accuracy in prediction.

In this project, the University of California, Berkeley (UCB) first conducted a systematic review on the impacts from air pollution, including criteria air pollutants and non-exhaust trace metals on respiratory disease outcomes. Second, UCB developed daily land use regression models (res: 30m) and surfaces (res: 100) for criteria pollutants nitrogen dioxide (NO_2), particulate matter with diameter less than 2.5 microns ($PM_{2.5}$) and ozone (O_3) for 2012-2019. Monthly models (res: 30m) and surfaces (res: 100m) were also developed for six trace metals including chromium (Cr), manganese (Mn), nickel (Ni), lead (Pb), selenium (Se) and zinc (Zn). A spatially and temporally-resolved rescue inhaler use dataset collected by ResMed and its sub-division Propeller Health using digital sensors for 3,386 patients across California were assigned air pollution exposure using the daily surfaces developed for the State and were used to identify the impacts of both on-road tailpipe emissions and non-tailpipe trace metals emissions on sub-acute respiratory disease symptoms from January 1, 2012 to December 31, 2019 through advanced linear mixed models taking into consideration of excessive days without rescue inhaler use and a random forest modeling technique. In the health outcome model with simultaneous exposures to NO_2, $PM_{2.5}$ and O_3, we found that all the three criteria pollutants were statistically significantly associated with rescue inhaler use after comprehensive control for confounding. Their respective exposure-response functions were 1.002482 (95% CI: 1.001255 – 1.003710), 1.008790 (95% CI: 1.007243 – 1.010340) and 1.005306 (95% CI: 1.003983 – 1.006630). The impact of trace metals on rescue inhaler use was found statistically non-significant, probably due to inefficient air quality monitoring and exposure assessment.

The results of this project will help the California Air Resources Board (CARB) characterize the health effects of regulatory programs and identify disproportionally exposed communities and related sources. More importantly, the study will provide the information needed to include respiratory disease exacerbations as a new endpoint for CARB's health analysis in regulatory processes.

Executive Summary

a. Background

Regulations and technological upgrades have resulted in a steady decline in vehicle tailpipe emissions in California. However, these positive developments have not been able to fully compensate for the rapid growth of the motor vehicle fleet, the accompanying peak traffic and vehicular congestion on California roadways, and the increased area-based contributions from on-road vehicles while off-road (such as shopping centers, parking lots, and freight distribution centers). Some communities in proximity to the heavy traffic emissions sources continue to be disproportionately exposed. Further, the emissions from non-exhaust tire and brake wear (TBW) are, by contrast, becoming increasingly important as the benefits of implementation of tailpipe emission regulations become more widespread. It has been shown that even with zero tailpipe emissions, traffic will continue to contribute to particulate matter (PM) through non-exhaust emissions.

When the location of exposure was provided, previous studies often used an individual's residential address in defining the location of air pollution exposure. Air pollution exposure can occur in the community, at work, at home, at school, and elsewhere; therefore, a residential address does not capture the full signature of exposure for an individual. Digital sensors fitted onto inhalers can capture the date, time, and location of rescue inhaler use. The sensors also send out "heartbeats" every 3-4 hours: a signal from the sensor that reports battery life (lasts about 18 months after a single charge) and records the Global Positioning System (GPS) locations when paired with a smartphone. Evaluating the signals of all heartbeats and rescue inhaler uses for an individual over time can help characterize an individual's exposure space over time.

b. Objective

The objective of this study was two folds: First, UCB conducted a systematic review on the impacts from air pollution, including criteria air pollutants and non-exhaust trace metals on respiratory disease outcomes. Second, UCB and ResMed aimed to quantify the relationship between on-road vehicle emissions, including on-road exhaust criteria pollutants and non-exhaust trace metals, and sub-acute respiratory disease symptoms, which may include chest pain, shortness of breath, coughing or wheezing, in major metropolitan areas of California. We used rescue inhaler use as a proxy for the symptoms. The goal was to identify the impacts of both tailpipe and non-tailpipe vehicle emissions on sub-acute respiratory disease symptoms represented by rescue inhaler use for health outcome data collected from January 1, 2012 to December 31, 2019 across California.

c. Methods

In the literature review, we assessed impacts from criteria pollutants including nitrogen dioxide (NO_2), particulate matter with diameter less than 2.5 microns ($PM_{2.5}$) and 10 microns (PM_{10}), ozone (O_3), sulfur dioxide (SO_2) and more than 10 trace metals considered toxic air contaminants on respiratory disease outcomes. The outcome measures were converted to odds ratios, and we standardized the effects of each criteria pollutant so comparisons could be made among different studies. Pooled analyses were also conducted to identify the overall size of impact of a pollutant on respiratory disease outcomes. In developing daily air pollutant surfaces, we applied a deletion/substitution/addition (D/S/A) V-fold cross-validation modeling technique that minimizes over-fitting to the data to maximize the probability that guarantees the models predict well at locations that have not been sampled. The air pollution surfaces were assigned to corresponding space-time activity space measured through digital sensors by patients. In health outcome modeling, we applied both linear mixed models with capability of dealing with excessive zeros and overdispersion in daily rescue inhaler use puffs and advanced machine learning algorithms to identify associations of daily air pollution exposure with daily rescue inhaler use in number of puffs.

d. Results

The literature review identified significant associations of respiratory disease outcomes with the criteria pollutants and trace metals. In air pollution modeling, we developed daily prediction models for the three criteria pollutants with adjusted R^2 of 79.6%, 65.3% and 93.6%, respectively, for NO_2, $PM_{2.5}$ and O_3. The models had greater performance than other daily models and had higher performance than some annual models. The prediction models for the six criteria pollutants were less effective compared to the criteria pollutants models and they lacked predictors that specifically targeted tire- and break-wear. In health outcome modeling, both advanced linear mixed models and random forest models identified that all the three criteria pollutants had significant ($p < 0.001$) and positive associations with daily rescue inhaler use. We identified, for example, in the linear mixed model with all the three criteria pollutants integrated in a single model, an effect of NO_2 on 1 ppb (Part per Billion) increase, $PM_{2.5}$ on 1 ug m^{-3} increase, and O_3 on 1 ppb increase, respectively, for a 0.25%, 0.88% and 0.53% increase in daily rescue puffs use.

The effect of exposure to trace metals were found statistically non-significant with daily rescue inhaler use.

e. Conclusion

We successfully conducted a literature review on the impacts from air pollution, including criteria air pollutants and non-exhaust trace metals on respiratory disease outcomes. In air pollution modeling, we are the first in literature that incorporated terabytes (TB) of comprehensive data sources for high spatial resolution daily (criteria pollutants) and monthly (trace metals) land use regression modeling development (res: 30m) and surfaces generation (res: 100m). Our research identified that exposures to NO_2, $PM_{2.5}$ and O_3 were significantly associated with daily rescue inhaler use but the association with trace metals should be further explored with better exposure data. The results of this project will help CARB characterize the health effects of regulatory programs and help identify disproportionally exposed communities and related sources. More importantly, the project will also provide the information needed to include respiratory disease symptoms as a new endpoint for CARB's health analysis in the regulatory processes.

Background

Regulations and technological upgrades have resulted in a steady decline in vehicle tailpipe emissions in California.[1,2] However, these positive developments have not been able to fully compensate for the rapid growth of the motor vehicle fleet, the accompanying peak traffic and vehicular congestion on California roadways, and the increased area-based contributions from on-road vehicles while off-road (such as shopping centers, parking lots, and freight distribution centers). Some communities in proximity to the heavy traffic emissions sources continue to be disproportionately exposed. Several studies have been published showing that communities exposed to on-road vehicle emissions are at greater risk for respiratory disease exacerbations.[3-6] Another important on-road vehicle pollution source is non-exhaust emissions from tire and brake wear, which will become increasingly important as the benefits of implementation of tailpipe emission regulations become more widespread. It has been shown that even with zero tailpipe emissions, traffic will continue to contribute to fine and ultrafine PM through non-exhaust emissions.[7]

Previous studies on the effects of air pollution on respiratory disease have relied on aggregated and infrequently-reported acute respiratory disease outcome measures, such as emergency department (ED) visits or hospitalizations, which lack temporal and spatial resolution due to annual aggregation and grouping to a zip code or county level.[8-10] Other studies have used patient self-reported data to assess the location and frequency of symptoms,[11] which could be fraught with missing data, errors, and are burdensome for the patients.[12-14] When the location of exposure was provided, previous studies often used an individual's residential address in defining the location of air pollution exposure. Air pollution exposure can occur in the community, at work, at home, at school, and elsewhere; therefore, a residential address does not capture the full signature of exposure for an individual. Significant exposure misclassification exists and health risks estimated from such data can lead to exposure measurement error and flawed findings.[15] Additionally, the effect of air pollution on respiratory disease has largely been assessed using single pollutant modeling approaches despite the fact that people are exposed to multiple pollutants simultaneously,[16] which may interactively influence respiratory disease symptoms. Further, current studies typically assume the existence of a predefined structure of association between a predictor and a respiratory disease outcome before a model was developed. However, an association could be linear, non-linear, or piecewise.

Digital sensors fitted onto inhalers can capture the date, time, and location of SABA use; thereby, offering an objective signal of respiratory disease symptoms in real-time. A previous feasibility study demonstrated such sensors can collect spatiotemporal data on the use of rescue medications.[17] The sensors also send out "heartbeats," which is a signal from the sensor that reports battery life and records the GPS locations when paired with a smartphone. Information from both SABA use and heartbeats is sent to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) compliant server for storage and analysis, and heartbeats of every 3-4 hours also report the battery status of the inhaler, which lasts about 18 months. The heartbeat indicates that the sensor is functioning properly, and able to transmit data; but that the inhaler has not been used since the last data transmission. Evaluating the signals of all heartbeats and rescue inhaler uses for an individual over time, can help characterize each individual's exposure space over time. UCB uses the heartbeat locations and locations of SABA use of a patient as his/her activity space for this project. However, the spatiotemporally rich SABA use data, location of activity space and extensive information on environmental exposure make modeling environmental impacts difficult issues to tackle. Recent advances in artificial intelligence and machine learning techniques have boosted the potential of post-analyzing the high dimension patient data.[18-20]

A recent systematic literature review[19] indicates that machine learning algorithms are increasingly being applied to air pollution epidemiology. One of the key applications is modeling the health effects of air pollution mixtures,[20-23] which could be highly correlated and with potential interaction, a challenge difficult to tackle with the traditional environmental epidemiology method. Random forest models, one of the most frequently used machine learning techniques in health-related research,[24] has an advantage over linear models because random forest models (1) make no assumption as to the underlying data structure, (2) are resistant to model overfit, multicollinearity and interaction among simultaneous multiple pollutant exposures, and (3) demonstrate enhanced prediction.[25] Random forest models have been applied in epidemiology studies for variable selection[26,27] and health outcome modeling[28-30] and typically perform better than traditional epidemiology models.[25,28]

The objective of this study was two folds: First, we conducted a systematic review on the impacts from air pollution, including criteria air pollutants and non-exhaust trace metals on respiratory disease outcomes. Second, we aimed to quantify the relationship between on-road vehicle emissions, including on-road exhaust criteria pollutants and non-exhaust trace metals, and sub-acute respiratory disease symptoms, which may include chest pain, shortness of breath, coughing or wheezing, in major metropolitan areas of California. We used rescue inhaler use as a proxy for the symptoms. The goal was to identify the impacts of both tailpipe and non-tailpipe vehicle emissions on sub-acute respiratory disease symptoms represented by rescue inhaler use for health outcome data collected from January 1, 2012 to December 31, 2019 across California.

Task 1: Literature Review

UCB conducted a systematic literature review, using peer-reviewed journal papers, to expand the literature citied in the background section of this study on impacts from on-road vehicle emissions, including on-road non-exhaust pollutants on sub-acute respiratory disease symptoms. The following inter-connected steps were used to complete the review:

A. Determine inclusion criteria that will include the:
   - Study population of children (<18 years), adults (≥18 years), and all ages.
   - Study intervention for individuals (1) exposed to air pollution, including NO_2, $PM_{2.5}$ and PM_{10}, O_3 and SO_2 (2) more than 10 trace metals considered toxic air contaminants

(including Aluminum, Iron, Magnesium, Sulfur, Nickel, Vanadium, Chromium, Arsenic, Manganese, Barium, Copper, Antimony, Zinc and Lead).

- Study outcome for respiratory disease in asthma and COPD such as coughing, wheezing, shortness of breath, ED visits, hospitalizations, exacerbations and mortality, respiratory infections, and lung cancer.

B. Identify the publications characteristics for studies:

- Published in peer-reviewed journals.
- Published between January 1, 2000, and January 1, 2021;
- Written in English.

C. Select the proper search databases and engines, including

- PubMed
- Medline
- Web of Science Core Collection; and
- Google Scholar

D. Decide the search terms and selection process.

The search category and search terms are listed in Table 1.

Table 1. Literature review categories and search terms

Category	Search Terms
Air pollutants	$PM_{2.5}$, PM_{10}, NO_2, O_3, SO_2 and trace metals
Disease	Asthma, COPD, respiratory disease
Years of publication	January 1st, 2000, through January 31st 2021
Publication type	Peer-reviewed journals
Publication language	English

The following steps were used to select scientific publications for the literature review:

1) Use one term from each category and combine them together (+) to create integrated search terms using the search databases and engines listed above.
2) Merge together the selected publications and remove the duplicates.
3) Obtain abstracts for the remaining publications selected from Step 2), screen and remove the publications that are not related to research topic.
4) Obtain full text for the remaining publications selected from Step 3), screen and remove the publications that are not related to the topic.

We identified 519 papers considering the search criteria given in Table 1. After screening titles, 137 papers were selected for an abstract search. We evaluated the abstract of each paper and selected 54 papers for full-text screening. After full-text screening, 36 papers were included in the final review.

UCB conducted the literature review with the final selected ones. The outcome measures were converted to odds ratios, and we standardized the effects of each criteria pollutant so comparisons could be made among different studies (see Appendix). Pooled analyses were also conducted to identify the overall size of impact of a pollutant on respiratory disease outcomes. We identified significant associations of respiratory disease outcomes with the criteria pollutants and trace metals. We further identified that the effects were greater for children exposed to NO_2, $PM_{2.5}$ and SO_2 while the effects were greater for adults exposed to PM_{10} and O_3.

We identified consistently significant associations of all the criteria pollutants and trace metals with a broad range of respiratory disease outcomes. After pooling all the effect estimates, we further identified that the associations were slightly greater for children for exposures to NO_2 (1.68 vs 1.14), $PM_{2.5}$ (1.58 vs 1.25) and SO_2 (1.38 vs 1.13), while the effects were slightly greater for adults for exposures to PM_{10} (1.10 vs 1.05) and O_3 (1.40 vs 1.17).

Table 2 summarizes the pooled effect estimates for all criteria pollutants for the three age categories we considered in the literature review: children, adults, and all.

Table 2. Pooled effect estimates

	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2
Children	1.58(1.04-2.41)	1.05(0.99-1.11)	1.68(1.14-2.63)	1.17(1.05-1.31)	1.38(1.09-1.78)
Adults	1.25(0.99-1.64)	1.10(0.94-1.33)	1.14(0.99-1.32)	1.40(1.14-1.81)	1.13(1.10-1.16)
All	1.48(1.13-2.14)	1.25(1.04-1.51)	1.18(1.01-1.42)	1.28(1.12-1.47)	1.10(1.02-1.19)

Outdoor NO_2 concentration is often considered a good marker of traffic-related air pollution (TRAP) and concentrations decline rapidly with increasing distance from highways and major roadways. $PM_{2.5}$ has smaller spatial gradients compared to NO_2 and regional sources contribute to $PM_{2.5}$ concentrations. However, a major source of $PM_{2.5}$, especially in high-income countries, is from traffic. SO_2 air pollution can be generated from traffic, especially from diesel vehicles, and industrial point sources such as coal-fired power plants, with a rapid decline with increasing distances from these point sources.

The relatively greater impact from NO_2, $PM_{2.5}$ and SO_2 on respiratory disease outcomes for children in our pooled analysis may be due to the greater probability of children living or having physical activity in high air pollution areas for these pollutants.[31-35] Because children's lungs are not completely developed, their exposures to high levels of air pollution can affect both their short-term and long-term respiratory health.[36] Policymakers and stakeholders should adopt strategies to help children, especially those living in vulnerable communities, to reduce exposures to those sources in their neighborhoods.

By contrast, the impact of O_3 on respiratory disease outcomes was found to be relatively greater for adults. Due to the chemical reaction between O_3 and traffic emission of nitrogen oxides that lead to generation of NO_2, NO_2 concentrations are higher near busy roadways while O_3 are higher at locations farther away from busy roadways. We suspect that, compared to children, more adults might have moved away from living close to traffic sources to communities with less traffic-related air pollution,[37] thus leading to greater O_3 exposure.

It should be addressed that due to the limitation in the number of studies in each type of outcomes for specific criteria pollutant, we pooled the results for all types of outcomes together for each criteria pollutant and did the comparison across different types of outcomes. It is also worth noting that the impact of trace metals on human respiratory health is understudied and only 3 studies[38-40] were included in the pooled analysis in this review.

Overall, this systematic review identified consistent and significant effects of ambient exposures to criteria pollutants and trace metals on a broad suite of respiratory disease outcomes. It contributes to better understanding of the size of the effect of the five criteria pollutants and trace metals that can be compared across various studies. The study also helped identify groups that are more vulnerable to adverse respiratory outcomes from air pollution exposure across available studies. The full literature review paper is included in the appendix of the project report.

Task 2: Exposure Assessment, Health Dataset, and Analysis of Health Effects

Our goal was to identify the impacts of both tailpipe and non-tailpipe vehicle emissions on sub-acute respiratory disease symptoms represented by rescue inhaler use for health outcome data collected from January 1, 2012 to December 31, 2019 across California. First, we applied D/S/A V-fold cross-validation land use regression modeling technique that minimizes over-fitting to the data to maximize the probability that guarantees the models predict well at locations that have not been sampled. The high spatial resolution air pollution surfaces were then generated using the

D/S/A modeling results and assigned to corresponding patients' space-time activity space measured through digital sensors. In health outcome modeling, we applied both linear mixed models with capability of dealing with excessive zeros and overdispersion in daily rescue inhaler use puffs and advanced machine learning algorithms to identify associations of daily air pollution exposure with daily rescue inhaler use in number of puffs.

Subtask 2.1: Development of Comprehensive Data Sources

UCB developed comprehensive data sources to be used in air pollution exposure modeling. The data sources include daily traffic data, daily remote sensing data, daily weather data, one time land use and land cover data, daily criteria air pollutants, trace metals, stationary sources of emissions and highway pavement conditions data.

Comprehensive Data Sources for Criteria Pollutants Modeling

Acquire and process potential criteria pollutants predictors

Daily traffic data: For daily traffic data, UCB used the data collected by the California Department of Transportation (CalTrans) Performance Measurement System (PeMS) (https://dot.ca.gov/programs/traffic-operations/mpr/pems-source). PeMS data are collected in real-time from nearly 40,000 individual detectors spanning the freeway system across all major metropolitan areas of the State of California and provide an archived data user service that provides over fifteen years of data for historical analysis. PeMS integrates a wide variety of information from Caltrans and other local agency systems including traffic flow, speed, occupancy, incident, toll charge, and other information. UCB used PeMS 5-minute road link/segment traffic flow data in the analysis. In PeMS, traffic flow (volume) is a quantity representing the number of vehicles that passed over each detector on the roadway in a given time period (i.e. 5-minute flow, hourly flow, etc.). The detector measured traffic flow covered 12.52% highway segments and we summed hourly traffic to daily traffic for all the stations across California. The following interconnected stages were then used to derive daily traffic for all the California highways for years 2012-2019:

1) For a road segment with station traffic measure for a day, use all the station traffic measures on that road segment to generate a daily mean traffic for that road segment for that day.
2) For those road segments without traffic measures for a day, assign them using the assigned segments from step 1 by matching route, county, district, route type and day, and find the

one with the smallest distance if having multiple matches. California has 58 counties which are included in one of the 12 CalTrans air districts (1 - Eureka, 2 - Redding, 3 - Marysville / Sacramento, 4 - Bay Area / Oakland, 5 - San Luis Obispo / Santa Barbara, 6 - Fresno / Bakersfield, 7 - Los Angeles, 8 - San Bernardino / Riverside, 9 - Bishop, 10 - Stockton, 11 - San Diego, 12 - Orange County). Highways in California are split into at least four different types of systems: Interstate Highways, U.S. Highways, state highways, and county highways.

3) For those road segments without traffic being assigned from steps 1 & 2, assign them using the assigned segments from steps 1 & 2 by matching route, district, route type and day, and find the one with the smallest distance if having multiple matches. In this step county was not used as a restricting factor in daily traffic assignment.

4) For those road segments without traffic being assigned from the above steps, assign them using the above assigned segments by matching route, county, district and route type, plus at most one day difference in data availability and find the one with the smallest distance if having multiple matches.

5) Identify those not assigned and assign them using the assigned segments from above steps by matching county, district, route type and day and find the one with the smallest distance if having multiple matches. Here we removed the restricting factor of route number.

6) Identify those not assigned and assign them using the assigned segments from the above steps by matching district, route type and day and find the one with the smallest distance if having multiple matches. Here we removed the restricting factors of route number and county.

7.1) Identify those not assigned and assign them using the assigned state highway segments from the above steps by matching district and day. Here we removed the restricting factors of route number, route type and county.

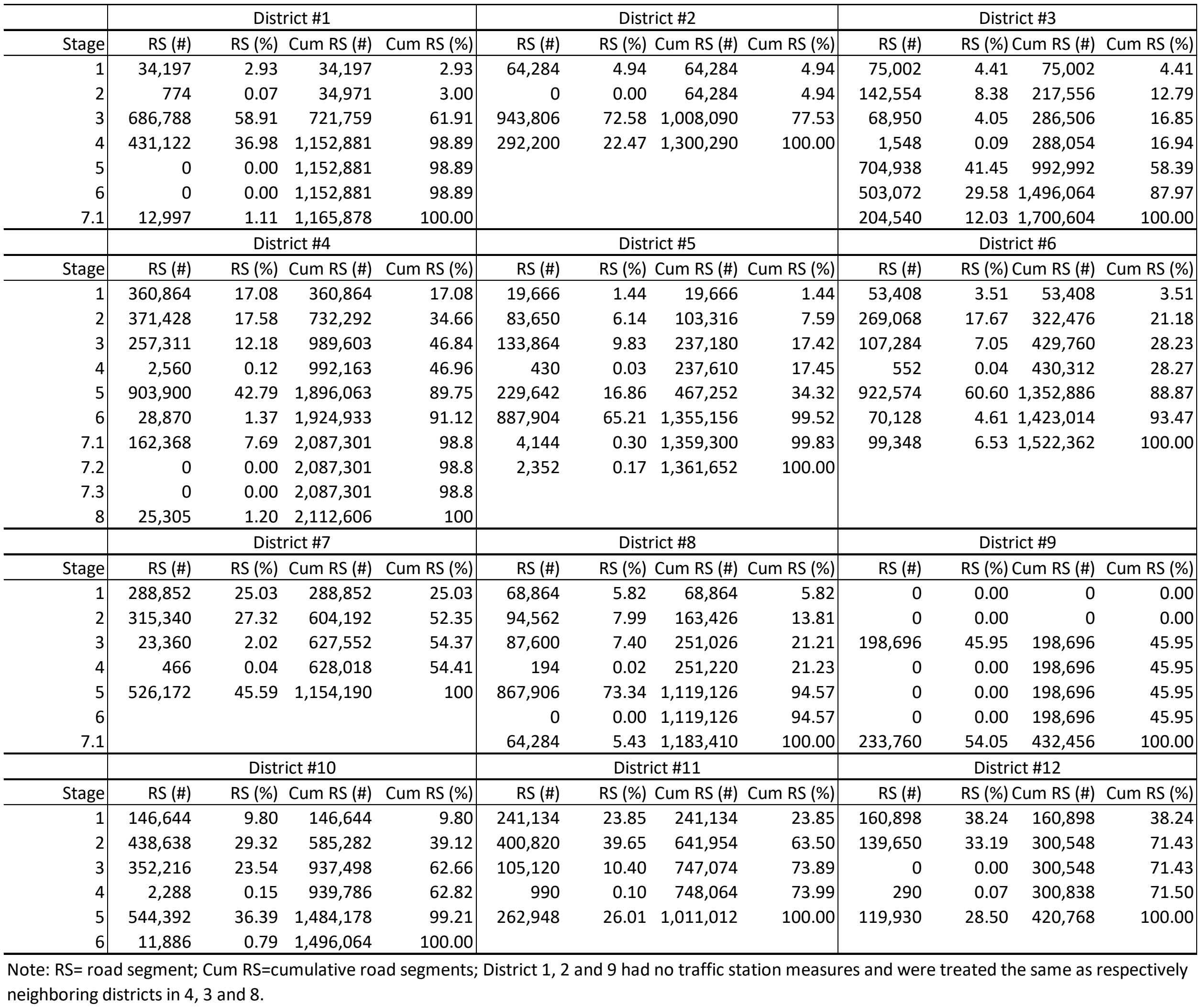
7.2) Identify those not assigned and assign them using the assigned U.S. highway segments from the above steps by matching district and day. Here we removed the restricting factors of route number, route type and county.

7.3) Identify those not assigned and assign them using the assigned interstate highway segments from the above steps by matching district and day. Here we removed the restricting factors of route number, route type and county.

8) Identify those not assigned and assign them using the assigned segments from steps 1-4 by matching district and season to find the one with the smallest distance if having multiple matches. Here route number, county and route type are not required to match.

Table 3 shows the daily traffic assignment statistics for the 12 California districts for years 2012-2019. Overall, 12.52% California highways had daily traffic measurements for the study period, with ranges being from 0% (district 9) to 38.24% (district 12). We found that the districts with great population (i.e., metropolitan areas) had more roadways and more traffic measures. Those districts thus had smaller proportions of roadways being assigned traffic from greatly relaxed conditions (e.g., by gradually relaxing matching criteria on route, county, district, route type or day). The roadways in the vastly rural districts were the ones with much less proportion of traffic measures. Greater proportion of roadways were thus assigned through greatly relaxed conditions for those rural districts.

Table 3. The daily road traffic assignment statistics for 12 Caltrans districts in California for years 2012-2019.



	District #1				District #2				District #3			
Stage	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)
1	34,197	2.93	34,197	2.93	64,284	4.94	64,284	4.94	75,002	4.41	75,002	4.41
2	774	0.07	34,971	3.00	0	0.00	64,284	4.94	142,554	8.38	217,556	12.79
3	686,788	58.91	721,759	61.91	943,806	72.58	1,008,090	77.53	68,950	4.05	286,506	16.85
4	431,122	36.98	1,152,881	98.89	292,200	22.47	1,300,290	100.00	1,548	0.09	288,054	16.94
5	0	0.00	1,152,881	98.89					704,938	41.45	992,992	58.39
6	0	0.00	1,152,881	98.89					503,072	29.58	1,496,064	87.97
7.1	12,997	1.11	1,165,878	100.00					204,540	12.03	1,700,604	100.00
	District #4				District #5				District #6			
Stage	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)
1	360,864	17.08	360,864	17.08	19,666	1.44	19,666	1.44	53,408	3.51	53,408	3.51
2	371,428	17.58	732,292	34.66	83,650	6.14	103,316	7.59	269,068	17.67	322,476	21.18
3	257,311	12.18	989,603	46.84	133,864	9.83	237,180	17.42	107,284	7.05	429,760	28.23
4	2,560	0.12	992,163	46.96	430	0.03	237,610	17.45	552	0.04	430,312	28.27
5	903,900	42.79	1,896,063	89.75	229,642	16.86	467,252	34.32	922,574	60.60	1,352,886	88.87
6	28,870	1.37	1,924,933	91.12	887,904	65.21	1,355,156	99.52	70,128	4.61	1,423,014	93.47
7.1	162,368	7.69	2,087,301	98.8	4,144	0.30	1,359,300	99.83	99,348	6.53	1,522,362	100.00
7.2	0	0.00	2,087,301	98.8	2,352	0.17	1,361,652	100.00				
7.3	0	0.00	2,087,301	98.8								
8	25,305	1.20	2,112,606	100								
	District #7				District #8				District #9			
Stage	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)
1	288,852	25.03	288,852	25.03	68,864	5.82	68,864	5.82	0	0.00	0	0.00
2	315,340	27.32	604,192	52.35	94,562	7.99	163,426	13.81	0	0.00	0	0.00
3	23,360	2.02	627,552	54.37	87,600	7.40	251,026	21.21	198,696	45.95	198,696	45.95
4	466	0.04	628,018	54.41	194	0.02	251,220	21.23	0	0.00	198,696	45.95
5	526,172	45.59	1,154,190	100	867,906	73.34	1,119,126	94.57	0	0.00	198,696	45.95
6					0	0.00	1,119,126	94.57	0	0.00	198,696	45.95
7.1					64,284	5.43	1,183,410	100.00	233,760	54.05	432,456	100.00
	District #10				District #11				District #12			
Stage	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)	RS (#)	RS (%)	Cum RS (#)	Cum RS (%)
1	146,644	9.80	146,644	9.80	241,134	23.85	241,134	23.85	160,898	38.24	160,898	38.24
2	438,638	29.32	585,282	39.12	400,820	39.65	641,954	63.50	139,650	33.19	300,548	71.43
3	352,216	23.54	937,498	62.66	105,120	10.40	747,074	73.89	0	0.00	300,548	71.43
4	2,288	0.15	939,786	62.82	990	0.10	748,064	73.99	290	0.07	300,838	71.50
5	544,392	36.39	1,484,178	99.21	262,948	26.01	1,011,012	100.00	119,930	28.50	420,768	100.00
6	11,886	0.79	1,496,064	100.00								

Note: RS= road segment; Cum RS=cumulative road segments; District 1, 2 and 9 had no traffic station measures and were treated the same as respectively neighboring districts in 4, 3 and 8.

Note: RS = road segment; Cum RS = cumulative road segments, or the total number of road segments being assigned traffic from previous stages. In situations when districts 1, 2, & 9 had no CalTrans traffic measures, we used their neighboring traffic measures, respectively, from districts 4, 3 & 8 to start traffic assignments.

NO_2 remote sensing data: For NO_2, UCB incorporated the tropospheric column densities of NO_2 from the Ozone Monitoring Instrument (OMI) NO_2 level-3 data (0.25°×0.25° resolution, which is about 27 kilometer [km] X 27km) to develop NO_2 daily surfaces.[41] The OMI satellite retrievals have nearly global coverage on a daily basis, with a local bypass time of 12:00 p.m. - 15:00 p.m. UCB applied quality screening criteria similar to Zhan et al.,[42] including terrain reflectivity < 30 percent, 0≤ solar zenith angle ≤85°, cloud fraction < 0.3, and a lack of row-anomalies. Temporal linear interpolation algorithms were then used to fill pixels with missing observations: (1) if observations existed at the location of the missing pixel one day before and after the missing pixel, the mean value from the two observations was used as the missing pixel value; (2) if only one observation existed at the location of the missing pixel one day before or after the missing pixel, the observed value was then used as the missing pixel value; (3) if the one day before and after search could not find an interpolated value, this process was extended to look for observations two days before and after. All the remaining missing pixels after a maximum of two days search were not interpolated and kept as missing observations. The predictors are resampled through a mean function to have a spatial resolution of 0.25°×0.25°. The prediction models will then be used to update NO_2 concentrations on tiles at dates without effective NO_2 measurements.

$PM_{2.5}$ remote sensing data: For $PM_{2.5}$, UCB obtained Aerosol Optical Depth (AOD) data from the Moderate Resolution Imaging Spectroradiometer instruments onboard the National Aeronautics and Space Administration Terra and Aqua satellites. The Multiangle Implementation of Atmospheric Correction algorithm was used to derive 1 km resolution AOD surfaces.[43] Due to extensive missing data presented at the 1 km resolution AOD surfaces, we further aggregated the daily AOD surfaces into monthly means.

O_3 remote sensing data: For daily O_3 remote sensing surfaces, UCB used the OMI/Aura Ozone (O_3) Differential Absorption Spectroscopy (DOAS) Total Column L3 1 day 0.25 degree x 0.25 degree V3 (OMDOAO3e) data.[44] This Level-3 global total column ozone product is derived from OMDOAO3 which is based on the Differential DOAS fitting technique that essentially uses the OMI visible radiance values between 331.1 and 336.1 nm. In addition to the total ozone column and its precision, OMDOAO3e also contains some ancillary parameters such as cloud fraction, cloud height, etc. Similar screening and interpolation algorithms as the one applied on NO_2 remote sensing data were used to reduce missing daily data.

Parcel-level land use data: UCB acquired statewide parcel data from CARB for 2017 for all the counties in California. The parcel data provides land use information at parcel level, such as agricultural, residential, commercial, industrial, government and institutions, open land, parks, and recreational facilities. For residential land use, the parcel data is further classed into single-family homes, town houses, condominiums, and high-rise apartment buildings. The parcel data also includes building characteristics, including building age, type and existence of fireplace, gas ranges, and other information that can be used to calculate building-specific factors to characterize the indoor infiltration of pollutants.

Land cover data: UCB acquired the land cover data for year 2016 from the National Land Cover Database (NLCD). The NLCD provides a synoptic nationwide classification of land cover into 16 classes at a spatial resolution of 30 m. The 16 land cover classes were aggregated into eight major land cover types including forest, herbaceous/grassland, shrubland, developed, agriculture, wetlands, water and other, which includes ice/snow, barren areas. UCB also acquired tree canopy and percent impervious surfaces for 2016.

Two-week interval vegetation index: UCB has acquired the 16-day interval (23 surfaces for a year) Normalized Difference Vegetation Index (NDVI) surfaces for California at a spatial resolution of 30 m[45] for years 2012 to May 2019 for the study.

GridMET meteorological data: UCB has acquired daily high-spatial resolution (~4-km, 1/24th degree) surface meteorological data covering the contiguous U.S. for years 2012-2019. Primary climate variables collected include maximum temperature, minimum temperature, precipitation accumulation, downward surface shortwave radiation, wind-velocity, humidity (maximum and minimum relative humidity and specific humidity).

Acquire and process other potential predictors

Digital elevation model (DEM) – in meters: UCB acquired the national elevation dataset for California from the U.S. Geological Survey (USGS) (https://www.usgs.gov/the-national-map-data-delivery) for 2011. The data included 45 1/3 arc-second (approx. 10 meters) raster DEM and were mosaicked into a single DEM raster for the entire State.

Distance to coast – in meters: The California shoreline was derived from The National Assessment of Shoreline Change: A Geographic Information System (GIS) Compilation of Vector Cliff Edges and Associated Cliff Erosion Data for the California Coast (http://pubs.usgs.gov/of/2007/1112). These data are integrated into the GIS mapping tool to produce a geographic view of topographical changes in California's coastline over time. The most recent view was created using data collected between 1998-2002.

Distance to roadways – in meters: UCB used Business Analysts 2010 Street Carto map layer provided by the Environmental Systems Research Institute (ESRI in Redlands, CA) to derive distance to nearest highway (defined as feature class classification or FCC A1 and A2), to nearest major roadway (FCC A3) and to nearest local roadway (FCC A4).

Distance weighted cargo volume – TEU/Km: UCB acquired monthly and annual cargos for the Oakland, Los Angeles (LA) and Long Beach (LB) ports from corresponding port authorities for 2003 – 2012 periods. The boundary layer of the three ports was acquired from CalTrans (California Department of Transportation) for 2011. We used twenty-foot equivalent units (TEU) statistics for the cargo volumes in the three ports. TEUs are a standardized maritime industry measurement used when counting cargo containers of varying lengths. Because of the adjacency of LA and LB ports, we merged monthly and annual cargos and corresponding spatial boundary and treated them as a single LA-LB port complex. TEUs were first adjusted by corresponding distances to the Oakland and LA-LB ports through TEUs/Km and then added up to create a single distance weighted cargo volume for each location of interest for each year from 2004 to 2010.

Location category – unitless: By using Business Analysts 2010 and the port boundary layer, UCB first separated the entire California roadway system into three parts: the first part includes locations within 500 m of FCC A1 or A2 allowing truck traffic, or within 500 m of any of the three ports (i.e., goods movement corridors or GMCs). The second part includes locations within 500 m of FCC A1 or A2 not allowing truck traffic or within 300 m of FCC A3 (i.e., non-goods movement corridors or NGMCs). The third part includes locations not encompassed in the first and second parts (i.e., control areas or CTRLs).

Acquire and process air pollution data (dependent variable in air pollution modeling)

NO_2 data from CalEPA monitoring: UCB downloaded and processed the California Environmental Protection Agency (CalEPA) daily pollutant concentrations of NO_2 for the period of 01/01/2012 – 12/31/2019 across the State of California (https://www.arb.ca.gov/aqmis2/aqdselect.php & https://aqs.epa.gov/aqsweb/airdata/download_files.html). The unique number of NO_2 monitoring stations for each year is listed in Table 4. The number of air quality monitoring stations for a specific year ranged from 106 to 113, with the minimum values below detection limit, the mean values close to 10 ppb and the maximum values over 60 ppb.

Table 4 CalEPA NO_2 monitoring stations statistics

Year	# Sites	Min (ppb)	Mean (ppb)	Max (ppb)
2012	109	1.00	10.61	66.65
2013	106	1.00	10.69	57.85
2014	109	1.00	10.16	63.69
2015	112	1.00	10	60.2
2016	113	1.00	9.71	143.62
2017	110	1.00	9.94	70.62
2018	107	1.00	9.82	71.67
2019	106	1.00	9.71	61.74

NO_2 data from research saturation monitoring: UCB also collected saturation-sampling data (Table 5) for NO_2 using Ogawa monitors (https://ogawausa.com/passive-sampler/) in the Los Angeles Metro for 2012-2013, in the Alameda County for 2012-2013, and in the Sacramento Metro for 2016. In total, we deployed NO_2 saturation sampling at 165 sites across three metropolitan areas throughout California. The Ogawa sampler is a high precision low-cost passive air monitoring sensor that has been widely used by scientific research to study air quality through saturation sampling technique.[46-48] The saturation monitoring was designed to densely sample air quality in a region, though in a shorter period of time compared to government continuous monitoring, for the purpose of measuring the small area variations in pollutant concentrations.

The minimum values were close to the detection limits and the higher ones ranged from 6.65 ppb to 33.16 ppb. Those two-week Ogawa measurements were disaggregated to the daily concentrations for each region by matching location category to the CalEPA NO_2 monitoring data:

$$Sat^{r}_{i,j,k} = Gov^{r}_{j,k} * Sat^{r}_{i,j} \,/\, Gov^{r}_{j} \qquad (1)$$

where $Sat^{r}_{i,j,k}$ refers to the derived NO_2 concentrations at site i, in location category j for day k in the two week sampling period in region r. $Gov^{r}_{j,k}$ and Gov^{r}_{j} refer to CalEPA all sites mean NO_2 concentrations in location category j, respectively, for day k and for two week period. $Sat^{r}_{i,j}$ refers to the Ogawa measured two-week NO_2 concentrations at site i and in location category j.

Table 5. The research saturation sampling for NO_2 conducted in Alameda, LA and Sacramento.

Region	Year	Month	# Sites	Min (ppb)	Mean (ppb)	Max (ppb)
Alameda County	2012	10	45	3.82	11.47	21.85
	2013	3	45	2.85	12.43	25.45
LA Metro	2012	10	57	4.72	21.33	32.17
	2013	3	59	4.28	20.74	33.16
Sacramento Metro	2016	5	61	0.54	3.24	6.65
	2016	12	64	3.64	8.68	12.51
Note: Saturation samplings were conducted using Ogawa for a period of 14 days.						

NO_2 data from Google Air mobile monitoring: UCB acquired Google Street Car Air monitoring data[49] which had measurements collected using four Google Street View vehicles equipped with the Aclima mobile measurement and data integration platform from 05/28/2015 to 06/07/2019. Not all four cars were actively mapping over the entire time frame. Table 6 lists the dates of operation for each of the four cars. Note that there may be gaps in the data when an individual car was not mapping due to operational, mechanical, or system difficulties.

Table 6: Start and end date each instrumented Street View cars

Car	Start Date	End Date
Car A	05/03/2016	04/30/2019
Car B	05/03/2016	06/08/2018
Car C	05/28/2015	06/07/2019
Car D	06/24/2015	11/05/2018

<u>Geographic location of data collection:</u> Data was collected in several geographic regions of California including the San Francisco Bay Area, Los Angeles, and the northern San Joaquin Valley. Mapping occurred in targeted neighborhoods or cities within these regions.

<u>Data Overview:</u> The dataset contains a table titled "California_Unified_2015_2019" which consists of the concentration of the pollutants O_3, NO_2, Nitrogen Monoxide (NO), Methane (CH_4), Carbon Dioxide (CO_2), Black Carbon (BC), $PM_{2.5}$, and Ultrafine Particles (UFP) measured using four Google Street View cars equipped with fast time-response, laboratory-grade instruments. The data was collected at 1-Hz time resolution from 05/28/2015 to 06/07/2019 for roads in three regions of California - the San Francisco Bay area, Los Angeles, and the northern San Joaquin Valley (Table 7). Specific areas mapped varied by region based on desired spatial data coverage and science questions. Each data point is geolocated with latitude and longitude as well as the identity and speed of the car.

Table 7. Date range during which each Street View car mapped in the different regions of California

Car ID	SF Bay Area	San Joaquin Valley	Los Angeles
Car A	May - Sept 2016 April - June 2017 Sept - Oct 2017 Jan 2018 - April 2019	Dec 2016 - April 2017 June 2017-Sept 2017 Nov-Dec 2017	Sept - Oct 2016
Car B	May - Sept 2016 April - June 2017 Sept - Oct 2017 Jan 2018 - June 2018	Dec 2016 - April 2017 June 2017-Sept 2017 Nov-Dec 2017	Sept - Oct 2016
Car C	June 2015 - June 2019		
Car D	June 2015 - Nov 2018		

Methods used for collection/generation of data: Four Google Street View vehicles were equipped with the Aclima proprietary mobile measurement and data integration platform. The vehicles were equipped with fast-response (1-Hz), laboratory-grade analyzers. NO_2 was measured using cavity-attenuation phase-shift spectroscopy and O_3 was measured using ultraviolet (UV) Absorption. Particle number concentration less than 2.5 micrometers in size ($PM_{2.5}$) was measured using an optical particle counter with particle counts reported in 5 size ranges - 0.3 to 0.5 µm, 0.5 to 0.7 µm, 0.7 to 1.0 µm, 1.0 to 1.5 µm, and 1.5 to 2.5 µm. Ultrafine particle counts were measured using a condensation particle counter.

Measurements and diagnostic parameters were logged from each instrument at 1-Hz via the onboard data acquisition and management system. Data was continuously streamed to backend servers where processing and storage occurred. Data collection occurred on weekdays during typical work hours, between ~9 AM to ~5 PM. Driving plans, dates and times of data collection varied by region.

Methods for processing the data: Timestamp adjustments to account for residence time in the tubing and instrument response was performed as documented in Apte, 2017 supplemental.[50] A snapping procedure was used to relocate each datapoint to the nearest road segment if needed.

Environmental/experimental conditions: No extreme weather conditions were encountered during this sampling period that prohibited safe data collection.

Quality-assurance procedures performed on the data: Algorithms developed at Aclima were used to identify and remove invalid data points using a combination of instrument status indicators, filtering, and other methods. These data points resulted from times when the instruments were performing internal processes designed to ensure overall instrument performance or if one of the instruments was operating out of specifications.

Selection of spatially non-correlated road segments for air pollution modeling:

The intensive sampling of air pollutants created a situation that the concentrations measured on the majority of road segments were geospatially auto-correlated. To reduce spatial auto-correlation, UCB applied location-allocation algorithm to select 150 road segments for each of the following four regions: Alameda and Contra Costa; San Francisco and San Mateo; Los Angeles; Central Valley. Each region has 50 road segments selected from GMCs, 50 selected from NGMCs and 50 selected from CTRLs.

For each region, we only maintained the road segments with more than 100 measurements (Figure 1 top left and bottom left). We then selected 50 road segments in the CTRLs (Figure 1 middle right and bottom right) through a location-allocation algorithm[51] that took into account variability in traffic pollution and the spatial distribution of residential land use. Briefly, the location-allocation algorithm involves a two-step algorithm that (1) builds a demand surface of spatial variation and (2) solves a constrained spatial optimization problem to determine locations for a pre-specified number of samplers – here 50 road segments. The demand surface was created using two criteria: first, samplers should be placed where the pollution surface is expected to exhibit high spatial variability (Figure 1 middle left), and second, population density should be relatively high (Figure 1 top right). To create an initial pollution surface across a region, we applicated the statewide land use regression (LUR) surfaces generated from our Health Effects Institute (HEI) grant.[52] We also required those 50 sites selected in CTRLs should be located in residential area (Figure 1 top right).

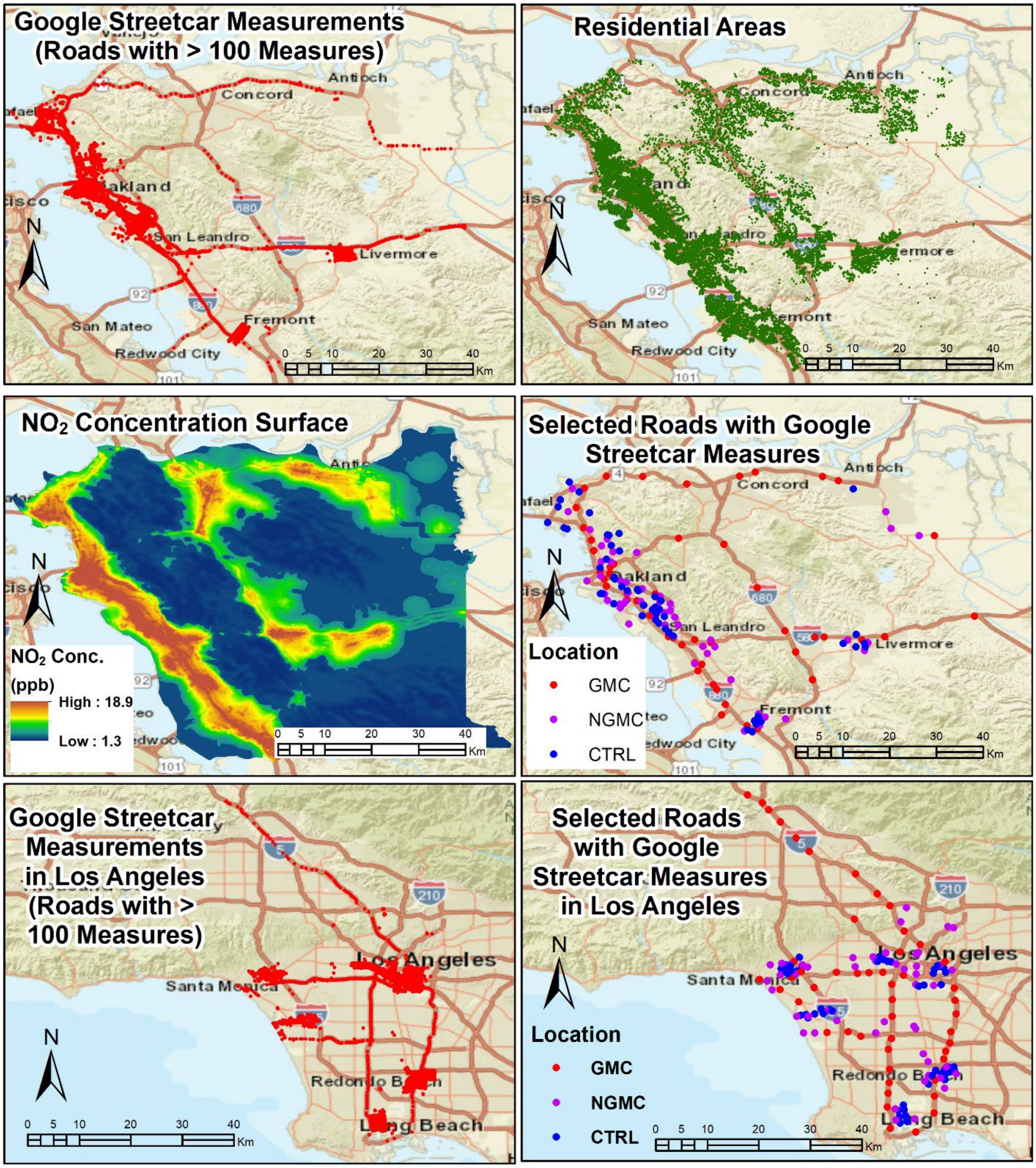

Google Streetcar Measurements (Roads with > 100 Measures)
Residential Areas
NO₂ Concentration Surface
NO₂ Conc. (ppb)
High : 18.9
Low : 1.3
Selected Roads with Google Streetcar Measures
Location
GMC
NGMC
CTRL
Google Streetcar Measurements in Los Angeles (Roads with > 100 Measures)
Selected Roads with Google Streetcar Measures in Los Angeles
Antioch
Concord
Oakland
San Leandro
Livermore
Fremont
San Mateo
Redwood City
Los Angeles
Santa Monica
Redondo Beach
Long Beach
N
0 5 10 20 30 40 Km

Figure 1. Location-allocation algorithm in selecting road segments with Google Streetcar air quality measurements for reduced spatial auto-correlation. The top and middle rows are for selection of 150 road segments with the Google Streetcar measures in the Alameda and Contra Costa counties. The bottom row is for selection of 150 road segments with the Google Streetcar measures in the Los Angeles region.

Those selected 50 road segments were then used in a second location-allocation algorithm as pre-determined fixed sites and we further selected another 50 road segments in the NGMCs (Figure 1d) with only one demand surface – the surface of NO_2 concentrations from the HEI grant – to reflect the spatial variability requirement. The selected 100 road segments were then further used in a third location-allocation algorithm as pre-determined fixed sites, and we further selected another 50 road segments in the GMCs (Figure 1d) with the surface of NO_2 concentrations from the HEI grant as the demand surface. A total 150 road segments were selected for each of the four regions (Figure 1e for Alameda-Contra Costa and Figure 1f for Los Angeles).

The Google Street View cars operated in a region followed similar vehicle driving pattern and sequence from one day to the other, creating a situation that some road segments, for example, were always sampled in the rush hours while others sampled in the non-rush hours. The measured concentrations do not reflect the daily mean pollutant concentrations generated on a road segment. Due to this reason, we applied detrending algorithm to convert Google Air measured hourly concentrations to corresponding daily concentrations through the following algorithm:

$$Adj_{i,j,k,d}^{r} = Gov_{j,d}^{r}/Gov_{j,k,d}^{r} * Google_{i,j,k,d}^{r} \qquad (2)$$

where $Adj_{i,j,k}^{r}$ refers to the daily detrended NO_2 concentrations at road segment i, in location category j during hour k of day d in region r. $Gov_{j,k,d}^{r}$ and $Gov_{j,d}^{r}$ refer to CalEPA all sites mean NO_2 concentrations in location category j, respectively, during hour k in day d and 24 hour mean in day d. $Google_{i,j,k,d}^{r}$ refers to the Google Air measured NO_2 concentrations on road segment i , in location category j during hour k in day d. Their adjusted concentrations on a road segment for a day were aggregated to daily concentrations means and used in our LUR modeling process together with those measured through CalEPA monitoring and through our research saturation sampling. Table 8 shows the statistics of the adjusted daily pollutant concentrations of NO_2 on the selected road segments. Some location categories do not have full 50 road segments, due to the restriction of the location-allocation algorithm in selecting a further road segment when the neighboring road segment is also chosen. We found that the mean concentrations measured by Google Air are similar to CalEPA monitored values, but with the maximum values much higher, probably due to the fact that Google Air measured NO_2 concentrations directly from close-range road vehicle emissions.

Table 8. The daily pollutant concentrations statistics of NO_2 on the selected road segments.

Region	Location	# Segments	Min (ppb)	Mean (ppb)	Max (ppb)
Alameda - Contra Costa	GMC	49	1.04	22.89	102.74
	NGMC	49	0.09	25.4	121.99
	CTRL	50	0.48	17.26	145.87
LA Metro	GMC	50	3.65	48.26	260.25
	NGMC	50	2.15	32.28	145.7
	CTRL	50	0.51	22.39	138.39
Central Valley	GMC	50	0.01	29.13	262.89
	NGMC	50	0.55	16.95	211.17
	CTRL	49	0.24	8.8	67.26
San Francisco - San Mateo	GMC	49	0.01	19.81	117.24
	NGMC	50	0.13	18.65	209.34
	CTRL	50	0.07	16.3	213.91

$PM_{2.5}$ data from CalEPA monitoring: UCB downloaded and processed CalEPA daily pollutant concentrations of $PM_{2.5}$ for the period of 01/01/2012 – 12/31/2019 across the State of California. The unique number of $PM_{2.5}$ monitoring stations for each year is listed in Table 9. The number of air quality monitoring stations for a specific year ranged from 112 to 116, with the minimum values below detection limit, the mean values close to 10 ug/m3 and the maximum values over 500 ug/m3. Though Google Air also measured $PM_{2.5}$ concentrations, they were measured by five binned particle counts, not mass. The Google Air $PM_{2.5}$ measurements were thus not used in our study.

Table 9. CalEPA $PM_{2.5}$ monitoring stations statistics

Year	# Sites	Min (ug/m3)	Mean (ug/m3)	Max (ug/m3)
2012	114	1.00	9.26	168.30
2013	112	1.00	10.36	167.30
2014	116	1.00	9.48	190.25
2015	112	1.00	9.35	270.17
2016	115	1.00	8.57	104.79
2017	116	1.00	9.85	557.08
2018	120	1.00	10.99	411.70
2019	117	1.00	7.48	98.92

O_3 data from CalEPA monitoring: UCB downloaded and processed CalEPA daily pollutant concentrations of O_3 for the period of 01/01/2012 – 12/31/2019 across the State of California. The unique number of O_3 monitoring stations for each year is listed in Table 10. The number of air quality monitoring stations for a specific year ranged from 178 to 194, with the minimum values below detection limit, the mean values close to 32 ppb and the maximum values over 100 ppb.

Table 10. CalEPA O_3 monitoring stations statistics

Year	# Sites	Min (ppb)	Mean (ppb)	Max (ppb)
2012	194	0.00	32.89	91.65
2013	187	0.00	31.69	90.06
2014	186	0.00	32.87	92.94
2015	186	0.00	32.70	100.82
2016	186	0.00	32.77	95.77
2017	185	0.00	33.57	99.00
2018	181	0.00	33.11	96.41
2019	178	0.00	32.63	94.53

O_3 data from research saturation monitoring: UCB also collected saturation-sampling data for O_3 using Ogawa monitors in the Sacramento Metro for May and December 2016 (Table 11). In total, we deployed O_3 saturation sampling at 39 sites, with lower values identified in December and higher identified in May. The measured two-week values ranged from 6.58 ppb in December to 37.77 ppb in May. Those two-week Ogawa measurements were disaggregated to the daily concentrations by matching location category to the CalEPA O_3 monitoring data in a way similar to eq. (1).

Table 11. The research saturation sampling for Sacramento Metro.

Region	Year	Month	# Sites	Min (ppb)	Mean (ppb)	Max (ppb)
Sacramento Metro	2016	5	39	17.86	31.53	37.77
	2016	12	38	6.58	9.95	28.12
Note: Saturation samplings were conducted using Ogawa for a period of 14 days.						

O_3 data from Google Air mobile monitoring: Google Air mobile monitoring data for O_3 was collected at the same time with those for NO_2. Please refer to section "NO_2 data from Google Air mobile monitoring" for detail. We applied a procedure similar to those of NO_2 in selecting road segments and applying diurnal data detrending for final air pollution modeling.

Table 12 list the measured O_3 statistics across four regions in California. Opposite to what we identified for NO_2, O_3 concentrations were normally greater in CTRL than in GMC. Overall, the low values were at near detection limit and high values were 50-60 ppb.

Table 12. The daily pollutant concentrations statistics of O_3 on the selected road segments.

Region	Location	# Sites	Min (ppb)	Mean (ppb)	Max (ppb)
Alameda - Contra Costa	GMC	48	1.23	20.50	45.20
	NGMC	42	1.18	18.92	44.31
	CTRL	35	1.62	26.40	41.94
LA Metro	GMC	50	0.32	16.60	55.59
	NGMC	49	0.99	19.14	42.13
	CTRL	50	3.26	22.94	59.01
Central Valley	GMC	47	2.78	20.08	48.61
	NGMC	48	5.54	22.79	48.27
	CTRL	46	9.29	27.21	52.75
San Francisco - San Mateo	GMC	50	0.74	18.77	48.26
	NGMC	50	0.92	19.64	56.88
	CTRL	47	0.42	23.98	62.42

Comprehensive Data Sources for Trace Metals Modeling

Acquire and process potential trace metals predictors

The data sources used as potential predictors for NO_2 modeling were also used in our trace metals modeling, including remote sensing data, land use and land cover data, road traffic data and meteorological data. Further we added road surface and slope gradient data in enhancing modeling of trace metals.

Surface roughness and pavement type: Given that surface roughness and pavement type have significant impacts on emissions of pollutants, especially for non-tailpipe vehicle emissions from tire- and brake-wear, UCB used the road International Roughness Index (IRI) and surface type data provided by the Highway Performance Monitoring System (HPMS) for all the public roadways in California for 2017 to model trace metals. IRI is obtained from measured longitudinal road profiles and calculated using a quarter-car vehicle math model and the response is accumulated to yield a IRI with units of slope (inches [in]/miles [mi] and m/km,). California roadways have a mean IRI of 80 in/mi with the maximum being 400 in/mi and the pavement surface types include unpaved local roadways, asphalt, jointed concrete, and continuously reinforced concrete.

Roadway slope gradient: Because uphill tire-wear and downhill brake-wear have a significant impact on emissions of trace metals, UCB also used the statewide digital elevation models (DEM) at 30 m resolution to derive road network slopes for highways and major roadways in California. Due to the stop-and-go nature at intersections, including highway ramps, UCB also generated intersection points of highway and major roadways across California and used it as a potential predictor for trace metals modeling.

Acquire and process trace metals pollutants data

Trace metals data from CalEPA monitoring: UCB has collected daily fine PM trace metals data, including Cr, Mn, Ni, Zn, Se and Pb from 48 CalEPA air quality monitoring stations for the 2012-2019 period. Because the trace metals were measured every 4th or 5th day in a week by CalEPA and there were significant number of measurements were zero (i.e., below detection limit), we opted to aggregate them into monthly means for our analysis. Table 13 displays the number of sites with effective measurements for each year and the monthly statistics in minimum, maximum and mean values.

Table 13. CalEPA trace metals monitoring statistics.

Year	# Sites	Min (ug/m3)	Mean (ug/m3)	Max (ug/m3)	Year	# Sites	Min (ug/m3)	Mean (ug/m3)	Max (ug/m3)
Chromium (Cr)					Nickel (Ni)				
2012	41	0	0.0012	0.0218	2012	41	0	0.0007	0.0150
2013	39	0	0.0015	0.0303	2013	39	0	0.0006	0.0111
2014	38	0	0.0009	0.0192	2014	38	0	0.0006	0.0061
2015	35	0	0.0009	0.0146	2015	35	0	0.0006	0.0064
2016	34	0	0.0008	0.0121	2016	34	0	0.0006	0.0074
2017	32	0	0.0010	0.0094	2017	32	0	0.0006	0.0122
2018	32	0	0.0013	0.0377	2018	32	0	0.0005	0.0117
2019	32	0	0.0011	0.0103	2019	32	0	0.0005	0.0026
Manganese (Mn)					Zinc (Zn)				
2012	41	0	0.0018	0.0131	2012	41	0	0.0074	0.1678
2013	39	0	0.0019	0.0122	2013	39	0	0.0057	0.0967
2014	38	0	0.0018	0.0076	2014	38	0	0.0056	0.1435
2015	35	0	0.0015	0.0130	2015	35	0	0.0090	0.2812
2016	34	0	0.0016	0.0113	2016	34	0	0.0087	0.1610
2017	32	0	0.0015	0.0180	2017	32	0	0.0083	0.6118
2018	32	0	0.0017	0.0110	2018	32	0	0.0093	0.2274
2019	32	0	0.0016	0.0128	2019	32	0	0.0076	0.1716
Selenium (Se)					Lead (Pb)				
2012	41	0	0.0005	0.0094	2012	41	0	0.0015	0.0218
2013	39	0	0.0005	0.0151	2013	39	0	0.0011	0.0110
2014	38	0	0.0005	0.0108	2014	38	0	0.0013	0.0316

2015	35	0	0.0004	0.0196	2015	35	0	0.0015	0.0354
2016	34	0	0.0004	0.0090	2016	34	0	0.0015	0.0192
2017	32	0	0.0004	0.0287	2017	32	0	0.0014	0.0417
2018	32	0	0.0004	0.0268	2018	32	0	0.0016	0.0258
2019	32	0	0.0002	0.0052	2019	32	0	0.0014	0.0234

Trace metals data from research saturation monitoring: UCB acquired trace metals data from the University of California, Los Angeles (UCLA) Dr. Michael Jerrett's research lab, which measured the six trace metals proposed in our research. The data were collected in September 2019 and February 2020 through gravimetric analysis, each for a period of 14 days. The data included both fine and coarse size fractions and we used the size of fine particles for our research. We identified that 24 sites of data were available for our analysis for the fall and 26 sites of data were available for the winter period (Table 14). Because the trace metals were sampled through gravimetric analysis, in a way different from traditional CalEPA daily measurements, all the measured two-week trace metals concentrations data were corrected to corresponding CalEPA monthly values by the ratios of research gravimetric analysis to the CalEPA real-time measurements through a site that was co-located with a CalEPA site.

Table 14. Trace metals saturation monitoring statistics.

Year	Month	Pollutant	# Sites	Min (ug/m3)	Mean (ug/m3)	Max (ug/m3)
2019	9	Cr	24	0.21	1.29	2.38
2019	9	Mn	24	0.55	2.78	7.19
2019	9	Ni	24	0.11	0.51	0.84
2019	9	Zn	24	1.22	6.86	12.49
2019	9	Se	24	0.01	0.3	0.56
2019	9	Pb	24	0.26	1.21	2.68
2020	2	Cr	26	0.92	1.74	2.88
2020	2	Mn	26	1.38	3.98	6.56
2020	2	Ni	26	0.22	0.62	1.51
2020	2	Zn	26	3.93	11.48	19.7
2020	2	Se	26	0.08	0.4	0.81
2020	2	Pb	26	1.12	2.65	10.59

Note: Concentrations were calculated through gravimetric analysis for a two-week period.

Subtask 2.2: Daily Hybrid LUR Models and Air Pollution Surfaces for Criteria Pollutants

UCB developed daily LUR models using the potential predictors described above for NO_2, $PM_{2.5}$, and O_3 across California at a spatial resolution of 30m (more than 1 gigabyte (GB) for a raster). To save storage space and be able to assign exposures to space-time rescue mediation use and activity space, the daily surfaces were built for a spatial resolution of 100m (about 400 megabytes (MBs)), which still maintain identifying the small area variations of pollutant concentrations in vulnerable communities.

Generate buffer statistics on 30 m spatial resolution potential predictors

A series of buffer statistics of 50-5000 m at an interval of 50 m were created for the potential spatial predictors with a spatial resolution of 30 m. They include daily traffic data, every two-week NDVI data, parcel-level land use data, NLCD land cover data, and NLCD % impervious and tree-canopy data. For each variable, e.g., industrial land use, a total of 100 buffered statistics (i.e., covariates) were generated. For all the potential predictors, with the inclusion of buffered and non-buffered variables, about 2500 covariates were identified for the prediction of a single pollutant concentrations at daily level. This increases the chance of identifying the optimal distance impact of a predictor and helps improve model performance. However, this also creates high-dimension covariates that are highly correlated. To solve this issue, we applied data reduction strategy to reduce the number of covariates used in predicting a pollutant concentration.

Apply data reduction strategy to reduce potential number of predictors

To reduce the number of covariates and avoid high correlations between them for LUR modeling, we first created a correlation coefficient matrix between a pollutant and all the covariates. The absolute correlation coefficients between the covariate of the highest correlation with the pollutant and all the remaining covariates were calculated, and the covariates with an absolute correlation coefficient greater than or equal to 0.9 were removed. This process continued until no absolute correlation coefficient was greater than 0.9 between any remaining covariates. After applying the data reduction strategy, we maintained the number of predictors in a LUR model to be less than 100.

Develop daily LUR models and surfaces for the criteria pollutants

In developing daily LUR models for the criteria pollutants, we aimed at developing the models at its finest spatial resolution: 30 m. We also aimed to identify the optimal distance of impact for a potential predictor and the models should be able to deal with multicollinearity among predictors and can reduce model overfit. Further, we wanted to avoid excessive number of predictors in the final models and allowed a maximum of 15 predictors (in addition to four Seasons) in a LUR model. Due to those considerations, we applied the D/S/A algorithm for our daily prediction models.[53,54] The D/S/A algorithm is an aggressive model search algorithm, which iteratively generates polynomial generalized linear models based on the existing terms in the current 'best' model and the following three steps: (1) a deletion step, which removes a term from the model, (2) a substitution step, which replaces one term with another, and (3) an addition step, which adds a term to the model. The search for the 'best' estimator starts with the base model specified with 'formula': typically, the intercept model except when the user requires number of terms to be forced in the final model. Before searching through the statistical model space of polynomial functions, the original sample is randomly partitioned into V equal size subsamples. Of the V subsamples, a subsample is retained as the validation data for testing the model, and the remaining V-1 subsamples are used as training data. The cross-validation process is then repeated V times, with each of the V subsamples used exactly once as the validation data. The advantage of this method over the leave-one-out cross-validation technique is that the prediction errors are less impacted by single outliers, and compared to repeated random sub-sampling, all observations in the V-folds are used for both training and validation, and each observation is used for validation once. With each iteration, an independent validation dataset is used to assess the performance of a model built using a training dataset. This technique, therefore, minimizes over-fitting to the data to maximize the probability that the models will predict well at locations that have not been sampled. The D/S/A algorithm can deal with both linear and non-linear associations. However, for simplicity of model development and for the clear interpretation of the predictors selected for a model, we limited the predictors to be only on linear terms (the maximum sum of powers in each variable to be 1) and disallowed any interaction except corridor by year.

For NO_2, the LUR was developed using the CalEPA daily NO_2 monitoring data, our research saturation monitoring data, and the Google Air data UCB collected as a response variable. To integrate the three types of air quality measurements, we divided each type of air quality monitoring data equally into 10-folds and then merged corresponding folds of data into a large 10-fold dataset, with each fold having equal presentation of the three types of air quality monitoring data. The predictors include daily traffic data, daily remote sensing pollutant data, daily weather conditions, every 2-week NDVI index, one-time land use and land cover, and other traditional geographic features like DEM, distance to highways, major roadways, and ports. Due to seasonal changes, weekday and weekend variations, and annual reductions trend for some pollutants, UCB also included weekend, season, and year as predictors. Optimal distance of impact was first estimated and LUR run with optimal variables selected through a D/S/A algorithm. Table 15 shows the daily LUR model developed for NO_2 for the State of California for years 2012-2019. The adjusted R^2 is 79.6%.

Due to the requirement of more than 3 GBs of storage space for a single statewide raster surface of spatial resolution of 30 m, we opted to build daily surfaces of NO_2 concentrations using a spatial resolution of 100 m. Rather than creating a series of daily surfaces using a storage space of 9 TBs for the 2922 days for years 2012-2019, the 100 m spatial resolution surfaces required a storage space of less than 1.5 TBs. The reduced size of the daily NO_2 surfaces also made exposure assignments feasible through an Amazon SageMaker cloud platform. The 100 m spatial resolution surfaces still maintain the ability to identify small area variations of pollutant concentrations, especially those heightened exposures endured by vulnerable communities. Figure 2 displays developed samples of daily and aggregated monthly, and annual NO_2 surfaces for the State of California for year 2012.

The daily $PM_{2.5}$ and O_3 models were developed separately; but, in a way, similar to the model used for deriving daily NO_2 concentrations, with $PM_{2.5}$ air quality data measured only from CalEPA monitoring data and O_3 data measured from CalEPA, our research saturation monitoring and Google Air monitoring. In developing daily $PM_{2.5}$ models and surfaces, monthly median rather than daily AOD values were used as a potential predictor due to extensive missing AOD values on daily measurements; while for both NO_2 and O_3, daily remote sensing data were applied as potential predictors. Table 16 & Table 17 show the daily LUR models developed, respectively, for $PM_{2.5}$ and O_3. The $PM_{2.5}$ and O_3 models explained, respectively, 65.3% and 93.6% variance in predicted concentrations. Figure 3 and Figure 4 display, correspondingly, samples of daily and aggregated monthly, and annual surfaces for $PM_{2.5}$ and O_3.

Table 15. Daily NO_2 model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	41.31525359	1.27380322	32.43456526	**<0.001**
Season [Spring]	37.88236015	1.27857987	29.62846605	**<0.001**
Season [Summer]	37.84886991	1.30048646	29.10362477	**<0.001**
Season [Winter]	42.05274917	1.25767472	33.43690420	**<0.001**
NDVI	-0.00018156	0.00001679	-10.81666470	**<0.001**
Week [Weekend]	-2.32441019	0.03671236	-63.31410266	**<0.001**
Distance to ports (m)	-0.00000584	0.00000024	-23.88315011	**<0.001**
NO2 from OMI	0.00000000	0.00000000	135.03776831	**<0.001**
VKT (350m)	0.00006147	0.00000071	86.71347177	**<0.001**
Developed high intensity (ha) (5000m)[†]	0.00017474	0.00000231	75.62641371	**<0.001**
Minimum relative humidity (%)	-0.12444801	0.00102569	-121.3315915	**<0.001**
Wind velocity at 10m (m/s)	-0.93918093	0.01122917	-83.63763975	**<0.001**
Roadway area (ha) (50m)	6.29933371	0.10347870	60.87565365	**<0.001**
Minimum temperature (K)	-0.09471578	0.00443069	-21.37721199	**<0.001**
Percent impervious (%) (50m)	0.01781697	0.00091568	19.45772955	**<0.001**
Developed low intensity (ha) (400m)	0.01218760	0.00020753	58.72813594	**<0.001**
Shrubs (ha) (3250m)	-0.00009070	0.00000304	-29.82997246	**<0.001**
Water (ha) (50m)	-1.93161136	0.07029621	-27.47817093	**<0.001**
Developed open space (ha) (50m)	-0.19145437	0.01075227	-17.80594787	**<0.001**
Residential (ha) (350m)	-0.07513870	0.00236946	-31.71130369	**<0.001**
Precipitation amount (mm, daily total)	0.04020234	0.00378600	10.61868762	**<0.001**

Wetlands (ha) (550m)	-0.02732793	0.00117326	-23.29238367	**<0.001**
Observations	162570			
R^2 / R^2 adjusted	0.796 / 0.796			

Note: NDVI = Normalized Difference Vegetation Index.

OMI = Ozone Monitoring Instrument.

VKT = Vehicle Km Traveled.

†: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.

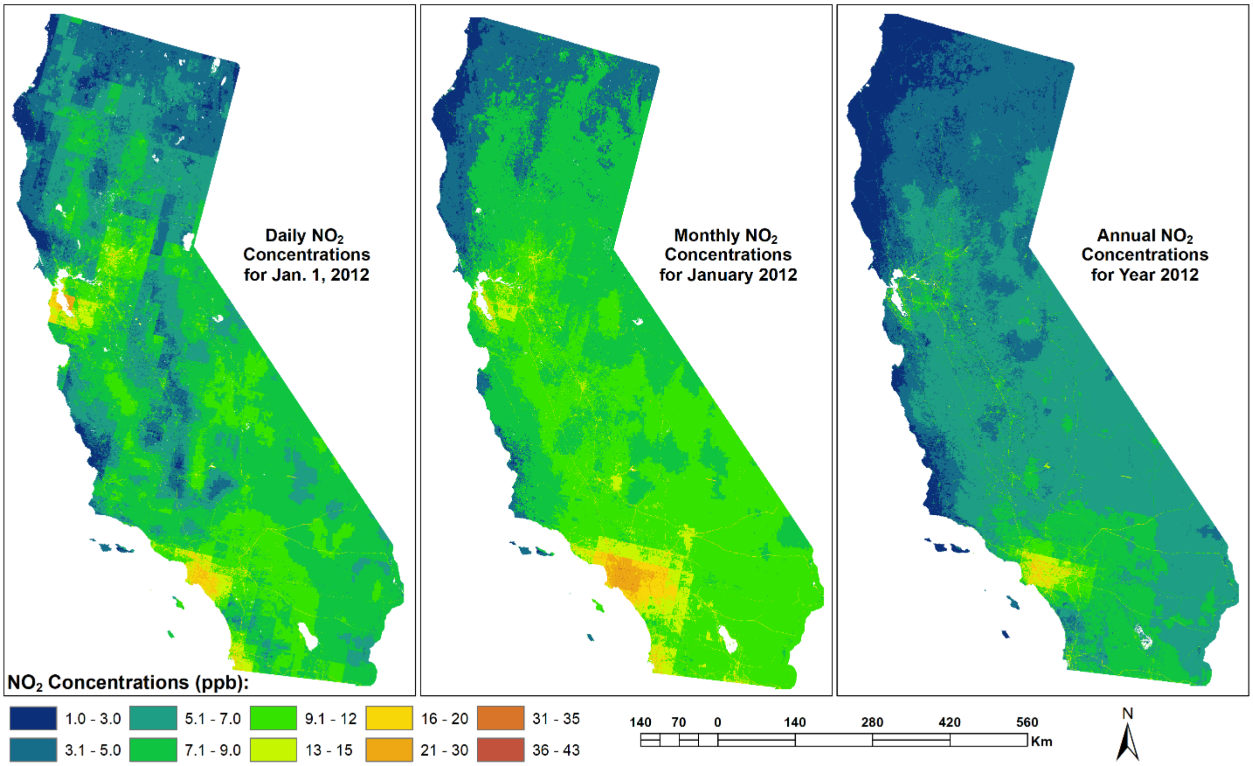

Daily NO₂ Concentrations for Jan. 1, 2012
Monthly NO₂ Concentrations for January 2012
Annual NO₂ Concentrations for Year 2012
NO₂ Concentrations (ppb):
1.0 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 9.0
9.1 - 12
13 - 15
16 - 20
21 - 30
31 - 35
36 - 43
140 70 0 140 280 420 560
Km
N

Figure 2. Example of daily, monthly and annual NO_2 concentration surfaces (ppb) for year 2012.

Table 16. Daily $PM_{2.5}$ model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	90.21122937	1.04163758	86.60519819	**<0.001**
Season [Spring]	88.15829132	1.04862676	84.07022866	**<0.001**
Season [Summer]	89.58297126	1.06433106	84.16833306	**<0.001**
Season [Winter]	90.66738819	1.02822904	88.17820237	**<0.001**
AOD[†]	0.03232083	0.00014388	224.64103333	**<0.001**
Wind velocity at 10m (m/s)	-0.91353396	0.00935806	-97.62006415	**<0.001**
Roadway area (ha) (5000m) [§]	0.00057177	0.00002919	19.58814640	**<0.001**
Minimum temperature (K)	-0.27202880	0.00361147	-75.32355385	**<0.001**
Minimum relative humidity (%)	-0.10749589	0.00109883	-97.82789738	**<0.001**
DEM (m)	-0.00355748	0.00006678	-53.26864451	**<0.001**
Industrial (ha) (1850m)	0.00997592	0.00032603	30.59788510	**<0.001**
Distance to ports (m)	0.00001155	0.00000027	42.05332038	**<0.001**
Residential (ha) (850m)	0.00904293	0.00040292	22.44333719	**<0.001**
VKT (350m)	0.00000772	0.00000073	10.62487610	**<0.001**
NDVI	-0.00035052	0.00001309	-26.76815385	**<0.001**
Barren land (ha) (3000m)	-0.00073488	0.00002057	-35.73111153	**<0.001**
Shrubs (ha) (200m)	-0.01737372	0.00087123	-19.94171252	**<0.001**
Location category[‡]	-0.39053840	0.02256090	-17.31041212	**<0.001**
Developed open space (ha) (4950m)	-0.00007838	0.00000264	-29.65769646	**<0.001**
Unknow land use (ha) (450m)	-0.04719305	0.00195384	-24.15395733	**<0.001**
Agricultural (ha) (50m)	-2.88221319	0.13664311	-21.09300113	**<0.001**
Observations	310720			
R^2 / R^2 adjusted	0.653 / 0.653			

Note: VKT = Vehicle Km Traveled.
NDVI = Normalized Difference Vegetation Index.

DEM = Digital Elevation Model.
[†]: AOD = Aerosol Optical Depth and monthly median values were used.
[‡]: Location category: 1 = GMC; 2 = NGMC and 3=CTRL.
[§]: The first paired parenthesis refers to unit of analysis and the second pair refers to circular buffer distance.

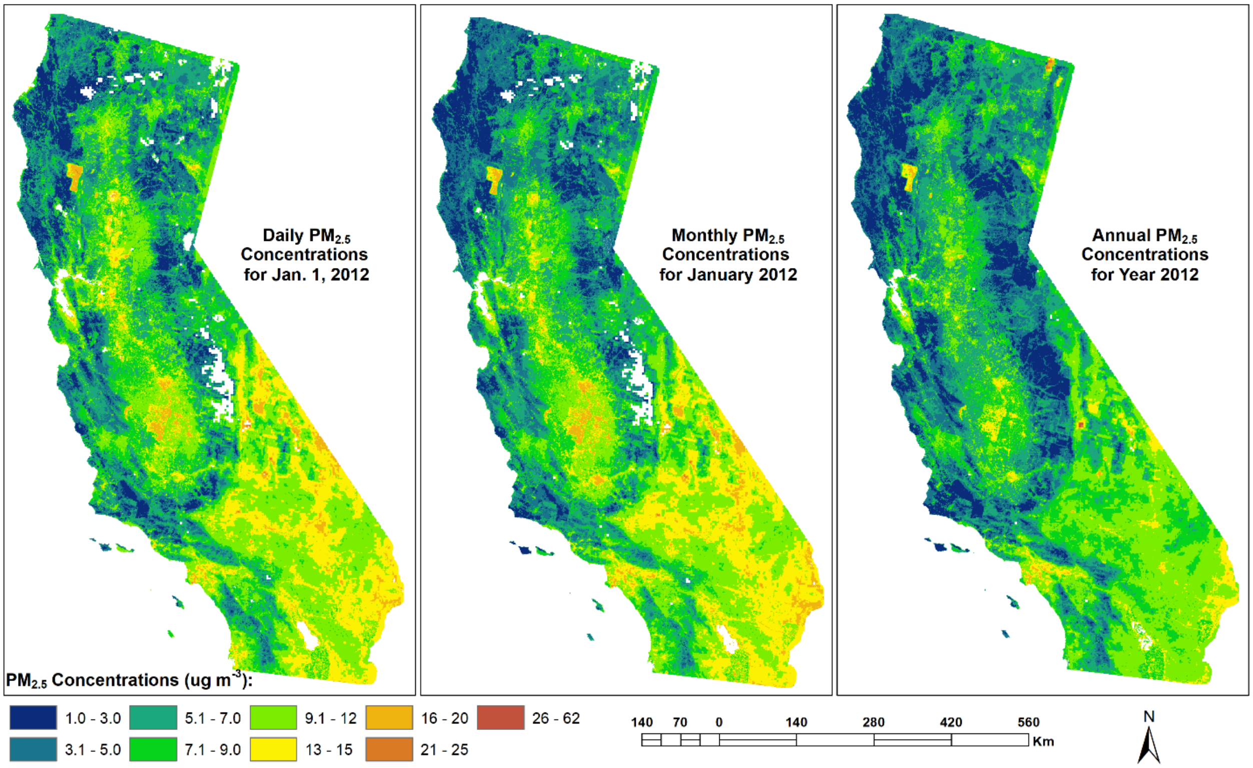

Daily $PM_{2.5}$ Concentrations for Jan. 1, 2012
Monthly $PM_{2.5}$ Concentrations for January 2012
Annual $PM_{2.5}$ Concentrations for Year 2012
$PM_{2.5}$ Concentrations (ug m^{-3}):
1.0 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 9.0
9.1 - 12
13 - 15
16 - 20
21 - 25
26 - 62
140 70 0 140 280 420 560
Km
N

Figure 3. Example of daily, monthly and annual $PM_{2.5}$ concentration surfaces (ug m^{-3}) in year 2012.

Table 17. Daily O_3 model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	-381.08768502	15.42996891	-24.69789066	**<0.001**
Season [Spring]	-374.82273040	15.42985672	-24.29204218	**<0.001**
Season [Summer]	-380.45822233	15.43082396	-24.65572955	**<0.001**
Season [Winter]	-381.66985531	15.42936058	-24.73659574	**<0.001**
Year	0.14050221	0.00769760	18.25271874	**<0.001**
Week [Weekend]	1.35484176	0.03871875	34.99187912	**<0.001**
VKT (350m)	-0.00006042	0.00000112	-53.96616312	**<0.001**
Vapor Pressure (kPa)	5.25136585	0.02702161	194.33947191	**<0.001**
DEM (m)	0.00988294	0.00005987	165.06931005	**<0.001**
O3 from OMI	0.04894654	0.00074407	65.78241812	**<0.001**
Minimum temperature (K)	0.37555432	0.00529035	70.98858639	**<0.001**
Water (ha) (700m)[†]	0.00765764	0.00021231	36.06740612	**<0.001**
Wind velocity at 10m (m/s)	0.67016975	0.01116061	60.04778299	**<0.001**
Barren land (ha) (250m)	-0.04829488	0.00222738	-21.68235173	**<0.001**
Crops (ha) (5000m)	-0.00003294	0.00000075	-44.11509667	**<0.001**
Developed high intensity (ha) (100m)	-0.16420518	0.00248647	-66.03947344	**<0.001**
Wetlands (ha) (1600m)	-0.00283386	0.00006554	-43.23768595	**<0.001**
Government & Institutional (ha) (1800m)	-0.00408827	0.00009946	-41.10571087	**<0.001**
Developed low intensity (ha) (200m)	-0.04221532	0.00079418	-53.15572993	**<0.001**
Developed medium intensity (ha) (150m)	-0.06929079	0.00096769	-71.60415741	**<0.001**
Commercial (ha) (3200m)	0.00284461	0.00014036	20.26675566	**<0.001**
Observations	258575			

R^2 / R^2 adjusted	0.936 / 0.936

Note: VKT = Vehicle Km Traveled.
OMI = Ozone Monitoring Instrument.
DEM = digital elevation model.
[†]: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.

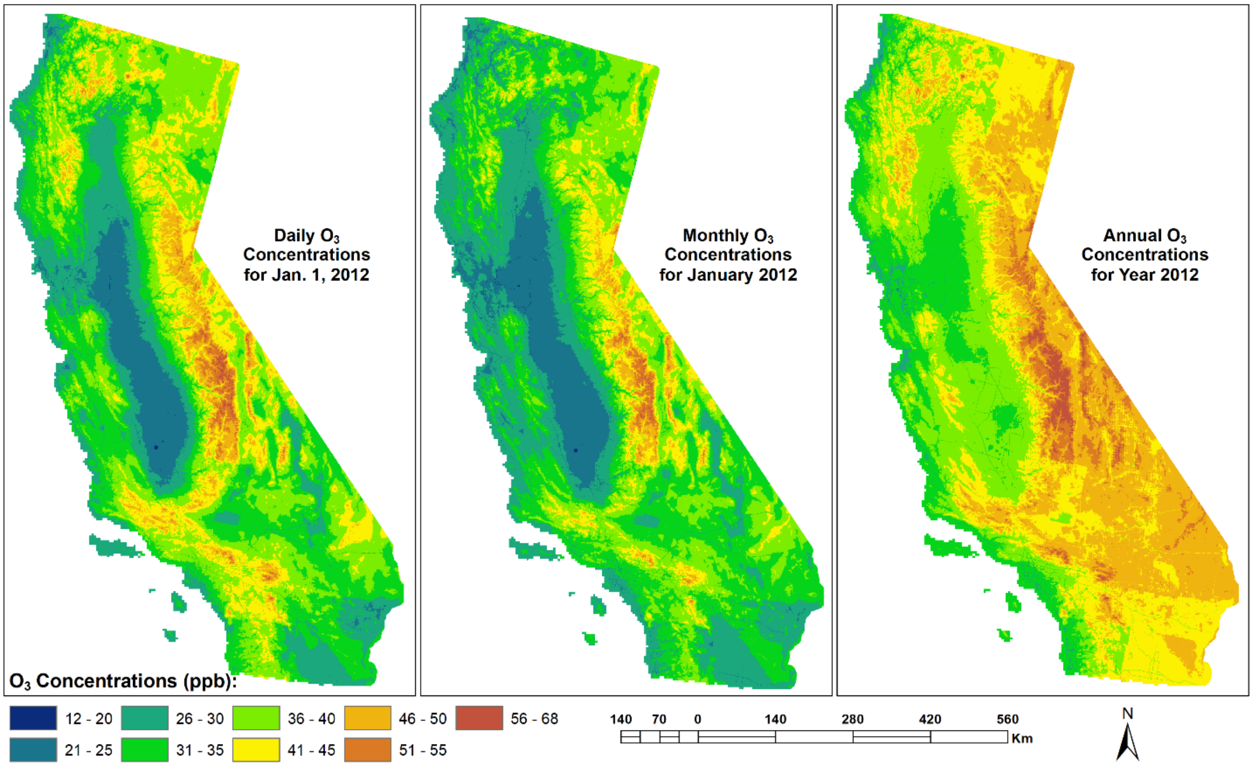

Daily O_3 Concentrations for Jan. 1, 2012
Monthly O_3 Concentrations for January 2012
Annual O_3 Concentrations for Year 2012
O_3 Concentrations (ppb):
12 - 20
21 - 25
26 - 30
31 - 35
36 - 40
41 - 45
46 - 50
51 - 55
56 - 68
140 70 0 140 280 420 560
Km
N

Figure 4. Example of daily, monthly and annual O_3 concentration surfaces (ppb) in year 2012.

Subtask 2.3: Monthly Hybrid LUR Models for On-Road Non-Exhaust Trace Metals

The number of CalEPA government sites in trace metals measurements ranged from 32 to 41 for a specific year. The measured concentrations were largely low, with mean monthly concentrations statewide (averaged from daily values) smaller than 0.002 ug m^{-3}, except for Zinc for a mean of 0.006-0.009 ug m^{-3}. Due to the fact that most of the trace metals were measured on the 4th or 5th day of a week and their concentrations measured had extensive zero days. Plus the fact that the saturation monitoring data were collected for a total of two weeks (see below), we opted to derive monthly concentrations for the trace metals in our modeling process.

The saturation monitoring of the trace metals was conducted in Los Angeles by Dr. Michael Jerrett's research group. After calibration of the gravimetric analysis from Dr. Jerrett's research group to the standard unit of measure in ug m^{-3} through a collocated sample site, we found that the measured concentrations of the saturation sampling sites in southern California were much higher than those averaged across the State, with the former ranged from 0.4 to 11.5 ug m^{-3} for a total of 26 effective measurement sites. Similar procedures to LUR modeling for criteria pollutants were applied to derive monthly LUR models for the six trace metals.

Generate buffer statistics on 30 m spatial resolution potential predictors

Similar to processing potential predictors for the criteria pollutants modeling, we applied circular buffer statistics with a range of 100 – 15000 m for an interval of 100 m. Those predictors were described in detail in modeling criteria pollutants. Further, we created buffer statistics for the road surface types including the total acres (ha) of a specific road surface type in a specific buffer distance. For road surface roughness and slope gradient, the buffer statistics applied only to road segments in calculating mean surface roughness and slope gradient for a specific distance of impact.

Apply data reduction strategy to reduce potential number of predictors

We applied the same data reduction strategies as those applied for the criteria pollutants modeling. The complete data sources included one-time land use and land cover information, monthly traffic flow, monthly weather conditions, monthly vegetation index, one time highway surface conditions (roughness and surface type), one-time roadway slope gradient, location category and distance to roadways. For those ones with buffer statistics, due to the buffer distance ranged from 100 m to 15000 m at an interval of 100, we created a total of 150 buffers for a single predictor. This process created more covariates than those used for criteria pollutants modeling. However, after data reduction, we still maintained the total number of potential predictors/covariates less than 100.

Develop monthly LUR models and surfaces for the trace metals

We integrated the data from the government regulatory continuous monitoring and the southern California saturation monitoring into a single modeling framework. The data from the southern California saturation monitoring were randomly separated into 10-folds of equal size. The government regulatory continuous monitoring data were also randomly separately into 10-

folds of equal size. The two 10-folds of data were merged into corresponding folds (fold #1, #2, …, #10) to form a dataset of 10-folds. The V-fold cross-validation technique was then used for LUR modeling. This modeling approach makes sure that data from southern California is not predominantly used for building a model and the two types of data (the government regulatory monitoring and the southern California saturation monitoring) have equal presentation in the modeling process. The monthly models developed for the six trace metals were presented from Table 18 to Table 23, with prediction powers being 52%, 77%, 67%, 63%, 44% and 51%, respectively, for Cr, Mn, Ni, Pb, Se and Zn.

Table 18. Monthly Chromium model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	-0.00044847	0.00015463	-2.90029132	**0.004**
Season [Spring]	-0.00076638	0.00016356	-4.68563157	**<0.001**
Season [Summer]	-0.00049386	0.00014799	-3.33706780	**0.001**
Season [Winter]	-0.00062043	0.00017005	-3.64845897	**<0.001**
Percent Impervious (%) (1000m) †	0.00004163	0.00000267	15.61843380	**<0.001**
Developed Low Intensity (ha) (700m)	0.00000108	0.00000012	8.66606414	**<0.001**
Commercial (ha)	-0.00000399	0.00000031	-12.69591453	**<0.001**
Residential (ha) (12900m)	0.00000008	0.00000001	9.96272230	**<0.001**
Industrial (ha) (1800m)	-0.00000505	0.00000111	-4.54466003	**<0.001**
Slope Gradient (degrees) (900m)	68.44675493	29.39658033	2.32839174	**0.020**
Crops (ha) (1200m)	0.00000137	0.00000033	4.11877984	**<0.001**
Wetlands (ha) (400m)	-0.00000624	0.00000240	-2.59807151	**0.009**
Maximum Relative Humidity (%)	0.00000550	0.00000210	2.62582687	**0.009**
Observations	3170			
R^2 / R^2 adjusted	0.523 / 0.521			

[†]: The content in the first pair of parentheses presents unit of analysis and the content in the second pair represents distance of buffer.
Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Table 19. Monthly Manganese model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	0.00210709	0.00018824	11.19348896	**<0.001**
Season [Spring]	0.00171809	0.00019030	9.02831619	**<0.001**
Season [Summer]	0.00153421	0.00020969	7.31670259	**<0.001**
Season [Winter]	0.00169150	0.00017864	9.46885920	**<0.001**
Maximum Relative Humidity (%)	-0.0000128	0.00000188	-6.81645176	**<0.001**
Developed Medium Intensity (ha) (100m) [†]	0.00003897	0.00000272	14.32512822	**<0.001**
Crops (ha) (4400m)	0.00000009	0.00000000	18.12103758	**<0.001**
Industrial (ha) (1200m)	0.00002702	0.00000129	20.98336051	**<0.001**
Developed Low Intensity (ha) (100m)	0.00001393	0.00000322	4.32090223	**<0.001**
Industrial (ha) (100m)	-0.0025273	0.00022516	-11.2245076	**<0.001**
Asphalt (ha) (100m)	0.00219813	0.00022618	9.71847803	**<0.001**
Vapor Pressure (kPa)	0.00040950	0.00004838	8.46491184	**<0.001**
Wind Velocity at 10m (m/s)	-0.0001397	0.00002469	-5.65945968	**<0.001**
Observations	3170			
R^2 / R^2 adjusted	0.775 / 0.774			

[†]: The content in the first pair of parentheses presents unit of analysis and the content in the second pair represents distance of buffer.
Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Table 20. Monthly Nickel model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	-0.00007103	0.00002635	-2.69582823	**0.007**
Season [Spring]	-0.00012495	0.00002755	-4.53595808	**<0.001**
Season [Summer]	-0.00007939	0.00002636	-3.01147435	**0.003**
Season [Winter]	-0.00009771	0.00002794	-3.49663493	**<0.001**
Industrial (ha) (600m) †	-0.00000672	0.00000222	-3.03501645	**0.002**
Crops (ha) (15000m)	0.00000000	0.00000000	11.84307201	**<0.001**
Commercial (ha) (100m)	-0.00036520	0.00003178	-11.49131134	**<0.001**
Commercial (ha) (1900m)	0.00000300	0.00000043	7.02392195	**<0.001**
Developed Low Intensity (ha) (500m)	0.00000170	0.00000011	14.88077912	**<0.001**
Asphalt (ha) (1000m)	0.00005032	0.00000369	13.63909698	**<0.001**
Industrial (ha) (3200m)	-0.00000143	0.00000014	-10.28884717	**<0.001**
Developed Low Intensity (ha) (3000m)	-0.00000004	0.00000001	-6.32253378	**<0.001**
Percent Impervious (%) (500m)	0.00001247	0.00000103	12.09922588	**<0.001**
Observations	3170			
R^2 / R^2 adjusted	0.667 / 0.666			

†: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.
Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Table 21. Monthly Lead model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	0.00089106	0.00014690	6.06571715	**<0.001**
Season [Spring]	0.00090684	0.00016395	5.53115887	**<0.001**
Season [Summer]	0.00077776	0.00014945	5.20414043	**<0.001**
Season [Winter]	0.00126383	0.00015556	8.12438783	**<0.001**
Developed High Intensity (ha) (1900m) †	0.00000083	0.00000005	18.24019787	**<0.001**
Crops (ha) (1100m)	-0.00000955	0.00000049	-19.43597477	**<0.001**
Asphalt (ha) (300m)	0.00019326	0.00007622	2.53566721	**0.011**
Crops (ha) (1700m)	0.00000226	0.00000024	9.26337546	**<0.001**
Crops (ha) (4100m)	0.00000032	0.00000002	14.73482053	**<0.001**
Developed High Intensity (ha) (700m)	-0.00000346	0.00000027	-12.97981443	**<0.001**
Developed Low Intensity (ha) (400m)	0.00000337	0.00000035	9.59858255	**<0.001**
Wind Velocity at 10m (m/s)	-0.00017268	0.00003864	-4.46916803	**<0.001**
Industrial (ha) (600m)	0.00004141	0.00000609	6.80289860	**<0.001**
Industrial (ha) (2800m)	-0.00000273	0.00000050	-5.41422849	**<0.001**
Observations	3170			
R^2 / R^2 adjusted	0.628 / 0.626			

†: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.
Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Table 22. Monthly Selenium model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	-0.00017963	0.00005583	-3.21738551	**0.001**
Season [Spring]	-0.00018029	0.00005114	-3.52530874	**<0.001**
Season [Summer]	-0.00015972	0.00007374	-2.16604349	**0.030**
Season [Winter]	-0.00003939	0.00004496	-0.87599184	0.381
Crops (ha) (4200m) †	0.00000015	0.00000001	26.53916117	**<0.001**
CRCP‡ (ha) (1400m)	0.00040647	0.00002126	19.11806703	**<0.001**
Crops (ha) (1000m)	-0.00001058	0.00000055	-19.39659272	**<0.001**
Crops (ha) (600m)	0.00002032	0.00000161	12.58813849	**<0.001**
Commercial (ha) (200m)	-0.00019384	0.00001432	-13.53211350	**<0.001**
Barren Land (ha) (200m)	0.00023405	0.00003214	7.28161895	**<0.001**
Developed Medium Intensity (ha) (200m)	0.00000691	0.00000073	9.42621853	**<0.001**
Road Roughness Index (m/km) (400m)	0.00005621	0.00000602	9.34223707	**<0.001**
Industrial (ha) (100m)	0.00089763	0.00018172	4.93975240	**<0.001**
Vapor Pressure (kPa)	0.00013107	0.00003334	3.93069924	**<0.001**
Industrial (ha) (2800m)	-0.00000113	0.00000026	-4.41585385	**<0.001**
Observations	3170			
R^2 / R^2 adjusted	0.447 / 0.444			

†: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.
Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Table 23. Monthly Zinc model for the State of California.

Coefficient	*Estimates*	*std. Error*	*Statistic*	*P-Value*
Season [Fall]	0.01258722	0.00153177	8.21745501	**<0.001**
Season [Spring]	0.01051563	0.00168071	6.25667163	**<0.001**
Season [Summer]	0.01235502	0.00155605	7.93999157	**<0.001**
Season [Winter]	0.01331568	0.00160805	8.28063112	**<0.001**
Crops (ha) (2800m) †	0.00001104	0.00000032	34.24752963	**<0.001**
Crops (ha) (1100m)	-0.00008087	0.00000634	-12.75167447	**<0.001**
Commercial (ha) (2900m)	0.00006806	0.00000426	15.97803598	**<0.001**
Commercial (ha) (200m)	-0.00431677	0.00027451	-15.72511322	**<0.001**
Tree canopy (%)	-0.00009186	0.00002536	-3.62183753	**<0.001**
Developed Open Space (ha) (6800m)	-0.00000035	0.00000004	-9.74164927	**<0.001**
Developed Low Intensity (ha) (100m)	0.00020686	0.00004702	4.39905289	**<0.001**
Wind Velocity at 10m (m/s)	-0.00152850	0.00036609	-4.17520276	**<0.001**
Asphalt (ha) (300m)	0.00315615	0.00054954	5.74325071	**<0.001**
Barren Land (ha) (200m)	0.00313139	0.00066102	4.73724301	**<0.001**
Distance to Ports (m)	-0.00000001	0.00000000	-3.88054474	**<0.001**
Wetlands (ha) (2300m)	-0.00000159	0.00000059	-2.71278511	**0.007**
Observations	3170			
R^2 / R^2 adjusted	0.509 / 0.507			

†: The content in the first pair of parentheses presents unit of analysis and the contents in the second pair represents distance of buffer.

Threshold for significance is 0.05 and bold p values indicate it's significant at 0.05 level.

Subtask 2.4: Development of Health Dataset

Collection of patient level data

Propeller Health, a sub-division in ResMed, led the development of a health cohort using data from the respiratory disease management platform. The patients have self-reported diagnosis of asthma or COPD and were recruited between January 1st, 2012 and December 31st, 2019. The digital medication sensors attached to patients' SABA inhaler(s) recorded space-time locations of inhaler use, sensor heartbeats and resets. If patients used controller medication sensors, we also included the space-time locations of controller medication use. Patients for this study came from both urban and rural regions of California, and were enrolled via major medical systems in California, as well as directly through social media campaigns. Up to December 31, 2019, 6,752 patients in California have been enrolled through 48 programs. After removing 3,366 patients that never synced (no rescue, heartbeat and reset events), 3,386 patients showed rescue, heartbeat and reset events (Table 24).

Table 24. Patient level statistics

Variable	N (# Patients)	Mean/%	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Disease	3386						
... Asthma	3034	89.6%					
... COPD	352	10.4%					
Sex	3386						
... Female	2358	69.6%					
... Male	1028	30.4%					
Age	3386	37.773	17.115	4	25	50	90
Plan w Controller (Y/N)‡	3386	67.7%	46.8%	0%	0%	100%	100%
Baseline Risk	3386						
... Very High	60	1.8%					
... High	722	21.3%					
... Medium	2286	67.5%					
... Low	318	9.4%					

‡: Plan w Controller (Y/N) indicates whether a patient had controller inhaler use plan.

Of those 3,386 effective patients participated in the study, 2,955 (87.27%) lived in urbanized areas (UAs: urban areas with population over 50,000), 244 (7.21%) lived in urban clusters (UCs: urban areas with population between 2,500 and 50,000) and 187 (5.52%) lived in rural areas. The cities with more than 15 patients spread across California (Figure 5) include Los Angeles (124), San Diego (112), Sacramento (97), Bakersfield (79), Stockton (71), Fresno (70), San Francisco (63), Modesto (47), Riverside (46), San Jose (43), San Bernardino (32), Long Beach (29), Lancaster (28), Oakland (28), El Cajon (27), Victorville (27), Hemet (24), Manteca (24), Anaheim (22), Tracy (22), Hesperia (21), Palmdale (21), Redding (19), Turlock (19), Chula Vista (18), Lodi (18), Oceanside (18), Fremont (17), Murrieta (17), Antioch (16), Clovis (16), Escondido (16), Garden Grove (16), Pomona (16), Visalia (16), Chico (15), Fontana (15), Moreno Valley (15), Sylmar (15). Across California, the cities include 473,501 rescue inhaler events, 158,748 controller inhaler events, 1,448,700 heartbeats and 14,377 resets.

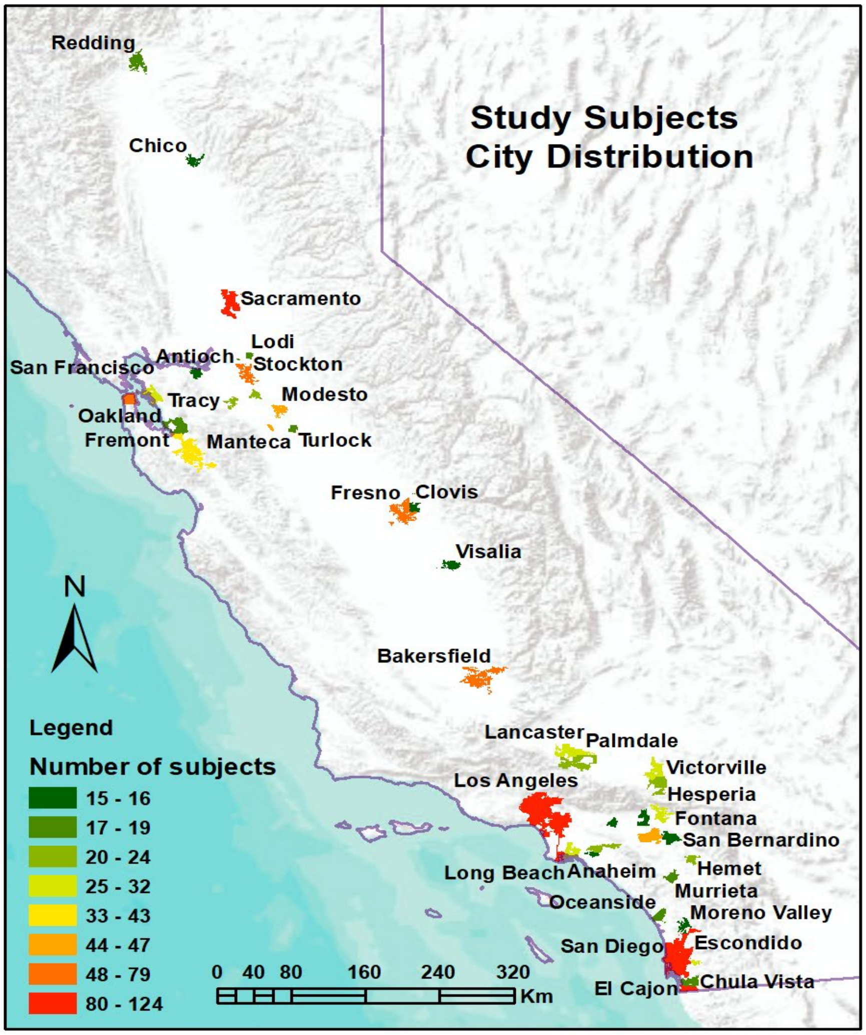


Figure 5. The spatial distribution of the study subjects in California cities for years 2012-2019. Those cities with subjects < 15 were not shown to preserve the privacy of the participants.

Participants had on average 99 days of participation from January 1, 2012 to December 31, 2019, with the 1st and 3rd quartile participating lengths being 33 and 253 days. Females represent 69.6% of the patients, and the mean age is 37.8 with 1st, median and 3rd quartile ages being 25, 35, and 50, respectively (Table 24). Based on the Center for Disease Control and Prevention (CDC) statistics, the prevalence rates in asthma for women and men are, respectively, 10.4% and 6.7% for California residents. These numbers are proportionate to the composition of women to men in our study for 69.6% to 30.4%. Race-ethnicity information is not mandatory at enrollment; thus, Propeller Health does not have complete information on race-sources for the 3,386 patients. For all the patients, Propeller Health also collected data on whether a patient has a controller medication prescribed as part of the therapy, and if so, the prescribed and actual number of controller medication uses taken per week. There were 2,294 (67.7%) patients who had controller medication as part of their therapy (Table 24).

Patients' baseline risks of rescue inhaler use were determined based on the enrollment period Asthma Control Test (ACT) score for asthma patients and COPD Assessment Test (CAT) score for COPD patients. The ACT is a patient self-administered survey of five items with a four-week recall on asthma symptoms and daily functions. The survey assesses the frequency of shortness of breath and general asthma symptoms, use of rescue medications, the effect of asthma on daily functioning, and overall self-assessment of asthma control. The CAT is a validated 8-item questionnaire widely used in clinical practice to quantify symptom burden and risk for exacerbation. The survey assesses respiratory symptoms (such as cough, sputum production, chest tightness and dyspnea), non-respiratory symptoms (such as lack of energy or sleep disturbance), and other manifestations of COPD on patient quality of life (such as limitations in doing activities at home or confidence leaving home). Ghobadi et al.[55] classified the health impact from CAT scores into four categories: those of low, medium, high and very high impact. A low score in CAT indicates COPD status as well controlled with low risk for health impact. The CAT scores are ranked in an opposite direction to the ACT scores.

To be consistent, we classified the ACT scores into four corresponding categories and the details about this categorization were presented in Table 25. We aimed to control for the potential impact of patients' health conditions during enrollment on model outcome: rescue inhaler use in number of puffs per day.

Table 25. Rescue inhaler use risks associated with ACT and CAT scores.

Inhaler Use Risk of Impact	ACT	CAT
Low Health Impact	≥ 20	< 10
Medium Health Impact	[10 – 19]	[10 – 19]
High Health Impact	[5 – 9]	[20 – 29]
Very High Health Impact	< 5	≥ 30

Among those 3,386 patients, 60 (1.8%) patients were categorized as having very high baseline risk, 722 (21.3%) patients having high impact baseline risk, 2286 (67.5%) patients having medium impact baseline risk, and 318 (9.4%) patients having low impact baseline risk (Table 24). All the patients signed an agreement upon joining the Propeller Health program, which explicitly enables the collection of location data and the use of de-identified and aggregated data for public health-oriented analyses.

Geocoding home address

Participants' home addresses were converted to geographical coordinates (latitude and longitude) to later fill missing event coordinates. Each geocoded home address has a confidence score, ranging from 0 to 1, where 0 reflects the inability to determine a confidence due to lack of a bounding box, and 1 indicates high precision. The score differs based on how close an address matches the standard address presented in the geocoding database. Non-standard address input from a patient or physician might reduce a geocoding score though the address input was correct. The extra information presented from an input like apartment number might also reduce the geocoding score. We considered geocoded home addresses as having high confidence if their confidence score is greater than 0.5. For those patients whose geocoded home addresses are less than or equal to 0.5, we used mathematic modes (cluster centroids) of latitude and longitude of all the events of corresponding patients in three decimal places as geographical coordinates of home address (latitude and longitude). If the centroid data were not available (because patients did not enable their GPS service), we used zip code centroids corresponding to patients' zip code as geographical coordinates of home address.

GPS determination for location-missing inhaler use events and days lack of inhaler use

The Propeller Health digital sensors passively and objectively monitor SABA use for asthma and COPD by capturing the date, time, and location of uses. The digital health platform also tracks the use of controller medications, which are used daily to prevent symptoms long-term. The SABA sensors also collect a regular "heartbeat" signal to track battery life, and a "reset" signal whenever a sensor software restarted. The locations recorded by SABA use, controller medication use, heartbeat and reset events, when combined, can be used to evaluate where people have spent time and the level of air pollution exposure after assigned from modeled surfaces. The sensors send SABA use, controller medication use, heartbeat and reset data via Bluetooth to a paired smartphone and transmit the information to HIPAA-compliant servers.

Among those participants transmitting data via a smartphone, geographical coordinate data were acquired for all medication use events, sensor heartbeats and sensor resets when available. For days with non-rescue inhaler events (controller inhaler, heartbeat, reset events) with geographical coordinates, we included those coordinates for exposure assignment. If there was a rescue inhaler event on a day without geographical coordinates, we assigned home address geographical coordinates to that rescue inhaler event. For non-rescue events without geographical coordinates, we did not assign any geographical coordinates. For days without any events and/or geographical coordinates, we assigned home address geographical coordinates to those days. This process enabled a detailed characterization of participant exposure through space and time. All participant data were stored in encrypted servers compliant with the HIPAA. Overall, there were 2,276,336 events, 100% of which were assigned date and time, and 1,855,663 (81.5%) captured event geographical coordinates. The remaining events without geographical coordinates (420,673 events, 18.5%) were retro-filled with home address locations.

Exposure assignment and generation of daily statistics

Each rescue inhaler use, controller medication use, heartbeat event or reset event were assigned daily concentrations of the three critical pollutants, monthly concentrations of the six trace metals, and daily maximum temperature and maximum relative humidity based on responding location and timestamp information. All the exposures within a day for a specific patient were aggregated into daily means and the number of puffs and inhaler use events were added up to form a daily total.

The daily mean (standard deviation) exposures to NO_2, $PM_{2.5}$ and O_3 among the 3,386 participants were, respectively, 8.3 (5.2) ppb, 10 (3.5) $\mu g\ m^{-3}$, and 35.4 (7.4) ppb (Table 26). The daily mean weather conditions for ambient temperature and relative humidity were, respectively, 297 (7.3) k and 80 (18.3) %. The daily mean number of rescue puffs and rescue events per person were, respectively, 0.9 (2.6) and 0.5 (1.4).

In typical clinic practices, asthma severity is assessed during the initial diagnosis of a patient while not taking long-term controller medication.[56] After the initial stage of assessment, follow-up visits are mostly concerned with the level of control achieved on the given step of therapy, including well controlled, not well controlled, or very poorly controlled, based on the most severe component of the patient's impairment.[56] We do not have data on impairments; however, because Propeller Health sensors also measure space-time SABA use events, we used patient SABA use frequency to identify potential patient-level weekly health impact risk that might contribute to the daily rescue inhaler use (other than air pollution) in that week.[57] When SABA use ≤ 2 days/week we considered impact on health outcome is low, 3-6 days/week considered as medium impact, and several times per day considered as high impact. This weekly health impact risk was used as a confounding control in our modeling the impact of air pollution on daily inhaler use. Of all the patient-day events, 34,467 (6.5%) were distributed in high health impact weeks, 86,699 (16.3%) were distributed in medium health impact weeks, and 409,972 (77.2%) were in low health impact weeks (Table 26). We also controlled for seasonal variation. Of all the patient-day events, 157,471 (29.6%) took place in Fall, 130,051 (24.5%) occurred in Spring, 128,050 (24.1%) were in Summer, and 115,566 (21.8%) happened in Winter.

Table 26. Patient-day exposure, inhaler use and health impact risks summary statistics

Variable	N†	Mean/%	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
NO_2 (ppb)	531138	8.303	5.222	1	4.388	11.394	52.004
PM_{25} (ug m^{-3})	531138	10.046	3.478	1	7.964	12.049	43.456
O_3 (ppb)	531138	35.354	7.373	11.897	29.841	39.929	75.923
Max Temperature (k)	531138	297.022	7.276	267.2	291.5	302.18	324
Max Humidity (%)	531138	79.972	18.329	7.9	71	94	100
# Puffs‡	531138	0.864	2.639	0	0	0	100
# Rescue Events‡	531138	0.503	1.408	0	0	0	20

Weekly Health Risk	531138						
... High	34467	6.5%					
... Medium	86699	16.3%					
... Low	409972	77.2%					
Season	531138						
... Spring	130051	24.5%					
... Summer	128050	24.1%					
... Fall	157471	29.6%					
... Winter	115566	21.8%					

[†]: N = total patient-day events, summarized at patient and day integrated levels.
‡: Puffs and rescue events refer to counts per person per day.

The maximum number of daily rescue puffs were 100 and the maximum number of daily rescue events were 20. Figure 6 shows the distribution of the daily rescue puffs (left) and rescue events (right) for days with at least one puff/event for a total of 115,057 person days. Due to the substantial number of person days having rescue puffs and events, the rare occurrence of number of rescue puffs of 100 or rescue events of 20 for a day did not change the overall statistics of the daily rescue puffs and events.

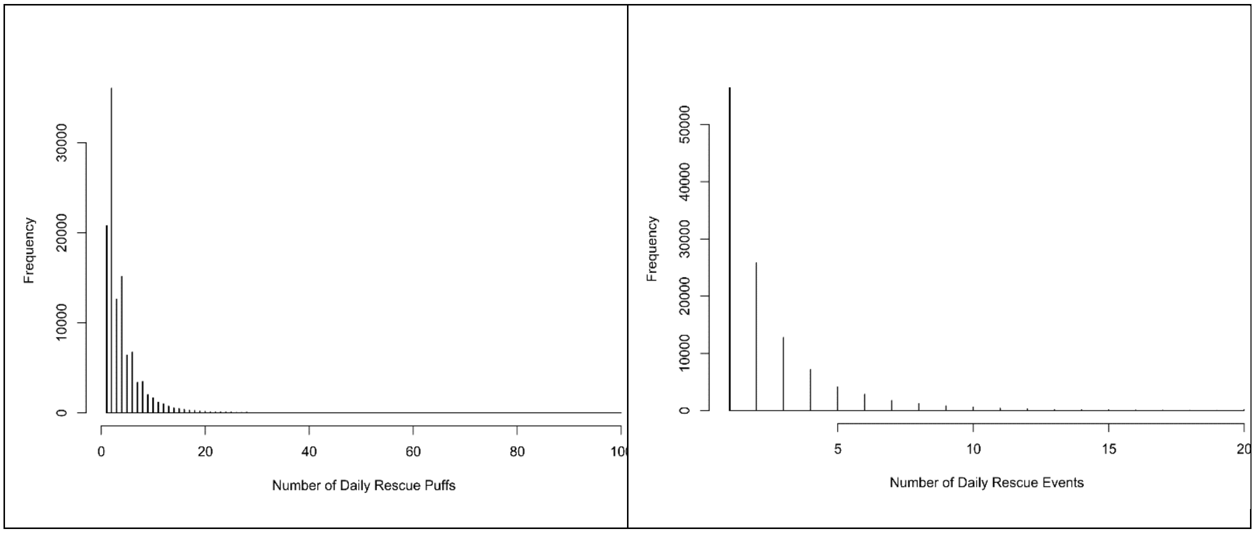


Figure 6. The distribution of daily rescue medication use puffs and events for the California participants in years 2012-2019.

Subtask 2.5: Statistical Analysis of Health Effects

UCB identified associations of sub-acute respiratory symptoms represented by SABA use with tailpipe air pollution from NO_2, $PM_{2.5}$, and O_3, and on-road non-exhaust trace metals Cr, Pb, Mn, Ni, Se, and Zn for California. All the exposures experienced by an individual across a day were averaged to daily means and the total number of puffs from SABA use was used as modeling outcome. We modeled the effects separately for the three criteria pollutants and the six trace metals.

Health effects of criteria pollutants through linear mixed model glmmTMB

Model diagnostics

The health outcome in this study is the number of rescue puffs experienced by an individual in a day: count data normally modeled through a Poisson regression. Before identifying the proper modeling framework, we assessed the outcome zero inflation and overdispersion. The Chi-square test showed a score of 318065 with a p value < 0.001, indicating significant zero inflation: extensive days without SABA use. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod] was applied to all the three pollutants and the weather conditions, and we found a dispersion score of 0.99. The score is lower than critical value of 1.2 as low overdispersion magnitude,[58] indicating our study subjects do not have significant overdispersion. R package Generalized Linear Mixed Models using Template Model Builder (glmmTMB version 1.1.2.3) was used to deal with zero-inflated mixed-effects for data in this study (days without any SABA use) (https://cran.r-project.org/web/packages/glmmTMB/index.html).

Linear mixed glmmTMB modeling of air pollution effect

The impact of individual air pollutants on SABA use was identified through a linear mixed model:

$$Y_{si} = \beta_0 + \beta_1 C_{si} + \beta_2 P_{si} + \beta_3 O_{si} + \gamma_s + \varepsilon_{si} \qquad (3)$$

Y_{si} and C_{si}, are respectively, the SABA use (i.e., total number of puffs) and air pollution (NO_2, $PM_{2.5}$ or O_3) for patient s in day i. P_{si} are the individual characteristics of patient s in day i such as baseline and weekly health impact. O_{si} are other confounding factors like weather. Season is included in the model due to the seasonal variations in respiratory disease symptoms. β_0 is the model constant; β_1 is the coefficient for air pollution exposure; β_2 and β_3 are respectively vectors of coefficients for the individual characteristics and other confounding factors. γ_s is the random effect of patient s, and ε_{si} is the error term of patient s for day i. The random effects approach is proven effective not only on dealing with panel data; but, also on processing unbalanced data when there are some individuals with one or a few observations.[80] The comprehensive control for impacts from confounding factors included individual characteristics and respiratory disease status (age and gender, baseline ACT and CAT health impact, weekly individual level health impact, and weekly controller plan) and weather conditions.

Due to the existence of excessive zeros, a zero-inflation component was also included in the glmmTMB modeling process. Due to the uncertainty of identifying which observation having greater zero-inflation, we applied the zero-inflated Poisson model with a single zero-inflation parameter applying to all observations (i.e., ziformula~1).

Associations of individual air pollutant exposure with outcome

For NO_2 (Table 27), we identified significant ($p < 0.001$) and positive associations of daily NO_2 exposure with daily SABA use, with 1 ppb increase in NO_2 exposure for a day, the associated puffs would increase by 0.35%. For $PM_{2.5}$ (Table 28), we also identified significant ($p < 0.001$) and positive associations of daily $PM_{2.5}$ exposure with daily SABA use, with 1 ug m^{-3} increase in $PM_{2.5}$ exposure for a day, the associated puffs would increase by 0.73%. For O_3 (Table 29), we identified positive associations of daily O_3 exposure with daily SABA use, with 1 ppb increase in O_3 exposure for a day, the associated puffs would increase by 0.002%; however, the association was found statistically non-significant ($p = 0.978$). Further, we identified that increasing age, having controller inhaler use, baseline (enrollment) with low and medium risks, and weekly risks of low and medium were all significantly associated with less daily rescue inhaler use. By contrast, being male and having COPD were associated with greater daily rescue inhaler use but the associations were not statistically significant. All models indicated that there was a zero inflation: existence of excessive days for a patient without rescue inhaler use.

Table 27. The effect of NO_2 on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	22.075035	4.487622	14.82044 – 32.880749	15.221878	**<0.001**
NO_2	1.003470	0.000486	1.002517 – 1.004423	7.149111	**<0.001**
Max. Temp.	0.996055	0.000366	0.995339 – 0.996772	-10.76647	**<0.001**
Max. Humid.	0.999572	0.000118	0.999341 – 0.999802	-3.641593	**<0.001**
Sex [Male]	1.035101	0.042699	0.954707 – 1.122266	0.836333	0.403
Age	0.992222	0.001260	0.989755 – 0.994695	-6.148450	**<0.001**
Plan Controller	0.908805	0.037307	0.838548 – 0.984947	-2.329433	**0.020**
Baseline Risk† [High]	0.927480	0.147261	0.679446 – 1.266060	-0.474154	0.635
Baseline Risk [Medium]	0.625112	0.101055	0.455359 – 0.858149	-2.906263	**0.004**
Baseline Risk [Low]	0.388067	0.067008	0.276649 – 0.544357	-5.481990	**<0.001**
Disease [COPD]	1.021780	0.084184	0.869415 – 1.200847	0.261514	0.794
Weekly Risk‡ [Low]	0.110116	0.000886	0.108393 – 0.111865	-274.2917	**<0.001**
Weekly Risk [Medium]	0.761164	0.003163	0.754991 – 0.767388	-65.68353	**<0.001**
Zero-Inflated Model					
(Intercept)	0.662723	0.004642	0.653686 – 0.671885	-58.72902	**<0.001**
Observations	531138				

†: Baseline risk is the potential impact of a patient's health at the initial enrollment stage on outcome. Please refer to the patient ACT and CAT health impact classifications described above for detail.

‡: Weekly risk is the potential impact of a patient's health in a specific week during the study period on outcome. Please refer to the patient's weekly health impact classification described above for detail.

Table 28. The effect of $PM_{2.5}$ on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	22.154870	4.517322	14.85633 – 33.039011	15.194193	**<0.001**
$PM_{2.5}$	1.007275	0.000632	1.006037 – 1.008516	11.545380	**<0.001**
Max. Temp.	0.995864	0.000366	0.995146 – 0.996583	-11.26248	**<0.001**
Max. Humid.	0.999758	0.000118	0.999527 – 0.999990	-2.043799	**0.041**
Sex [Male]	1.026518	0.042557	0.946406 – 1.113410	0.631303	0.528
Age	0.991928	0.001273	0.989436 – 0.994426	-6.314811	**<0.001**
Plan Controller	0.914008	0.037635	0.843143 – 0.990829	-2.183721	**0.029**
Baseline Risk[†] [High]	0.931593	0.147389	0.683213 – 1.270271	-0.447875	0.654
Baseline Risk [Medium]	0.626522	0.101020	0.456761 – 0.859377	-2.899856	**0.004**
Baseline Risk [Low]	0.390942	0.067305	0.278977 – 0.547843	-5.455327	**<0.001**
Disease [COPD]	1.030996	0.086203	0.875160 – 1.214582	0.365090	0.715
Weekly Risk[‡] [Low]	0.109843	0.000897	0.108098 – 0.111616	-270.3257	**<0.001**
Weekly Risk [Medium]	0.760876	0.003162	0.754704 – 0.767099	-65.76154	**<0.001**
Zero-Inflated Model					
(Intercept)	0.660736	0.004924	0.651155 – 0.670457	-55.60959	**<0.001**
Observations	531138				

[†]: Baseline risk is the potential impact of a patient's health at the initial enrollment stage on outcome. Please refer to the patient ACT and CAT health impact classifications described above for detail.

[‡]: Weekly risk is the potential impact of a patient's health in a specific week during the study period on outcome. Please refer to the patient's weekly health impact classification described above for detail.

Table 29. The effect of O_3 on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	20.416802	4.281919	13.535342 – 30.79684	14.382427	**<0.001**
O_3	1.000015	0.000533	0.998971 – 1.001059	0.027381	0.978
Max. Temp.	0.996487	0.000444	0.995617 – 0.997357	-7.903062	**<0.001**
Max. Humid.	0.999307	0.000119	0.999073 – 0.999540	-5.814957	**<0.001**
Sex [Male]	1.033074	0.042514	0.953021 – 1.119852	0.790698	0.429
Age	0.992163	0.001254	0.989709 – 0.994623	-6.226291	**<0.001**
Plan Controller	0.909947	0.037333	0.839641 – 0.986140	-2.300144	**0.021**
Baseline Risk† [High]	0.929722	0.147328	0.681504 – 1.268348	-0.459846	0.646
Baseline Risk [Medium]	0.626139	0.101021	0.456392 – 0.859020	-2.901848	**0.004**
Baseline Risk [Low]	0.389878	0.067188	0.278125 – 0.546535	-5.465749	**<0.001**
Disease [COPD]	1.019627	0.083798	0.867931 – 1.197837	0.236501	0.813
Weekly Risk‡ [Low]	0.109859	0.000898	0.108113 – 0.111634	-270.195692	**<0.001**
Weekly Risk [Medium]	0.761084	0.003163	0.754911 – 0.767308	-65.702235	**<0.001**
Zero-Inflated Model					
(Intercept)	0.661419	0.004769	0.652138 – 0.670832	-57.331417	**<0.001**
Observations	531138				

†: Baseline risk is the potential impact of a patient’s health at the initial enrollment stage on outcome. Please refer to the patient ACT and CAT health impact classifications described above for detail.

‡: Weekly risk is the potential impact of a patient’s health in a specific week during the study period on outcome. Please refer to the patient’s weekly health impact classification described above for detail.

Associations of simultaneous air pollutant exposure with outcome

To evaluate the simultaneous environmental exposure impact, we first identified whether significant multicollinearities existed among the three criteria pollutants through a variance inflation factors (VIF) analysis. Multicollinearity inflates the variance of predictors and leads to biased coefficient estimation and a loss of power.[59,60] A VIF score was calculated for each pollutant by doing a linear regression of the predictor on all other pollutants, and then obtaining the variance being explained value (R^2) from that regression. Literature suggests that a VIF score lower than 2.5, which corresponds to an R^2 of 0.60, indicates lack of significant multicollinearity. For the three criteria pollutants NO_2, $PM_{2.5}$ and O_3, we identified their respective VIF scores were 1.59, 1.46 and 1.11. This indicates that the three criteria pollutants do not have significant multicollinearity in their simultaneous exposure, and we included all those three criteria pollutants in a single glmmTMB model. Based on the integrated model, we identified that all the three criteria pollutants had significant ($p < 0.001$) and positive associations with daily rescue inhaler use (Table 30), with the effect of NO_2 on 1 ppb increase for a 0.25% increase in daily puffs, the effect of $PM_{2.5}$ on 1 ug m^{-3} increase for a 0.88% increase in daily puffs, and the effect of O_3 on 1 ppb increase for a 0.53% increase in daily puffs.

Table 30. The multipollutant effect on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	40.208631	8.671167	26.348382 – 61.35989	17.129640	**<0.001**
NO_2	1.002482	0.000626	1.001255 – 1.003710	3.966870	**<0.001**
O_3	1.005306	0.000675	1.003983 – 1.006630	7.880484	**<0.001**
$PM_{2.5}$	1.008790	0.000790	1.007243 – 1.010340	11.172772	**<0.001**
Max. Temp.	0.992891	0.000517	0.991878 – 0.993905	-13.69339	**<0.001**
Max. Humid.	1.000460	0.000148	1.000171 – 1.000750	3.114965	**0.002**
Sex [Male]	1.034575	0.042757	0.954078 – 1.121864	0.822473	0.411
Age	0.992191	0.001263	0.989718 – 0.994671	-6.156177	**<0.001**

Plan Controller	0.911073	0.037494	0.840472 – 0.987605	-2.263033	**0.024**
Baseline Risk[†] [High]	0.932345	0.147766	0.683394 – 1.271986	-0.442000	0.658
Baseline Risk [Medium]	0.629191	0.101529	0.458594 – 0.863250	-2.871255	**0.004**
Baseline Risk [Low]	0.391304	0.067446	0.279126 – 0.548565	-5.443638	**<0.001**
Disease [COPD]	1.020539	0.083992	0.868509 – 1.199182	0.247029	0.805
Weekly Risk[‡] [Low]	0.110128	0.000899	0.108381 – 0.111904	-270.3552	**<0.001**
Weekly Risk [Medium]	0.761435	0.003165	0.755257 – 0.767662	-65.57978	**<0.001**
Zero-Inflated Model					
(Intercept)	0.662611	0.004820	0.653232 – 0.672125	-56.58228	**<0.001**
Patient-day Events	531138				

[†]: Baseline risk is the potential impact of a patient's health at the initial enrollment stage on outcome. Please refer to the patient ACT and CAT health impact classifications described above for detail.

[‡]: Weekly risk is the potential impact of a patient's health in a specific week during the study period on outcome. Please refer to the patient's weekly health impact classification described above for detail.

Health effects of criteria pollutants through random forest machine learning

Using the random forest R package (version 4.6.14),[61] we modeled daily rescue inhaler use in total number of puffs using the same parameters as those specified in the linear mixed glmmTMB models. Both glmmTMB and random forest models included variables to account for confounding such as individual level age, gender, baseline risks, controller on plan plus, potential weekly health impact and daily weather conditions. Based on the penalized partial dependence plots from random forest modeling (Figure 5), the same-day exposure impact from a unit increase using the linear trend for a pollutant was estimated to increase by 0.0058 (95% CI: 0.0053-0.0063), 0.0058 (95% CI: 0.0050-0.0066) and 0.0029 (95% CI: 0.0022-0.0035), respectively, for NO$_2$, PM$_{2.5}$ and O$_3$. The equivalent incidence rate ratios (through exponentiality) were respectively, 1.0058 (95% CI: 1.0053-1.0063), 1.0058 (95% CI: 1.0050-1.0066) and 1.0029 (95% CI: 1.0022-1.0035), similar to the glmmTMB modeling results.

Health effects of trace metals through linear mixed model glmmTMB

Similar to model the criteria pollutants, we also conducted model diagnostics and found that, in addition to the condition of outcome of zero inflation and lack of overdispersion, the six trace metals did not have significant multicollinearity. The associations of individual and simultaneous trace metals air pollutant exposure with daily rescue inhaler use in number of puffs are presented in Table 31–Table 37. Across all the trace metals, we found that their associations were largely statistically non-significant after comprehensive control for confounding. Further, we identified that increasing age, having controller inhaler use, baseline (enrollment) with low and medium risks, and weekly risks of low and medium were all significantly associated with less daily rescue inhaler use. By contrast, being male and having COPD were associated with greater daily rescue inhaler use but the associations were not statistically significant. All models indicated that there was a zero inflation: existence of excessive days for a patient without rescue inhaler use.

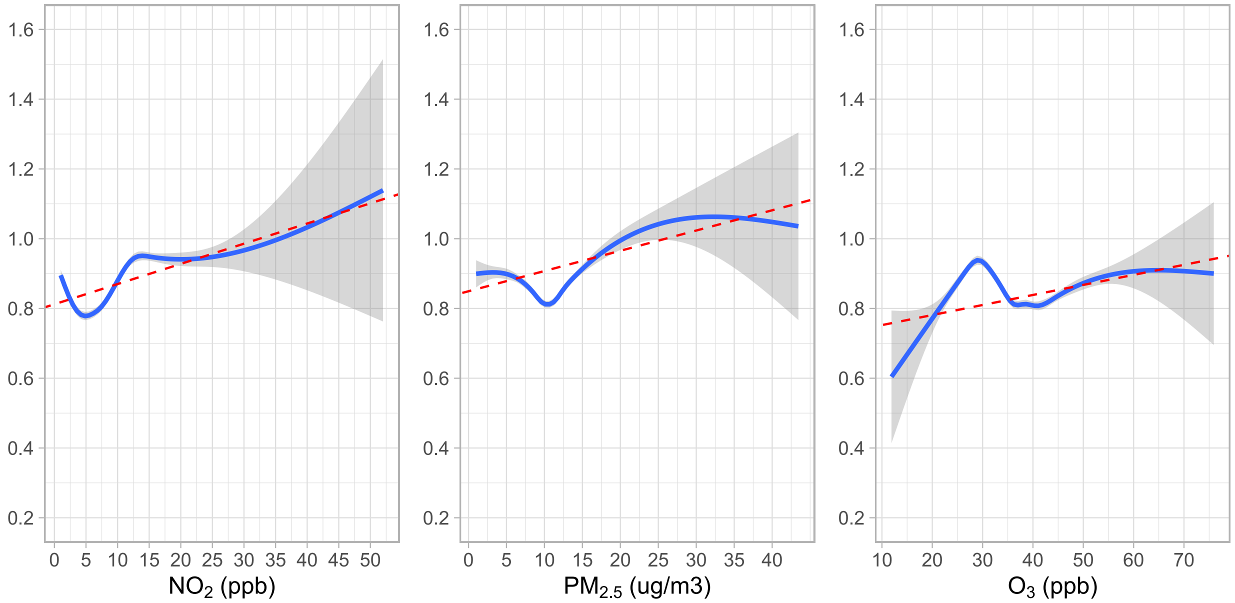


Figure 7. The penalized partial dependence plots of NO_2, $PM_{2.5}$ and O_3 with daily rescue inhaler use in number of puffs (Y-axis), modeled through a random forest model and adjusted for potential confounding from patient individual characteristics, potential weekly health impacts and weather conditions.

Table 31. The effect of Cr on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	19.127199	3.919608	12.800307 – 28.581326	14.401056	**<0.001**
Cr	0.291586	1.442333	0.000018 – 4734.103705	-0.249150	0.803
Max. Temp.	0.996348	0.000375	0.995614 – 0.997083	-9.724210	**<0.001**
Max. Humid.	0.999354	0.000117	0.999124 – 0.999583	-5.522074	**<0.001**
Sex [Male]	1.042966	0.044718	0.958902 – 1.134400	0.981172	0.327
Age	0.992306	0.001314	0.989734 – 0.994884	-5.834267	**<0.001**
Baseline Risk [High]	0.950448	0.149386	0.698459 – 1.293349	-0.323349	0.746
Baseline Risk [Medium]	0.643655	0.102982	0.470396 – 0.880729	-2.753765	**0.006**
Baseline Risk [Low]	0.396669	0.067813	0.283734 – 0.554557	-5.408733	**<0.001**
Disease [COPD]	1.007374	0.082998	0.857157 – 1.183918	0.089178	0.929
Weekly Risk [Low]	0.111073	0.000991	0.109146 – 0.113033	-246.1877	**<0.001**
Weekly Risk [Medium]	0.762513	0.003183	0.756300 – 0.768777	-64.95515	**<0.001**
Zero-Inflated Model					
(Intercept)	0.667437	0.005845	0.656079 – 0.678992	-46.16761	**<0.001**
Observations	523699				

Table 32. The effect of Mn on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	18.548667	3.717154	12.523677 – 27.472207	14.572838	**<0.001**
Mn	0.000000	0.000000	0.000000 – 47.697289	-1.611333	0.107
Max. Temp.	0.996595	0.000368	0.995873 – 0.997317	-9.228304	**<0.001**
Max. Humid.	0.999230	0.000121	0.998993 – 0.999468	-6.352827	**<0.001**
Sex [Male]	1.035093	0.044188	0.952010 – 1.125426	0.807945	0.419
Age	0.992151	0.001310	0.989586 – 0.994723	-5.965616	**<0.001**
Baseline Risk [High]	0.945414	0.148406	0.695032 – 1.285994	-0.357589	0.721
Baseline Risk [Medium]	0.640333	0.102319	0.468157 – 0.875830	-2.789692	**0.005**
Baseline Risk [Low]	0.396832	0.067753	0.283974 – 0.554544	-5.413335	**<0.001**
Disease [COPD]	1.009874	0.083057	0.859528 – 1.186519	0.119471	0.905
Weekly Risk [Low]	0.108807	0.000948	0.106965 – 0.110681	-254.5926	**<0.001**
Weekly Risk [Medium]	0.760776	0.003163	0.754602 – 0.767001	-65.76537	**<0.001**
Zero-Inflated Model					
(Intercept)	0.669537	0.005687	0.658483 – 0.680776	-47.23193	**<0.001**
Observations	535154				

Table 33. The effect of Ni on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	19.054731	3.833831	12.845180 – 28.266071	14.648611	**<0.001**
Ni	0.000000	0.000000	0.000000 – 6436482.17	-1.244378	0.213
Max. Temp.	0.996462	0.000362	0.995753 – 0.997171	-9.758211	**<0.001**
Max. Humid.	0.999342	0.000115	0.999118 – 0.999567	-5.727562	**<0.001**
Sex [Male]	1.035196	0.044192	0.952106 – 1.125538	0.810287	0.418
Age	0.992152	0.001310	0.989587 – 0.994724	-5.965409	**<0.001**
Baseline Risk [High]	0.945541	0.148426	0.695126 – 1.286166	-0.356737	0.721
Baseline Risk [Medium]	0.640430	0.102335	0.468228 – 0.875963	-2.788741	**0.005**
Baseline Risk [Low]	0.396935	0.067770	0.284047 – 0.554687	-5.411834	**<0.001**
Disease [COPD]	1.009937	0.083061	0.859584 – 1.186589	0.120231	0.904
Weekly Risk [Low]	0.108813	0.000947	0.106972 – 0.110685	-254.8164	**<0.001**
Weekly Risk [Medium]	0.760840	0.003163	0.754665 – 0.767065	-65.74645	**<0.001**
Zero-Inflated Model					
(Intercept)	0.669535	0.005681	0.658492 – 0.680764	-47.27738	**<0.001**
Observations	535154				

Table 34. The effect of Pb on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	18.970522	3.900548	12.678354 – 28.385443	14.312884	**<0.001**
Pb	1.053431	23.244717	0.000000 – 6.382E+18	0.002359	0.998
Max. Temp.	0.996367	0.000376	0.995630 – 0.997104	-9.646693	**<0.001**
Max. Humid.	0.999346	0.000114	0.999124 – 0.999568	-5.759468	**<0.001**
Sex [Male]	1.042950	0.044716	0.958889 – 1.134379	0.980832	0.327
Age	0.992306	0.001314	0.989734 – 0.994884	-5.834386	**<0.001**
Baseline Risk [High]	0.950430	0.149383	0.698447 – 1.293323	-0.323467	0.746
Baseline Risk [Medium]	0.643631	0.102978	0.470379 – 0.880695	-2.754010	**0.006**
Baseline Risk [Low]	0.396654	0.067810	0.283723 – 0.554535	-5.408969	**<0.001**
Disease [COPD]	1.007335	0.082994	0.857124 – 1.183871	0.088707	0.929
Weekly Risk [Low]	0.111073	0.000991	0.109146 – 0.113033	-246.1889	**<0.001**
Weekly Risk [Medium]	0.762512	0.003183	0.756299 – 0.768776	-64.95557	**<0.001**
Zero-Inflated Model					
(Intercept)	0.667436	0.005845	0.656079 – 0.678991	-46.17061	**<0.001**
Observations	523699				

Table 35. The effect of Se on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	18.772892	3.801058	12.623605 – 27.917658	14.482781	**<0.001**
Se	0.508799	0.936423	0.013802 – 18.756279	-0.367138	0.714
Max. Temp.	0.996461	0.000369	0.995738 – 0.997184	-9.578335	**<0.001**
Max. Humid.	0.999317	0.000115	0.999092 – 0.999542	-5.949706	**<0.001**
Sex [Male]	1.035822	0.044338	0.952467 – 1.126472	0.822234	0.411
Age	0.992161	0.001312	0.989594 – 0.994735	-5.953202	**<0.001**
Baseline Risk [High]	0.945491	0.148420	0.695086 – 1.286105	-0.357064	0.721
Baseline Risk [Medium]	0.640492	0.102347	0.468270 – 0.876056	-2.788061	**0.005**
Baseline Risk [Low]	0.396845	0.067757	0.283979 – 0.554568	-5.412979	**<0.001**
Disease [COPD]	1.009550	0.083064	0.859196 – 1.186215	0.115514	0.908
Weekly Risk [Low]	0.108884	0.000954	0.107030 – 0.110770	-253.0195	**<0.001**
Weekly Risk [Medium]	0.760846	0.003163	0.754671 – 0.767071	-65.74239	**<0.001**
Zero-Inflated Model					
(Intercept)	0.670291	0.005749	0.659118 – 0.681654	-46.64269	**<0.001**
Observations	535154				

Table 36. The effect of Zn on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	18.622445	3.731495	12.574079 – 27.580188	14.594387	**<0.001**
Zn	1.151632	1.247090	0.137898 – 9.617681	0.130374	0.896
Max. Temp.	0.996476	0.000365	0.995760 – 0.997192	-9.634225	**<0.001**
Max. Humid.	0.999308	0.000111	0.999090 – 0.999526	-6.216065	**<0.001**
Sex [Male]	1.035078	0.044187	0.951997 – 1.125410	0.807622	0.419
Age	0.992154	0.001310	0.989588 – 0.994725	-5.964042	**<0.001**
Baseline Risk [High]	0.945767	0.148459	0.695295 – 1.286468	-0.355217	0.722
Baseline Risk [Medium]	0.640453	0.102337	0.468247 – 0.875990	-2.788565	**0.005**
Baseline Risk [Low]	0.396930	0.067769	0.284045 – 0.554678	-5.411969	**<0.001**
Disease [COPD]	1.009863	0.083056	0.859520 – 1.186504	0.119337	0.905
Weekly Risk [Low]	0.108810	0.000947	0.106969 – 0.110683	-254.7867	**<0.001**
Weekly Risk [Medium]	0.760812	0.003163	0.754638 – 0.767036	-65.75569	**<0.001**
Zero-Inflated Model					
(Intercept)	0.669517	0.005682	0.658472 – 0.680746	-47.27387	**<0.001**
Observations	535154				

Table 37. The effect of simultaneous trace metals air pollution exposure on daily rescue inhaler use in number of puffs after comprehensive confounding control.

Coefficient	*Incidence Rate Ratios*	*std. Error*	*95% CI*	*Statistic*	*P-Value*
Count Model					
(Intercept)	19.688438	4.129801	13.051644 – 29.700060	14.207021	**<0.001**
Cr	0.083437	0.427900	0.000004 – 1934.991596	-0.484294	0.628
Mn	0.000000	0.000000	0.000000 – 0.664038	-1.995169	**0.046**
Ni	0.000000	0.000000	0.000000 – 0.042967	-2.070728	**0.038**
Pb	333.89359	7471.042	0.000000 – 3.7127E+21	0.259696	0.795
Se	2.491577	5.054353	0.046746 – 132.801005	0.450028	0.653
Zn	6.015382	7.957908	0.449973 – 80.415604	1.356326	0.175
Max. Temp.	0.996397	0.000396	0.995620 – 0.997173	-9.078667	**<0.001**
Max. Humid.	0.999332	0.000132	0.999073 – 0.999591	-5.052439	**<0.001**
Sex [Male]	1.042932	0.044742	0.958824 – 1.134417	0.979847	0.327
Age	0.992308	0.001314	0.989736 – 0.994886	-5.831945	**<0.001**
Baseline Risk [High]	0.950356	0.149387	0.698370 – 1.293264	-0.323929	0.746
Baseline Risk [Medium]	0.643620	0.102987	0.470356 – 0.880708	-2.753838	**0.006**
Baseline Risk [Low]	0.396694	0.067823	0.283742 – 0.554609	-5.407861	**<0.001**
Disease [COPD]	1.007252	0.082996	0.857039 – 1.183793	0.087699	0.930
Weekly Risk [Low]	0.111066	0.000992	0.109138 – 0.113028	-245.9894	**<0.001**
Weekly Risk [Medium]	0.762475	0.003183	0.756262 – 0.768740	-64.96192	**<0.001**
Zero-Inflated Model					
(Intercept)	0.667429	0.005856	0.656049 – 0.679006	-46.08068	**<0.001**
Observations	523699				

Discussion and Conclusion

Literature Review

In this systematic review, we assessed the impact of exposures to criteria pollutants (NO_2, $PM_{2.5}$, PM_{10}, O_3 and SO_2), and more than 10 trace metals that can be considered toxic air contaminants on respiratory disease outcomes, including airway inflammation, coughing, wheezing, exacerbations, ED visits, hospitalizations and mortality. We converted all the study outcomes in rate ratios (RRs) and percentage changes to odds ratios (ORs) and standardized the impact of exposure increase interval to make the studies comparable (see Appendix A). We further pooled the effects, separately, for children, adults, and all ages, for each pollutant of interest. We identified consistently significant associations of all the criteria pollutants and trace metals with a broad range of respiratory disease outcomes. The relatively greater impact from NO_2, $PM_{2.5}$ and SO_2 on respiratory disease outcomes for children may be due to the greater probability of children living or having physical activity in high air pollution areas for these pollutants.[31-35] Because children's lungs are not completely developed, their exposures to high levels of air pollution can affect both their short-term and long-term respiratory health.[36] Policymakers and stakeholders should adopt strategies to help children, especially those living in vulnerable communities, to reduce exposures to those sources in their neighborhoods. By contrast, the impact of O_3 on respiratory disease outcomes was found to be relatively greater for adults. Due to the chemical reaction between O_3 and traffic emission of nitrogen oxides that lead to generation of NO_2, NO_2 concentrations are higher near busy roadways while O_3 are higher at locations farther away from busy roadways. We suspect that, compared to children, more adults might have moved away from living close to traffic sources to communities with less traffic-related air pollution,[37] thus leading to greater O_3 exposure.

It should be addressed that due to the limitation in the number of studies in each type of outcomes for specific criteria pollutant, we pooled the results for all types of outcomes together for each criteria pollutant and did the comparison across different types of outcomes. It is also worth noting that the impact of trace metals on human respiratory health is understudied and only 3 studies[38-40] were included in the pooled analysis in this review. Overall, this systematic review identified consistent and significant effects of ambient exposures to criteria pollutants and trace metals on a broad suite of respiratory disease outcomes. It contributes to better understanding of the size of the effect of the five criteria pollutants and trace metals that can be compared across various studies. The study also helped identify groups that are more vulnerable to adverse respiratory outcomes from air pollution exposure across available studies.

Exposure Assessment

It is well known in environmental justice research that impacted communities (i.e., those of low income or visible minorities in a block group or zip code) face higher air pollution burden than advantaged communities. It is thus crucial to develop high spatial resolution air pollution surfaces to help identify exposure disparities those impacted communities experience in assessing their air pollution exposure. Large majority of exposure assessment uses air pollutant concentrations measured from the nearest government monitoring sites, which are normally sparsely distributed and reflect the background concentrations of a region. Using nearest government site as a location of exposure assumes the land use and land cover information of a community of interest to be the same as those of the government site, which could be incorrect. Some exposure assessment used government monitoring data through kriging or inverse distance weighting (IDW) techniques to interpolate exposure for a community. Those techniques used more government sites for exposure assessment; however, they still assumed the same land use and land cover information of the community to those of government sites.

Land use regression applies land use, land cover and other related information to model pollutant concentrations measured at monitoring sites; and the developed model is used to predict pollutant concentrations in communities through those communities' land use, land cover and other information. LUR modeling advances exposure assessment by utilizing land use, land cover and other related information at community of interest. Traditional LUR modeling approach, however, has two key shortcomings: (1) missing potential optimal predictors, including buffered distance of impact and (2) model overfit. All the LUR models applied in literature, except the ones developed by Dr. Su, used limited number of predictors with optimal distances of impact subjectively selected from 4-5 categories (e.g., distances of 250, 500, 750 and 1000 m). For a predictor in land use (e.g., commercial, industrial, and residential), land cover (e.g., forest, shrublands and developed), % tree canopy or % impervious, we derived buffer statistics for each predictor at distances of 50-5000 m at an interval of 50 m. When all the predictors and their associated buffered statistics were calculated, we normally had 2000-3000 potential covariates in our modeling process to make sure we captured the optimal distance of impact from a predictor.

The large majority of LUR models used all the data for both model construction and model effectiveness assessment. This creates a situation that model performance is only valid for the data used in model construction, not for the data that have never seen in the modeling process. It creates model overfit.[62] In our modeling process, instead, we used a v-fold cross-validation strategy through LUR D/S/A machine learning algorithm to improve model fit: i.e., using a subset of the original dataset for model development [53,62]. In our D/S/A V-fold cross-validation modeling, the original sample was randomly partitioned into V equal size subsamples. Of the V subsamples, a subsample was retained as the validation data for testing the model, and the remaining V-1 subsamples were used as training data. The cross-validation process was then repeated V times, with each of the V subsamples used exactly once as the validation data. This technique, therefore, minimizes over-fitting to the data to maximize the probability that guarantees the models predict well at locations that have not been sampled.

In this research, we used very comprehensive data sources for air pollution exposure modeling. The data sources included statewide daily traffic data on highways, daily remote sensing data, daily weather data, parcel level land use and detailed land cover data, every two-week vegetation index and tree canopy data. For traffic, we used road type category criteria of the nearest neighbor to derive daily highway roadway traffic, from the measured 4.6% roadways, for the entire California and those derived roadway traffic data were converted into daily traffic surfaces of 30 m spatial resolution across California, plus buffer statistics 50-5000 m at an interval of 50 m through R programming. We also incorporated parcel level land use data for 40 million people living in the state of California from 58 counties in our modeling process using a spatial resolution of 30 m. The parcel level land use data included agricultural, residential, commercial, industrial, government and institutions, open land, parks, and recreational facilities. We also included comprehensive land cover (16 classes such as forest, shrubland, developed) data, every two week of vegetation index, tree canopy and impervious surface at a spatial resolution of 30 m and their corresponding buffer statistics 50-5000 m at an interval of 50 m through Google Earth Engine scripting. We also included daily remote sensing data in Ozone Monitoring Instrument for NO_2 and O_3 at 25 km spatial resolution and daily Aerosol Optical Depth for $PM_{2.5}$ at 1 km spatial resolution through R programming. Further, we included daily meteorological conditions data of

4 km spatial resolution through Google Earth Engine scripting. They were maximum and minimum temperature, precipitation accumulation, downward surface shortwave radiation, wind-velocity, maximum and minimum relative humidity, and specific humidity. Other potential predictors of 30 m spatial resolution included elevation (digital elevation model), distance to coast, distance to ports, distance to highway roadways. Further, we were the first in literature to incorporate data from multiple air pollution measurement instruments into a single modeling framework, including those from government continuous monitoring across California, our research saturation monitoring in Los Angeles, Alameda and Sacramento counties, and Google Street Car mobile monitoring across San Francisco Bay (counties of Alameda, San Francisco and San Mateo), Los Angeles County and central valley regions. The capability of integrating data sources from various platforms into a single modeling framework enabled us to deal with TBs of data at statewide level in our LUR modeling process.

Due to our LUR modeling process incorporating thousands of predictors and covariates, we are the only research team developing LUR models with data reduction strategies. The data reduction strategy typically reduces the number of covariates from 2000-3000 to 80-100 but still maintain the optimal and least collinear covariates in the modeling process. The random forest LUR modeling process ensures least collinearity of feature selection in the modeling process; however, it cannot reduce the number of predictors in the modeling process. Random forest modeling technique is thus predominantly used for situations like traditional LUR modeling with limited number of predictors. Further, our D/S/A machine learning algorithm can specify the number of covariates/predictors in the final model output after data reduction and the final selected models maintained only the most effective predictors. No other LUR algorithm in literature has that capability. Data reduction is necessary if a comprehensive data source in predictors is available and research wants to make sure that the final selected predictors are interpretable (e.g., green vegetation is associated with reduced air pollution and higher traffic is associated with greater level of air pollution). In our research, the final daily prediction models for the three criteria pollutants, had adjusted R^2 of 79.6%, 65.3% and 93.6%, respectively, for NO_2, $PM_{2.5}$ and O_3, greater performance than other daily models and higher than some annual models.

Though we already generated TBs of data in exposure surfaces due to the high spatial resolution implemented in the study (30m for modeling and 100m for surface building), we did not take peak hourly exposures into consideration. No hourly pollutant surfaces were developed.

This limitation could be resolved in the future when support from computer clusters is available with hundreds of TBs of storage space plus great computing power.

Health Outcome Modeling

Previous studies on the effects of air pollution on respiratory disease relied on aggregated and infrequently-reported acute respiratory disease outcome measures, such as emergency department visits or hospitalizations, which lacked temporal and spatial resolution due to annual aggregation and grouping to a zip code or county level.[8-10] Other studies used patient self-reported data to assess the location and frequency of symptoms,[11] which could be fraught with missing data, errors, and are burdensome for the patients.[12-14] When the location of exposure was provided, previous studies often used an individual's residential address in defining the location of air pollution exposure. Air pollution exposure can occur in the community, at work, at home, at school, and elsewhere; therefore, a residential address does not capture the full signature of exposure for an individual. Significant exposure misclassification exists and health risks estimated from such data can lead to exposure measurement error and flawed findings.[15] Digital sensors fitted onto inhalers can capture the date, time, and location of rescue inhaler medication use and identify activity space through sensor "heartbeats" — sensor checking into battery life every 3-4 hours with location information; thereby, offering an objective signal of respiratory disease symptoms and exposure space in real-time. The combination of the best available spatiotemporal air pollution surfaces developed in our research and the activity space detected through sensor technology enabled us to have individual spatiotemporal exposures with the least misclassification.

The spatiotemporally rich data in rescue inhaler use, locations of activity space, and extensive information on environmental exposure, however, raised methodological challenges in modeling the impacts of environmental exposures on respiratory disease symptoms. Traditional linear mixed models might not be sufficient in dealing with the frequent zeros in modeling health outcome data, as was seen here in rescue inhaler use puffs per person per day measure. We used glmmTMB, a linear mixed model capable of processing excessive zeros and overdispersion, to address these issues by creating two models: one dealt with normal count data with a Poisson function and the second dealt with the excessive zeros through a logit function. The random forest model, like other complex nonparametric models (e.g., neural networks, support vector machines and super learners), is becoming more common in predictive analytics, especially when dealing with large observational datasets that do not adhere to the strict assumptions imposed by traditional statistical techniques (e.g., multiple linear regression assumes linearity, homoscedasticity, and normality).[63] In our health outcome analysis, we found both glmmTMB and random forest modeling technique successfully predicted positive and significant impact of daily air pollution on daily rescue inhaler use in number of puffs after comprehensive control for confounding.

Additionally, the effect of air pollution on respiratory disease was, similar to other health outcome analysis, largely assessed using single pollutant modeling approaches despite the fact that people were exposed to multiple pollutants simultaneously,[16] which might interactively influence respiratory disease symptoms. Due to our ability to identify small area variations in pollutant concentrations and detect difference in environmental exposure burden between different communities, we found, contradictory to most health outcome studies, that the air pollution exposures from the criteria pollutants were not collinear. We thus were able to integrate exposure to all the three criteria pollutants into a single health outcome modeling framework and identified the marginal effect of each criteria pollutant on daily rescue inhaler use.

Based on the separate air pollutant health effect modeling results, the exposure-response (daily air pollution exposure – daily rescue inhaler use puffs) functions were 1.003470 (95% CI: 1.002517 – 1.004423), 1.007275 (95% CI: 1.006037 – 1.008516) and 1.00005 (95% CI: 0.998971 – 1.001059), respective, for NO_2, $PM_{2.5}$ and O_3 in per unit increase in air pollutant exposure. The impact of O_3 exposure was shown not statistically significant. The integrated rescue inhaler use model with simultaneous exposure to the three criteria pollutants identified corresponding exposure-response functions of 1.002482 (95% CI: 1.001255 – 1.003710), 1.008790 (95% CI: 1.007243 – 1.010340) and 1.005306 (95% CI: 1.003983 – 1.006630). All the three criteria pollutants were statistically significantly associated with rescue inhaler use after comprehensive control for confounding. Among the three criteria pollutants, we found that exposure to $PM_{2.5}$ had the greatest impact on rescue inhaler use. For example, per 1 $\mu g\ m^{-3}$ increase in $PM_{2.5}$ exposure was associated with EXP(1.008790 - 1) = 0.00879% increase in rescue inhaler use puffs per person per day in the integrated model. In our study, more than 80% study subjects are adults. Those effects identified from this research may seem trivial; however, the corresponding 10 ppb, 10 ug m^{-3} and 30 ppb increase in exposure to NO_2, $PM_{2.5}$ and O_3 would result in, respectively, a 2.5% (OR = (1.002482)^10 = 1.025), 9.1% (OR = (1.008790)^10 = 1.091) and 17.2% (OR = (1.005306)^30 = 1.172) increase in daily rescue puffs.[64] These effects seemed low compared to the pooled results identified in our literature research: with 10 ppb, 10 ug m^{-3} and 30 ppb increase in exposure to NO_2, $PM_{2.5}$ and O_3, we saw a corresponding 14% (95% CI: -1%-32%), 25% (95% CI: -1%-64%) and 40.0% (95% CI: 14%-81%) increase in asthma symptoms for adults. During the study period, a patient sometimes did not have any rescue inhaler use or heartbeat events for a specific day. We attributed those days having exposure from home location. Because those days had zero rescue inhaler use (outcome of value 0), the impact of air pollution on rescue inhaler use identified from this study is thus the marginal exposure effect for days exposure being above typical home-based exposure from which a rescue inhaler use did not occur. In our previous research, we identified effects of a respective increase of 15.3%, 13.1% and 11.3% in daily rescue medication use from an IQR (NO_2: 9.44 ppb, $PM_{2.5}$: 5.8 µg m^{-3} and O_3: 15.65 ppb) increase in exposure to NO_2, $PM_{2.5}$ and O_3.[65] Those effects would be equal to an increase of rescue medication use of 16.3%, 23.6% and 22.8% from a respective increase in NO_2, $PM_{2.5}$ and O_3 exposure of 10 ppb, 10 ug m^{-3} and 30 ppb. In that research, days for participants' activity participation in the study but without activity space being identified were removed from analysis, rather than using home address as location of exposure as in the current CARB project. Though potential differences in population characteristics, air pollution exposure metrics and other impacting factors, we believe that our study findings were consistent with the literature research findings.

Based on our D/S/A daily land use regression modeling results, The main sources contributing to elevated NO_2 concentrations included traffic and highly developed urban areas (Impervious surfaces account for 80% to 100% of the total cover including apartment complexes, row houses and commercial/industrial land use). For $PM_{2.5}$, the main sources of air pollution included those from traffic, industrial and residential land use. For both NO_2 and $PM_{2.5}$, open and vegetative spaces were the sinks of air pollution. For O_3, however, the high concentrations were associated with greater open and high vegetation coverage. The near roadway traffic and developed urban land contributed to reduced O_3 concentrations in a way opposite to the impact from NO_2 and $PM_{2.5}$.

Given that most impacted communities live near roadways with high traffic, close to industrial sources or population densely distributed urban cores, these communities have great air pollution impact from NO_2 and $PM_{2.5}$. Greater air pollution from NO_2 and $PM_{2.5}$ tend to have greater impact on rescue inhaler use for those vulnerable communities than for other communities.

Regionwide, we expect that air pollution has the greatest impact on rescue inhaler use in Central Valley due to its high air pollutant concentrations from all the three criteria pollutants. Southern California has higher concentrations of NO_2 and $PM_{2.5}$ than San Francisco Bay, and we expect their impact on rescue inhaler use was much higher for those living in Southern California than those living in the San Francisco Bay Area.

We believe the associations identified of the impact of air pollution on rescue inhaler use are more accurate than the ones that relied only on home or school address as location of exposure. We understand that severity of asthma or COPD has direct impact on rescue inhaler use. To identify the real impact of air pollution on rescue inhaler use, we included comprehensive confounding control. In addition to integrating typical confounding variables (e.g., age and gender, baseline ACT and CAT scores), we also controlled for the impact of daily weather conditions on rescue inhaler use. We also included control for whether a patient had controller medication use plan for every week while being in the study. Based on the most severe component of impairment,[56] an patient's asthma and COPD severity could be identified as well controlled, not well controlled, or very poorly controlled. We do not have data on the impairments from rescue inhaler use. However, because Propeller Health sensors also measure space-time rescue inhaler use events, we used patient weekly rescue inhaler use frequency to identify potential patient-level weekly health impact risk that might contribute to the daily rescue inhaler use (other than air pollution) in that week.[57] When rescue inhaler use ≤ 2 days/week we considered the health impact on outcome is low, 3-6 days/week considered as medium impact, and several times per day considered as high impact. This weekly health impact risk was used as a confounding control in our modeling the impact of air pollution on daily inhaler use. We understand that this strict confounding strategy might over control the impact of air pollution on rescue inhaler use: A week of high rescue inhaler use (e.g., several times a day throughout a week) might be due to the impact of weekly high air pollution from extreme events such as wildfires but we contributed those impacts to patient's high weekly health risk rather than air pollution. This process, however, guaranteed that the impact of a patient's high weekly health risk from his/her own conditions was always controlled for in assessing associations between air pollution and rescue inhaler use.

In summary, our literature review found that exposures to air pollutants NO_2, $PM_{2.5}$ and O_3 were significantly associated with respiratory systems and associated ED visits, hospitalizations and even mortality. Using rescue medication use as an endpoint, our study confirmed that instantaneous air pollution exposures are significantly associated with adverse asthma symptoms. Our study further added to literature the significant association of air pollution with asthma symptoms including those for ED visits and hospitalizations. Previous studies typically used home address as location of exposure or using short-term (e.g., 1 week) mobile monitoring as exposure space, locations of exposure for those study subjects could be significantly biased. Combining high spatiotemporal exposure surfaces, year-round activity space and locations of rescue medication use, we were able to accurately identify space-time air pollution exposure and precisely estimate the impact of air pollution on asthma symptoms. Further, most studies identified separate effects of individual air pollutants on asthma symptoms, not taking into consideration that study subjects were simultaneous exposed to multiple pollutants. Our study was able to integrate three types of air pollutants into a single health outcome model, effectively identifying the marginal effects of each pollutant on rescue medication use. Further, we also included the number of days per week with rescue medication use as a proxy for ongoing disease status, aiming to adjust for each subject's individual variability in disease severity to better assess the impact of air pollutant exposure alone. We believe that our study design had the least exposure misclassification throughout the literature on identifying the associations of air pollution with asthma symptoms.

We could not focus the study on children due to the requirement of ownership of a smartphone though alternative techniques exist. We did not have enough children in our study and most of the study subjects were adults. If we could have enough children in our study, we would expect seeing greater rescue inhaler use per person per day due to the greater physical activity children normally have and the fact that their lungs are more sensitive to environmental exposure impacts.

Our goal was to identify the impact of space-time air pollution exposure on rescue inhaler use. We assumed that higher air pollution exposure was associated with greater rescue inhaler use. It does not matter whether an exposure happened during an earlier or later study period. The air pollution exposure for a space-time rescue inhaler use or heartbeat was based on the daily air pollution surfaces developed for the study. If it was true that long-term trend of air pollution exposure did exist, we could safely say, based on our modeling results, that later years of exposure incurred less rescue inhaler use.

Due to the fact that the large majority of patients lived in urban areas or urban clusters (and their activity space occurred also largely in urban areas), We did not differentiate between urban and rural on association between air pollution and rescue inhaler use. However, based on our modeling results, we expected that those lived in urban areas had greater impact from traffic, urban development (e.g., densely apartment complex and commercial land use) and industrial land use, and they experienced higher NO_2 and $PM_{2.5}$ exposure impact. In rural area, patients were more impacted by open space and vegetation distribution for higher O_3 concentrations.

For trace metals, we identified that most of the trace metals were measured on the 4th or 5th day in a week and their measured concentrations were largely low, with mean monthly concentrations statewide (averaged from daily values) smaller than 0.002 ug m^{-3}, except for Zinc for a mean of 0.006-0.009 ug m^{-3}. We also found that their concentrations measured had extensive days with a value of 0. Since most of the measured concentrations were < 0.005 ug m^{-3}, the variations measured through government and saturation monitoring might reflect the random oscillation around instrument detection limits. Still, we applied the most comprehensive data sources to build monthly land use regression models for the six trace metals, with prediction powers between 0.4 to 0.7, and we anticipated less prediction power than the models developed for the criteria pollutants. Further, our aim was to identify the impact of trace metals from tire- and break-wear. We expected that road roughness index, surfaces types, road slope gradients and traffic would be significant predictors for trace metals concentrations; however, this largely did not happen and the trace metal surfaces built thus might not be the real trace metals space-time concentration distributions. This partially contributed to the non-significant association between modeled trace metals exposures and rescue inhaler use.

Appendix: Literature Review Method

We investigated the effect of increased exposure for different criteria pollutants including $PM_{2.5}$, PM_{10}, NO_2, O_3, SO_2, and trace metals including Aluminum, Iron, Magnesium, Sulfur, Nickel, Vanadium, Chromium, Arsenic, Manganese, Barium, Copper, Antimony, Zinc and Lead. For health outcome measures, we included coughing, wheezing, shortness of breath, ED visits, hospitalizations, exacerbations, and mortality for respiratory diseases, including asthma, chronic obstructive pulmonary disease (COPD), respiratory infections and lung cancer.

Conversion of Outcomes to OR

Not all studies included in this review reported the results in the same scale. Estimated outcomes associated with an increase of exposure to the air pollutants are reported in terms of either OR, RR or percentage increase and 95% Confidence Intervals (CI). We converted all reported results to ORs plus associated CIs for the purpose of comparing results from different studies and, moreover, pulled all the effects together and identified the overall size of impact of a pollutant on respiratory disease outcomes.

To convert the results reported in percentage increase to OR, we exponentiated the reported results and lower/upper CI values from the corresponding studies, as presented in the following set of formulas:

$$OR = \exp\,(\%\, increase)$$

$$CI_{lower} = \exp\,(\%\, CI_{lower})$$

$$CI_{upper} = \exp\,(\%\, CI_{upper})$$

Also, in order to convert the results reported in RR to OR, the following conversion was used:

$$OR = (1 - risk_0) \times \frac{RR}{1 - risk_0 \times RR}$$

where $risk_0$ is the risk of having a positive outcome in the control or unexposed group. Similarly, the associated lower and upper CIs can be calculated as:

$$CI_{lower} = (1 - risk_0) \times \frac{RR_{lower}}{1 - risk_0 \times RR_{lower}}$$

$$CI_{upper} = (1 - risk_0) \times \frac{RR_{upper}}{1 - risk_0 \times RR_{upper}}$$

Standardization of Exposure Increase Ranges

Furthermore, not all studies used the same exposure increase interval to assess association with respiratory diseases. We standardized all the reported results so that the results were comparable between various studies. To make the results comparable between different studies, we standardized the OR's through the following conversions:

$$OR(x_s) = OR(x_g)^{\frac{x_s}{x_g}}$$

where *OR(x_s)* is the standardized OR for pollutant *x* when its exposure increase interval is set at x_s. Moreover, x_g and $OR(x_g)$ represent standardized exposure increase interval and associated OR. Based on the potential exposure levels from current studies, we used interquartile ranges of 10 ppb, 10 µg m^{-3}, 10 µg m^{-3}, 10 ppb and 30 ppb, respectively, for NO_2, $PM_{2.5}$, PM_{10}, SO_2 and O_3 in standardizing exposure level of increase.

Tables A1-A6 list all the effect estimates and corresponding standardized ORs.

Pooled Analysis

We excluded the studies that were deemed significant outliers, such as OR greater than 4.0. After standardizing the reported results from studies, effect pooled analysis was performed for each pollutant in three different age categories including children (<18 years), adults (≥18 years), and all ages. The Meta-Analysis package with R was used to pool the effects, separately, for children, adults, and all ages, separately, for NO_2, $PM_{2.5}$, PM_{10}, SO_2, O_3 and trace metals. Studies with mixed ages were evaluated separately from those studies focused on just children or adults to avoid the potential repeated counting of children and adults in the all-ages category (i.e., we did not pool the effects from children, adults, and all age groups to form a category of all subjects).

Table A1: $PM_{2.5}$ associations with respiratory disease outcomes

Papers	Study group	Outcome	Reported OR/RR(CI)	Standardized OR(CI)
Hansel et al. 2019	children	uncontrolled asthma	1.59(1.26-2.00)	2.44(1.56-3.79)
Fan et al. 2016	children	asthma ED	1.04(1.02-1.05)	1.04(1.02-1.05)
Brauer et al. 2007	children	wheezing	1.20(1.00-1.40)	1.74(1-2.77)
Brauer et al. 2007	children	doctor-diagnosed asthma	1.30(1.00-1.70)	2.21(1.00-4.99)
Brauer et al. 2007	children	ear/nose/throat infections	1.20(1.00-1.30)	1.74(1.00-2.21)
Brauer et al. 2007	children	flu	1.20(1.00-1.40)	1.74(1.00-2.77)
Mar et al. 2004	children	trouble breathing	1.13(0.86-1.48)	1.13(0.86-1.48)
Mar et al. 2004	children	coughing	1.17(0.98-1.40)	1.17(0.98-1.40)
Mar et al. 2004	children	sputum production	1.06(0.92-1.22)	1.06(0.92-1.22)
Huang et al. 2019	adults	susceptibility to COPD	1.29(1.01-1.65)	1.30(1.01-1.68)
Lamichane et al. 2018	adults	reduced lung function due to COPD	1.34(0.89-2.02)	1.34(0.89-2.02)
Mirabelli et al. 2016	adults	any asthma symptoms	1.03(1.01-1.06)	1.40(1.12-1.77)
Fan et al. 2016	adults	asthma ED	1.02(1.01-1.03)	1.02(1.01-1.03)
Cortez-Lugo et al. 2015	adults	COPD cough	1.39(1.05-1.99)	1.39(1.05-1.99)
Cortez-Lugo et al. 2015	adults	COPD phlegm	1.26(1.02-1.72)	1.26(1.02-1.72)
Mar et al. 2004	adults	wheezing	1.04(0.86-1.26)	1.04(0.86-1.26)
Yu et al. 2020	all	respiratory mortality	1.02(1.01-1.03)	1.25(1.15-1.30)
Fan et al. 2016	all	asthma ED	1.01(1.01-1.02)	1.01(1.01-1.02)
Rasschou-Nielsen et al. 2013	all	lung cancer	1.18(0.96-1.46)	1.39(0.92-2.13)
Rasschou-Nielsen et al. 2013	all	adenocarcinoma	1.55(1.05-2.29)	2.40(1.10-5.24)

Kloog et al. 2013	all	PM-related mortality	1.60(1.50-1.80)	1.60(1.50-1.80)
Katanoda et al. 2011	all	mortality due to lung cancer and respiratory diseases	1.24(1.12-1.37)	1.24(1.12-1.37)

Table A2: PM_{10} associations with respiratory disease outcomes

Papers	**Study Group**	**Outcome**	**Reported OR/RR(CI)**	**Standardized OR(CI)**
Weinmayr et al. 2010	children	asthma symptoms	1.03(1.01-1.05)	1.03(1.01-1.05)
Weinmayr et al. 2010	children	coughing	1.01(1.00-1.03)	1.01(1.00-1.03)
Mar et al. 2004	children	trouble breathing	1.04(0.95-1.15)	1.04(0.95-1.15)
Mar et al. 2004	children	coughing	1.09(1.02-1.16)	1.09(1.02-1.16)
Mar et al. 2004	children	sputum production	1.08(0.98-1.17)	1.08(0.98-1.17)
Magzamen et al. 2018	adults	inhaler use due to COPD	1.07(1.03-1.10)	1.09(1.04-1.13)
Lamichane et al. 2018	adults	reduced lung function due to COPD	1.39(0.85-2.25)	1.39(0.85-2.25)
Mar et al. 2004	adults	wheezing	1.01(0.93-1.09)	1.01(0.93-1.09)
Mar et al. 2004	adults	trouble breathing	1.02(0.96-1.08)	1.02(0.96-1.08)
Mar et al. 2004	adults	sputum production	1.01(0.92-1.12)	1.01(0.92-1.12)
Rasschou-Nielsen et al. 2013	all	lung cancer	1.22(1.03-1.45)	1.22(1.03-1.45)
Rasschou-Nielsen et al. 2013	all	adenocarcinoma	1.51(1.10-2.08)	1.51(1.10-2.08)
Analitis et al. 2006	all	respiratory mortality	1.00(1.00-1.01)	1.00(1.00-1.01)

Table A3: NO_2 associations with respiratory disease outcomes

Papers	Study Group	Outcome	Reported OR/RR(CI)	Standardized OR(CI)
Hasunuma et al. 2016	children	persistence of asthmatic symptoms	1.02(0.99-1.06)	1.22(0.90-1.79)
Belanger et al. 2013	children	asthma severity	1.37(1.01-1.89)	1.88(1.02-3.57)
Belanger et al. 2013	children	wheezing	1.49(1.09-2.03)	2.22(1.19-4.12)
Belanger et al. 2013	children	night symptoms due to asthma	1.52(1.16-2)	2.31(1.34-4.00)
Belanger et al. 2013	children	medication use due to asthma	1.78(1.33-2.38)	3.17(1.77-5.66)
Takenoue et al. 2012	children	asthma development	1.13(1.03-1.25)	1.13(1.03-1.25)
Takenoue et al. 2012	children	wheezing	1.05(1.02-1.08)	1.05(1.02-1.08)
Weinmayr et al. 2010	children	asthma symptoms	1.03 (1.00-1.06)	1.06(1.00-1.12)
Migliaretti and Cavallo 2004	children	asthma hospitalizations	1.03(1.01-1.05)	1.05(1.01-1.10)
Magzamen et al. 2018	adults	inhaler use due to COPD	1.04(1.01-1.08)	1.05(1.01-1.09)
Lamichane et al. 2018	adults	reduced lung function due to COPD	1.14(1.00-1.30)	1.28(1.00-1.64)
De Mrco et al. 2002	adults	asthma attack	1.13(0.98-1.32)	1.13(0.98-1.33)
De Mrco et al. 2002	adults	chest tightness	1.11(0.98-1.26)	1.11(0.98-1.27)
De Mrco et al. 2002	adults	wheezing	1.11(0.96-1.28)	1.11(0.96-1.29)
Ghozikali et al. 2016	all	COPD hospitalizations	1.01(1.00-1.02)	1.01(1.00-1.03)
Li et al. 2016a	all	COPD exacerbations	1.02(1.00-1.02)	1.04(1.01-1.03)
Katanoda et al. 2011	all	mortality due to lung cancer and respiratory diseases	1.17(1.10-1.26)	1.17(1.10-1.26)
Sunyer et al. 2002	all	mortality due to asthma	1.63(0.93-2.86)	1.49(0.94-2.37)

Table A4: O_3 associations with respiratory disease outcomes

Papers	**Study Group**	**Outcomes**	**Reported OR/RR(CI)**	**Standardized OR(CI)**
Pepper et al. 2020	children	asthma rescue inhaler use	1.12(1.07-1.20)	1.22(1.13-1.38)
Gent et al. 2003	children	chest tightness(1-hr)	1.26(1-1.48)	1.15(1-1.26)
Gent et al. 2003	children	shortness of breath(1hr)	1.22(1.02-1.45)	1.13(1.01-1.25)
Gent et al. 2003	children	chest tightness(8-hr)	1.33(1.09-1.62)	1.19(1.05-1.33)
Gent et al. 2003	children	shortness of breath(8hr)	1.30(1.05-1.61)	1.17(1.03-1.33)
Pepper et al. 2020	adults	asthma rescue inhaler use	1.09(1.07-1.12)	1.16(1.12-1.22)
Day et al. 2017	adults	pulmonary inflammation	1.20(1.05-1.40)	1.72(1.14-2.73)
Silverman and Ito 2010	adults	asthma HA	1.22(1.2-1.34)	1.31(1.16-1.49)
Khaniabadi et al. 2017	all	cardiopulmonary mortality	1.06(1.02-1.10)	1.42(1.15-1.76)
Khaniabadi et al. 2017	all	COPD hospitalization	1.04(1.02-1.06)	1.28(1.16-1.44)
Ghozikali et al. 2016	all	COPD hospitalization	1.02(1.01-1.03)	1.13(1.05-1.20)

Table A5: SO_2 associations with respiratory disease outcomes

Papers	Study Group	Outcomes	Reported OR/RR(CI)	Standardized OR(CI)
Greenberg et al. 2016	children	asthma severity	1.10(1.05-1.16)	1.24(1.11-1.38)
Smargiassi et al. 2009	children	asthma emergency department visits	1.10(1.00-1.22)	1.16(1.00-1.37)
Smargiassi et al. 2009	children	asthma hospitalization	1.42(1.10-1.82)	1.74(1.16-2.58)
Mercan et al. 2020	adults	asthma hospitalization	1.07(1.06-1.08)	1.19(1.17-1.21)
Mercan et al. 2020	adults	COPD hospitalization	1.07(1.06-1.08)	1.18(1.15-1.21)
Li et al. 2016b	adults	COPD mortality	1.04(1.01-1.06)	1.09(1.03-1.16)
Kan et al. 2010	adults	respiratory mortality	1.01(1.00-1.02)	1.04(1.02-1.06)
Ghozikali et al. 2016	all	COPD HA	1.01(1.00-1.01)	1.01(1.00-1.03)
Li et al. 2016a	all	COPD exacerbations	1.01(1.00-1.02)	1.03(1.00-1.06)
Katanoda et al. 2011	all	mortality due to lung cancer and respiratory diseases	1.26(1.07-1.48)	1.26(1.07-1.48)

Table A6: Trace metal associations with respiratory disease outcomes

Papers	metal	Exposure increase	Study Group	Outcomes	Reported OR/RR(CI)
Wu et al. 2019	Lead	-	children	risk for active asthma	1.24(1.08-1.42)
Wu et al. 2019	Lead	-	children	Wheezing	1.19(1.04-1.38)
Mao et al. 2018	Copper	-	all	asthma susceptibility	1.04(0.20-1.87)
Mao et al. 2018	Iron	-	all	asthma susceptibility	2.83(0.47-5.18)
Pollitt et al. 2016	Aluminum	54.2 ng/m^3	children	airway inflammation due to asthma	1.04(0.99-1.10)
Pollitt et al. 2016	Iron	59.4 ng/m^3	children	airway inflammation due to asthma	1.01(0.96-1.06)
Pollitt et al. 2016	Magnesium	15.1 ng/m^3	children	airway inflammation due to asthma	1.02(0.97-1.09)
Pollitt et al. 2016	Sulfur	179.0 ng/m^3	children	airway inflammation due to asthma	1.03(0.98-1.09)
Pollitt et al. 2016	Nickel	0.9 ng/m^3	children	airway inflammation due to asthma	1.01(0.97-1.05)
Pollitt et al. 2016	Vanadium	2.18 ng/m^3	children	airway inflammation due to asthma	1.05(0.97-1.14)
Pollitt et al. 2016	Chromium	2.98 ng/m^3	children	airway inflammation due to asthma	1.01(0.99-1.04)
Pollitt et al. 2016	Arsenic	0.79 ng/m^3	children	airway inflammation due to asthma	1.04(0.94-1.16)
Pollitt et al. 2016	Manganese	2.21 ng/m^3	children	airway inflammation due to asthma	0.99(0.95-1.02)
Pollitt et al. 2016	Barium	9.42 ng/m^3	children	airway inflammation due to asthma	1.09(1.03-1.17)
Pollitt et al. 2016	Copper	10.4 ng/m^3	children	airway inflammation due to asthma	1.02(1.00-1.04)
Pollitt et al. 2016	Antimony	23.4 ng/m^3	children	airway inflammation due to asthma	1.02(0.86-1.24)
Pollitt et al. 2016	Zinc	18.2 ng/m^3	children	airway inflammation due to asthma	0.99(0.96-1.03)

References

1 Haugen, M. J. & Bishop, G. A. Long-Term Fuel-Specific NO x and Particle Emission Trends for In-Use Heavy-Duty Vehicles in California. *Environmental science & technology* **52**, 6070-6076 (2018).

2 Haugen, M. J., Bishop, G. A., Thiruvengadam, A. & Carder, D. K. Evaluation of Heavy-and Medium-Duty On-Road Vehicle Emissions in California's South Coast Air Basin. *Environmental science & technology* **52**, 13298-13305 (2018).

3 Khreis, H. & Nieuwenhuijsen, M. J. Traffic-related air pollution and childhood asthma: recent advances and remaining gaps in the exposure assessment methods. *International journal of environmental research and public health* **14**, 312 (2017).

4 Kwon, J.-W. *et al.* Emergency Department visits for asthma exacerbation due to weather conditions and air pollution in Chuncheon, Korea: a case-crossover analysis. *Allergy, asthma & immunology research* **8**, 512-521 (2016).

5 Magzamen, S. *et al.* in *ISEE Conference Abstracts.*

6 Carlsten, C. *et al.* Diesel exhaust augments allergen-induced lower airway inflammation in allergic individuals: a controlled human exposure study. *Thorax* **71**, 35-44 (2016).

7 Kumar, P., Pirjola, L., Ketzel, M. & Harrison, R. M. Nanoparticle emissions from 11 non-vehicle exhaust sources–a review. *Atmospheric Environment* **67**, 252-277 (2013).

8 Hasunuma, H. *et al.* Association between traffic-related air pollution and asthma in preschool children in a national Japanese nested case-control study. *BMJ open* **6**, e010410, doi:10.1136/bmjopen-2015-010410 (2016).

9 Gorai, A. K., Tchounwou, P. B. & Tuluri, F. Association between Ambient Air Pollution and Asthma Prevalence in Different Population Groups Residing in Eastern Texas, USA. *Int J Environ Res Public Health* **13**, doi:10.3390/ijerph13040378 (2016).

10 Zhang, S., Li, G., Tian, L., Guo, Q. & Pan, X. Short-term exposure to air pollution and morbidity of COPD and asthma in East Asian area: A systematic review and meta-analysis. *Environmental research* **148**, 15-23, doi:10.1016/j.envres.2016.03.008 (2016).

11 Joseph, C. L. *et al.* Identifying students with self-report of asthma and respiratory symptoms in an urban, high school setting. *Journal of urban health : bulletin of the New York Academy of Medicine* **84**, 60-69, doi:10.1007/s11524-006-9121-y (2007).

12 Jordan, K., Jinks, C. & Croft, P. Health care utilization: measurement using primary care records and patient recall both showed bias. *J Clin Epidemiol* **59**, 791-797, doi:10.1016/j.jclinepi.2005.12.008 (2006).

13 Rockenbauer, M. *et al.* Recall bias in a case-control surveillance system on the use of medicine during pregnancy. *Epidemiology* **12**, 461-466, doi:Doi 10.1097/00001648-200107000-00017 (2001).

14 de Marco, R. *et al.* Incidence and remission of asthma: A retrospective study on the natural history of asthma in Italy. *J Allergy Clin Immun* **110**, 228-235, doi:Unsp 1/81/125600
10.1067/Mai.2002.125600 (2002).

15 Guarnieri, M. & Balmes, J. R. Outdoor air pollution and asthma. *Lancet* **383**, 1581-1592 (2014).

16 Levy, I., Mihele, C., Lu, G., Narayan, J. & Brook, J. R. Evaluating Multipollutant Exposure and Urban Air Quality: Pollutant Interrelationships, Neighborhood Variability, and Nitrogen Dioxide as a Proxy Pollutant. *Environ Health Persp* **122**, 65-72, doi:10.1289/ehp.1306518 (2014).

17 Su, J. G. *et al.* Feasibility of Deploying Inhaler Sensors to Identify the Impacts of Environmental Triggers and Built Environment Factors on Asthma Short-Acting Bronchodilator Use. *Environ Health Perspect* **125**, 254-261, doi:10.1289/EHP266 (2017).

18 Zhou, C. Y., Jia, Y., Motani, M. & Chew, J. W. Learning Deep Representations from Heterogeneous Patient Data for Predictive Diagnosis. *Acm-Bcb' 2017: Proceedings of the 8th Acm International Conference on Bioinformatics, Computational Biology,and Health Informatics*, 115-123, doi:10.1145/3107411.3107433 (2017).

19 Bellinger, C., Jabbar, M. S. M., Zaïane, O. & Osornio-Vargas, A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health* **17**, 907 (2017).

20 Hamra, G. B. & Buckley, J. P. Environmental exposure mixtures: questions and methods to address them. *Current epidemiology reports* **5**, 160-165 (2018).

21 Huang, H. *et al.* Cumulative risk and impact modeling on environmental chemical and social stressors. *Current environmental health reports* **5**, 88-99 (2018).

22 Luo, L., Hudson, L. G., Lewis, J. & Lee, J.-H. Two-step approach for assessing the health effects of environmental chemical mixtures: application to simulated datasets and real data from the Navajo Birth Cohort Study. *Environ Health-Glob* **18**, 46 (2019).

23 Liu, S. H. *et al.* Modeling the health effects of time-varying complex environmental mixtures: Mean field variational Bayes for lagged kernel machine regression. *Environmetrics* **29**, e2504 (2018).

24 Beeler, C. *et al.* Assessing patient risk of central line-associated bacteremia via machine learning. *Am J Infect Control* **46**, 986-991, doi:10.1016/j.ajic.2018.02.021 (2018).

25 Breiman, L. Random forests. *Mach Learn* **45**, 5-32, doi:Doi 10.1023/A:1010933404324 (2001).

26 Vicendese, D. *et al.* Bedroom air quality and vacuuming frequency are associated with repeat child asthma hospital admissions. *Journal of Asthma* **52**, 727-731 (2015).

27 Schneeweiss, S. *et al.* Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology (Cambridge, Mass.)* **28**, 237-248 (2017).

28 Mansiaux, Y. & Carrat, F. Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. *BMC medical research methodology* **14**, 99 (2014).

29 Katsis, Y. *et al.* in *Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies.* 222-231 (IEEE Press).

30 Zhang, K., Li, Y., Schwartz, J. D. & O'Neill, M. S. What weather variables are important in predicting heat-related mortality? A new application of statistical learning methods. *Environ Res* **132**, 350-359, doi:https://doi.org/10.1016/j.envres.2014.04.004 (2014).

31 Shirinde, J., Wichmann, J. & Voyi, K. Association between wheeze and selected air pollution sources in an air pollution priority area in South Africa: a cross-sectional study. *Environ Health-Glob* **13**, 1-12 (2014).

32 Tabaku, A., Bejtja, G., Bala, S., Toci, E. & Resuli, J. Effects of air pollution on children's pulmonary health. *Atmos Environ* **45**, 7540-7545 (2011).

33 Salvi, S. Health effects of ambient air pollution in children. *Paediatric respiratory reviews* **8**, 275-280 (2007).

34 Fisher, J. E. *et al.* Physical activity, air pollution, and the risk of asthma and chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine* **194**, 855-865 (2016).

35 Christian, H. *et al.* Traffic exposure, air pollution and children's physical activity at early childhood education and care. *International Journal of Hygiene and Environmental Health* **240**, 113885 (2022).

36 Bateson, T. F. & Schwartz, J. Children's response to air pollutants. *Journal of Toxicology and Environmental Health, Part A* **71**, 238-243 (2007).

37 Liu, Z. & Yu, L. Stay or leave? The role of air pollution in urban migration choices. *Ecological Economics* **177**, 106780 (2020).

38 Wu, K.-G., Chang, C.-Y., Yen, C.-Y. & Lai, C.-C. Associations between environmental heavy metal exposure and childhood asthma: A population-based study. *Journal of Microbiology, Immunology and Infection* **52**, 352-362 (2019).

39 Mao, S., Wu, L. & Shi, W. Association between trace elements levels and asthma susceptibility. *Respiratory medicine* **145**, 110-119 (2018).

40 Pollitt, K. J. G. *et al.* Trace metal exposure is associated with increased exhaled nitric oxide in asthmatic children. *Environmental Health* **15**, 1-11 (2016).

41 Team, O. N. A. Vol. 3.

42 Zhan, Y. *et al.* Satellite-Based Estimates of Daily NO2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model. *Environmental science & technology* **52**, 4180-4189 (2018).

43 Kloog, I. *et al.* A new hybrid spatio-temporal model for estimating daily multi-year PM2. 5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment* **95**, 581-590 (2014).

44 Oude Nijhuis, A. *The influence of stratospheric temperature changes on ozone trends: Analysis of OMI ozone products and improvements for the differential optical absorption spectroscopy (DOAS) technique that is applied to OMI satellite measurements*, (2012).

45 Robinson, N. P. *et al.* A dynamic Landsat derived normalized difference vegetation index (NDVI) product for the conterminous United States. *Remote Sensing* **9**, 863 (2017).

46 Su, J. G. *et al.* Identification of effects of regulatory actions on air quality in goods movement corridors in California. *Environmental science & technology* **50**, 8687-8696 (2016).

47 Yu, Y. *et al.* Ozone Exposure, Outdoor Physical Activity, and Incident Type 2 Diabetes in the SALSA Cohort of Older Mexican Americans. *Environ Health Persp* **129**, 097004 (2021).

48 Su, J. G., Jerrett, M., Meng, Y.-Y., Pickett, M. & Ritz, B. Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. *Sci Total Environ* **506**, 518-526 (2015).

49 Solomon, P. A. *et al.* Mobile-platform measurement of air pollutant concentrations in California: performance assessment, statistical methods for evaluating spatial variations, and spatial representativeness. *Atmospheric Measurement Techniques* **13**, 3277-3301 (2020).

50 Apte, J. S. *et al.* High-resolution air pollution mapping with Google street view cars: exploiting big data. *Environmental science & technology* **51**, 6999-7008 (2017).

51 Kanaroglou, P. S. *et al.* Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmos Environ* **39**, 2399-2409 (2005).

52 Su, J. G. *et al.* Predicting differential improvements in annual pollutant concentrations and exposures for regulatory policy assessment. *Environment International* **143**, 105942 (2020).

53 Beckerman, B. S. *et al.* A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States. *Environ Sci Technol* **47**, 7233-7241, doi:10.1021/es400039u (2013).

54 Su, J. G., Jerrett, M., Meng, Y. Y., Pickett, M. & Ritz, B. Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. *Sci Total Environ* **506**, 518-526, doi:10.1016/j.scitotenv.2014.11.022 (2015).

55 Ghobadi, H., Ahari, S. S., Kameli, A. & Lari, S. M. The relationship between COPD assessment test (CAT) scores and severity of airflow obstruction in stable COPD patients. *Tanaffos* **11**, 22 (2012).

56 Fornadley, J. A. Stepwise treatment of asthma. *Otolaryngologic Clinics of North America* **47**, 65-75 (2014).

57 Su, J. G. *et al.* Identifying impacts of air pollution on subacute asthma symptoms using digital medication sensors. *International Journal of Epidemiology* (2021).

58 Payne, E. H., Gebregziabher, M., Hardin, J. W., Ramakrishnan, V. & Egede, L. E. An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. *Communications in Statistics-Simulation and Computation* **47**, 1722-1738 (2018).

59 Yoo, W. *et al.* A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology* **4**, 9 (2014).

60 Dormann, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27-46 (2013).

61 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22 (2002).

62 Su, J. G. *et al.* Modeling particulate matter concentrations measured through mobile monitoring in a deletion/substitution/addition approach. *Atmos Environ* **122**, 477-483, doi:10.1016/j.atmosenv.2015.10.002 (2015).

63 Greenwell, B. M. pdp: an R Package for constructing partial dependence plots. *The R Journal* **9**, 421-436 (2017).

64 Katz, M. H. Multivariable analysis: a primer for readers of medical research. *Annals of internal medicine* **138**, 644-650 (2003).

65 Su, J. G. *et al.* Feasibility of deploying inhaler sensors to identify the impacts of environmental triggers and built environment factors on asthma short-acting bronchodilator use. *Environ Health Persp* **125**, 254-261 (2017).