Health Effects Institute

NUMBER 226
MARCH 2025

RESEARCH REPORT

Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models

Gerard Hoek, Femke Bouma, Nicole Janssen, Joost Wesseling, Sjoerd van Ratingen, Jules Kerckhoffs, Ulrike Gehring, Wouter Hendricx, Roel Vermeulen, and Kees de Hoogh

INCLUDES A COMMENTARY BY THE INSTITUTE'S IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL

www.healtheffects.org

Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models

Gerard Hoek, Femke Bouma, Nicole Janssen, Joost Wesseling, Sjoerd van Ratingen, Jules Kerckhoffs, Ulrike Gehring, Wouter Hendricx, Roel Vermeulen, and Kees de Hoogh

with a Commentary by the HEI Improved Exposure Assessment Studies Review Panel

Research Report 226
Health Effects Institute
Boston, Massachusetts

Trusted Science · Cleaner Air · Better Health

Publishing history: This report was posted at *www.healtheffects.org* in March 2025.

Citation for report:

Hoek G, Bouma F, Janssen N, Wesseling J, van Ratingen S, Kerckhoffs J, et al. 2025. Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models. Research Report 226. Boston, MA: Health Effects Institute.

 Compositor: Progressive Publishing Services, York, PA. Library of Congress Catalog Number for the HEI Report Series: WA 754 R432.

ISSN: 2688-6855 (online)

CONTENTS

ABOUT HEI

The Health Effects Institute is a nonprofit corporation chartered in 1980 as an independent research organization to provide high-quality, impartial, and relevant science on the effects of air pollution on health. To accomplish its mission, the Institute

- identifies the highest-priority areas for health effects research
- competitively funds and oversees research projects
- provides intensive independent review of HEI-supported studies and related research
- integrates HEI's research results with those of other institutions into broader evaluations
- communicates the results of HEI's research and analyses to public and private decision-makers.

HEI typically receives balanced funding from the US Environmental Protection Agency and the worldwide motor vehicle industry. Frequently, other public and private organizations in the United States and around the world also support major projects or research programs. HEI has funded more than 390 research projects in North America, Europe, Asia, and Latin America, the results of which have informed decisions regarding carbon monoxide, air toxics, nitrogen oxides, diesel exhaust, ozone, particulate matter, and other pollutants. These results have appeared in more than 275 comprehensive reports published by HEI, as well as in more than 2,500 articles in peer-reviewed literature.

HEI's independent Board of Directors consists of leaders in science and policy who are committed to fostering the public–private partnership that is central to the organization. The Research Committee solicits input from HEI sponsors and other stakeholders and works with scientific staff to develop a Five-Year Strategic Plan, select research projects for funding, and oversee their conduct. The Review Committee or Panel, which has no role in selecting or overseeing studies, works with staff to evaluate and interpret the results of funded studies and related research.

All project results and accompanying comments by the Review Committee or Panel are widely disseminated through HEI's website (*www.healtheffects.org*), reports, newsletters, annual conferences, and presentations to legislative bodies and public agencies.

ABOUT THIS REPORT

Research Report 226, *Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models*, presents a research project funded by the Health Effects Institute and conducted by Dr. Gerard Hoek at Utrecht University, the Netherlands, and his colleagues. The report contains three main sections:

The HEI Statement, prepared by staff at HEI, is a brief, nontechnical summary of the study and its findings; it also briefly describes the Review Panel's comments on the study.

The Investigators' Report, prepared by Hoek and colleagues, describes the scientific background, aims, methods, results, and conclusions of the study.

The Commentary, prepared by members of the Review Panel with the assistance of HEI staff, places the study in a broader scientific context, points out its strengths and limitations, and discusses the remaining uncertainties and implications of the study's findings for public health and future research.

This report has gone through HEI's rigorous review process. When an HEI-funded study is completed, the investigators submit a draft final report presenting the background and results of the study. Outside technical reviewers first examine this draft report. The report and the reviewers' comments are then evaluated by members of the Review Panel, an independent panel of distinguished scientists who are not involved in selecting or overseeing HEI studies. During the review process, the investigators have an opportunity to exchange comments with the Review Panel and, as necessary, to revise their report. The Commentary reflects the information provided in the final version of the report.

Although this report was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not necessarily reflect the views of the Agency; no official endorsement by it should be inferred. The contents of this report also have not been reviewed by private party institutions, including those that support the Health Effects Institute, and may not reflect the views or policies of these parties; no endorsement by them should be inferred.

CONTRIBUTORS

RESEARCH COMMITTEE

David A. Savitz, Chair *Professor of Epidemiology, School of Public Health, and Professor of Obstetrics and Gynecology and Pediatrics, Alpert Medical School, Brown University, USA*

Benjamin Barratt *Professor, Environmental Research Group, School of Public Health, Imperial College London, United Kingdom*

David C. Dorman *Professor, Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, USA*

Christina H. Fuller *Associate Professor, School of Environmental, Civil, Agricultural and Mechanical Engineering, University of Georgia College of Engineering, USA*

Marianne Hatzopoulou *Professor, Civil and Mineral Engineering, University of Toronto, Research Chair in Transport Decarbonization and Air Quality, Canada*

Heather A. Holmes *Associate Professor, Department of Chemical Engineering, University of Utah, USA*

Markey Johnson* *Air Health Science Division, Health Canada, Ottawa, Ontario, Canada*

Neil Pearce *Professor of Epidemiology and Biostatistics, London School of Hygiene and Tropical Medicine, United Kingdom*

Evangelia (Evi) Samoli *Professor of Epidemiology and Medical Statistics, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece*

Alexandra M. Schmidt *Professor of Biostatistics, School of Population and Global Health, McGill University, Canada*

Neeta Thakur *Associate Professor of Medicine, University of California San Francisco, USA*

Gregory Wellenius *Professor, Department of Environmental Health, Boston University School of Public Health and Director, BUSPH Center for Climate and Health, USA*

IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL

Jana Milford, Chair *Professor Emerita, Department of Mechanical Engineering and Environmental Engineering Program, University of Colorado Boulder, Colorado, USA*

Susanne Breitner-Busch *Senior Scientist, IBE Chair of Epidemiology, LMU Munich and Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany*

Jamie Hart *Associate Professor, Brigham & Women's Hospital, Harvard Medical School, and Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA*

Anna Oudin *Associate Professor, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden, and Division of Occupational and Environmental Medicine, Lund University, Lund, Sweden*

John Volckens *Professor, Department of Mechanical Engineering, Walter Scott Jr. College of Engineering, Colorado State University, Fort Collins, Colorado, USA*

Shu Yang *Associate Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA*

Jeff Yanosky *Associate Professor, Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania, USA*

HEI PROJECT STAFF

Dan Crouse *Senior Scientist (Study Oversight)*

Hanna Boogaard *Consulting Principal Scientist (Report Review)*

Kristin C. Eckles *Senior Editorial Manager*

Hope Green *Editorial Project Manager*

Tom Zaczkiewicz *Consulting Editor*

*Consultant.

PREFACE

HEI's Research to Assess Health Effects of Traffic-Related Air Pollution and to Improve Exposure Assessment for Health Studies

INTRODUCTION

Traffic emissions are an important source of urban air pollution and have been linked to various adverse health outcomes (Atkinson et al. 2018; Health Canada 2016; HEI 2010; HEI 2022a; Huangfu and Atkinson 2020; US Environmental Protection Agency [US EPA] 2016). Over the last several decades, air quality regulations and improvements in vehicular emission control technologies have steadily decreased emissions from motor vehicles. As a result, ambient concentrations of several major traffic-related air pollutants have decreased in most high-income countries, even as vehicle miles traveled and economic activity have increased and older or malfunctioning vehicles have remained on the roads (HEI 2022a; US EPA 2023).

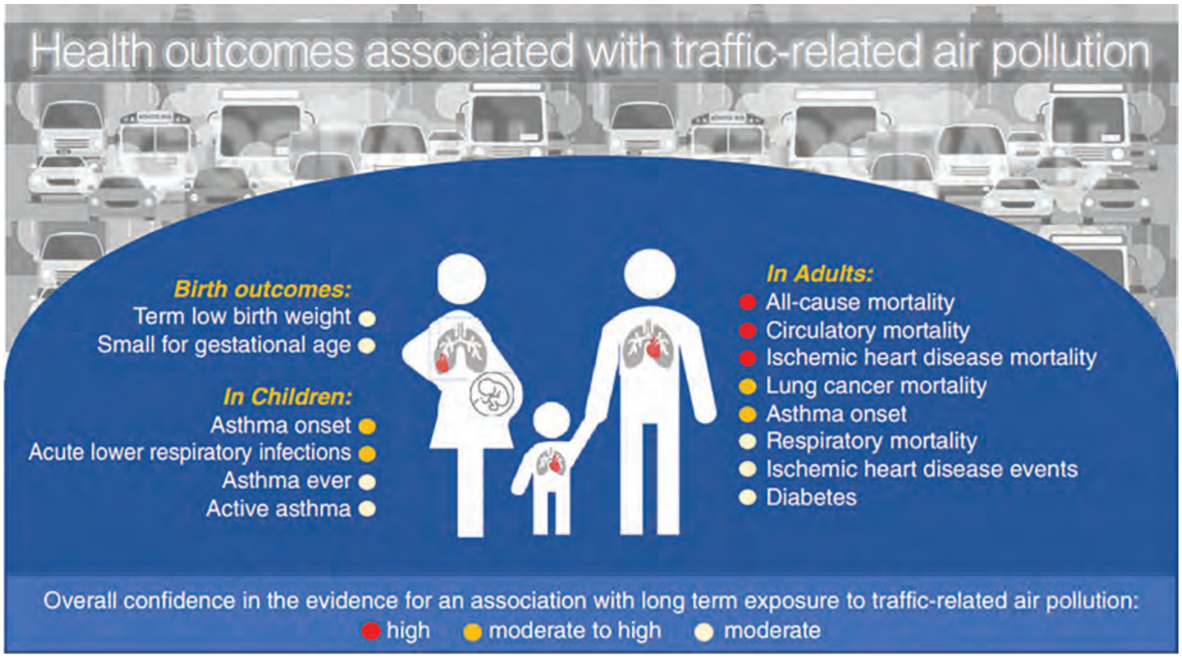
Following HEI's widely cited 2010 Report (HEI 2010), HEI published *Special Report 23*, a systematic review of more than 350 epidemiological studies on the health effects of long-term exposure to emissions of primary traffic-related air pollutants (HEI 2022a). The 2022 report found a high level of confidence that strong connections exist between traffic-related air pollution and early death due to cardiovascular diseases. A strong connection was also found between traffic-related air pollution and lung cancer mortality, asthma onset in children and adults, and acute lower respiratory infections in children (**Preface Figure**). The confidence in the evidence was considered moderate, low, or very low for the other selected outcomes, such as coronary events, diabetes, and adverse birth outcomes.

Although traffic-related emissions have decreased over the past decades, further research is warranted in several areas. Emerging evidence suggests that transportation can affect health through many intertwined pathways such as collisions, noise, climate change, temperature, stress, and the lack of physical activity and green space (Glazener et al. 2021). As tailpipe emissions from internal combustion engines decrease and electric vehicles increase market share, more studies are needed to quantify human exposures to nontailpipe particulate matter better and to assess the health effects associated with those exposures. Relatively few studies evaluate how influential factors such as green space, heat exposure, noise pollution, and physical activity interact with or modify air pollution health effects. Evaluation of those factors and exposures are critical because they reflect real-world conditions and might further advance our understanding of the implications of transportation activities on traffic-related air pollution and health (Khreis et al. 2020).

Moreover, better understanding is needed of the role of specific pollutants including nitrogen dioxide (NO_2) and ultrafine particles (UFPs), the health effects of short-term exposures versus long-term exposures, the effects on a broader range of health outcomes (such as neurological and birth outcomes) that have not been extensively examined, and the ways in which marginalized communities are affected. However, a challenge for exposure assessment of traffic-related air pollution is that traffic emits a complex mixture of pollutants in particulate and gaseous forms, many of which are also emitted by other sources. In addition, traffic-related air pollution is characterized by high spatial and temporal variability, with the highest concentrations occurring at or near major roads. Therefore, it has been difficult to identify an appropriate exposure metric that uniquely indicates traffic-related air pollution and to model the distribution of exposure at a sufficiently high degree of spatial and temporal resolution.

Various air quality models — such as dispersion, land use regression, and hybrid models — have been developed to estimate long-term exposure to air pollution (HEI 2022a; Hoek 2017; Jerrett et al. 2005). Recent developments in measurement technologies and approaches to modeling long-term exposure to air pollution have increasingly been used to provide air pollution estimates at fine spatial scales for epidemiological studies of large populations. Advances include novel air pollution sensors, mobile monitoring, satellite data, hybrid models, and machine-learning approaches (Hoek 2017).

Moreover, many improvements in exposure models have occurred over time with the advance of geographic information system approaches and the application of more sophisticated statistical methods; see, for example, several studies previously funded by HEI: Apte et al. 2024, Barratt et al. 2018,



Preface Figure. Overall confidence in the evidence for an association between long-term exposure to traffic-related air pollution and selected health outcomes. Health outcomes for which the overall confidence in the evidence was low to moderate, low, or very low are not in the figure. Reproduced from HEI 2022a.

Batterman et al. 2020, Frey et al. 2022, and Sarnat et al. 2018. However, the usefulness of exposure estimates still depends on the model assumptions and input data quality, and there remain limitations and challenges when predicting air pollution exposure, particularly for such pollutants as UFPs, NO_2, and black carbon (BC) that vary highly in space and time. Few studies have compared the performance of different models and evaluated exposure measurement error and possible bias in health estimations.

Thus, HEI issued complementary requests for applications in 2017 (RFA 17-1) and 2019 (RFA 19-1) to evaluate traffic-related health effects in the context of spatially correlated factors — specifically traffic noise, socioeconomic status, and green space — and to improve exposure assessment for health studies.

OBJECTIVES OF THE RFAs

OBJECTIVES OF RFA 17-1

RFA 17-1, *Assessing Adverse Health Effects of Exposure to Traffic-Related Air Pollution, Noise, and Their Interactions with Socioeconomic Status*, solicited studies that sought to assess adverse health effects from exposure to traffic-related air pollution and to disentangle the effects from spatially correlated confounding or modifying factors — most notably, traffic noise, socioeconomic status, and the built environment, including green space. The RFA had five major objectives:

1. In the proposed health studies, develop, validate, and apply improved exposure assessment methods and models suitable for estimating exposure to traffic-related air pollution that take into account other air pollution sources in urban areas (such as airports, [sea]ports, industries, and other local point sources) and that would be able to distinguish between tailpipe and nontailpipe traffic emissions.
2. Propose ways in these studies to disentangle the relationship of adverse health effects of traffic-related air pollution and traffic noise.
3. Develop, evaluate, and apply indicators of socioeconomic status at the individual and community level in the proposed health studies; if such indicators are novel, compare with socioeconomic status indicators commonly used in the literature.
4. Explore the role of other factors that might confound or modify the health effects of traffic-related air pollution at the individual (e.g., age, smoking status, diet, physical activity, and health status) and community level (e.g., presence of green space, other factors related to the built environment, and walkability).
5. Investigate — to the extent that the measurements and patterns of a range of different indicators of traffic-related air pollution allow it (e.g., NO_2, UFPs, BC, and indicators of nontailpipe emissions) — whether one or more of them can be shown to have health effects independent of the other pollutants.

OBJECTIVES OF RFA 19-1

RFA 19-1, *Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies*, solicited studies to assess exposures to air pollution using new and conventional exposure assessment approaches, to evaluate quantitatively exposure measurement error to determine the added value of the novel approaches, and to apply the exposure estimates in epidemiological analyses to evaluate the potential effect of exposure measurement error on chronic health estimates. The RFA had four major objectives:

1. Conduct a new monitoring campaign designed to determine long-term exposure to outdoor air pollutants with high spatial and temporal variability by using sensors, mobile monitoring, location tracking, or other approaches.
2. Develop several exposure assessment approaches suitable to estimate long-term exposure to air pollution at relevant spatial and temporal scales for use in an ongoing or future health study.
3. Quantify exposure measurement error by evaluating and comparing the performance of models of long-term air pollution exposure developed under this RFA to the performance of previous models.
4. Apply the various exposure estimates in an ongoing health study to evaluate the potential impact of exposure measurement error in health estimates or explain how the exposure assessments would be directly applicable to future health studies.

DESCRIPTION OF THE RESEARCH PROGRAM

Three 4-year studies were funded under RFA 17-1 and five 3-year studies were funded under RFA 19-1 to cover the various RFA objectives; they are summarized below (**Preface Table**). The study by Hoek and colleagues described in this report (Research Report 226) is the third to be published.

STUDIES FUNDED UNDER RFA 17-1

HEI funded two studies in Europe and one study in the United States to evaluate various aspects of the association between long-term traffic-related air pollution and health by using existing cohorts (Denmark, USA) and a newly recruited cohort (Spain). Two studies focused on health outcomes during pregnancy (Dadvand) and childhood (Franklin), and one study focused on cardiometabolic outcomes in adults (Raaschou-Nielsen).

"Traffic-Related Air Pollution and Birth Weight: The Roles of Noise, Placental Function, Green Space, Physical Activity, and Socioeconomic Status (FRONTIER)," Payam Dadvand and Jordi Sunyer, Barcelona Institute for Global Health (ISGlobal), Spain Dadvand, Sunyer, and colleagues established a new cohort, named Barcelona Life Study Cohort (BiSC) of 1,080 healthy pregnant women in Barcelona, Spain, in 2018. They estimated exposure to various traffic-related pollutants by using hybrid models that included dispersion models, land use data, time-activity data, and personal and home-outdoor air pollution monitoring data. They linked the exposure to various birth outcomes including birth weight, small for gestational age, and fetal growth trajectories. They evaluated the role of traffic noise and green space and also took into account socioeconomic status and maternal stress (in review).

"Intersections as Hot Spots: Assessing the Contribution of Localized Non-Tailpipe Emissions and Noise on the Association between Traffic and Children's Respiratory Health," Meredith Franklin, University of Southern California, Los Angeles Franklin and colleagues developed novel exposure models of tailpipe and nontailpipe air pollutants and noise and applied those models to children's respiratory health in a large Southern California cohort that was also studied in a previous HEI-funded study led by Frank Gilliland; see *HEI Research Report 190*. They made use of the most recent Children's Health Study (CHS) cohort that was initiated in 2003 and included about 2,000 children in eight communities. Longitudinal data on asthma and lung function were collected at various time points (2008–2012) at ages 11 through 16. Air pollution models were supported by particulate matter filters at more than 200 locations in the eight Southern California communities (in review).

"Cardiometabolic Health Effects of Air Pollution, Noise, Green Space and Socioeconomic Status: The HERMES Study," Ole Raaschou-Nielsen, Danish Cancer Society Research Center, Copenhagen, Denmark Raaschou-Nielsen and colleagues evaluated effects of traffic-related air pollution, traffic noise, lack of green space, and other factors on myocardial infarction, stroke, diabetes, and related biomarkers in three cohorts, including an administrative cohort of about 2.6 million Danish adults in the period 2005–2017. They assessed traffic-related air pollution using a chemical transport model for various pollutants, including UFPs and NO_2. In addition, they assessed noise, household- and neighborhood-level socioeconomic status, and various residential green space exposure metrics (*Research Report 222*).

STUDIES FUNDED UNDER RFA 19-1

HEI funded five studies in North America and Europe to evaluate different aspects of improvements to exposure assessment and the application of different exposure assessment approaches to existing cohorts. Three studies are focused on combining novel methods for measuring air pollution and diverse exposure assessment approaches to improve exposure assignment, including machine learning and mobile monitoring (Weichenthal and Hoek) and mobility (de Hoogh). Two studies are testing the added value of incrementally more complex statistical modeling approaches to improving exposure assessment in London (Katsouyanni) and Seattle (Sheppard) and applying their findings to estimating health effects in epidemiological studies.

"Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada," Scott Weichenthal, McGill University, Montreal, Canada Weichenthal and colleagues estimated associations between long-term exposures to UFPs, BC, and other pollutants and mortality in Toronto and Montreal, Canada, using several exposure modeling approaches. They conducted mobile monitoring campaigns in both cities and used those newly collected data to develop various high-resolution exposure models, including land use regression and machine learning. They then evaluated how the effect estimates for nonaccidental and cause-specific mortality in the Canadian Census Health and Environment Cohort (CanCHEC) are influenced by different exposure models (*Research Report 217*).

"Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models," Gerard Hoek, Utrecht University, The Netherlands Hoek and colleagues compared the performance of a suite of long-term exposure assessment methods in the Netherlands for four air pollutants. The predictions of the exposure models were compared at 20,000 random residential addresses in the Netherlands and tested on existing and new validation data over a 20-year period. They applied the various models to three major cohorts in the Netherlands — an administrative cohort of 10.8 million adults (DUELS), the European Prospective Investigation into Cancer and Nutrition Netherlands (EPIC-NL), and the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) birth cohort — to evaluate how they influence health effect estimates in epidemiological studies (current report).

"Accounting for Mobility in Air Pollution Exposure Estimates in Studies on Long-Term Health Effects (MOBI-AIR)," Kees de Hoogh, Swiss Tropical and Public Health Institute, Basel, Switzerland de Hoogh and colleagues used agent-based modeling to model mobility patterns in Switzerland and the Netherlands based on travel survey information. They used location tracking using a mobile phone application and GPS units for about 700 individuals in the Netherlands and Switzerland. They then compared exposure estimates accounting for individual mobility to those accounting only for home addresses in association with health effects estimates in three major cohorts: the Study on Air Pollution and Lung Disease in Adults (SAPALDIA) in Switzerland, participants in the European Prospective Investigation into Cancer and Nutrition Netherlands (EPIC-NL), and the Swiss National Cohort (SNC) (in review).

"Investigating the Consequences of Measurement Error of Gradually More Sophisticated Long-Term Personal Exposure Models in Assessing Health Effects: The London Study (MELONS)," Klea Katsouyanni, Imperial College, United Kingdom Katsouyanni and colleagues evaluated whether successively more detailed estimates of long-term exposure to outdoor air pollution can be used to produce more accurate and realistic estimates of the health effects associated with exposure than are produced by other, less detailed approaches. They leveraged personal exposure data from four earlier studies in London. They compared predictions from various exposure models that accounted for exposure to indoor sources and mobility by using several types of air pollution models (dispersion, land use regression, machine learning, and hybrid models). Finally, exposures were applied to the London segment of the UK Biobank study with about 62,000 participants to evaluate associations with several health outcomes (in review).

"Optimizing Exposure Assessment with Application to Cognitive Function," Lianne Sheppard, University of Washington, Seattle Sheppard and colleagues compared the performance of different exposure assessment study design features on long-term exposure and health estimates in Seattle, for example fewer visits per site, fewer days of the week, fewer times of day, and fewer seasons. They leveraged detailed air pollution data, including a mobile monitoring campaign of UFPs, and cognitive function data from about 5,000 participants in the Adult Changes in Thought (ACT) Air Pollution study. The investigators used either the full dataset or subsets of measurements to develop annual average exposure estimates using a suite of models, including universal kriging, and machine learning models. In particular, they used statistical techniques to assess the bias and precision of health effect estimates and provided practical guidance on future mobile monitoring campaigns (in review).

FURTHER RESEARCH UNDERWAY

Given the large number of people exposed to traffic-related air pollution — both in and beyond the near-road environment — exposures to traffic-related air pollution remain an important public health concern and deserve greater attention from the public and from policymakers.

Although emissions from automobile exhaust systems have decreased in recent years, emissions from the use and wear of brakes, tires, and other nontailpipe sources now contribute a higher fraction of the particulate emissions. Therefore, HEI funded two ongoing studies funded under *RFA 21-1*, *Quantifying Real-World Impacts of Non-Tailpipe Particulate Matter Emissions*. The two studies involve measurements of mass and composition of ambient particles from nontailpipe motor vehicle sources to disentangle nontailpipe and tailpipe pollution and better understand how each effects human health. One study is measuring concentrations of nontailpipe particulate matter across Toronto, Canada, to determine how much nontailpipe pollution people might breathe in everyday life and how to improve measurement of these exposures in the future. The other study is a panel study in which asthmatic adults rode stationary bicycles on sidewalks in three different exposure environments in London, United Kingdom, to measure how exposure to traffic with different mixtures of nontailpipe and tailpipe emissions affects lung function.

Building on its prior and ongoing research and the recommendations from its systematic traffic review, HEI issued *RFA 23-1*, *Assessing Health Effects of Traffic-Related Air Pollution in a Changing Urban Transportation Landscape*. Investigators funded under RFA 23-1 will conduct epidemiological and health impact assessment studies to assess current and potential future population-level health effects and health burdens associated with current and future transportation systems and traffic-related air pollution. The studies began in late spring 2024. HEI also publishes reports on the State of Global Air to communicate the relationship between air quality and health around the world; see, for example, a recent report on cities and NO_2 (HEI 2022b).

Preface Table. Key Characteristics of HEI's Research to Assess Health Effects of Traffic-Related Air Pollution and to Improve Exposure Assessment for Health Studies

Principal Investigator	Study Name	Location	Study Period	Study Population	Sample Size	Outcomes	Main Air Pollutants	Monitoring Data	Exposure Assessment
RFA 17-1, Assessing Adverse Health Effects of Exposure to Traffic-Related Air Pollution, Noise, and Their Interactions with Socioeconomic Status									
Dadvand	FRONTIER (BiSC)	Barcelona, Spain	2018–2022	Newborns	1,080	Birth weight, small for gestational age, fetal growth trajectories, and placental function	BC, NO_2, $PM_{2.5}$, Cu, Fe, and Zn	Personal, indoor, and outdoor home measurements	LUR, dispersion, and hybrid models
Franklin	CHS	Southern California	2008–2012	Children	2,000	Asthma and lung function	PM_{coarse}, $PM_{2.5}$, Cu, Fe, Zn, and many other elemental components	Outdoor home and school measurements and measurements near road intersections	Machine learning and LUR models
Raaschou-Nielsen	HERMES (DK-POP, DNHS, DCH–NG)	Denmark	2005–2017	Adults	2.9 million	Myocardial infarction, stroke, and diabetes	UFPs, EC, NO_2, $PM_{2.5}$	NA	Chemical transport model
RFA 19-1, Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies									
Weichenthal	CanCHEC	Montreal and Toronto, Canada	1991–2016	Adults	1.5 million	Mortality	UFPs, BC	Mobile	Machine learning and LUR models
Hoek	DUELS, EPIC-NL, PIAMA	Netherlands	1993–2019	Children and adults	10 million	Mortality, stroke, coronary events, lung function, and asthma	UFPs, NO_2, BC and $PM_{2.5}$	Mobile, outdoor low-cost sensors, regulatory monitors	LUR, dispersion, machine-learning, and hybrid models

continued on next page

Principal Investigator	Study Name	Location	Study Period	Study Population	Sample Size	Outcomes	Main Air Pollutants	Monitoring Data	Exposure Assessment
de Hoogh	MOBI-AIR (EPIC-NL, SAPALDIA, SNC)	Netherlands, Switzerland	1991–2018	Adults	3.5 million	Mortality, stroke, coronary events, lung function, and blood pressure	NO_2, $PM_{2.5}$	Personal measurements, location tracking	Agent-based, LUR, and machine-learning models
Katsouyanni	MELONS (BLW, COPE, DEMiSt, PASTA, London segment of UK Biobank)	London, UK	2006–2024	Adults	62,000	Mortality, asthma, COPD, myocardial infarction, stroke, and dementia	BC, NO_2, $PM_{2.5}$, O_3	Personal measurements, regulatory monitors	LUR, dispersion, machine learning, and hybrid models
Sheppard	ACT Air pollution study	Seattle	1994–2020	Older adults	5,400	Cognitive function	UFPs, $PM_{2.5}$, and NO_2	Mobile, outdoor low-cost sensors	Universal kriging, spatiotemporal models, and machine-learning models

ACT = Adult Changes in Thought; BiSC = Barcelona Life Study Cohort; BLW = Breathe London Wearables; CanCHEC = Canadian Census Health and Environment Cohort; CHS = Children's Health Study; COPE = Characterisation of COPD Exacerbations using Environmental Exposure Modelling; DCH–NG = Diet, Cancer and Health–Next Generations cohort; DEMiSt = Driver Diesel Exposure Mitigation Study; DK–POP = Danish Population cohort; DNHS = Danish National Health Survey; DUELS = Dutch Environmental Longitudinal Study; EPIC–NL = European Prospective Investigation on Cancer and Nutrition–Netherlands; PIAMA = Prevention and Incidence of Asthma and Mite Allergy; NA = not applicable; PASTA = Physical Activity through Sustainable Transport Approaches; SAPALDIA = Swiss Study on Air Pollution and Lung Disease in Adults; SNC = Swiss National Cohort.

Looking ahead, HEI continues to support improvements in exposure assessment via the use of new technologies, such as satellite remote sensing data. HEI held a *workshop* to discuss applications of high-quality satellite remote sensing data, which have opportunities for increased use in large epidemiological studies, studying the health effects of wildfires, and addressing environmental justice concerns. Challenges include the complexities of data assimilation and accessibility, and current data and algorithmic limitations. HEI recently issued RFA 25-1, *Advancing Satellite-Derived Air Quality Data and Approaches for Use in Health Studies.* The overarching goal of this RFA is to develop a resource for health research that links satellite-derived air quality products to quantified uncertainties and strengthens the understanding of the implications of such uncertainties for exposure, epidemiological, and health assessment research.

REFERENCES

Apte JS, Chambliss SE, Messier KP, Gani S, Upadhya AR, Kushwaha M, et al. 2024. Scalable Multipollutant Exposure Assessment Using Routine Mobile Monitoring Platforms. Research Report 216. Boston, MA: Health Effects Institute.

Atkinson RW, Butland BK, Anderson HR, Maynard RL. 2018. Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies. Epidemiology 29:460–472, *doi:10.1097/EDE.0000000000000847.*

Barratt B, Lee M, Wong P, Tang R, Tsui TH, Cheng W, et al. 2018. A Dynamic Three-Dimensional Air Pollution Exposure Model for Hong Kong. Research Report 194. Boston, MA: Health Effects Institute.

Batterman S, Berrocal VJ, Milando C, Gilani O, Arunachalam S, Zhang KM. 2020. Enhancing Models and Measurements of Traffic-Related Air Pollutants for Health Studies Using Dispersion Modeling and Bayesian Data Fusion. Research Report 202. Boston, MA: Health Effects Institute.

Frey HC, Grieshop AP, Khlystov A, Bang JJ, Rouphail N, Guinness J, et al. 2022. Characterizing Determinants of Near-Road Ambient Air Quality for an Urban Intersection and a Freeway Site. Research Report 207. Boston, MA: Health Effects Institute.

Glazener A, Sanchez K, Ramani T, Zietsman J, Nieuwenhuijsen MJ, Mindell JS, et al. 2021. Fourteen pathways between urban transportation and health: A conceptual model and literature review. J Transport Health 21:101070, *https://doi.org/10.1016/j.jth.2021.101070*.

Health Canada. 2016. Human Health Risk Assessment for Ambient Nitrogen Dioxide. Ottawa, Ontario, Canada: Water and Air Quality Bureau.

HEI Panel on the Health Effects of Traffic-Related Air Pollution. 2010. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. HEI Special Report 17. Boston, MA: Health Effects Institute.

HEI Panel on the Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. 2022a. Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute.

HEI. 2022b. Air Quality and Health in Cities: A State of Global Air Report 2022. Boston, MA: Health Effects Institute. Available: *https://www.stateofglobalair.org/resources/health-in-cities*.

Hoek G. 2017. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. Curr Environ Health Rep 4:450–462, *doi:10.1007/s40572-017-0169-5*.

Huangfu P, Atkinson R. 2020. Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: A systematic review and meta-analysis. Environ Int 144:105998, doi:10.1016/j.envint.2020.105998.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. 2005. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 15:185–204, *doi:10.1038/sj.jea.7500388*.

Khreis H, Nieuwenhuijsen M, Zietsman J, Ramani T (eds.). 2020. Traffic-Related Air Pollution (1st edition). Waltham, MA: Elsevier.

Sarnat JA, Russell A, Liang D, Moutinho JL, Golan R, Weber RJ, et al. 2018. Developing Multipollutant Exposure Indicators of Traffic Pollution: The Dorm Room Inhalation to Vehicle Emissions (DRIVE) Study. Research Report 196. Boston, MA: Health Effects Institute.

US EPA. 2016. Integrated Science Assessment for Oxides of Nitrogen–Health Criteria. EPA/600/R-15/068. Washington, DC: US EPA.

US EPA. 2023. Our Nation's Air Trends Through 2022. Available: *https://gispub.epa.gov/air/trendsreport/2023/* [accessed 22 February 2024].

HEI STATEMENT

Synopsis of Research Report 226

Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models

BACKGROUND

Important limitations and challenges remain when estimating long-term exposure to ambient air pollution for use in epidemiological studies. In 2019, the Health Effects Institute issued the Request for Applications 19-1 to develop and apply novel, scalable approaches to improve assessments of long-term exposures to outdoor air pollutants that vary widely in space and time.

Dr. Hoek was one of the five investigators funded under this Request for Applications. Dr. Hoek and colleagues compared the performance of a suite of air pollution exposure assessment methods in the Netherlands, including a comparison of health effects estimates among different methods. They assessed four pollutants — ultrafine particles, black carbon, fine particulate matter, and nitrogen dioxide — and used existing data and models from previous collaborative projects where available.

What This Study Adds

- The study compared the performance of a suite of long-term exposure assessment methods in the Netherlands for four air pollutants: ultrafine particles, black carbon, fine particulate matter, and nitrogen dioxide.
- The predictions of the exposure models were compared at 20,000 random residen tial addresses in the Netherlands and tested on existing and new validation data over a 20-year period.
- The various exposure models generally resulted in moderately to highly correlated exposure predictions for all pollutants except fine particulate matter.
- The study included a comparison of health effects estimates of the various exposure methods in three cohort studies.
- Similar findings were reported on the presence of an association with various mortality and morbidity outcomes among the various exposure estimates, albeit with sometimes notable differences in the magni tude of the associations.

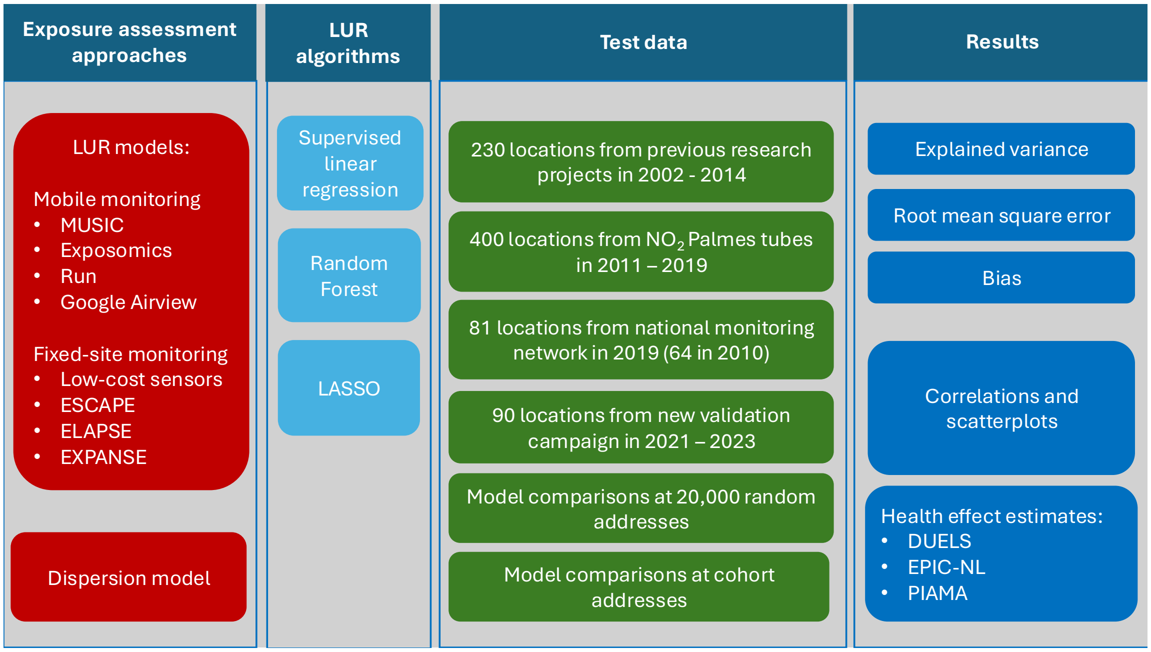
APPROACH

The investigators evaluated annual average outdoor air pollution concentrations using a suite of exposure assessment methods that differ in their monitoring data and modeling approaches, such as land use regression models and air pollution dispersion models (see **Statement Figure 1**). In addition, to better accommodate nonlinearity and complex interactions in the data, they tested various model development algorithms beyond linear regression, including machine learning methods.

The investigators compared predictions of the exposure models at 20,000 random residential addresses in the Netherlands and used existing and new validation data spanning a 20-year period at residential locations where appropriate. The investigators reported multiple measures to test the performance of the exposure models, including correlations, explained variance, and bias.

They conducted epidemiological analyses in three population-based cohort studies (DUELS, EPIC-NL, and PIAMA) to compare health effect estimates using various exposure assessment methods. For all cohort participants, the various exposure estimates were assigned to the residential address at the time of recruitment to the cohort. The investigators used single-pollutant Cox proportional hazard models for mortality and morbidity outcomes in DUELS and EPIC-NL, and linear regression and discrete-time hazard models for lung function and asthma incidence in PIAMA. All results were adjusted for age, sex, and individual and area-level socioeconomic status. EPIC-NL and PIAMA results were further adjusted for various lifestyle factors. The investigators expressed the effect estimates using fixed increments across methods and method-specific interquartile ranges of exposure. In further explorative analyses, the investigators examined if and why effect estimates differed among exposure assessment methods.

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Gerard Hoek at Utrecht University, the Netherlands, and his colleagues. Research Report 226 contains the detailed Investigators' Report and a Commentary on the study prepared by the HEI Improved Exposure Assessment Studies Review Panel.



Statement Figure 1. Schematic overview of the study design. LUR = land use regression

KEY RESULTS

The various exposure models generally resulted in moderately to highly correlated exposure predictions for black carbon, nitrogen dioxide, and ultrafine particles at residential sites across the Netherlands (correlation coefficient $R > 0.7$). In some cases, the predicted exposure levels and exposure contrast varied widely between methods. For fine particulate matter, the correlations between the different models were generally lower ($R < 0.4$), particularly for the poorly performing mobile monitoring models, and exposure contrast was comparatively low. In most comparisons, the investigators observed only small differences among the three land use regression algorithms.

Exposure models explained a low to moderate amount of spatial variance in ultrafine particles and nitrogen dioxide measurements at the new validation sites, and a positive bias (overestimation) was reported for all models. Most models explained a moderate amount of spatial variance in past exposures at existing validation sites for black carbon and nitrogen dioxide and explained less variance for ultrafine particles. Most models predicted past exposures of fine particulate matter poorly.

The application of the various exposure assessment approaches in health studies led to similar findings on the presence of an association with various mortality and morbidity outcomes, albeit with sometimes notable differences in the magnitude of the associations. Positive (adverse) associations were observed most clearly in DUELS and to a lesser extent in PIAMA. Null or sometimes even negative associations were reported in the EPIC-NL cohort. There were no consistent differences in effect estimates between exposure assessment approaches based on mobile monitoring, fixed-site monitoring, or dispersion models.

Factors that explained some of the heterogeneity of effect estimates included the performance of the model at external validation sites and the predicted exposure contrast. The year of the exposure model did not explain the heterogeneity, and the investigators documented stable spatial contrasts over a 10-year period in a subset of models that had multiyear predictions available.

INTERPRETATION AND CONCLUSIONS

In its independent review of the study, the Panel thought the study was well-motivated and effectively leveraged a wealth of air pollution and health data. The comparison of a large suite of exposure models commonly used in epidemiological studies was notable, making the results relevant and widely applicable. The extensive validation efforts and the reporting of multiple measures to test the performance of the exposure models were considered additional strengths. Applying those models in relation to various health outcomes in three different cohorts was another strength. In particular, the health analysis for a very large population

(10.8 million in DUELS) that included all Dutch adult citizens aged 30 years or older was considered informative.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results. The comparison of the different exposure approaches using new validation data was hampered by the short duration of the measurements, the nonsimultaneous sampling, and the small number of repeated measurements for the different seasons. Those issues limited the ability to extrapolate the measurements reliably to annual mean exposures and then compare them with the various predictions from annual average air pollution models. Some comparisons using existing validation data were limited by similar temporal coverage issues.

Substantial heterogeneity was found in the magnitude — but not the direction — of the air pollution associations of the different exposure assessment methods within and across studies. The Panel thought the heterogeneity was not fully explained by the various factors considered in the study and warrants further examination.

Overall, the comprehensive report includes many findings that will be of broad interest and value to a wide audience.

INVESTIGATORS' REPORT

Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models

Gerard Hoek[1], Femke Bouma[1], Nicole Janssen[2], Joost Wesseling[2], Sjoerd van Ratingen[2], Jules Kerckhoffs[1], Ulrike Gehring[1], Wouter Hendricx[2], Roel Vermeulen[1,3], Kees de Hoogh[4,5]

[1]*Institute for Risk Assessment Sciences, Utrecht University, the Netherlands;* [2]*National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands;* [3]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands;* [4]*Swiss Tropical and Public Health Institute, Basel, Switzerland;* [5]*University of Basel, Switzerland*

ABSTRACT

Introduction Assessment of long-term exposure to outdoor air pollution remains a major challenge for epidemiological studies. One of these challenges is characterizing fine-scale spatial variation of the ambient concentrations of key traffic-related air pollutants — including ultrafine particles (UFPs*), black carbon (BC), and nitrogen dioxide (NO_2). Epidemiological studies have used widely different approaches to address these challenges, including empirical land use regression (LUR) models based on fixed-site routine or targeted monitoring, low-cost sensor networks, mobile monitoring, and deterministic dispersion models. Little information is available about the relative performance of these different approaches for assessing long-term exposure to traffic-related air pollution. Different methods may result in heterogeneity in health effect estimates from epidemiological studies applying different exposure–assessment approaches.

The Specific Aims of the study:

1. Develop long-term ambient air pollution exposure estimates for selected epidemiological studies based on low-cost sensors, mobile and fixed-site monitoring, and deterministic dispersion modeling.
2. Compare different exposure assessment methods in terms of their ability to predict spatial variation of long-term average concentrations using external validation data.
3. Compare different exposure assessment methods in terms of air pollution effect estimates in selected epidemiological studies.

We assessed UFPs, NO_2, BC, and particulate matter ≤2.5 µm in aerodynamic diameter ($PM_{2.5}$).

Methods We evaluated annual average air pollution concentrations across the Netherlands using a suite of different exposure models, which differed in modeling approach (empirical LUR, deterministic dispersion models) and monitoring data used (low-cost sensors, mobile monitoring, nationwide and Europewide routine monitoring, and study-specific targeted monitoring). For empirical models, we tested three model development algorithms: supervised linear regression (SLR), Random Forest, and least absolute shrinkage and selection operator (LASSO). The predictions of the models were compared at 20,000 addresses across the Netherlands. The performance was also tested on external validation data, which were obtained from a new campaign (2021–2023) and existing data from different years, allowing assessment of how well recent models predict past air pollution exposure. Epidemiological analyses in three cohort studies were conducted to compare health effect estimates of the different exposure models. We assessed associations of air pollution in a national administrative cohort with natural-cause and cause-specific mortality, in a cohort study that had detailed lifestyle data with natural-cause mortality and incidence of stroke and coronary events, and in a mature birth cohort with lung function and asthma incidence.

Results Exposure predictions at residential sites from the dispersion model and the Europewide hybrid LUR models were available for multiple years in the period 2010–2019. For these models, exposure predictions of different years in the period 2010–2019 were highly correlated for BC, NO_2, and $PM_{2.5}$ (Correlation coefficient $R > 0.9$). Consistently, the year of the exposure model did not affect the presence of an association with mortality and morbidity. Small differences in hazard ratios (HR) were related to exposure contrast for different years. The HR for the association of NO_2 with natural-cause mortality was 1.026 (95% confidence interval [CI]: 1.022–1.031) for the 2010 exposure estimate and 1.030 (1.024–1.035) for the 2019 exposure estimate of the Europewide LUR model, expressed per 10 µg/m^3.

This Investigators' Report is one part of Health Effects Institute Research Report 226, which also includes a Commentary by the HEI Improved Exposure Assessment Studies Review Panel and an HEI Statement about the research project. Correspondence concerning the Investigators' Report may be addressed to Dr. Gerard Hoek, Institute for Risk Assessment Sciences, Utrecht University, Netherlands; email: *g.hoek@uu.nl*. No potential conflict of interest was reported by the authors.

Although this report was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not necessarily reflect the views of the Agency; no official endorsement by it should be inferred. The contents of this report also have not been reviewed by private party institutions, including those that support the Health Effects Institute, and may not reflect the views or policies of these parties; no endorsement by them should be inferred.

* A list of abbreviations and other terms appears at the end of this volume.

The exposure models generally resulted in highly to moderately correlated exposure predictions at residential sites across the Netherlands ($R > 0.7$ for BC, NO_2, and UFPs; $R > 0.5$ for $PM_{2.5}$). The predicted level of exposure and exposure contrast could differ substantially between models and algorithms within models; for example, the interquartile range (IQR) for BC for each of the various models at the 20,000 residential locations ranged between 0.1 and 2.2 μg/m^3. Mobile monitoring studies generally resulted in modestly higher BC concentrations and exposure contrasts compared to other exposure models. Small differences were found between the different models in explaining the spatial variation of air pollution concentrations at the new and existing validation sites. Models explained historical exposure patterns at external sites covering more than 10 years moderately well, especially for BC ($R > 0.7$) and NO_2 ($R > 0.7$), and moderately so for UFPs ($R > 0.5$). Most models predicted the small concentration contrasts of $PM_{2.5}$ relatively poorly.

Consistent with the high correlation of the different exposure models, the application of these models generally resulted in similar conclusions on the presence of associations with natural-cause, respiratory, and lung cancer mortality in the large nationwide cohort, and with asthma incidence and lung function in the birth cohort. However, the effect estimates differed substantially; for example, the HR for natural-cause mortality in the nationwide administrative cohort for a 1 μg/m^3 increase in BC ranged from 1.01 (95% CI: 0.99–1.02) to 1.09 (1.07–1.10). For the outcomes with small effect estimates and the smaller cohort studies, differences in conclusions related to the exposure assessment method were more distinct.

Differences in exposure assessment may contribute substantially to the observed heterogeneity of effect estimates in systematic reviews of epidemiological studies. High heterogeneity was indicated by the commonly used heterogeneity measure I^2, where the value was above 80% for a meta-analysis of the different effect estimates for natural-cause mortality in the nationwide cohort.

Validation of long-term exposure models for the nonroutinely monitored pollutants BC and especially UFPs was challenging, despite generally successful monitoring. The new external validation monitoring campaign resulted in rather unstable estimates of the long-term average spatial contrast, both across sites and where affected by temporal variation, especially for BC and $PM_{2.5}$.

No consistent differences were found in the model performance of SLR, Random Forest, and LASSO, both in internal cross-validation of model building and on external validation sites not used in model building. Exposure predictions from the three algorithms were generally highly correlated and resulted in similar associations with health. However, for individual models, occasionally large differences were found in exposure contrast, validation statistics, and associations with mortality and morbidity outcomes.

There was little benefit in using low-cost sensors for NO_2 and $PM_{2.5}$. The addition of low-cost sensor data did not improve NO_2 estimates in models that combined dispersion model estimates and data from the national monitoring network data.

Conclusions The main conclusions of the project:

- Exposure predictions of BC, NO_2, and $PM_{2.5}$ for different years between 2010–2019 were highly correlated, documenting stable spatial contrast patterns. Consistently, the year of the exposure model did not affect the presence of an association with mortality and morbidity outcomes.
- Models explained historical exposure patterns at external sites covering more than 10 years moderately well, especially for BC.
- Different exposure models generally resulted in highly to moderately correlated exposure predictions. The predicted level of exposure and exposure contrast could differ substantially between models. Small differences were found between the different models in explaining spatial variation at validation sites.
- Application of different exposure models resulted in similar conclusions about the presence of associations with health outcomes, but effect estimates differed substantially in magnitude between individual exposure models. No consistent differences in effect estimates were found between groups of mobile, dispersion, and fixed-site LUR models.
- Differences in exposure models may therefore contribute substantially to the observed heterogeneity of effect estimates in systematic reviews of epidemiological studies. Factors that explained some of the heterogeneity of effect estimates included the performance of the model at external validation sites and the predicted exposure contrast.
- Exposure predictions from the three algorithms were generally highly correlated and resulted in similar associations with health. No consistent differences were found in their model performances.

1. INTRODUCTION

In urban areas, motorized traffic is an important source of major health-relevant air pollutants (HEI 2022). Assessment of long-term exposure to traffic-related outdoor air pollution remains a major challenge for epidemiological studies. One of these challenges is the characterization of the spatial variation of the ambient concentrations of key traffic-related air pollutants including ultrafine particles (UFPs), black carbon (BC), and nitrogen dioxide (NO_2) as these pollutants vary on a fine spatial scale related to the intensity of and proximity to combustion sources (HEI 2022).

A range of different approaches have been developed to address these challenges (Hoek 2017; Jerrett et al. 2005). Exposure assessment methods applied in epidemiological studies include empirical Land Use Regression (LUR) models based on fixed-routine or study-specific monitoring, and more recently, mobile monitoring (especially for UFPs and BC) and low-cost sensors ($PM_{2.5}$, NO_2). Increasingly sophisticated hybrid models incorporating routine surface monitoring, satellite, chemical transport, and land use data have been used (Brauer et al. 2022; Di et al. 2016; Hoek 2017). Epidemiological studies have furthermore used exposure assessment based on dispersion or chemical transport modeling (Carey et al. 2013; Hvidtfeldt et al. 2019; Ostro et al. 2015). In some studies, multiple methods are used for different pollutants, such as mobile monitoring for UFPs while $PM_{2.5}$ and NO_2 exposures were derived from fixed-site LUR models (Downward et al. 2018).

Very little information is available about the relative performance of these different approaches to assess long-term exposure to traffic-related air pollution. Differences in performance may affect conclusions drawn from epidemiological studies applying different exposure assessment approaches. Jerrett and colleagues illustrated substantial differences in relative risks of $PM_{2.5}$ for mortality in the American Cancer Society study for different exposure models, particularly between exposure estimated exclusively from satellites compared to ground-based estimates (Jerrett et al. 2017). Differences in exposure assessment methods have also been hypothesized to contribute to heterogeneity in systematic reviews of the health effects of air pollution (Boogaard et al. 2023a; Chen et al. 2020). An added complexity is that models based on recent (mobile or fixed-site) monitoring data are often applied to cohort studies with a follow-up starting several years before the monitoring year (Strak et al. 2021). An important issue is therefore to assess how well models reflect spatial variability in the past. This affects virtually all UFP models, which are derived from relatively recent monitoring data.

Significant progress has been made in applying more sophisticated algorithms to develop empirical models, including machine learning methods such as Random Forest and regularization methods such as LASSO. We recently published a comparison of the performance of different algorithms to develop LUR models based on mobile and short-term monitoring of UFPs (Kerckhoffs et al. 2019), and European models of NO_2 and $PM_{2.5}$ (Chen et al. 2019). Approaches integrating deterministic and empirical models have been developed in recent years (Hoek 2017). de Hoogh and colleagues developed European models for NO_2 and $PM_{2.5}$ by integrating satellite and chemical transport data in a LUR framework (de Hoogh et al. 2016, 2018).

This project addressed these challenges by comparing a suite of air pollution exposure models, both in terms of performance of assessing spatial variation and in air pollution effect estimates in selected epidemiological studies in the Netherlands. The project has combined a wealth of models already developed in previous projects and newly designed monitoring and modeling efforts.

2. SPECIFIC AIMS

The specific aims of the project:

- Develop long-term ambient air pollution exposure estimates for selected epidemiological studies based on low-cost sensors, mobile and fixed-site monitoring, and deterministic dispersion modeling
- Compare different exposure assessment methods in terms of their ability to predict the spatial variation of long-term average concentrations using external validation data
- Compare different exposure assessment methods in terms of air pollution effect estimates in selected epidemiological studies.

3. STUDY DESIGN AND METHODS

3.1 OVERALL DESIGN

We generated and evaluated annual average air pollution maps using a suite of different exposure assessment models. We used existing models from previous collaborative projects where available. The models differed in the modeling approach (deterministic versus empirical) and monitoring data used (mobile monitoring, fixed-site routine monitoring, or study-specific monitoring). The empirical models used three common algorithms selected from distinct families: supervised linear regression (SLR), Random Forest, and LASSO (Chen et al. 2019; Kerckhoffs et al. 2019). We limited the comparison between these three algorithms and did not include study-specific exposure methods, such as the mixed-effect model for mobile monitoring and geographically weighted regression (Kerckhoffs et al. 2021, 2022; Shen et al. 2022).

The exposure predictions of the models were compared at 20,000 randomly selected residential addresses across the Netherlands, of which 10,000 were in the four major cities (Amsterdam, Rotterdam, Utrecht, and The Hague), with 5,000 in Amsterdam specifically. We further tested the performance of the models using external validation monitoring data, both for current and historical spatial surfaces based on existing validation data. We collected new validation data for 90 residential sites across the Netherlands because long-term average exposure data with sufficient spatial resolution did not exist for UFPs and BC, which are important traffic-related pollutants. Epidemiological analyses in three cohort studies were conducted to compare health effect estimates of the different exposure assessment models. Two adult cohorts, namely a nationwide administrative cohort of 10 million adults (DUELS) and a cohort with detailed lifestyle data for 33,000 adults (EPIC-NL) were selected, and one mature birth cohort (PIAMA). **Figure 1** illustrates the different exposure models and the overall design of the study. Different models were based on data from different years, so an evaluation of the impact of temporal trends was a key element in the analysis of the data.

Section 3.2 describes the exposure models, with an especially detailed description of the new low-cost sensor monitoring effort. The external validation data are described in section 3.3, with a focus on the new validation monitoring effort. In section 3.4, the epidemiological studies and outcome definitions are explained.

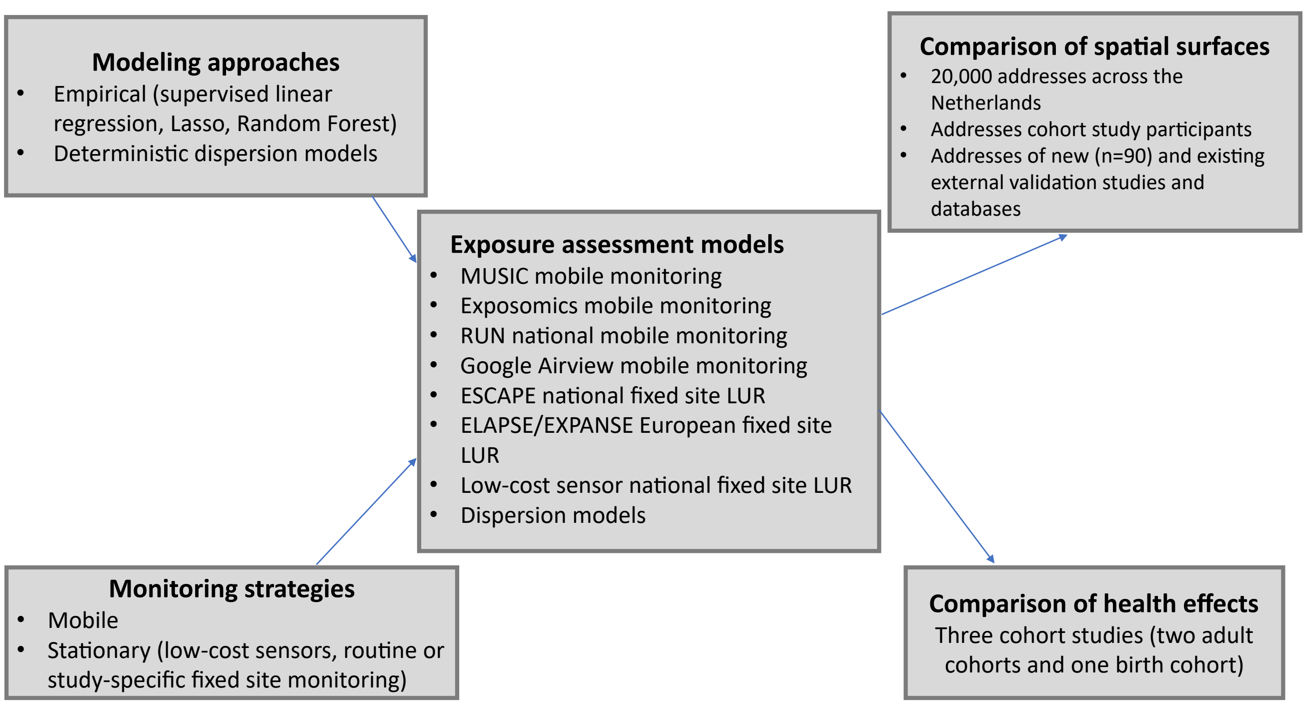


Figure 1. Design of the project comparing the various exposure-assessment models. Pollutants assessed are UFPs, NO_2, BC, and $PM_{2.5}$. For the empirical models, three algorithms were applied. Table 1 explains the models.

3.2 EXPOSURE MODELS

Table 1 lists the exposure models used and further developed in this project. The models differ in their modeling approach, monitoring base, the year for which predictions are made, and the monitoring domain. Specifically, four models were empirical LUR models based on mobile monitoring. Three models were LUR models based on fixed-site monitoring with more temporally rich data, including one based on new low-cost sensor monitoring. One model was based on dispersion modeling.

The UFP models in particular are more recent. All new and existing air pollution monitoring was conducted outdoors. In the analyses, we took into account differences between years such that we can attempt to disentangle reasons for potential differences in predictions between the models. For the dispersion modeling and the Europewide fixed-site model EXPANSE, we were able to produce exposure maps for multiple years to assess changes in spatial patterns between 2010 and 2019 (the years spanning the time window of the full suite of models). The empirical models were each developed with three algorithms: SLR, Random Forest, and LASSO (section 4.2). We separately document and discuss the variability of exposure prediction and effect estimates related to models and algorithms within models.

Not every model provided exposure estimates for all four pollutants. For BC, the largest diversity of models was available (dispersion, all mobile monitoring campaigns, and fixed-site national and European LUR), lacking estimates only from the low-cost sensor model. The BC models differed in the metric used, including BC, elemental carbon (EC), and absorbance of $PM_{2.5}$. Concentrations expressed in these different metrics are generally very highly correlated but differ up to about 30% in absolute levels (HEI 2022). For NO_2, we had estimates from all fixed-site models, one mobile monitoring campaign, and dispersion models. For $PM_{2.5}$, we had estimates from all models, including from mobile monitoring models that we developed for this specific project. For UFPs, we only had mobile monitoring models, from different monitoring domains and years (2013–2019).

Sections 3.2.1 to 3.2.8 describe the data underlying the different models. Section 4.1 describes the modeling approaches, including the algorithms for the empirical models.

3.2.1 Low-Cost Sensor Model

We have performed low-cost sensor measurements during two half-year periods. During each measurement period, approximately 50 citizen home addresses were selected as measurement locations for the low-cost sensor kit. All air pollution monitoring was conducted outdoors. The sensor kits were connected to the existing citizen science low-cost sensor measurement platform across the Netherlands, coordinated by RIVM (Wesseling et al. 2019). From these selected 100 sites, continuous monitoring was performed at only 84 sites (46 sites in period 1 and 38 in period 2) due to sensor malfunctions or issues with participants. The combined measurement periods resulted in more than 300,000 hourly measurements of NO_2, $PM_{2.5}$, and PM_{10} (Particulate matter ≤10 µm in aerodynamic diameter).

The individual sensor kits that were assimilated for this project combine three low-cost sensors for NO_2, $PM_{2.5}$, and $PM_{10,}$ respectively. The NO_2 (Alphasense B43F) measurements and $PM_{2.5}$ (SPS30) were only sparsely present on the citizen science platform at the start of this project. Currently, only about 10 NO_2 sensors from other projects are connected. There are about 2500 SDS011 (PM_{10}, $PM_{2.5}$) sensors operational, which makes it by far the most used sensor. Because the $PM_{2.5}$ measurements of the SPS30 seem to have higher quality than the SDS011 and are still affordable, the SPS30 was added to the sensor kit.

Site Selection The low-cost sensors (combined into single sensor kits) used in the project were distributed throughout the Netherlands in an attempt to achieve sufficient national coverage. Therefore, RIVM put out a call among people interested in citizen science and air quality, asking citizens to participate in this project. It was made clear that only 100 locations (plus some spare locations) would be selected. The subscription period ended in February 2021, yielding 450 applications.

Site Selection Criteria In this section, considerations regarding the choice of low-cost sensor locations are further explained. When focusing on the NO_2 measurement, there were a few factors that had to be taken into account in the selection process of these locations:

- The set of low-cost sensor locations had to sufficiently reflect population density, meaning sensor density was the highest near the area in the Netherlands called “Randstad,” which contained several large cities.
- The sensors were used as input for a fusion model, and the locations had to be chosen in such a way that the fusion model could utilize as much information as possible from the sensor measurements. In other words, there had to be enough temporal variations in the concentrations at the locations, and locations near high-density traffic were preferred as these are the largest local causes of variations in NO_2 concentrations.
- The number of sensor sets was limited. Fifty low-cost sensors were available for two consecutive half-year measurement periods.
- To be able to distinguish between spatiotemporal effects that are real and effects that possibly originate from artifacts of the calibrated sensors and/or fusion model, there had to be some spatial overlap between the location sets chosen in the first and second measurement periods. In practice, this meant that we covered each of the twelve provinces in the Netherlands with at least one sensor kit in both measuring campaigns.

Table 1. Exposure Models for Long-Term Air Pollution Exposures Evaluated in This Study

Exposure Model	Pollutants and Instruments	Monitoring Data	Temporal	Modeling Approach	Predictors	Domain	Time Period Monitoring	Reference
Low-cost sensor	$PM_{2.5}$, NO_2 with low-cost sensors SPS30 and Alphasense 43F	Long-term monitoring at 100 residential sites across the Netherlands Traffic sites overrepresented New monitoring for this project	Two rounds of 50 sites measured for 6 months simultaneously Continuous NAQMN data used to adjust differences between two rounds	Data fusion with dispersion and routine monitoring Empirical LUR using SLR, RF, and Lasso	Hourly NAQMN data and dispersion data National GIS data, including traffic intensity	Netherlands	2021–2023	New for the current project
MUSIC major city mobile monitoring model	UFPs, BC, $PM_{2.5}$ using CPC3007, AE51, DustTrak	Short-term monitoring: at 81 sites in Amsterdam and 80 sites in Rotterdam Mobile monitoring: 2964 road segments in between sites	Short-term: three times 30 minutes per site in different seasons Mobile: ~20 seconds per street segment Correction for temporal differences with a central reference site in Utrecht Weekdays 9–16 hr sampling (rush hour avoided)	Land use regression models using SLR, RF, and Lasso	National GIS data, including traffic intensity	Major cities: Amsterdam and Rotterdam	2013: winter and spring	Kerckhoffs et al. 2016; Klompmaker et al. 2015; Montagne et al. 2015
Exposomics major city mobile monitoring	UFPs, BC, $PM_{2.5}$ using CPC3007, AE51, DustTrak	Short-term monitoring: 240 sites in Amsterdam, Maastricht, Utrecht Mobile monitoring: 5,236 road segments	Short-term: three times 30 minutes per site in different seasons Mobile: ~25 seconds per street segment Correction for temporal differences with a central reference site in Utrecht Weekdays 9–16 hr sampling (rush hour avoided)	Land use regression models using SLR, RF, and Lasso	National GIS data, including traffic intensity	Major cities: Utrecht, Amsterdam, Maastricht	2014–2015: winter, spring, summer	Kerckhoffs et al. 2017; van Nunen et al. 2017

Lasso = least absolute shrinkage and selection operator; NAQMN = national air quality monitoring network; RF = Random Forest; SLR = supervised linear regression.

Exposure Model	Pollutants and Instruments	Monitoring Data	Temporal	Modeling Approach	Predictors	Domain	Time Period Monitoring	Reference
RUN national mobile monitoring	UFPs, BC, $PM_{2.5}$ using CPC3007, AE51, DustTrak	Short-term monitoring: at 400 sites across the Netherlands Mobile monitoring: 14,392 road segments across the Netherlands Regional background monitoring at 20 sites, 3 times 14 days.	Short-term: three times 30 minutes per site in different seasons Mobile: ~43 seconds per street segment Correction for temporal differences with a central reference site in Utrecht Weekdays 9–16 hr sampling (rush hour avoided)	Land use regression models using SLR, RF, and Lasso	National GIS data, including traffic intensity	Netherlands	June 2016–November 2017	Kerckhoffs et al. 2021; van de Beek et al. 2021 (for UFP mobile; remainder newly developed)
Google Air-view mobile monitoring	UFPs, BC, NO_2, $PM_{2.5}$ using EPC3783, AE33, CAPS Aerodyne, DustTrak	Mobile monitoring of all streets in Amsterdam; 46,664 street segments	Five- to ten repeats in a 10-month period On average 7 repeats with 8 seconds of monitoring per street segment per repeat (total ~1 minute) Reference site to correct for temporal variation Monitoring from 8 to 22 hr, weekdays.	Land use regression models using SLR, RF, and Lasso	National GIS data, including traffic intensity	Amsterdam	May 25, 2019–March 15, 2020	Kerckhoffs et al. 2022
ESCAPE Nationwide LUR	$PM_{2.5}$, BC, NO_2 using Harvard impactor, reflectance measurement on $PM_{2.5}$ filter and Ogawa passive sampler	Long-term study-specific monitoring: 40 sites for $PM_{2.5}$ (and BC) and 80 residential sites for NO_2 across the Netherlands	Three 14-day average integrated samples in three seasons Ten sites simultaneous in each session Correction for temporal differences with a central reference site in Utrecht	Land use regression models using SLR, RF, and Lasso	National GIS data, including traffic intensity	Netherlands	2009–2010: three seasons	Beelen, et al. 2013; Eeftens et al. 2012

Lasso = least absolute shrinkage and selection operator; NAQMN = national air quality monitoring network; RF = Random Forest; SLR = supervised linear regression.

continued on next page

Table 1. (*continued*)

Exposure Model	Pollutants and Instruments	Monitoring Data	Temporal	Modeling Approach	Predictors	Domain	Time Period Monitoring	Reference
European hybrid model (ELAPSE and EXPANSE)	$PM_{2.5}$, BC, NO_2 using various reference monitors	AirBase routine monitoring 549 sites for $PM_{2.5}$ and 2,400 sites for NO_2 across Europe (2010) More sites in 2019. ELAPSE: 436 sites for BC across Europe from ESCAPE	Annual average from continuous monitoring if 75% of the period had valid measurements. BC three times 14-day averages and correction with a continuous reference site	Land use regression models using SLR, RF, and Lasso	European GIS data, road length instead of traffic intensity; satellite and chemical transport model for background.	Europe	2010–2019	de Hoogh, et al. 2016, 2018; Shen et al. 2022
Dispersion model	$PM_{2.5}$, BC, NO_2	Not applicable	Model developed hourly and for project aggregated to an annual average	Dispersion modeling at the address level	National and foreign sources and hourly weather data	Netherlands	2010–2019	Velders et al. 2021

Lasso = least absolute shrinkage and selection operator; NAQMN = national air quality monitoring network; RF = Random Forest; SLR = supervised linear regression.

As stated in the second bullet, the fusion model needs a sufficient number of locations that are close to busy roads. However, NO_2 levels at these types of locations are typically higher than urban background levels where the larger part of the population lives, and so do not necessarily represent the average exposure of the overall population very well. Measurements at other types of locations were therefore also needed. NO_2 measurements at urban and rural background locations usually only show variations in concentration due to regional/national effects, as is the case for $PM_{2.5}$ and PM_{10} concentrations.

A fusion model that included the low-cost sensor data, dispersion model estimates, and reference measurements yielded results covering the whole country (contrary to results at just the sensor locations). We investigated whether the addition of low-cost sensor data improves the quality of NO_2 predictions at locations relevant to assessing the exposure of the general population.

Furthermore, the fusion model is an hourly model, and locations close to a road are only qualified as "road-contribution" during hours in which a sensor is situated downwind of this road. Wind data are obtained from the nearest weather station. For the remaining hours, the sensor participates in the estimate for the background concentration.

These considerations have led us to select approximately 75% of the sensor locations close to busy roads and 25% at urban background locations. Five additional sensor kits were used for calibration by circulating these between reference locations and selected participants.

Selection of Locations Close to Roads To determine whether a location is impacted by traffic, or more specifically a location where the emissions associated with traffic are substantial, we relied on the official Dutch air quality management system (*https://rwsenvironment.eu/subjects/air/air-quality/*, *https://www.cimlk.nl*). Using the results from that system, the modeled yearly average traffic contributions and background concentrations were estimated at all potential sensor locations (Additional Materials 1 Figure S1). Subsequently, we selected about 75 locations throughout the Netherlands that are close to a road, as well as 25 locations that typically have a very low road contribution (background locations).

Equipment Each sensor kit contains three low-cost sensors; one for NO_2, one for $PM_{2.5}$, and one for PM_{10}. The total cost of each sensor kit was 500 Euro, including the three sensors, communication hardware, and cover. We decided to use the Sensirion SPS30 particle counter for $PM_{2.5}$ (in addition to what we intended in the application), and the Nova Fitness SDS011 for PM_{10}. The Sensirion SPS30 does not provide usable PM_{10} concentrations, but it produces useful $PM_{2.5}$ measurements that seem to be less sensitive to humidity than other known sensors. Furthermore, it provides particle number and mass concentration in several categories, such as 0.3–0.5 µm, 0.3–1.0 µm, and 0.3–2.5 µm, although these were not used in the analysis. The disadvantage is that the SPS30 is approximately 25 euros more expensive than the well-known SDS011. RIVM presently does not use the particle number concentration reported by the sensors. Additional Materials 1 Figure S2 shows the sensor box assimilated by RIVM, containing the low-cost NO_2, PM_{10}, and $PM_{2.5}$ sensors.

Data Communication In 2021, RIVM started using NarrowBand Internet of Things as a way of communicating data from the sensors to the servers. The biggest advantages of this method are the plug-and-play usage, the coverage (in principle, all the Netherlands), and the availability of an onboard GPS.

Quality Assurance and Quality Control The measuring equipment (sensor kit) was sent to the participants to maintain the low-cost aspect of the model. Detailed installation guides were provided for the volunteers to ensure proper installation of the equipment. Additional Materials 1 Section 3.2.1 provides the instruction in both translated English and the local language version (Dutch) that the participants received.

We applied the validation procedures as described in (Wesseling 2019). This includes individual calibration of the NO_2 sensors to derive a sensor-specific correction for ozone (O_3) and temperature before deployment of the sensor. O_3 concentrations were derived from the nearest official monitoring station. If no nearby measurement was available, the official hourly O_3 concentration map was used to estimate the local concentration. Calibration was performed before the first and second 6-month period. To account for the potential drift of the sensor within the 6-month period, nighttime measurements were used to adjust differences between sensors daily (night calibration). This assumes that local sources around the sensor kits have a relatively small impact during nighttime hours, yielding a rather homogeneous concentration field. Data from nearby reference monitors of the national monitoring network were used for this night calibration. We used five low-cost sensors to circulate between several measurement locations (selected from among the volunteers) and relocated calibration locations. We calibrated and checked them for at least 4 weeks near official measurements, and used them in the field for 4 weeks.

For the SDS011 PM_{10} sensor, a calibration scheme using groups of sensors close to official PM_{10} measurements was used to determine the ratios between reference measurements and close-by sensors. These correction factors at the locations of the official measurements were subsequently interpolated over the country, providing an estimated correction factor for all other sensors that were further away from the reference measurement locations. The interpolation was performed assuming that the differences between sensors and official measurements mainly originate from meteorological (relative humidity in particular) circumstances (Wesseling et al. 2019). During the project, there was not enough information available to check if a similar calibration was needed for the SPS30 $PM_{2.5}$ sensor. Hence, no calibration was performed for the $PM_{2.5}$ sensor, and no corrections to the $PM_{2.5}$ sensor data were made.

Scheduling of Sensor Measurements From the perspective of calibration of the NO_2 sensors, we had to avoid calibrating solely during summer, as calibration during other periods of the year gives better overall results (Ratingen et al. 2021). This led to the following schedule for the sensor measurements:

Feb–May 2021	Assembly of sensors
May–September 2021	Calibration
Oct 2021–April 2022	Measurements at 1st set of 50 selected locations
April 2022–July 2022	Return/check/recalibrate sensors
July 2022–March 2023	Measurements at 2nd set of 50 selected locations

As of March 2021, all necessary equipment had been ordered. However, because of COVID-19 restrictions and other circumstances, some delays were encountered in development and delivery. Therefore, the calibration started in May and the measurement campaign needed to be delayed until autumn 2021, as reliable calibration could not be performed over the summer months due to the low dynamic range of NO_2 concentrations.

3.2.2 MUSIC and Exposomics Major City Models

In the framework of the MUSIC and Exposomics campaigns, LUR models were developed and published based on mobile (on-road) and short-term roadside monitoring for selected major cities in the Netherlands (Kerckhoffs et al. 2016, 2017; Montagne et al. 2015; van Nunen et al. 2017). In MUSIC, measurements were performed in Amsterdam and Rotterdam city. In Exposomics, measurements were performed in Amsterdam, Utrecht and surroundings, and Maastricht and surroundings, in agreement with the study areas included in the cohorts in the project (EPIC-NL primarily, section 3.4). Except for Maastricht, all cities are located in the west of the country within about 50 km of Utrecht (where continuous monitoring took place). Maastricht is located about 150 km away in the south of the country. In these projects, mobile monitoring of UFPs, BC, and $PM_{2.5}$ was performed using CPC3007, micro-aethalometer AE51, and the DustTrak, respectively. All equipment was carried by an electric vehicle. In addition to the mobile (on-road) monitoring, mobile (short-term roadside) monitoring was performed: three times 30-minute sampling in different seasons supplemented with a continuous reference site. Mobile (on-road) monitoring involved measurements while the car was driving between the roadside monitoring sites. In both campaigns, measurements were performed during weekdays only. Measurements were performed between 9:00 and 16:00 hours, deliberately avoiding the rush hour. Rush hour was avoided to increase comparability between measurements, which are made at different locations and times. On every monitoring day, short-term roadside monitoring sites were selected in different parts of the city including both traffic and background locations. This design was chosen to reduce the correlation between spatial and temporal variation. Traffic locations were streets carrying at least 10,000 motor vehicles per day. Measurements were performed during the full study periods at a reference site, which was a regional background location just outside the city of Utrecht in the center of the country that used the same equipment as the monitoring vehicle. We used the *difference* method, which first calculates the overall mean concentration at the reference site over the entire campaign. Next, using the concentrations at the reference site, the average reference concentration from around 30 minutes of the time of sampling was subtracted from the overall mean. This difference was then added to the measured concentration at the residential sites of that time and date.

To develop the models, national geographic information system (GIS) variables were obtained, including traffic intensity, road length, population density, and land use in multiple buffer sizes (50–5,000 meters in radius), following the ESCAPE study (Eeftens et al. 2012). We developed separate models based on the short-term roadside monitoring component of the study and the on-road mobile monitoring, the latter covering more street segments but with a shorter monitoring period compared to the short-term roadside sites. Predictions of UFPs from models based on mobile measurements were highly correlated with the short-term roadside models but predicted about 30% higher concentrations (Kerckhoffs et al. 2016, 2017).

3.2.3 RUN Nationwide Models

In the framework of the RUN project, mobile (on-road) and short-term roadside monitoring across the Netherlands for UFPs, BC, and $PM_{2.5}$ was performed for 14 months in 2016–2017 using an electric vehicle. Measurements were performed in major cities, and also in smaller towns, covering towns of participants of cohort studies in the Netherlands including EPIC-NL and PIAMA (section 3.4). The mobile monitoring design closely followed the earlier MUSIC and Exposomics campaigns (Klompmaker et al. 2015; van Nunen et al. 2017), as described in the previous section. The mobile monitoring was supplemented with longer-term regional background monitoring at 20 sites across the Netherlands (van de Beek et al. 2021). Measurements were performed with the DiskMini following the ESCAPE method: three times 14-day average periods in different seasons. Because of limited equipment, only four sites were measured simultaneously. A continuous reference site in the surroundings of Utrecht city was used to adjust for temporal variation. Kriging was used to develop regional background maps of the Netherlands, used as input for LUR models. Empirical UFP models have been developed previously (Kerckhoffs et al. 2021). We used a mobile platform to monitor UFPs, BC, and $PM_{2.5}$ concentrations across the Netherlands for a period of 14 months (2016–2017), with a total of 14,392 road segments sampled three times on average. Average concentrations per road segment (about 43 seconds of data) and extracted GIS variables were used to develop LUR models using SLR, as well as LASSO and Random Forest algorithms. In addition, a model based on a deconvolution approach was developed, separating models for different spatial scales (Kerckhoffs et al. 2021). Differences in model performance between the different modeling approaches (algorithms) were small, as judged by comparison with the Exposomics validation data. The deconvolution model was not further used in the current project, as it was not applied in any of the other exposure assessment approaches that we evaluated.

3.2.4 Google Airview

The Google Airview project involved mobile (on-road) monitoring of UFPs, NO_2, BC, and $PM_{2.5}$ of all streets in the Amsterdam area. All streets were monitored repeatedly following the successful monitoring design in Oakland (Apte et al. 2017; Messier et al. 2018). The project involved two cars equipped with research-grade instruments (Table 1). Compared to the previous mobile monitoring campaigns, higher quality instruments were applied, especially for BC (AE33 aethalometer instead of the AE-51 portable aethalometer). NO_2 was measured for the first time with a mobile monitoring campaign by our group. All instruments sampled 1-second values. Short-term roadside monitoring was not performed. Monitoring started in May 2019 and finished in March 2020, because of the lockdown related to the COVID-19 epidemic. All streets of Amsterdam were driven from 5 to 10 times. Additional Materials 1, Figure S3 documents instrumentation. Detailed QA-QC was conducted, including zero checks, frequent calibration, and co-location with Dutch routine monitoring sites. Models have been developed with different approaches including SLR, Random Forest, other machine learning methods, and methods combining models and measurements (Kerckhoffs et al. 2022). The mixed-effect model developed in the Google Airview project was not further used in the current study, as it was not commonly applied in the other exposure assessment approaches.

In the major city campaigns and the RUN nationwide campaign, monitoring took place on weekdays and between 9:00 and 16:00 hours, deliberately avoiding the rush hour to increase comparability across road segments. Furthermore, the design was built on short-term roadside sites with mobile (on-road) monitoring in between. In the Google Airview campaign, mobile (on-road) monitoring was performed on weekdays between 8:00 and 22:00 hours, though mostly between 10:00 and 20:00 hours. In addition, all roads were repeatedly monitored.

3.2.5 ESCAPE Nationwide LUR Models

In the ESCAPE project, monitoring of $PM_{2.5}$, BC, and NO_2 was performed across the Netherlands in 2009–2010. $PM_{2.5}$ was measured at 40 sites and NO_2 at 80 sites, distributed over regional and urban background and traffic sites (Eeftens et al. 2012; Beelen et al. 2013). Harvard impactors were used to obtain integrated 14-day average $PM_{2.5}$ samples on Teflon filters. The reflection of the $PM_{2.5}$ filters was measured (absorbance) as a surrogate of BC. Ogawa passive samplers were used to measure NO_2. All sites were residential locations. Monitoring involved three times 14-day average concentrations in three seasons per site. In each season, measurements were performed simultaneously at 10 $PM_{2.5}$ and 20 NO_2 sites, as well as a continuous reference site. Each group included the different site types to avoid spatiotemporal bias. This approach is especially useful for model development, as limited or no bias is introduced in the comparison across site types. Adjustment for temporal variation using data collected at a continuous reference site near Utrecht was performed, using the difference method. A national LUR for $PM_{2.5}$, BC, and NO_2 was developed using SLR methods, allowing only predictors with an a priori–defined direction, which is positive for traffic intensity (Beelen et al. 2013; Eeftens et al. 2012). Land use and traffic variables were used as predictors to develop LUR models. Traffic intensity separated into heavy- and light-duty vehicles on all roads was available, by modeling at most roads.

3.2.6 ELAPSE and EXPANSE Europewide LUR Models

In the HEI-funded ELAPSE project, European LUR models were developed using long-term monitoring for $PM_{2.5}$, BC, and NO_2 (Brunekreef et al. 2021; de Hoogh et al. 2018). Monitoring data primarily for the year 2010 from routine monitors from the European Airbase database were used for $PM_{2.5}$ and NO_2. Airbase is the database of routine monitoring across Europe using national networks. Reference monitors were used, though were not necessarily the same across countries. BC models were developed based on the ESCAPE data combined in one Europewide database. In ESCAPE, the same campaign was performed in 20 $PM_{2.5}$ study areas across Europe in 2008–2010. In addition to land use and traffic variables, chemical transport and satellite data at a large spatial scale (≥ 10 × 10 km) were used as predictors to develop hybrid LUR models. Traffic intensity was not available publicly Europewide. To represent traffic, (major) road length was used in various buffers. The models produced 100 × 100-meter grids of predicted annual average air pollution concentrations. Back-extrapolation based on a chemical transport model was used to develop pollution maps for years other than 2010.

In the EXPANSE project (*https://expanseproject.eu/*), the European models for $PM_{2.5}$ and NO_2 were further developed. Specifically, annual average LUR models were developed for $PM_{2.5}$ and NO_2 based on Airbase monitoring data for all years between 2000 and 2019, as opposed to the back-extrapolation performed in ELAPSE (Shen et al. 2022). The spatial resolution was improved to 25 × 25-meter grids of predicted annual average concentration. The number of monitoring sites increased substantially in this period, especially for $PM_{2.5}$. The main model was based on geographically weighted regression, which slightly outperformed the SLR and Random Forest models used in the current project (Shen et al. 2022). We did not have the resources to develop LASSO models for the EXPANSE data. The availability of models for multiple years allowed us to assess the sensitivity of our comparisons to year of model.

3.2.7 Nationwide Dispersion Model

Fine-scale dispersion modeling was performed using the Dutch national model. With this model, annual average concentrations of $PM_{2.5}$, EC, and NO_2 at the address level have been calculated across the Netherlands. EC is a different metric from the BC in that it is assessed in other models, but the two metrics are highly correlated and differ by less than about 30% (HEI 2022). The model consists of two stages: First, large-scale concentration maps of $PM_{2.5}$, NO_2, and EC at a spatial resolution of 1 × 1 kilometer are produced, using the Operational Priority Substances dispersion model that incorporates a wide range of Dutch and foreign sources and measurements from the National Air Monitoring network. Second, the local traffic contribution is added to these background grids at a much higher spatial resolution using dedicated models (Velders et al. 2016; Wesseling et al. 2014, 2016). For this study, local contributions were calculated for the exact coordinates of the Dutch address database. The availability of models for multiple years (2010, 2013, 2016, and 2019) allowed us to assess the sensitivity of our comparisons to year of model. For EC, the first model available was in 2013. No deterministic UFP model was available.

3.2.8 Approaches to Address the Different Years for the Models

We used two approaches to address temporal issues in the comparison across models that differ both in methodology and year. The first approach is to document how large the changes in levels of measured concentrations have been in the period in which monitoring data have been collected (2009–2021). This is feasible for components measured in the National Air Quality Monitoring Network (NAQMN) (NO_2, $PM_{2.5}$, and BC). The second approach is that we were able to produce exposure maps for multiple years to assess changes in spatial patterns between 2010 and 2019. This is feasible for dispersion modeling and the Europe-wide empirical modeling (ELAPSE and EXPANSE) based on routine monitoring data. For these models that differ substantially in principle, we can therefore disentangle the model and year.

3.3 VALIDATION DATA

We specifically collected independent validation data, because 10-fold cross-validation statistics cannot be compared between models. This is especially problematic when comparing models based on mobile and longer-term exposure data, because of the larger random error in the mobile data. Larger random error of the average per site leads to lower R^2 values for models based on shorter time periods than those based on longer time periods. This was clearly demonstrated in a recent publication where the cross-validated R^2 of the mobile monitoring model was only 0.15, whereas when validated using a longer-term external exposure dataset (the Exposomics data residential outdoor data) the model R^2 increased to 0.57 (Kerckhoffs et al. 2017), confirming previous observations (Klompmaker et al. 2015; Montagne et al. 2015).

3.3.1 Existing External Validation Data

The following existing data was available to test the predictions of the various exposure models. Coordinates of all measurement sites were available in high resolution (meters). These data were not used in model development (hence external validation data):

1. **Concentration data for UFPs, BC, and $PM_{2.5}$ collected in the Exposomics project in the major cities of Utrecht and Amsterdam (van Nunen et al. 2017)** — Three times 24-hour monitoring per site was performed at 42 sites (about 50% traffic, 50% urban background) in different seasons. Sites were measured in groups of two (a traffic and a background site) along with measurements at a regional background reference site. Average concentrations per site were calculated after temporal adjustment using the difference method. Measurements were conducted in 2014. UFPs, BC, and $PM_{2.5}$ were measured with the DiscMini and the integrated Harvard impactor with subsequent measurement of reflectance of the filter, as used in the new validation campaign (section 3.3.2). Reflectance is transformed into an absorbance metric, a surrogate for BC.
2. **Historical data from an earlier study (RUPIOH) in Amsterdam for UFPs, BC, and $PM_{2.5}$ (48 sites, 2002–2004 period; Hoek et al. 2011)** — The UFP data, in particular, provides an important asset to our knowledge; there are no other datasets with spatially resolved UFP data available for a large number of sites in the Netherlands or other countries because of the typical lack of routine monitoring. UFPs, BC, and $PM_{2.5}$ were measured with the condensation particle counter CPC3022a and the integrated Harvard impactor with subsequent measurement of the reflectance of the filter. Measurements were made for a single week at each home, and each home was measured on a different week. Adjustment for temporal variation using a continuous urban background monitor in the city of Amsterdam was performed. Adjustment for temporal variation made a large difference, as evidenced by a small positive temporal correlation between $PM_{2.5}$ and UFPs at the reference site (R = 0.19) and a strong positive correlation for the adjusted concentrations of 0.66 (Hoek et al. 2011). The adjustment was further supported by the correlation of 24-hour average concentrations at the residential sites and the continuous reference site for all pollutants (Hoek 2011). The median correlation was 0.72, 0.94, and 0.98 for UFPs, BC, and $PM_{2.5,}$ respectively.
3. **Annual average NO_2 concentrations across the Netherlands for 144 sites in 2007 from a previous study (TRACHEA), four times 1-week average sample (Eeftens et al. 2011)** — Residential sites included 26 regional background sites, 78 urban background sites, and 40 traffic sites. NO_2 was measured with passive Ogawa badges in the same weeks for all sites, precluding the need for adjustment for temporal variation.

4. **NAQMN data for BC, NO_2, and $PM_{2.5}$** — Annual average concentrations are available using European reference monitors at an increasing number of sites across the country, but with a small number of sites per city. The sites include regional background, urban background, and traffic locations. Sites differ in location from the previous targeted research studies, as because of the required space, public rather than private residential homes are often used. Traffic sites are often closer to the road than in our targeted campaigns. BC was measured with the multiangle absorption photometer at the onset of monitoring, and the AE33 aethalometer in 2019. NO_2 was measured with chemiluminescence continuous monitors. $PM_{2.5}$ was measured with integrated samplers (Derenda and Leckel samplers). The network is used for multiple purposes, including air quality limit value compliance, trend analysis, and finetuning of the dispersion model.
5. **Passive sampling data for NO_2 using Palmes tubes across the Netherlands in multiple years up to 2019, mostly in the major cities** — The Palmes tube data of Amsterdam (n = 82) have been used extensively as external validation in the Google Airview model evaluation (Kerckhoffs et al. 2022). Monthly average samples calibrated by comparison with routine monitors are obtained every month. Additionally, data from 2011 and onward were available for Amsterdam (and surroundings), Utrecht, and a few other cities across the Netherlands. These data have been compared with the national dispersion model.

3.3.2 New External Validation Monitoring Data

We collected new validation data because there was insufficient data for UFPs and BC at a sufficiently large number of locations across the Netherlands for validation purposes. A previous validation study collected data in 2014 at 42 sites in the cities of Utrecht and Amsterdam as part of the Exposomics project. While these data are useful for evaluating Amsterdam and major city mobile monitoring models, the data are less useful for judging national models. Furthermore, 42 sites limited the detection of differences in performance between models (Kerckhoffs et al. 2019).

Against this backdrop, we aimed to perform outdoor air pollution measurements at 100 sites across the Netherlands. At each site, we aimed to collect two 1-week samples in different seasons. We monitored UFPs, BC, $PM_{2.5}$, and NO_2. In the previous Exposomics project, we observed that three times 24-hour monitoring per site in different seasons at 42 sites of UFPs was useful for external validation of UFP mobile LUR models (Kerckhoffs et al. 2019; van Nunen et al. 2017). We now collected more data per site to better approximate the long-term average. The new validation data were considered particularly useful for BC and especially UFPs, for which limited or no routine monitoring data are available. $PM_{2.5}$ and NO_2 were added for completeness.

In the design of the study, we could not perform measurements at all sites simultaneously. We did not have sufficient instruments and field worker capacity to achieve simultaneous measurements as in the TRACHEA passive sampler study. Based on the experiences of the RUPIOH, ESCAPE, and Exposomics campaigns included in this project and the earlier TRAPCA study (Hoek et al. 2002), we anticipated that with a continuous reference site, we would be able to assess the spatial contrast in annual average concentrations, with two repeats of 1-week. We were able to measure at four sites simultaneously, including the reference site.

We monitored UFPs with the DiscMini (van Nunen et al. 2017), BC with the MA200 micro-aethalometer, $PM_{2.5}$ with the Harvard impactor, and NO_2 with the Ogawa badge (Cyrys et al. 2012). **Table 2** provides an overview of the monitors used.

QA-QC Detailed QA-QC plans were developed including standard operating procedures, zeroing of instruments, co-location with reference monitors in the NAQMN, co-location of the four instrument sets every 3–5 measurement weeks, and field blanks and duplicates of the integrated samplers (Additional Materials 1 Section 3.3).

Table 2. External Validation New Campaign: Measured Pollutants and Measurement Devices

Pollutant	Measurement Device	Type of Monitor
UFPs	DiscMini	Continuous monitor, set at 1-second resolution
$PM_{2.5}$	Harvard impactor with collection of PM on a Teflon filter Pre- and post-weighing of filter in the standard weighing room	Integrated weekly sampler
BC	MA200	Continuous portable monitor, set at 1-minute resolution
$PM_{2.5}$ absorbance (proxy for BC)	Filter reflectance on $PM_{2.5}$ filter	Integrated weekly sampler
NO_2	Ogawa badge, followed by spectrophotometric analysis	Integrated weekly sampler

Site Selection The planned 100 sites for the validation study were divided into 38 traffic sites, 50 urban background sites, and 12 regional background sites. This distribution was chosen to be broadly representative of the proportion of the population living near major roads, urban areas, and regional areas, with an over-representation of major roads due to their high variation in air pollution concentrations. The sites were spread out over the Netherlands, weighted by degree of urbanization.

The six largest cities (Amsterdam, Rotterdam, The Hague, Utrecht, Groningen, and Eindhoven) were each allocated two traffic sites, and the three largest of these (Amsterdam, Rotterdam, and The Hague) were allocated two urban background sites as well. All other cities were allocated one urban background site. The twelve regional background sites were distributed among each of the twelve provinces in the Netherlands. We list below the following site selection criteria:

- **Traffic sites —** Locations were chosen with a traffic intensity of at least 10,000 vehicles per day. A closed garden or a balcony on the first floor on the street side of the building had to be available to perform measurements safely. There must be no bus stop or service road in front of the building. Gas stations, traffic lights, large intersections, tunnels, and parking garages in the vicinity were avoided. To minimize the strong impacts of wind direction, large open streets such as highways were not selected. Wherever possible, traffic sites were selected with variations in traffic type and intensity, land use, and road length and width.
- **Urban background sites** — These were placed at least two kilometers away from corresponding traffic sites, wherever possible. This was to ensure that at the traffic sites, all factors that contribute to air pollution concentrations rather than just the local traffic contribution were estimated, including those that vary on a larger scale such as population density, land use, and road length. Major roads or other sources (such as parking lots and construction sites) should not be present within a radius of 100 meters of the urban background site. Large industrial sources should be further away.
- **Regional background sites —** These were located in small villages with a maximum of 10,000 inhabitants. Major highways should be at least 5 kilometers away from the site, wherever possible.
- **Reference site —** This was set up near the city of Utrecht, at a regional background site.

Using these criteria, sites were selected using Google Airview Earth data. Additional Materials 1 Figure S4 illustrates the distribution of the selected sites. Sites were finally selected during a physical visit to assess the suitability of a site and obtain permission of the homeowner. All sites were residential locations.

The micro-environmental criteria for the sites were checked during the enrollment stage of the site selection. These criteria ensured that the airflow around the sampler inlets was unrestricted. The sampler inlet should be placed:

- At least 20 centimeters from any vertical surface (such as a wall), and if possible at least twice the distance from an obstacle as the height of that obstacle
- At least 2 meters from a high-volume sampler inlet, 1 meter from a medium-volume sampler (such as the Harvard impactor), and 50 centimeters from a NO_x measurement
- At least 5 meters from exhaust flues or vents from homes or other buildings and air conditioning
- At least 1.5 meters (preferably 2 meters) above the surface on which the sampler is placed and preferably between 1.5 and 3 meters above the ground
- Away from the drip line of trees

The site has to be safe from vandalism or accidental damage. Electricity has to be available and the site should be accessible for field workers.

Measurement Schedule Recruitment of participants who would be willing to host the outdoor air pollution monitors started in November 2020. Due to the COVID-19 pandemic and the increased measures taken by the Dutch government from mid-December 2020 onward, recruitment and measurement activities for the external validation study were suspended until June 2021. As mobility was restricted in the Netherlands due to the COVID-19 pandemic measures, traffic intensity decreased. Measurements of air pollution taken during this period would likely have been unrepresentative for normal circumstances. Because the measurements were used to validate and compare models that have been developed under normal circumstances, this would likely have led to biased results. We therefore decided to postpone the external validation measurement campaign until September 2021 when most measures were lifted.

After the start of the measurement campaign, each week the equipment was set out on 1 day, collected 1 week later on the same day, and set out again at different sites some days after collection. The original plan of deploying the instruments to a new set of sites the day after was not feasible, because we observed that the DiscMini required cleaning of the corona wire after each monitoring week. The number of DiscMini and MA200 devices available to us was limited, thereby limiting the number of measurement kits that could be set out each week to three, in addition to the continuous reference site.

The sites where measurements took place each week were spread out across the country as much as possible to avoid clustering of measurements during certain periods or seasons. In addition, at least two different types of sites (traffic, urban background, regional background) were included in each week's measurements, to avoid space-time bias.

3.4 EPIDEMIOLOGICAL STUDIES

We calculated long-term average air pollution exposure at the residential address for three selected epidemiological studies using the different exposure models. For the empirical models, different algorithms were specified. The three cohort studies that we used were selected because previous studies in these cohorts showed associations between long-term air pollution and the selected health outcomes. We selected two cohorts of adults (DUELS and EPIC-NL) and one mature birth cohort (PIAMA). **Table 3** presents some basic characteristics of the studies including the outcomes studied. Additional Materials 1 Table S1 presents the definition of the outcome for the adult cohort analyses.

The Dutch Environmental Longitudinal Study (DUELS) cohort is a large national administrative cohort (Fischer et al. 2015; Klompmaker et al. 2021; Stafoggia et al. 2022). The cohort was analyzed extensively in the ELAPSE study. Primary health outcomes were natural-cause and cause-specific mortality. Briefly, population statistics based on digital municipal registers are combined by Statistics Netherlands into a longitudinal file for each individual registered in the municipal registration. Changes in demographic attributes (e.g., death, address, marital status, emigration, region of origin) are updated yearly. In these files, the individual identification number is replaced by an encrypted unique identification number. This identification number is used to enrich the individual files with information from other central data sources available at Statistics Netherlands, including the tax registry and morbidity and mortality registries. In previous analyses, we used all Dutch inhabitants ≥30 years old on January 1, 2004, and used a national LUR based on the national monitoring network (Fischer et al. 2015). For the ELAPSE study, we included a follow-up from January 1, 2008, to January 1, 2013 (Klompmake et al. 2021; Stafoggia et al. 2022). Most analyses were based on the European ELAPSE exposure model. Klompmaker and colleagues compared the ELAPSE, ESCAPE, and the national dispersion model in the DUELS cohort (Klompmaker et al. 2021).

For the current project, Dutch inhabitants ≥30 years old on January 1, 2013, were included with a follow-up until December 31, 2019. The rationale for starting in 2013 is to have more recent data, and because the mortality cause of death coding changed from manual to automatic in 2013, which has resulted in significant shifts in the coding of specific causes of death.

The Prevention and Incidence of Asthma and Mite Allergy (PIAMA) birth cohort was set up in The Netherlands with two main aims: to investigate the effect of mite-allergen avoidance on the incidence of childhood asthma and allergies (Intervention study), and to assess lifestyle and environmental risk factors for childhood asthma and allergies (Natural History study) (Wijga et al. 2014). The PIAMA study provided important insights into the impact of outdoor air pollution on the development of asthma and lung function in children (Gehring et al. 2015, 2020; Milanzi et al. 2018). The baseline study population consisted of 3,963 children born in 1996–1997. Questionnaires were completed by parents during pregnancy, when the child was 3 months old, and then annually from 1 to 8 years. Then at 11, 14, 17, and 20 years, participants themselves completed the questionnaires. Data were obtained on child and family characteristics, a wide range of environmental and lifestyle exposures, and asthma and other allergic and respiratory outcomes. Home visits and clinical examinations were conducted in subgroups at age 1, 4, 8, 12, 16, and 20 to obtain objective measures. These included anthropometric measures, measurements of lung function, bronchial hyper-responsiveness (at age 8), blood pressure (age 12 and up), house dust (up to age 8), and blood samples. Blood samples were used to determine total and specific IgEs, cholesterol, HbA1c, and to extract DNA.

The primary health outcomes evaluated in the current study were lung function (Gehring et al. 2013; Milanzi 2018) and asthma incidence (Gehring et al. 2020). Following our most recent papers assessing associations between UFPs and health in the PIAMA study, we analyzed asthma incidence using data from age 1 to 20 (Yu et al. 2022), and lung function at age 16 (Yu et al. 2021).

The EPIC-NL study is the Dutch contribution to the EPIC study (European Prospective Investigation into Cancer and Nutrition). The EPIC-NL study consists of two cohorts: the Monitoring Project on Chronic Disease Risk Factors (MORGEN) cohort and the Prospect cohort. The MORGEN

Table 3. Selected Epidemiological Studies

	Study Population	Recruitment	Follow-Up	Endpoints Analyzed	References
DUELS	Administrative cohort of 10 million adults above 30 years old	2013	2013–2019	Natural-cause, cardiovascular, respiratory, and lung cancer mortality	Fischer et al. 2015; Klompmaker et al. 2021; Stafoggia et al. 2022
PIAMA	Birth cohort study of almost 4000 children	1996–1997	1996–present	Lung function at age 16 and asthma incidence at ages 1–20 years	Yu et al. 2021, 2022
EPIC-NL	Adult cohort of 40,000 subjects with detailed lifestyle information available	1993–1997	1993–2013	Natural-cause mortality and incidence of coronary events and stroke	Strak et al. 2021; Wolf et al. 2021

cohort consists of 22,654 men and women aged 20–65 randomly selected from the Dutch population in three cities or towns in the Netherlands (Amsterdam, Doetinchem, and Maastricht). The Prospect cohort consists of 17,357 women aged 50–69, from the Dutch city of Utrecht or its vicinity, who participated in a breast cancer screening program. For both cohorts, participants were recruited between 1993 and 1997. Follow-up was until December 31, 2013. A baseline questionnaire was administered with detailed information on lifestyle: including smoking status, intensity, and duration; body mass index; and food and alcohol intake. Primary health outcomes were natural-cause mortality and incidence of (specific) cardiovascular disease, following the ELAPSE study of which EPIC-NL was part (Strak 2021; Wolf 2021). We specifically assessed the incidence of stroke and coronary events, following ELAPSE definitions (Wolf et al. 2021). We used the health and covariate database of the ELAPSE study to link all exposures. No further update of the follow-up was feasible for this study. Additional Materials 1 Table S1 presents the definitions of the outcome variables for the DUELS and EPIC-NL study, following the ELAPSE study.

4. DATA ANALYSIS

In section 4.1, the statistical methods to develop empirical exposure models are discussed. In section 4.2 we discuss the comparison of predictions of the different exposure models at randomly selected comparison sites and external validation sites. In section 4.3 we discuss the epidemiological analysis methods. In section 4.4 we discuss the comparison of effect estimates in the epidemiological studies.

4.1 EXPOSURE MODEL DEVELOPMENT

4.1.1 Modeling of Low-Cost Sensor Data

The newly collected $PM_{2.5}$ and NO_2 annual average concentrations were used to produce maps of annual average concentrations in two ways:

1. The main approach was combining data from the low-cost NO_2 sensor network, dispersion model estimates, and the routine monitoring network using data fusion. Presently, an increasing number of low-cost PM_{10} and $PM_{2.5}$ sensors are being combined with official measurements in the Netherlands into experimental hourly maps of air quality. The number of PM sensors feeding this map varies throughout the year for different reasons, which makes it too difficult to translate the results into long-term averages. Hence, the data fusion model was only developed for NO_2.
2. Development of empirical LUR models using land use and traffic data for $PM_{2.5}$ and NO_2.

Modeling Using Data Fusion Within the study, the data fusion method(s) focused on optimizing the prediction of NO_2 concentrations, because NO_2 is considered to be a good proxy for monitoring traffic emissions in general. Within the data fusion model, we tried to distinguish between background concentrations and concentrations originating from sources close by roads to improve the quality of the predicted total concentrations.

About 30–50% of the 50 sensor kits within the project were installed in the direct vicinity of roads, and could in principle be used to improve model predictions for road contributions to the concentration of NO_2. The remaining part was used to improve background concentration predictions. The sensors were also used to improve the calibration within the data fusion. The added value of the low-cost sensors is in providing more detailed concentration gradients in the area being investigated.

Data Fusion on Finer Scale At present, RIVM already uses data fusion methods that make use of reference measurements. The result of this data fusion method takes the form of a nationwide (keeping in mind the relatively small surface area of the Netherlands) field of correction factors for the traffic contribution and background concentrations. The general idea can be summarized in Equation 1:

$$calc\_tot\_cor(x,y) = calc\_bg(x,y) + calc\_bg_{cor} + calc\_traf(x,y) * calc\_traf_{cor} \quad [1]$$

with

calc_tot_cor(x,y): The corrected (fused) calculated total concentration field,

calc_bg(x,y): The uncorrected calculated background concentration field, calculated with an interpolation of reference measurements at the background (i.e., away from busy roads) locations in the Netherlands. The interpolation model additionally uses land cover data and correlations between reference stations based on historical data.

calc_traf(x,y): The uncorrected calculated traffic contribution field, calculated with a Gaussian plume model that takes wind direction, wind speed, traffic emissions, and road characteristics as model input.

The correction (fusion) factors are calculated as follows:

$$calc\_bg_{cor} = \tfrac{1}{N}\sum_{i=1}^{n}\left(m\_bg\_loc_i - calc\_bg\_loc_i\right) \quad [2]$$

$$calc\_traf_{cor} = \frac{\sum_{j=1}^{m}\left(m\_traf\_loc_j - (calc\_bg\_loc_j + calc\_bg_{cor})\right)}{\sum_{j=1}^{m} calc\_traf\_loc_j} \quad [3]$$

with

m_bg_loc$_i$: measurement at background location *i,*

calc_bg_loc$_i$: background calculation at background location *i,*

m_traf_loc$_j$: measurement at traffic location *j,*

calc_traf_loc$_j$: calculation of traffic contribution at traffic location *j.*

The definition of background now differs slightly from the definition above. Roadside stations, upwind of a (busy) road, are additionally classified for that particular hour as "background" instead of "roadside."

All factors are calculated hourly. The initial model results are used to determine which measurement locations are influenced by emissions from nearby traffic. Only those locations are used in the above summation for *calc_traf_loc*$_j$. The other locations are used to estimate the bias in background concentrations.

Note that the correction factors could be calculated in lots of different ways (e.g., different types of regressions). Furthermore, considering that typical meteorological circumstances and traffic intensities (which we are trying to correct with the data fusion) can also vary significantly on a smaller scale, we investigated if the sensors could be used to apply this method at different scales. This is especially relevant because there may not be enough reference measurements to do so.

Improving Current Fusion Methods In the present national Dutch approach, correction factors are calculated every hour, and there is no forced relation between consecutive hours. We investigated introducing damping factors between hours to anticipate the larger measurement uncertainties of the sensors as well as the lower number of available measurements (even including sensors) when performing data fusion on a finer scale. The results (Additional Materials 2 Figure S16) show that introducing damping factors based on varying weights between the current and preceding hour has a negligible effect on the performance of the fusion model.

4.1.2 Empirical Exposure Model Development

We extended our earlier empirical model algorithm comparison work (Chen et al. 2019; Kerckhoffs et al. 2019) where we evaluated algorithms beyond linear regression, including machine learning methods such as Random Forest and neural networks, as well as regularization methods such as LASSO and elastic net. We focused on three algorithms from the large set of algorithms evaluated by Kerckhoffs et al. (2019) and Chen et al. (2019) covering three major families of modeling approaches (generalized, regularized, machine learning): SLR, LASSO, and Random Forest. We applied these algorithms for all empirical models to develop annual average air pollution models. For the three algorithms, the *R* scripts applied in our two comparison papers were used.

The SLR algorithm was used in multiple LUR models including ESCAPE, ELAPSE, and short-term roadside and mobile (on-road) monitoring models (de Hoogh et al. 2018; Eeftens et al. 2012; Kerckhoffs et al. 2016, 2017, 2019; Montagne et al. 2015; van Nunen et al. 2017). In this algorithm, a univariate linear regression model was run for each potential predictor to choose the model with the highest adjusted R^2 as the starting point. Additional predictor variables were allowed to enter the model if they added to the adjusted R^2 of the previous model step, and only if they adhered to a plausible direction of effect (e.g., positive for traffic intensity). The script includes checks on influential observations (Cooks' D) and variance inflation factors. Variables in the final model with a $P > 0.10$ are excluded from the model.

Regularization or shrinkage algorithms are used to estimate reliable predictor coefficients when the predictors are highly correlated. We used LASSO instead of the originally planned elastic net models to limit the number of models we compared, which became excessive. The script that we specified:

```
cv.glmnet(X,Y,alpha=1, nfolds=10, nlambda=100,
    useMin=TRUE,loss= "deviance")
```

with X=set of predictors, Y=dependent variable (measured concentration)

Machine learning algorithms can model nonlinearity and the potentially complex interactions among predictors. We used Random Forest, a relatively frequently used decision tree method. Random Forest performed well compared to, for example, neural networks in the two algorithm comparison papers (Chen et al. 2019; Kerckhoffs et al. 2019). The script that we specified:

```
randomForest(Y,X,mtry=ncol(X)/3, ntree=1000, nodesize=5,
    maxnodes=NULL)
```

with X=set of predictors, Y=dependent variable (measured concentration)

The predictors offered in all the empirical procedures are summarized in Additional Materials 1 Table S2. They are based on the GIS variables developed in the ESCAPE study and further developed in subsequent mobile monitoring studies (Eeftens et al. 2012; Kerckhoffs et al. 2021).

For the European models, traffic intensity variables were not available. Instead, the length of all roads — and specifically major roads — was used to represent motorized traffic emissions. To assess the background across Europe, predictors from satellite observations and chemical transport models were offered.

4.1.3 Data Fusion Approaches

In addition to the data fusion used in the low-cost sensor model development, we used an ensemble approach to combine predictions from the different models. This method is applied especially if there is uncertainty about which model performs best. We applied the simplest method by taking the median and mean of predictions from different models, building on the MACC ensemble model of chemical transport models used in the European hybrid model (de Hoogh et al. 2018). Calculation of a weighted average of different models was also performed, with the weight determined by the coefficient of determination (R^2), which represents model performance (Jerrett et al. 2017). We only used model averaging in assessing whether this procedure improved the prediction of external validation data.

4.1.4 Exposure Assignment

We used the 2019 version of the standard Dutch database of geo-coded addresses (BAG) as a basis for exposure assignment to cohorts and validation sites. By extracting all geographical predictors to the about 10 million residential addresses from this database, we obtained model predictions for most empirical models. For the national monitoring network validation sites, we additionally extracted predictor variables, as these sites may deviate in exact location from residential addresses. For the European models developed using grids, we used an overlay of the coordinates of the addresses and validation sites with the 100-meter grids (ELAPSE) and 25-meter grids (EXPANSE). Dispersion model estimates were made for all BAG residential addresses. For assignment to network sites, we used the nearest residential site. Distance between network sites and the nearest residential sites was generally small (up to 30 m) for urban sites, mostly parallel to roads (e.g., a monitor located in front of a nonresidential building was assigned the modeled value for the neighboring home). Some traffic monitoring sites were located a few meters closer to the road than to the nearest home, which will have affected comparisons. For some urban background sites (e.g., those in parks), differences were larger (up to 100 m). For regional background sites, distances larger than 100 meters were found. For background sites, these distances are unlikely to have affected concentrations given their sphere of representativeness. Hence when comparing validation statistics for network sites between dispersion and empirical models, some underestimation may occur for the dispersion models. As routine monitoring network sites have been used (in the past) to evaluate models, these sites have limited value when used to compare empirical and dispersion models. We therefore did not perform the more appropriate specific dispersion model calculations for the network sites. The underestimation of agreement for dispersion does not apply to the specifically selected validation sites including the new validation sites, RUPIOH, Exposomics, and TRACHEA, as we selected residential addresses to perform targeted monitoring. These sites are also truly independent, meaning they were not used in model development.

The BAG database was then used to link exposures to the national cohort in the secure Central Bureau of Statistics environment and to the two other cohort studies using tailor-made solutions to safeguard privacy. Exposure data were assigned to the residential address at recruitment of the study. The data fusion procedure integrating low-cost sensor measurements was too computationally intensive to apply to all 10 million addresses, as the model operates on an hourly basis. We applied the data fusion method to the 20,000 randomly selected BAG addresses used for exposure prediction comparison (section 4.2), the about 9,000 addresses of the PIAMA birth cohort study (birth and current addresses of the different follow-up years), and the new external validation sites.

4.2 EXPOSURE MODEL PREDICTION COMPARISON AT RESIDENTIAL AND VALIDATION SITES

We compared the different exposure models in two ways. First, we compared model predictions at 20,000 residential addresses across the Netherlands, of which 10,000 were residential addresses in the four major cities of the Netherlands (Amsterdam, Rotterdam, Utrecht, and The Hague), and 5,000 of those were addresses in Amsterdam specifically. We also compared model predictions at addresses of the three epidemiological studies included in the project. This analysis informs on the correlation and possible differences in level between the different exposure models but does not inform on differences in predictive performance. Second, we compared models in their ability to predict newly collected and existing external validation data for UFPs, BC, $PM_{2.5}$, and NO_2. In addition to comparing the different exposure models (Table 1), we compared differences between the three algorithms within each of the empirical models.

Comparison of Predictions at 20,000 Residential Addresses The comparison addresses were randomly selected from the Dutch address database of 2019 and stratified by degree of urbanization. Of the 20,000 addresses, 10,000 addresses were in the four major cities (all located in the west of the country within 100km of each other); the remaining 10,000 were from less urbanized towns. Of the 10,000 addresses in major cities, 5,000 were selected from Amsterdam. We made comparisons for three spatial domains: the entire Netherlands, major cities, and Amsterdam because the monitoring area differs between exposure models (Table 1). Furthermore, the comparison between models may depend on the spatial scale of the study domain, for example, whether the domain includes only within-urban differences in exposure or also urban-rural and regional differences.

We compared predictions of the different models by preparing scatterplots and calculating the Pearson correlation between predictions of the different models. For epidemiological studies, ranking of exposures is critically important, even when the absolute level is biased. We also calculated Spearman correlations, which were similar (not further reported for conciseness). We further prepared Bland-Altman plots to allow the identification of any systematic difference between the measurements (i.e., fixed difference) or trends with the level of exposure. We further compared the predicted exposure contrast (IQR of the exposure distribution across sites), as a smaller exposure contrast may lead to a higher effect estimate in epidemiological studies when expressed for fixed increments.

We additionally calculated the correlation of predictions for multiple pollutants for the different models. High correlations between the predictions of two pollutants complicate the assessment of independent effects in epidemiological studies. We evaluated two-pollutant models only in the large DUELS cohort. We selected dispersion modeling and Europe-wide hybrid models as the models covering most pollutants.

Comparison of Model Predictions with External Validation Data We assessed model performance in terms of R^2, root mean square error (RMSE), and bias (model minus measurement) by comparison of model predictions with independent external validation data. Validation data have been described in section 3.3. The R^2 is important for epidemiological studies, as ranking of subjects in low and high exposures is essential. Bias is important as epidemiological studies are used in risk assessment to determine the level of exposure at which potential air quality standards are set. The predicted exposure contrast is important, as it may affect the magnitude and precision of the health effect estimates.

We compared annual average concentrations predicted by the models and the validation sites, as we were interested in how well long-term exposure is predicted by the models.

4.3 EPIDEMIOLOGICAL ANALYSIS

We analyzed the association between selected health outcomes and exposure in the epidemiological studies following the analytical design of the original papers (Stafoggia et al. 2022; Strak et al. 2021; Wolf et al. 2021; Yu et al. 2021, 2022). No new analysis methods were applied.

Briefly, we used Cox proportional hazards models to assess the association between air pollution exposure and natural-cause and cause-specific mortality in DUELS, as well as natural-cause mortality and incidence of stroke and coronary events in EPIC-NL. We built on the methods developed within the ELAPSE project (Stafoggia et al. 2022; Strak et al. 2021). We applied Cox proportional hazards models with varying control for individual- and area-level covariates to analyze the associations between air pollution and mortality and morbidity. Age was used as the time scale because of evidence that this results in better adjustment for potential confounding by age (Strak et al. 2021). No assumptions are needed with respect to the shape of the association between age and mortality. Time to event is based on age at the event and baseline. Censoring occurs at the time of the event of interest, death from other causes, emigration, loss to follow-up for other reasons, or at the end of follow-up, whichever comes first. We specified three confounder models a priori, with increasing level of adjustment:

1. Model 1 includes only age (time axis) and sex (strata).
2. Model 2 added individual-level variables: In DUELS, only information on individual-level socioeconomic status was available, while no data on lifestyle was available. Therefore, the included covariates in this cohort were standardized household income, region of origin (Dutch, Western, non-Western, Morocco, Turkey, Suriname, Antilles Netherlands), and marital status (married, widowed, divorced, single). In EPIC-NL we added smoking status (never, former, current), smoking intensity (number of cigarettes per day), smoking duration (years of smoking), fruit and vegetable intake, alcohol consumption (low, medium, high), body mass index (categories defined by WHO, <18, 18–25, 25–30, >30), educational level (low, medium, high), employment and occupational status (employed, unemployed, housekeeping, retired), and marital status (single, married or with partner, divorced, widowed). Detailed documentation of the variables is included in previous papers (Stafoggia et al. 2022; Strak et al. 2021).
3. Model 3 added (correspondingly) area-level socioeconomic status variables to the final determined Model 2. Area-level in DUELS was represented by both the neighborhood and the regional scale (COROP), following the ELAPSE study. The philosophy is that we prefer to adjust for established risk factors for mortality and morbidity (socioeconomic variables) at different spatial scales compared to specifying indicator variables of the region of the country. We included mean income per income recipient, social assistance per 1,000 inhabitants, % low education, % non-Western immigrants, and unemployment per 1,000 inhabitants as variables characterizing socioeconomic status, largely following ELAPSE. In EPIC-NL — a smaller cohort with a narrower geographical distribution than DUELS — we included mean income, ethnicity, and unemployment rate at the neighborhood level. Model 3 was a priori selected as the main model, following the ELAPSE study.

Additionally, selected two-pollutant models for BC, NO_2, and $PM_{2.5}$ were specified, which compared the dispersion model and the Europewide models for 2016. These models were applied in the analysis of natural-cause mortality in the DUELS cohort.

In PIAMA, associations between air pollution and lung function were analyzed by linear regression and associations with asthma incidence using discrete-time hazard models. Using the methods of the original papers that assessed the effects of UFPs on health by applying the nationwide RUN model for UFPs (Yu et al. 2021, 2022), we analyzed asthma incidence using the observations from ages 1–20 as in (Yu et al. 2022) and for age 16 only for lung function as in (Yu et al. 2021). Associations were assessed with an adjustment for the same confounder sets as in the original papers. For lung function, we adjusted for sex, log-transformation of age, weight, and height, parental education, maternal atopy, paternal atopy, breastfeeding, respiratory infections in the last 3 weeks (before the medical examination), Dutch nationality, indoor tobacco smoke exposure in the home, maternal smoking in pregnancy, furry pets in the home, mold in the home, gas cooking, and average air pollution concentrations for the 7 days preceding the lung function measurement. For asthma, we adjusted for sex, maternal and paternal asthma, hay fever, native nationality, parental education, breastfeeding, older siblings, daycare attendance, maternal smoking during pregnancy, parental smoking at home, mold or dampness at home, pets, use of gas for cooking, and active smoking (from age 14).

4.4 COMPARISON OF EFFECT ESTIMATES IN EPIDEMIOLOGICAL STUDIES

We compared the effect estimates and confidence intervals for the different exposure models primarily from single pollutant models. In the main analysis, we expressed effect estimates for a fixed exposure increment, specifically 5 μg m^3 for $PM_{2.5}$, 10 μg/m^3 for NO_2, 1 μg/m^3 for BC, and 5,000 p cm^3 for UFPs. Additionally, we compared effect estimates expressed per IQR for each model to investigate how much differences in exposure contrast contribute to differences in effect estimates. Next, we compared effect estimates in two pollutant models, as correlations between pollutants may differ depending on the exposure model. Two pollutant models were specified for the dispersion model and the Europewide model, the two models with the most pollutants predicted. We hypothesized that correlations between pollutants might be different between deterministic and empirical models.

The main comparisons were performed for the entire Netherlands or the full study area of the three cohorts. The domain of some models (mainly mobile monitoring models) was limited to specific major cities, whereas the other models applied to the whole of the Netherlands. To investigate the influence of these differences on health effect estimates, epidemiological analyses were performed for subsections of the large DUELS cohort living in Amsterdam in addition to analyses within the entire cohort. The PIAMA and EPIC_NL cohorts were too small to allow this analysis.

The comparison of effect estimates does not shed light on which effect estimate is closer to the true value. We do not think that the model with the highest effect estimate is the best model, because, for example, some exposure may induce more confounding. One perspective could be that the effect estimates from the model with the best performance in predicting outdoor exposure should be judged as the best approximation of the true effect. We consider this a reasonable starting point but need to consider some limitations of that approach (Szpiro et al. 2011). Models may differ in the correlation between predicted exposures and individual and especially area-level confounders. We therefore compared effect estimates from different models in the crudely adjusted model 1, the model adjusted for individual covariates (model 2), and the most-adjusted model 3 to judge how sensitive differences between models were for adjustment for confounders. If the type of exposure measurement error is different in the various models, the impact of measurement error on epidemiological effect estimates may differ. Specifically, if in some models the Berkson error has a larger contribution than classical error, there may be less bias. We did not further evaluate measurement error.

We investigated why effect estimates differed between models. We evaluated several factors that could explain differences in effect estimates. We formally evaluated whether the performance of the exposure model at external validation sites was associated with the epidemiological effect estimate by preparing scatterplots and linear regression analysis. This analysis was conducted in the large DUELS cohort for adults and the PIAMA birth cohort. The hypothesis was that models with higher performance would result in higher effect estimates. We further evaluated the predicted exposure contrast, hypothesizing that a lower predicted exposure contrast would result in a higher effect estimate when expressed for fixed increments. We calculated effect estimates both per fixed increment and model-specific IQR. We further evaluated the type of model (deterministic versus empirical), algorithm, and year of the model. We furthermore assessed consistency across the three cohort studies, albeit more qualitatively. We assessed whether we found specific models to result in lower or higher effect estimates.

We explored whether the characteristics of Bland-Altman plots were related to differences in HRs between models. We focused on hazard ratios of BC and NO_2 with natural-cause mortality, as the pollutants with the largest diversity of exposure models. We calculated differences between the HRs of all combinations of models. We also produced Bland-Altman plots of all combinations of models, aligning the differences between the two models. We then performed linear regression analyses of the difference in exposure predictions versus the mean for each Bland-Altman plot. We extracted the intercept, slope, and RMSE from each model. In addition, we obtained the mean difference between the two models. We then regressed the difference in HR on the statistics from the Bland-Altman plot.

To illustrate the degree of heterogeneity potentially introduced in meta-analyses by using different exposure models, we performed a random-effects meta-analysis with the different exposure models. We calculated the I^2 statistic to express heterogeneity, following the previous meta-analysis of outdoor air pollution (Boogaard et al. 2023a). This additional analysis was performed for the DUELS adult cohort and the PIAMA children's cohort.

5. RESULTS

In section 5.1, the internal model performance of the different empirical exposure models is presented. In section 5.2, the models based on the new low-cost monitoring campaign are discussed, using data fusion and standard LUR modeling approaches. Because the different exposure models were developed based on data from different years in the 2009–2019 period, in section 5.3 we first document air pollution time trends in the Netherlands and provide comparisons of the two exposure models with data from multiple years at 20,000 randomly selected sites across the Netherlands (dispersion modeling and Europewide hybrid models). In section 5.4, we compare predictions of the suite of models and the algorithms within the empirical models at the 20,000 randomly selected residential sites. In sections 5.5 and 5.6, we compare the exposure predictions of the different models with measurements at 90 new external validation sites and existing validation sites covering different years, allowing us to assess how well recent models can predict current and historical exposure contrasts. In sections 5.7 to 5.9, effect estimates from applying the different exposure models in three epidemiological cohort studies are compared. In section 5.10, we evaluate whether specific methods or design issues of the models are related to higher or lower effect estimates or confidence intervals, across the three cohorts.

5.1 EMPIRICAL EXPOSURE MODEL PERFORMANCE

For BC and $PM_{2.5}$, the largest diversity of models was available (Table 1). For UFPs, only mobile monitoring-based models were available. For NO_2, all types of models were available, although mobile monitoring was only available for the Amsterdam-based Google Airview campaign.

Table 4 presents the performance of the empirical models, indicating which models were developed specifically within the current project and which were taken from previous studies using the same algorithms. For some studies, such as the ESCAPE modeling, analyses were redone, as the internal cross-validation approach was not comparable to the current approach. The leave-one-out cross-validation in ESCAPE with a fixed model structure was shown to be overly optimistic as a validation method (Wang et al. 2012). Both model and ten-fold cross-validation R^2 are presented. Table 4 only shows internal validation; validation on external sites is discussed in sections 5.5 and 5.6. Internal validation statistics cannot be compared across models, as the monitoring data differs between models. We can compare statistics between algorithms within a specific model.

In most comparisons, we noted only small differences (<0.1 difference in R^2) between the three algorithms. The machine learning algorithm Random Forest did not perform consistently better than the SLR approach and the LASSO approach, although for BC and UFPs, Random Forest performed slightly better and LASSO slightly worse. The small differences in performance in internal model performance confirm the previous, more comprehensive algorithm comparison studies of models based on mobile monitoring data (Kerckhoffs et al. 2019) and Europewide routine monitoring data (Chen et al. 2019). External validation data were used in these studies, which is a more solid basis for comparing different models. We added the comparison for the nationwide ESCAPE study and the low-cost sensor data, both with fewer sites than the European model, but with fairly temporally stable monitoring data per site.

The low cross-validation R^2 for mobile monitoring and short-term roadside campaigns cannot be interpreted as indicating poor performance of spatial predictions, because the measured concentrations at individual sites include a large random error due to the short measurement time per site (Kerckhoffs et al. 2017; Montagne et al. 2015). Consistently, the cross-validation R^2 for models based on short-term roadside monitoring is generally higher than for the mobile monitoring in the same study.

We noted the generally poor performance of mobile monitoring-based $PM_{2.5}$ models, with most LASSO models containing only an intercept.

We developed SLR models for all empirical models and Random Forest for all except the ELAPSE European BC model. LASSO models were attempted for all mobile monitoring campaigns and ESCAPE. LASSO models were unavailable for multiple years of the EXPANSE project, and we did not have the resources to develop these models. We consider the comparison between the easy-to-interpret SLR and the more sophisticated Random Forest machine-learning model as the most important algorithm comparison.

Additional Materials 2 Table S1 lists the structure of the SLR models of the different empirical exposure models, to illustrate qualitatively which variables are included in the models. A quantitative assessment of the importance of each predictor in all models is beyond our scope. Most models included traffic variables with small (<100 m) and large buffer sizes (1,000 m and larger). Population or address density was also included in most models. Models also included sources beyond road traffic, including port, industry, and aircraft (e.g., in Google Airview UFP model). These sources were mostly included with large buffer sizes, often a 5,000-meter radius, reflecting the area-source nature of these sources. The inclusion of 5,000-meter buffers for population density and source-related variables implies that models may explain differences in air pollution within urban areas as well as between areas with different degrees of urbanization or the presence of large source areas.

The European models included different variables related to the availability of the data at a European scale. The models included road length but not traffic intensity, whereas all Dutch models did include traffic intensity. Furthermore, the European models included satellite measurements and large-scale chemical transport model estimates for NO_2, BC, and $PM_{2.5}$, to represent differences in background concentrations across Europe.

Table 4. Comparison of Internal Model Performance of the Different Empirical Models and Algorithms

				SLR		Random Forest		Lasso	
Pollutant	**Model**	**Model type**	***N* sites**	**Model R^2**	**CV R^2**	**Model R^2**	**CV R^2**	**Model R^2**	**CV R^2**
UFPs	Google	Mobile	48,594	0.18[a]	0.18	0.21	0.21	0.18	0.18
	MUSIC_m	Mobile	2,964	0.12[a]	0.12	0.34	0.19	0.12	0.10
	MUSIC_st	Mobile	128	0.39[a]	0.32	0.83	0.27	—	—
	EXPO_m	Mobile	5,236	0.15[a]	0.15	0.31	0.23	0.15	0.14
	EXPO_st	Mobile	240	0.47[a]	0.41	0.78	0.40	0.45	0.31
	RUN_m	Mobile	14,887	0.18[a]	0.17	0.29	0.21	0.18	0.16
	RUN_st	Mobile	389	0.31	0.28	0.72	0.22	0.24	0.04
BC	Google	Mobile	48,337	0.37[a]	0.37	0.42	0.41	0.39	0.39
	MUSIC_m	Mobile	2,313	0.13[a]	0.12	0.44	0.23	0.12	0.09
	MUSIC_st	Mobile	141	0.28[a]	0.20	0.64	0.15	—	—
	EXPO_m	Mobile	5,186	0.12[a]	0.12	0.34	0.21	0.11	0.10
	EXPO_st	Mobile	240	0.44[a]	0.38	0.77	0.42	0.41	0.32
	RUN_m	Mobile	15,239	0.12	0.12	0.32	0.27	0.12	0.11
	RUN_st	Mobile	389	0.33	0.32	0.63	0.31	0.32	0.27
	ESCAPE	LUR	40	0.92[a]	0.78	0.96	0.65	0.91	0.78
	ELAPSE	LUR	436	0.54[a]	0.51[a]	X	X	X	X
NO_2	Google	Mobile	47,266	0.49[a]	0.48	0.53	0.53	0.51	0.51
	ESCAPE	LUR	80	0.88[a]	0.76	0.94	0.73	0.94	0.77
	ELAPSE	LUR	2,399	0.59[a]	0.58[b]	0.95[a]	0.61[b]	0.61[a]	0.59[a]
	EXPANSE	LUR	3,176[c]	NR	0.62[a]	NR	0.65[a]	X	X
$PM_{2.5}$	RUN_m	Mobile	15,015	0.02	0.01	0.28	0.19	—	—
	RUN_st	Mobile	389	0.03	0.02	0.74	0.05	—	—
	EXPO_m	Mobile	3,711	0.05	0.04	0.67	0.62	0.27	0.24
	EXPO_st	Mobile	234	0.03	0.02	0.88	0.02	—	—
	MUSIC_m	Mobile	1,627	0.03	0.02	0.49	0.09	—	—
	MUSIC_st	Mobile	153	0.09	0.05	0.81	0.12	—	—
	ESCAPE	LUR	40	0.66[a]	0.36	0.94	0.39	0.59	0.44
	ELAPSE	LUR	543	0.62[a]	0.60[b]	0.96[a]	0.63[b]	0.64[a]	0.61[a]
	EXPANSE	LUR	1,433	NR	0.75[a]	NR	0.68[a]	X	X

Models explained in Table 1: CV = 10-fold cross-validation, can only be compared between algorithms. SLR = supervised linear regression. Sites refer to street segments in mobile monitoring. For the mobile monitoring-based models, _m refers to on-road mobile monitoring, and _st to short-term roadside measurements. LUR refers to models based on fixed-site monitoring (ESCAPE national, ELAPSE, and EXPANSE European).

— = Model with only intercept; NR = not reported (given the large N, nearly identical to CV R^2); X = no model developed.

[a] Model previously developed.

[b] 5-fold cross-validation.

[c] Numbers for 2019 are lower than in earlier years (Shen et al. 2022).

5.2 LOW-COST SENSOR NETWORK MODELS

The data of the low-cost sensor measurements were used primarily in a data fusion approach with dispersion models and NAQMN measurements (section 5.2.1). LUR models based on only the average concentration of the low-cost sensor measurements were developed, consistent with other empirical models (section 5.2.2).

5.2.1 Data Fusion Model

Quality Assessment and Control of PM Sensors Some of the distributed sensor kits have never been installed by participants or had connection issues. Therefore, we were not able to collect measurements with reasonable or sufficient temporal coverage at all of the 100 selected sites. We were able to retrieve data from 84 locations, with an average data coverage of 75%. The co-location of six sensor kits alongside reference measurements was used to estimate the data quality. Based on analyses of the co-located sensors we expect R^2 values for $PM_{2.5}$ sensor measurements to be around 0.85, with a slight positive bias (**Table 5**). The results shown in Table 5 are similar for PM_{10}, although the quality of sensor measurements is slightly less (see Additional Materials 2 Table S2), with R^2 around 0.75. Furthermore, the data show that PM sensors behave in the same way with high intercomparability ($R^2 > 0.9$). See Additional Materials 2 Figures S1 and S2.

If we consider all $PM_{2.5}$ measurements of all the sensor kits for the 84 aforementioned sites, we can conclude that the distribution of sensor measurements shows a slightly stronger positive skewness (**Figure 2**), compared to the nearest NAQMN-measurements (similar for PM_{10}; not shown). For NO_2 the results are vice-versa (Additional Materials 2 Figure S3).

Quality Assessment and Control of NO_2 Sensors Measured NO_2 and O_3 concentrations at nearby reference stations during hours with expected low spatial gradients of NO_2 are used to recalibrate the sensors at the home addresses of the participants. Hours with low NO_2 gradients usually occur in the late evening and at nighttime (night calibration). During these hours, the NO_2 concentrations at the sensor sites are expected to be approximately equal to the concentrations at the nearest official measurement station. In general, this assumption will become less justifiable with increasing distances to the nearest official station. As the O_3 concentration is also used to calibrate and correct the NO_2 sensor signal and spatial

Table 5. Metrics for All Hourly $PM_{2.5}$ Measurements (μg/m³) For Co-located Sensors at a Traffic Site (NL10636, Kardinaal de Jongweg) from the National Ambient Air Quality Network (NAQMN), with n = Number of Hourly Observations (Measurements from October 2021 to March 2023)

Kit-ID	n	$Mean_{sensors}$	$SD_{sensors}$	$Mean_{NAQMN}$	SD_{NAQMN}	Bias	RMSE	R^2
SSK_LH008	1,493	17.00	14.59	14.87	9.72	2.14	5.18	0.85
SSK_LH028	2,660	16.85	13.78	15.54	9.91	1.31	5.03	0.86
SSK_LH039	1,446	18.54	17.22	14.88	9.85	3.66	5.19	0.85
SSK_LH052	1,493	19.20	17.19	14.87	9.72	4.34	4.97	0.86
SSK_LH023	1,532	18.43	17.01	13.68	10.33	4.75	5.60	0.84
SSK_LH025	1,531	17.78	17.22	13.68	10.33	4.10	5.85	0.82

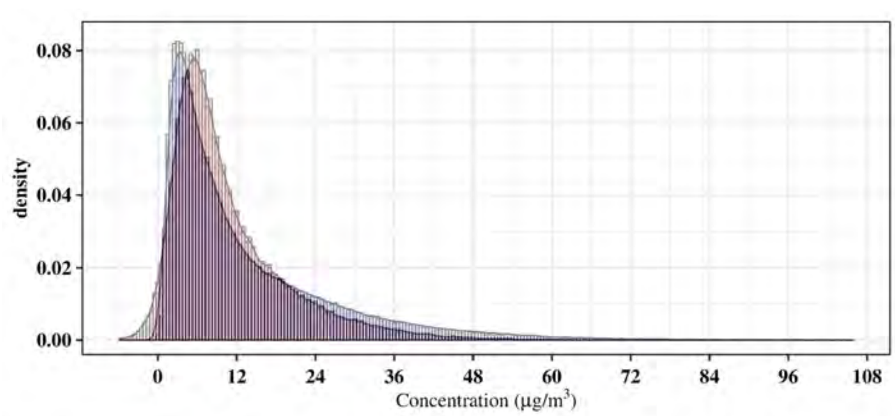


Figure 2. Distribution of all hourly measurements of $PM_{2.5}$ during both field campaigns for the sensors (blue) and stations (red) of the NAQMN.

gradients of O_3 are typically always small, we assume that the measured O_3 concentration at the reference station, nearest to the sensor, can be used to recalibrate and correct the sensors. Every Monday, the measurements from the nearest official measurement station during the low-gradient hours of the last 8 weeks are used to update the factors in the calibration formula for the NO_2 sensors. To quantify the quality of this calibration, some sensors have been rotating between reference and field locations. The calibration formula for these sensors is determined using the night calibration method at a participant's home address during a period of a few weeks. The same sensor was also placed at an official measurement site during the week before or after the measurements at the home address and calibrated using the night calibration, derived at the home address location. These calibrated NO_2 concentrations measured by the sensors were validated against the co-located official measurements.

As mentioned above, the quality of the night calibration is expected to decrease with increasing distance to the reference measurement that is used for the night calibration. To quantify this effect, multiple night calibrations, corresponding to different distances to the official measurement stations used in the night calibration, were derived for the rotating sensors during placement at the home addresses. The different night calibration formulas were additionally applied to the sensor signal of the rotating sensor during placement of the sensor at the reference station, and the resulting NO_2 concentration predictions were validated using the co-located reference measurement.

Figure 3 shows that in general, with increasing distance from a reference station that is used for calibrating and correcting the sensor signal, the correlation indeed decreases, and the RMSE increases (though less consistently). The bias between the calibrated sensor and co-located reference measurement (right figure) is sometimes still problematic. For example, sensor LH023 was validated in the middle of December 2022. This week was very cold compared to the weeks before which were used to calibrate the sensor. This shows that the night calibration is less able to account for (sudden) changes in meteorological circumstances in terms of bias. The correlation coefficient is less affected. The figure also shows that including the NO_2 sensor signal in the calibration increases the quality of the NO_2 prediction with respect to a calibration using only (close by) O_3 and temperature. Additional tests and validation of the calibrated NO_2 sensors are given in Additional Materials 2 Figures S17–S20. Figure 3 documents that the predictive power of the calibrated NO_2 sensor does not solely derive from the other predictors (O_3 & T) that are also used in the prediction/calibration. Therefore, we also included a calibration (green) that only uses T & O_3. If the blue triangles and green circles were on top of each other, the NO_2 sensor would have no added value in the prediction.

Spatiotemporal Sensor Patterns The positive bias of $PM_{2.5}$ (and PM_{10}) concentrations measured by the sensors can also be seen in the spatial representation of the average sensor measurements (**Figure 4**). For the first field campaign, the average $PM_{2.5}$ concentrations are displayed for all sensors and all NAQMN stations throughout the Netherlands. The

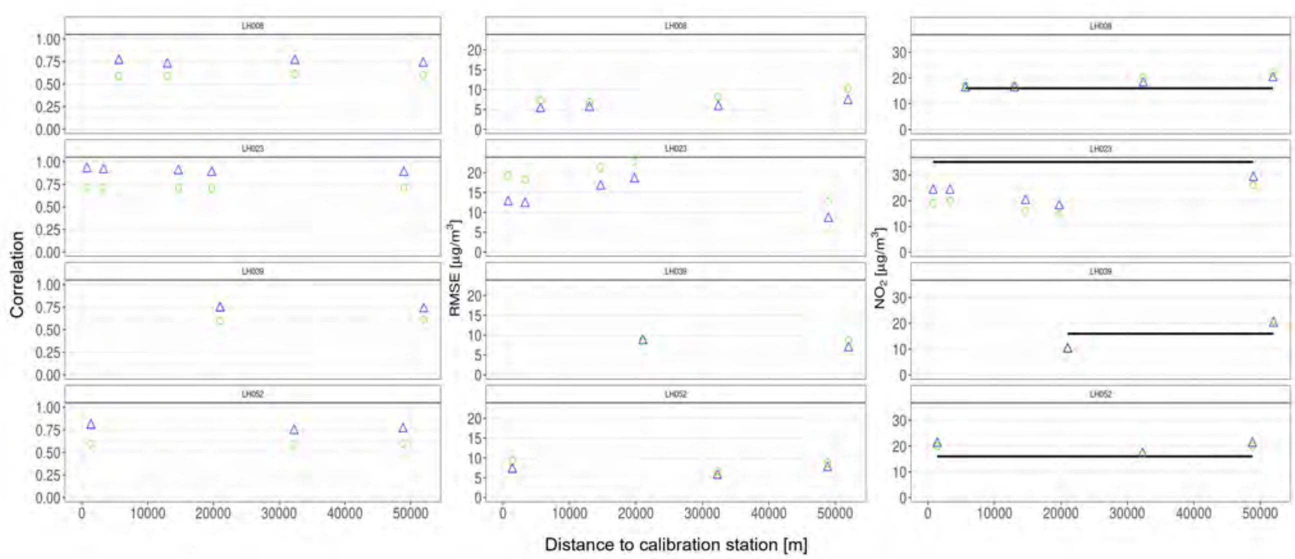


Figure 3. Validation metrics for the at-distance night-time calibration at participant locations, validated at NAQMN station Kardinaal de Jongweg. Left figure: Pearson correlation. Middle figure: Root Mean Square Error (µg/m³). Right figure: Average concentration (µg/m³) including the reference concentration in black. The x-axis shows the distance (m) to the reference station that was used to recalibrate. To estimate the quality of a calibration that does not use the raw NO_2 sensor signal as a predictor, the green circles indicate a prediction (linear regression) using only T and O_3 as predictor variables. The blue triangles show the improvement in validation statistics for the NO_2 sensor calibration which uses the raw NO_2-sensor signal, T, and O_3 as predictor variables.

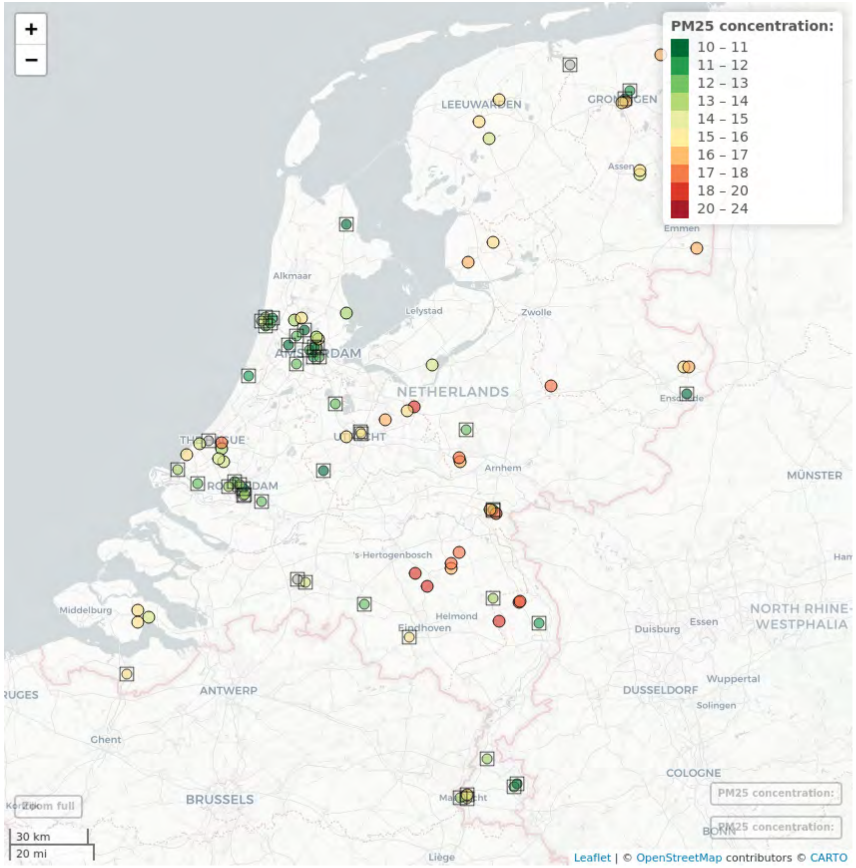


Figure 4. Map of the average $PM_{2.5}$ concentrations (μg/m³), for 10/2/2021–03/31/2022, based on all sensors (dots) and all their nearest NAQMN station average (squared dots).

sensors show relatively higher concentrations in the central part of the country although the differences are moderate (up to 5 μg/m³). The spatial patterns for PM_{10} and NO_2 (Additional Materials 2 Figures S4 and S5) are different, with relatively higher concentrations in the Randstad (western part of the country), with the highest concentrations in Rotterdam.

It is interesting to investigate whether the sensors can measure diurnal variation in concentrations due to, for example, rush hours. Because a significant amount of sensor kits are deployed near high-density traffic, we expected the sensors to show a diurnal pattern, at least for NO_2. This is confirmed by **Figure 5**, which shows that concentrations peak around 07:00–10:00 local (winter) time and a somewhat lower peak late in the afternoon and early evening. This pattern is less pronounced on weekend days and even less for $PM_{2.5}$ (**Figure 6**) and PM_{10} in general (Additional Materials 2 Figure S8). The diurnal pattern for $PM_{2.5}$ is probably a result of varying meteorological circumstances (lower boundary layer, higher humidity). For comparison with diurnal patterns, measured with reference equipment, see Additional Materials 2 Figures S6, S7, and S9. These showed similar diurnal patterns as the low-cost sensors.

Fusion of Calculations with Sensor Measurements The fusion of the NO_2 sensors with a calculation model was tested for a few fusion model variants and compared to a standard fusion model version that uses only measurements from the reference network. The variants consisted of running the fusion model in smaller regions and optionally forcing correlation between fusion models for consecutive hours. **Figure 7** shows the scatter of fusion model results on a week-average basis when compared to measurements of the validation campaign. It can be seen that including the sensors in the model results in increased overprediction for relatively low concentrations (5–15 μg/m³) measured in the external validation campaign. The nationwide model, using only reference measurements (top left figure), shows the best agreement with the validation campaign.

Additional Materials 2 Figures S13 and S14 also show a comparison of the investigated data fusion methods with roadside reference measurements, based on hourly averages and half-year average concentrations. From these combined analyses it is apparent that, in general, the standard fusion model, which does not use sensors, is hard to improve. At some individual roadside locations, however, the inclusion of sensors in the fusion model can decrease the RMSE slightly.

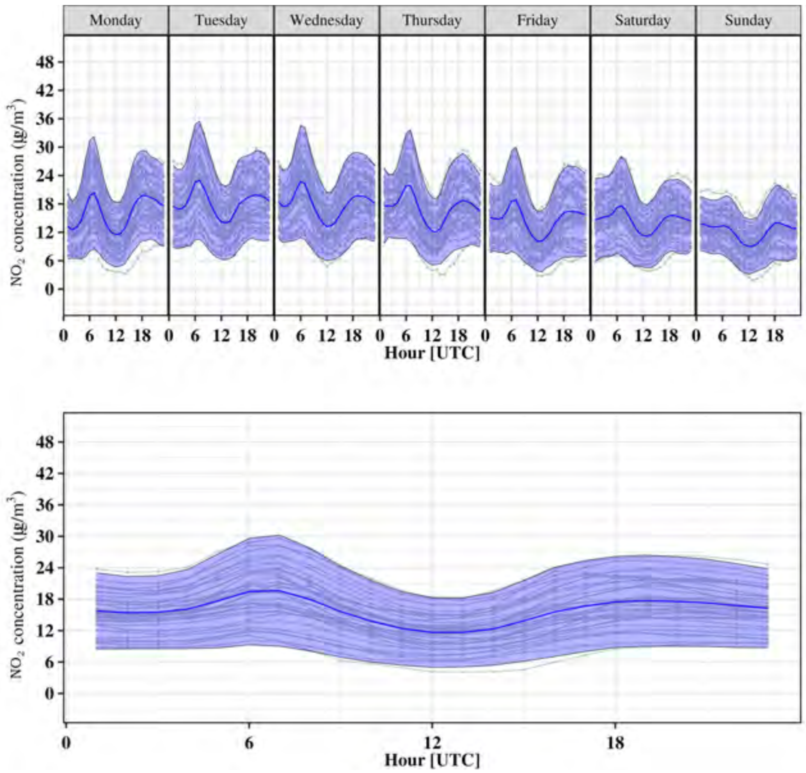


Figure 5. Weekly (upper) and daily (bottom) pattern of NO_2 concentrations for 10/2/2021–03/31/2022, based on all sensors (grey lines) and their average (blue line). The shaded area resembles the ± 2 times standard deviation.

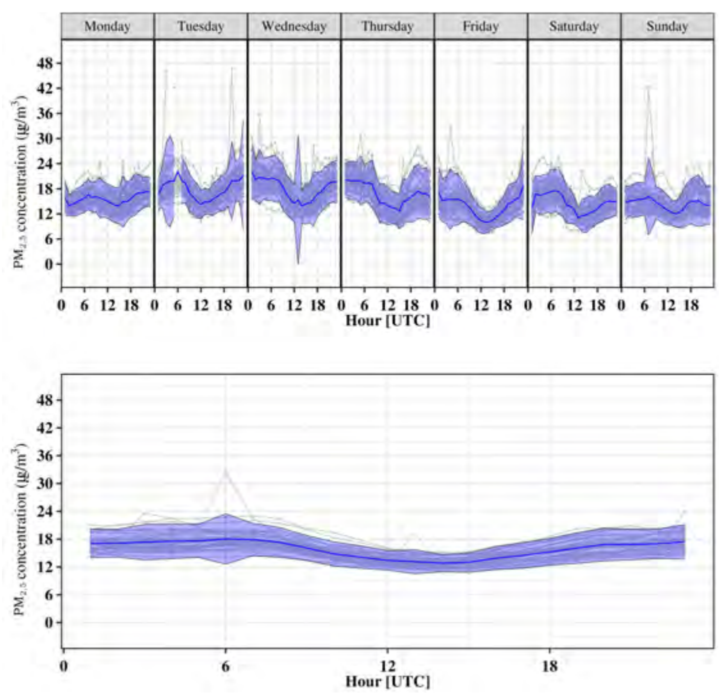


Figure 6. Weekly (upper) and daily (bottom) pattern of $PM_{2.5}$ concentrations for 10/2/2021–03/31/2022, based on all sensors (grey lines) and their average (blue line). The shaded area resembles the ± 2 times standard deviation.

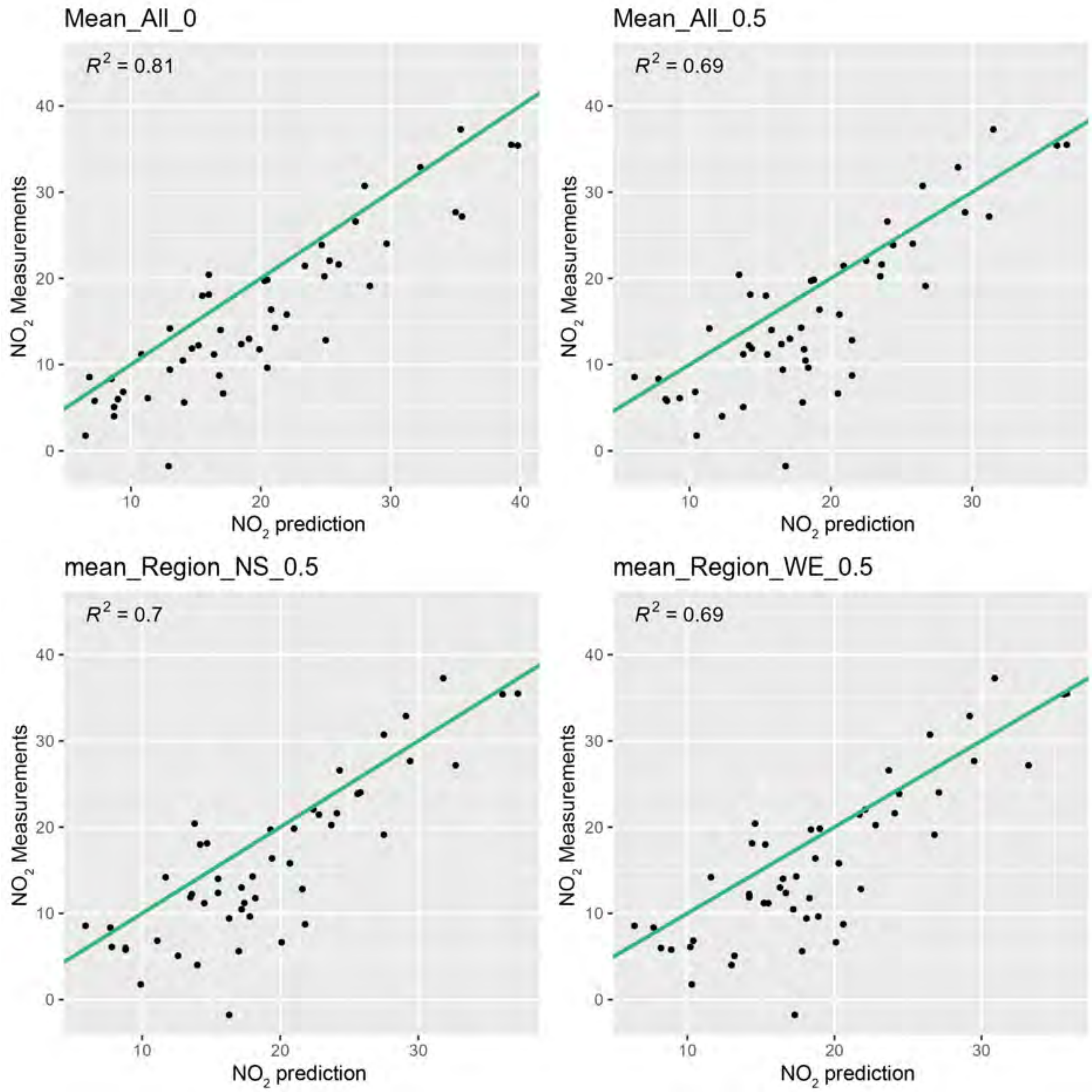


Figure 7. Weekly average NO_2 concentration of several model variants compared with concentrations (μg/m³) from the new external validation campaign (section 5.5). Model variants: Standard RIVM model with countrywide fusion of reference measurements (Mean_All_0), countrywide fusion of model with reference and sensor measurements (Mean_All_0.5), regional (North–South) fusion of model with reference and sensor measurements (mean_Region_NS_0.5), and regional (West–East) fusion of model with reference and sensor measurements (mean_Region_WE_0.5). The green line is the 1:1 line.

Additional Materials 2 further shows distributions of modeled road contribution and modeled total concentrations for PM_{10} and NO_2 (Additional Materials 2 Figure S12). It is shown that for PM_{10}, the road contributions are very small with respect to the background concentrations. This means that any correction on the estimated road contribution from PM_{10} (and also $PM_{2.5}$) will still have a very small effect on the total concentration. Therefore, only the results of the fusion model for NO_2 have been used for further data fusion analyses.

5.2.2 LUR Models Based on Low-Cost Sensor Data

Additional Materials 2 Figures S21 and S22 document the large variation of the annual average concentration of both pollutants at the 84 locations (from about 5 to 20 μg/m³ for $PM_{2.5}$ and 5 to 30 μg/m³ for NO_2). Figure 4 and Additional Materials 2 Figures S4 and S5 document that the variability in concentrations from the sensors was larger than that derived from the national monitoring network. There were only small differences in NO_2 concentrations measured in the low-cost sensors network between periods 1 and 2 (1.3 µg/m^3 higher concentrations in period 1 for NO_2). $PM_{2.5}$ concentrations differed more substantially (6.0 µg/m^3 higher concentrations in period 1 for $PM_{2.5}$) (Differences were corrected before modeling. The differences for $PM_{2.5}$ could be related to higher relative humidity and atmospheric pressure: mean relative humidity was 83% and 79% in periods 1 and 2, mean pressure (hPa) were 1,018 and 1,014 hPa, and mean temperature was 7.3 and 12.4°C.

For NO_2, models with reasonable performance were developed using the three different algorithms (**Table 6**). The model performance was similar for the three algorithms. The SLR model included three traffic-related variables (two near-roadway predictors for heavy-duty trucks and a 5,000-meter buffer), urban green in a 300-meter buffer, address density and port (both in a 5,000-meter buffer), and agriculture in a 1,000-meter buffer (Additional Materials 2 Table S1).

For $PM_{2.5}$, models with poor performance were developed in the three different algorithms (Table 6). The model performance was similar for the three algorithms. The SLR model included traffic load in a 50-meter buffer, urban green in a 5000-meter buffer, and agriculture in a 1000-meter buffer. Poor performance was also found when we offered Period as a predictor variable and used unadjusted concentrations.

Table 6. Performance Statistics of LUR Models Based on Low-Cost Sensor Measurement Locations Across the Netherlands

		SLR	Random Forest	Lasso
NO_2	Model R^2	0.53	0.38	0.39
	CV R^2	0.42	0.43	0.35
$PM_{2.5}$	Model R^2	0.13	0.00	—
	CV R^2	0.10	0.01	—

— = no model identified; CV = internal 10-fold cross-validation.

5.3 EVALUATION OF THE IMPACT OF TRENDS IN TIME OF AIR POLLUTION IN THE NETHERLANDS

Additional Materials 2 Figure S23 shows temporal trends of $PM_{2.5}$, NO_2, and BC concentrations based on routine monitoring within the NAQMN. Concentrations of NO_2, BC, and $PM_{2.5}$ have decreased substantially during 2010–2021. The downward trend is fairly continuous and not explained fully by the COVID-19 pandemic and associated reduced mobility (especially in 2020). Between 2010 and 2019, concentrations in urban areas have decreased by approximately 10, 5, and 1 µg/m^3 for NO_2, $PM_{2.5}$, and BC, respectively. The BC change is uncertain as the number of sites was small in 2010. BC concentrations have been reduced between 2016 and 2021 by 43% at traffic stations, 33% at urban background, and 29% at regional background stations (*https://www.clo.nl/indicatoren/nl0624-roet*). The spatial contrast in BC exposure has thus been reduced. For UFPs, routine monitoring data do not exist in the Netherlands.

The correlation of exposure estimates from different years at the 20,000 residential sites across the Netherlands is shown in **Figure 8** for dispersion modeling and Additional Materials 2 Figure S24 for the Europewide empirical model EXPANSE. With all models, we found a very high correlation between spatial contrasts in the 2010–2019 period. The correlation between years for $PM_{2.5}$ was lower than for BC and NO_2, for both models. For $PM_{2.5}$, the correlation decreased modestly with larger time intervals between models. We did observe a downward trend in pollution in this period, as evidenced by the orientation of the scatterplot relative to the 1:1 line. This high correlation implies that in the comparison of different methods, the year is not a critical factor for the correlation of exposure predictions from methods derived from different years. However, the predicted concentration level and the exposure contrast may differ.

5.4 COMPARISON OF EXPOSURE MODEL PREDICTIONS AT RESIDENTIAL SITES

5.4.1 Comparison of Exposure Model Predictions From Different Models

Scatterplots including the correlation of concentration predictions from different models at the 20,000 residential sites are shown in **Figure 9** (nationwide comparison), Additional Materials 2 Figure S25 (four major cities combined), and Additional Materials 2 Figure S26 (Amsterdam). For most mobile monitoring models (MUSIC, Exposomics, Google Airview), the application to the Netherlands exceeds the monitoring domain. In these figures, only the SLR algorithm is shown for conciseness. Additionally, for the European and dispersion models, only the year 2015 is shown because it's the year in the middle of the mobile monitoring campaigns. The predicted exposure distribution is shown in **Appendix Table S1** and Additional Materials 2 Table S4. Bland-Altman plots of the comparison for the three spatial scales are shown in **Figure 10** for selected comparisons and Additional Materials 2 Figure S27 for all comparisons. These plots also present the mean difference between the predictions of the two models for the three spatial scales. The closest year was chosen from dispersion and Europewide models when compared with the single-year mobile monitoring campaigns. We will interpret correlations below 0.4 as low, between 0.4 and 0.7 as moderate, between 0.7 and 0.8 as high, and above 0.8 as very high.

In general, predictions from the models were moderately to highly correlated in the nationwide comparison, except for $PM_{2.5}$. Correlations were, however, not so high, that identical

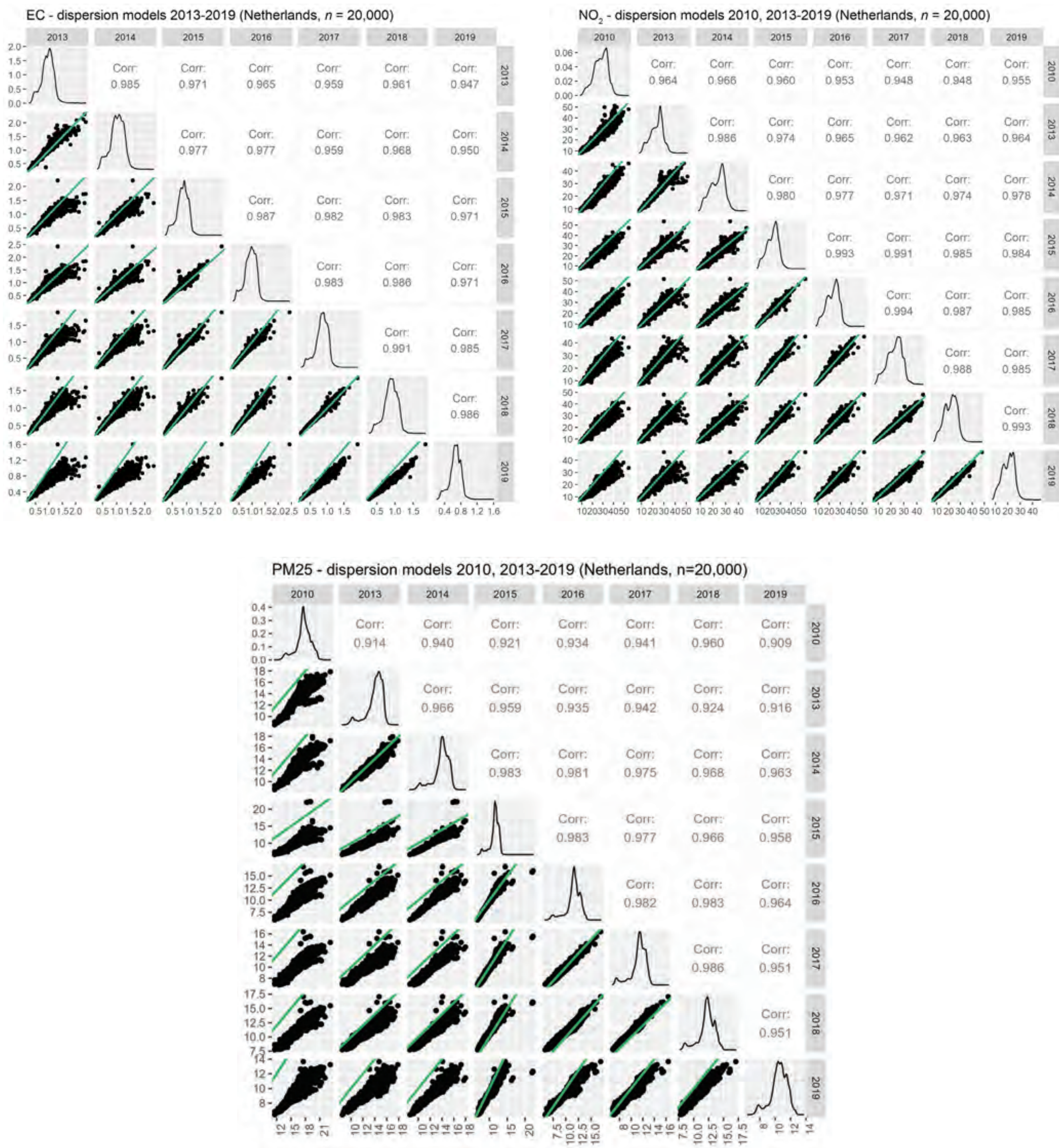


Figure 8. Correlation of exposure estimates of multiple years at the 20,000 residential sites from dispersion modeling for BC, NO_2, and $PM_{2.5}$. The green line is the 1:1 line. The scale of the y-axis of the top-left graph refers to the density plot (density times concentration difference of an interval is the probability of that interval), the remaining y-axes to absolute concentrations (in μg/m³).

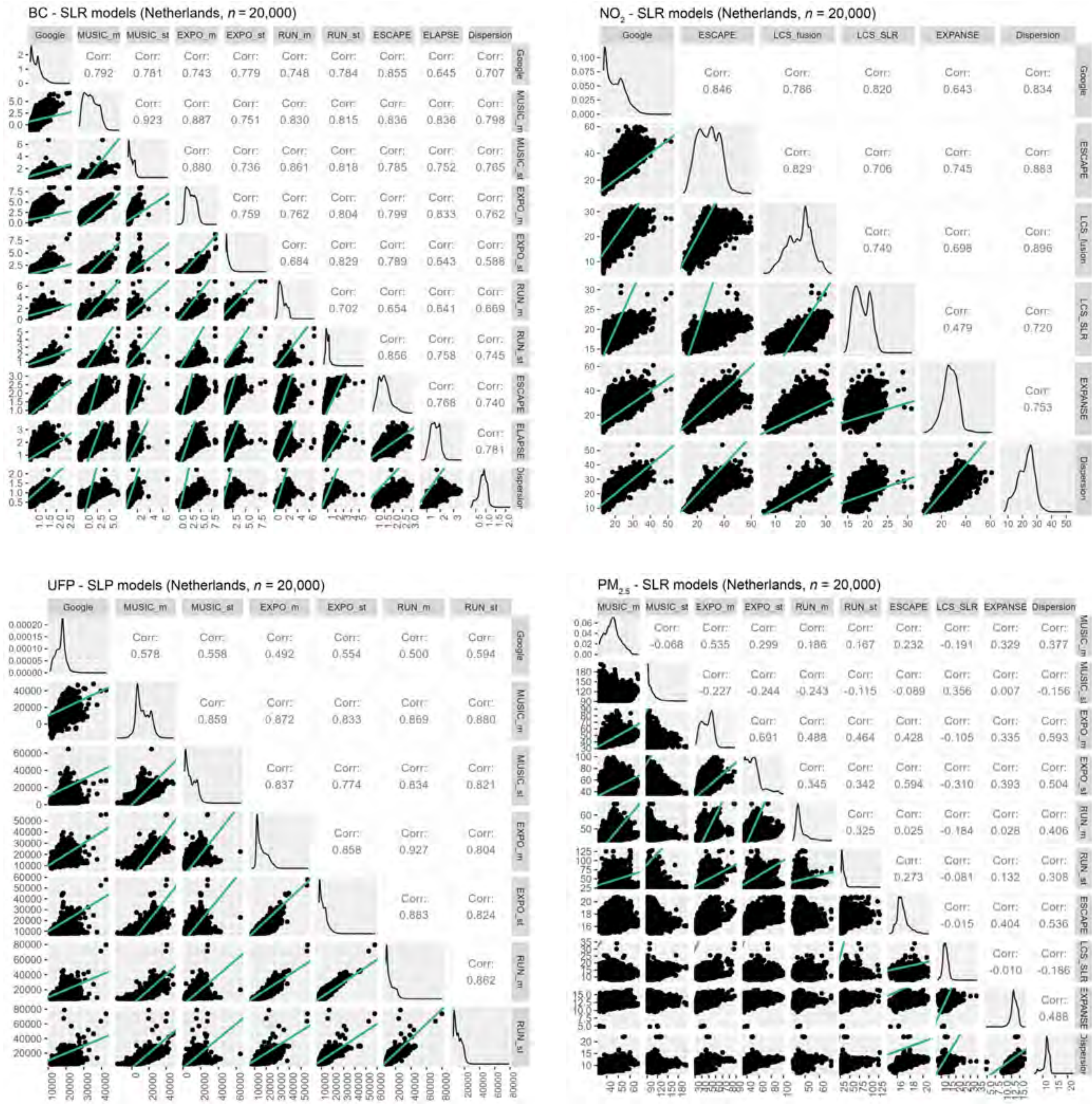


Figure 9. Correlation of exposure estimates from different models at 20,000 residential sites across the Netherlands for BC, NO_2, UFPs, and $PM_{2.5}$. The green line is the 1:1 line. The scale of the y-axis of the top-left graph refers to the density plot (density times concentration difference of an interval is the probability of that interval); the remaining y-axes to absolute concentrations (in µg/m³, except UFPs in p/cm³).

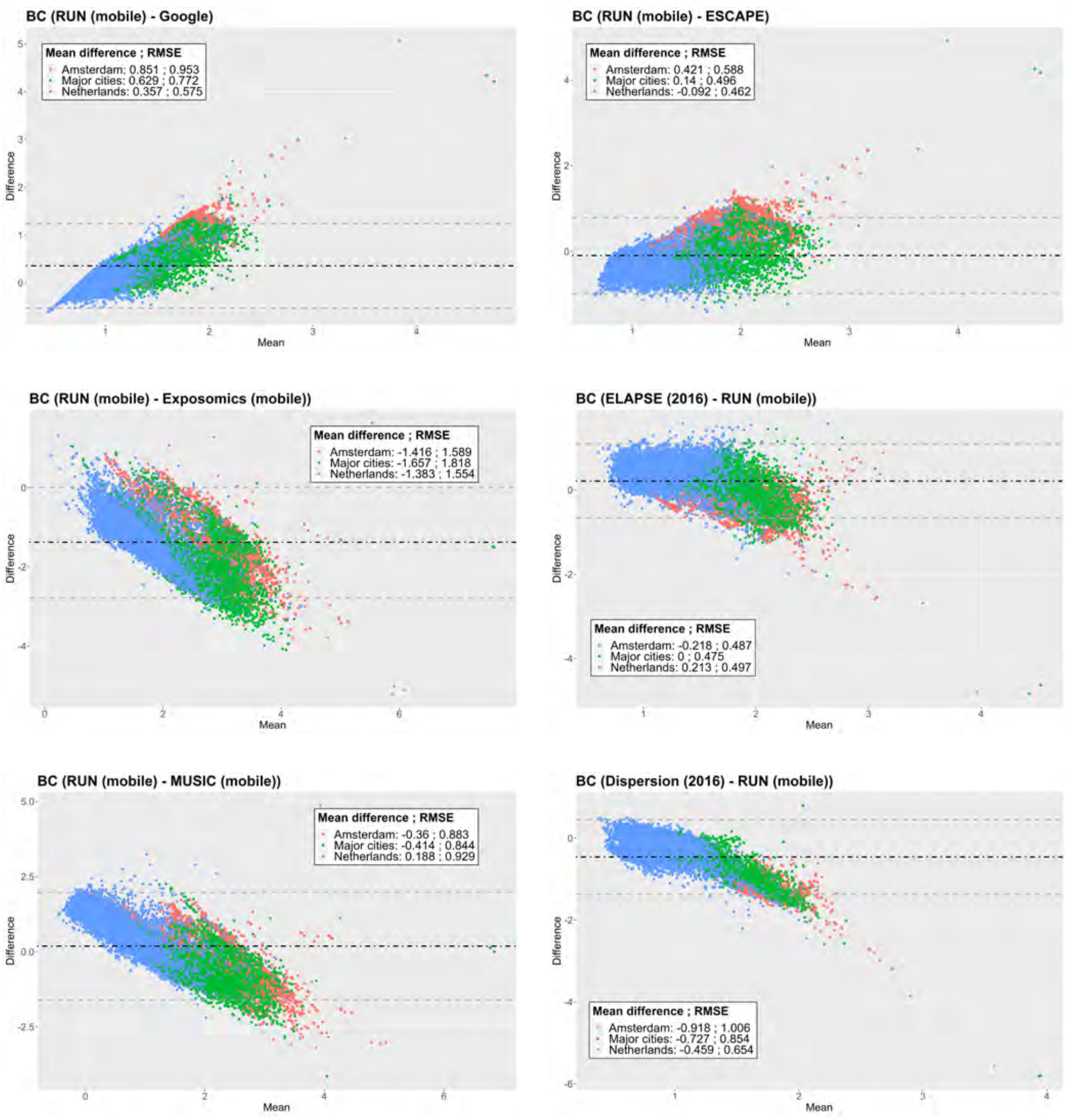


Figure 10. Bland-Altman plots of predicted annual average concentrations of exposure models at residential address across the Netherlands (blue), green (major cities), and Amsterdam (orange) for BC, $PM_{2.5}$, and NO_2 (µg/m^3). Lines represent mean difference and 95% confidence intervals for the Netherlands.

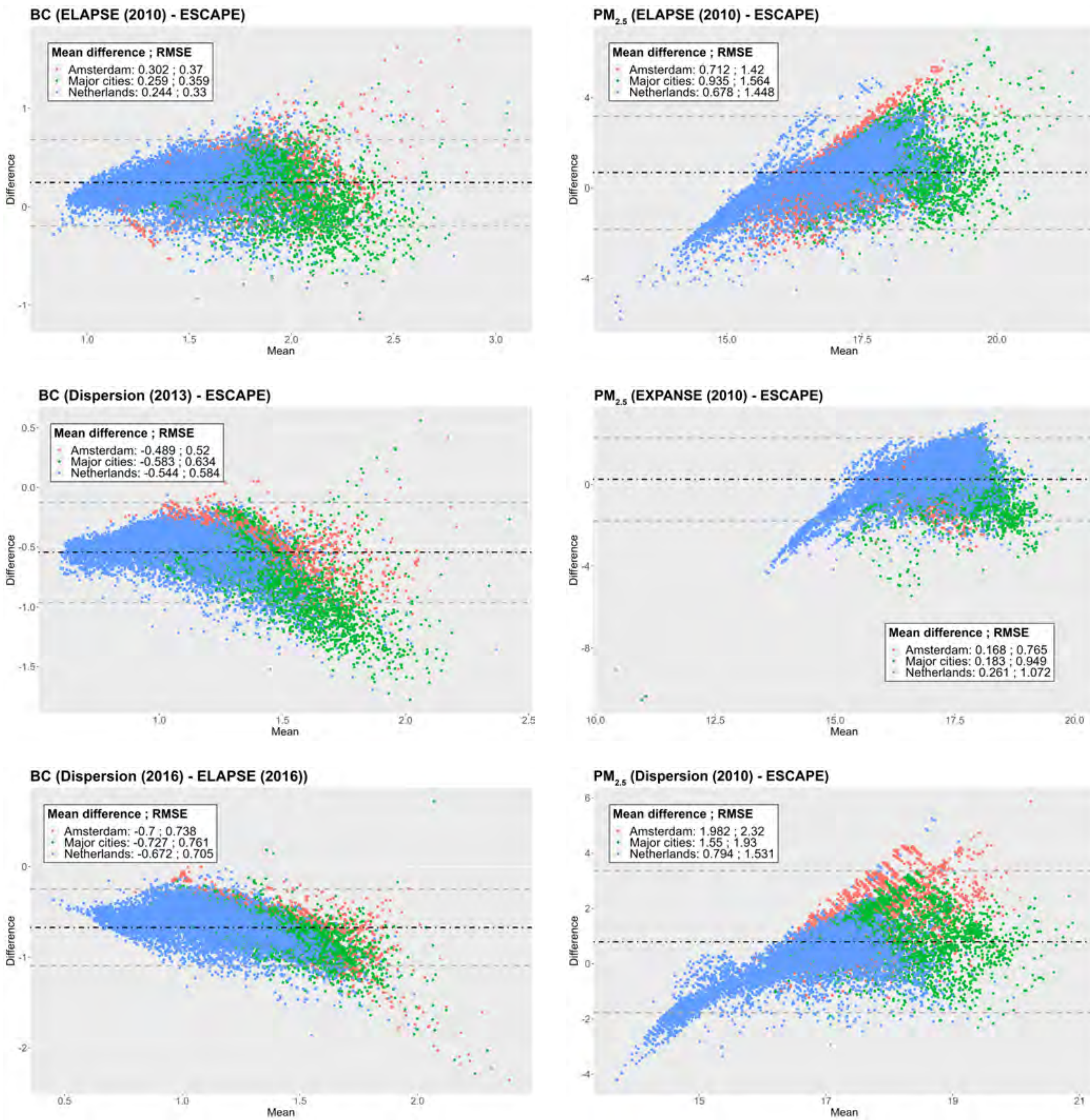


Figure 10. (*continued*)

continued on next page

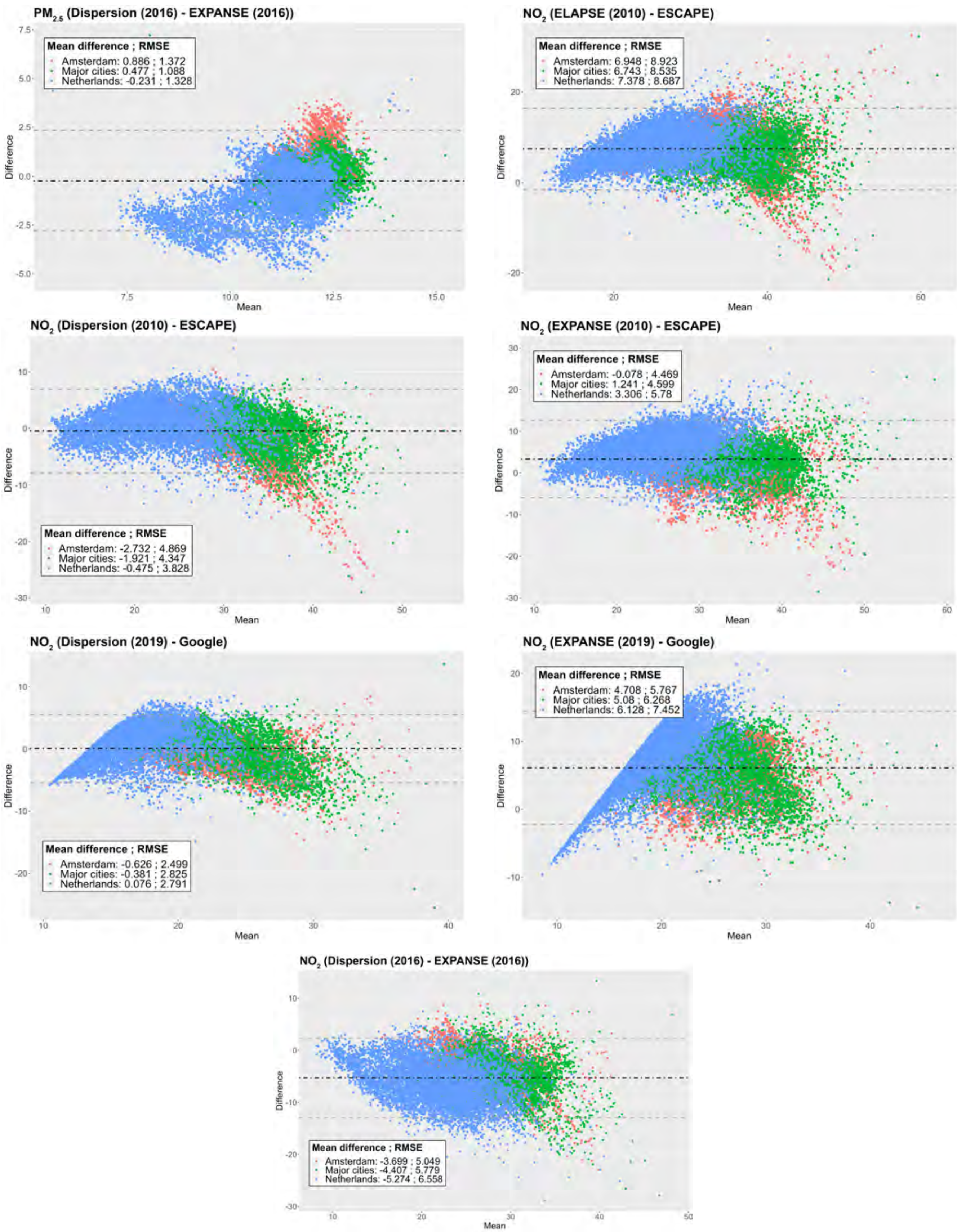


Figure 10. (continued)

results in the epidemiological analyses would be expected. Correlations were generally lower when the comparison was performed in the combined major cities or the Amsterdam area only. Sizable differences in predicted concentration levels and contrast were found between models and algorithms within models, that were larger than explained by differences in years of monitoring. Temporal trends have been documented in section 5.3 and are included in the dispersion model and European LUR models. A modest decrease in spatial contrast (IQR) was found between 2010 and 2019, based on the dispersion model and the Europewide hybrid LUR models. Bland-Altman plots showed that the difference between predictions of a sizable number of models differed substantially with the exposure level, resulting in varying mean differences across the Netherlands, four major cities, and Amsterdam.

For BC, correlations were generally between 0.7 and 0.8, with no pattern of higher or lower correlations between different types of models (mobile, fixed-site LUR, dispersion). The correlation was modestly lower in Amsterdam and especially the four major cities. The absolute value of the exposure estimates, however, varied substantially between models, as did the predicted exposure contrast (IQR). Bland-Altman plots documented that the difference between models varied over the range of exposure. The IQR for the SLR-based models ranged from 0.16 for the dispersion model for 2019 to 2.18 $\mu g/m^3$ for the MUSIC mobile monitoring campaign in 2013. In general, the mobile monitoring models predicted higher BC levels and contrasts at residential sites, probably related to the on-road monitoring. The dispersion model predicted lower concentrations for 2013 than the ESCAPE and ELAPSE models for 2010, also taking into account the difference in units (HEI 2022). The IQR of these three models was however comparable. The IQR from dispersion modeling decreased from about 0.30 to 0.16 $\mu g/m^3$ between 2013 and 2019. The IQR of short-term roadside models was smaller than from the corresponding mobile monitoring campaign and similar to the other exposure models, likely related to the larger distance from the road.

Most Bland-Altman plots for BC showed that differences depend strongly on the mean level, indicating systematic differences in model predictions not clearly apparent in the scatterplot. For example, the Exposomics model predicted higher BC levels than the Google Airview model at higher concentrations and lower levels at the lowest concentrations. This applied both to the Netherlands and to the city of Amsterdam. Most mobile models versus dispersion and fixed-site monitoring models followed this pattern for BC. Consequently, the difference between mobile monitoring and other models differed between the Netherlands versus major cities or Amsterdam domain. This pattern was least present for the RUN mobile campaign, with mobile monitoring across the country including smaller towns. Most Bland-Altman plots showed a wider scatter of differences at higher concentrations, such as the dispersion model versus ELAPSE and ESCAPE models for BC. Consistently, the difference between the dispersion model and empirical models was similar for the Netherlands, major cities, and Amsterdam domain. The RMSE was fairly large when comparing two mobile models (up to 1.8 $\mu g/m^3$ of BC), and smaller when comparing other models. RMSE was substantially larger than for $PM_{2.5}$ when compared to the mean exposure.

For NO_2, correlations were generally between 0.7 and 0.8, with no clear pattern of higher correlations between different types of models including one mobile, two fixed-site LUR, one low-cost network, and dispersion models. The Google Airview model developed in Amsterdam correlated well with other models developed in a larger domain. Correlations between predictions were moderately lower in the four major cities and Amsterdam, with correlations around 0.6. This also applied to models developed in Amsterdam, especially the Google Airview model. A moderate correlation of 0.48 was found between the Europewide EXPANSE model and the LUR model based on the new low-cost sensor campaign. The high correlation between dispersion models and low-cost sensor fusion largely results from using dispersion models in data fusion as well.

The difference in absolute levels of NO_2 also reflects the difference in the year, for example, while Google Airview and ESCAPE model predictions were highly correlated, the 2009 ESCAPE models predict consistently higher values compared to the 2019–2020 Google Airview data. The dispersion model and Europewide models both show a 7 $\mu g/m^3$ decrease in median NO_2 concentration between 2010 and 2019, albeit at higher levels for the Europewide model (Additional Materials 2, Table S4). This difference is nearly identical to the difference between the Google Airview and ESCAPE models. The IQR from the different SLR-based models ranged from about 3 to 13 $\mu g/m^3$. The largest contrast was found for the ESCAPE study, likely related to a larger historical contrast. Both the Europewide empirical and the dispersion model show a decrease in IQR of about 2 $\mu g/m^3$ between 2010 and 2019 (Appendix Table S1). Most Bland-Altman plots showed a fairly consistent difference across levels of concentration, certainly relative to BC. Consequently, the difference between mobile monitoring and other models differed only moderately between the Netherlands versus major cities or the Amsterdam domain. RMSE was lowest comparing the dispersion model with the Google Airview mobile monitoring model (up to 2.8 $\mu g/m^3$ of NO_2). The largest RMSE was found comparing ELAPSE and ESCAPE for the same year (up to 8.9 $\mu g/m^3$ of NO_2).

For UFPs, correlations were about 0.8 between three of the mobile monitoring campaigns and about 0.5 between the Google Airview campaign and the other three campaigns, perhaps because of the difference in monitoring equipment between Google Airview and the other three campaigns. The correlation was lower than for BC and NO_2. The contrast in predicted exposure differed widely between models; for example, the IQR for the SLR-based models ranged from 3,451 to 12,509 p/cm^3. The median exposure level differed moderately between models, ranging from about 8,000 to 17,000 p/cm^3 (Appendix Table S1). In general, exposure levels and contrasts were modestly larger for the models based on mobile (on-road) monitoring compared to short-term roadside monitoring, as reported previously for the major city mobile monitoring campaigns MUSIC and Exposomics. This difference was smaller for the national mobile monitoring campaign RUN. Like BC, most Bland-Altman plots showed that differences depend strongly on the mean level; for example, the MUSIC model predicted higher UFP levels than the Google Airview model at higher concentrations and lower levels at the lowest concentrations. Consequently, the difference between mobile monitoring and other models differed between the Netherlands versus major cities or Amsterdam domain. Most Bland-Altman plots showed a wider scatter of differences at higher concentrations.

For $PM_{2.5}$, correlations were generally low to essentially zero between mobile monitoring-based models and the other models. As documented in section 5.4.1, models based on mobile monitoring had poor performance. The mobile monitoring-based models resulted in unrealistically high predicted concentrations, likely related to the instrument used (DustTrak). Moderate correlations of about 0.4 to 0.5 were found between the ESCAPE, EXPANSE, and dispersion modeling methods. In Amsterdam, correlations were lower than for the Netherlands as a whole.

The dispersion model and EXPANSE Europewide model agree very well in predicted median concentration and IQR. Both models show a decrease in IQR between 2010 and 2019: from 1.7 to 1.3 µg/m^3 (dispersion) and 2.0 to 0.9 µg/m^3 (Europewide model). The low-cost sensor model resulted in slightly higher median concentrations compared to the dispersion model and the EXPANSE model (2019), with a larger predicted contrast (IQR 2.3 µg/m^3) (Appendix Table S1). The predictions of the low-cost sensor model were essentially uncorrelated with the other exposure models. Most Bland-Altman plots showed a fairly consistent difference across concentration levels, certainly relative to BC. Consequently, the difference between mobile monitoring and other models differed between the Netherlands versus major cities or Amsterdam domain. Exceptions were the comparison with mobile monitoring models, which predicted higher concentrations at the highest exposed sites than the fixed-site and dispersion models. RMSEs of comparisons between dispersion and fixed site-LUR models were fairly small at 1–2 µg/m^3, which were much smaller than for BC when compared to the mean exposure.

5.4.2 Comparison of Exposure Predictions of Three Algorithms Within Empirical Models

In general, moderate to high correlations were found between different algorithms within the same model (Additional Materials 2 Figure S28). However, exposure contrasts often differed substantially between algorithms (Appendix Table S1), especially for the MUSIC mobile monitoring campaign. Differences in level and contrasts for the models based on routine monitoring (ESCAPE, ELAPSE/EXPANSE) were small.

For BC, correlations were high to very high (R > 0.8) between SLR and LASSO, and moderate with Random Forest models. The correlation was generally fairly similar for the entire Netherlands compared to Amsterdam or four major cities. The correlation between algorithms differed substantially between campaigns, with the highest correlation found for the ESCAPE campaign. The IQR was generally substantially larger for the models developed with SLR than for Random Forest or LASSO, especially for the mobile monitoring models. The IQR for the ESCAPE models was only slightly larger for the SLR algorithm. Median predicted exposure levels were generally similar across algorithms, except for the MUSIC mobile models.

For NO_2, correlations between algorithm predictions within models were higher for the EXPANSE and ESCAPE models and lower for the Google Airview model. This was probably because the Google Airview model was based on Amsterdam data only. Indeed, correlations are higher when focusing the comparison on the city of Amsterdam for the Google Airview model. The correlation between LASSO and SLR was generally the highest.

The IQR was substantially larger for the SLR models compared to the Random Forest and LASSO models for the single mobile monitoring campaign (Google Airview) and the new low-cost sensor monitoring campaign. For the models based on fixed-site monitoring, smaller differences were typically found across algorithms.

For UFPs, correlations were high to very high (R > 0.8) between SLR and LASSO and moderate with Random Forest models. In general, correlations were lower than for BC. The correlation was generally higher for the entire Netherlands compared to Amsterdam and four major cities, except for the nationwide RUN campaign. The correlation between algorithms differed substantially between campaigns, with the highest correlation found for Exposomics. The IQR was generally substantially larger for the models developed with SLR than for Random Forest or LASSO.

For $PM_{2.5}$, correlations between predictions from the three algorithms from mobile monitoring models were generally low. For the ESCAPE campaign, correlations were moderately high. For the EXPANSE model, correlations were low (Amsterdam) to moderate (Netherlands) between SLR and Random Forest.

5.4.3 Correlation Between Pollutants for Dispersion and Empirical Models

Additional Materials 2 Figure S29 presents the correlation of predicted exposures of different pollutants at residential addresses across the Netherlands. The correlation between the different pollutants differs quite substantially between exposure models. For example, the correlation between $PM_{2.5}$ and NO_2 ranges from 0.85 (dispersion for the year 2016) to 0.42 (EXPANSE 2016). When comparing more or less the same years, correlations between $PM_{2.5}$ and NO_2 are sizably larger for dispersion modeling (2016) compared to the Europewide EXPANSE model (2016) and the nationwide ESCAPE model (2009–2010). Correlations between NO_2 and BC were about 0.9 for dispersion modeling and 0.8 for the European and nationwide LUR models. Correlations between BC and $PM_{2.5}$ were also higher for dispersion modeling. High correlations limit disentangling the health effects of different pollutants. UFPs were highly correlated with BC, especially from models of the RUN, MUSIC, and Exposomics campaigns and less so for the Google Airview campaign.

Correlations between *measured* annual average concentrations were 0.84 (BC–$PM_{2.5}$), 0.93 (BC–NO_2), and 0.75 ($PM_{2.5}$–NO_2) in the ESCAPE measurement campaign of 2009–2010 across the Netherlands (Eeftens et al. 2012). Correlations were 0.85 (UFPs–BC) and 0.66 (UFPs–$PM_{2.5}$) for the RUPIOH campaign in 2002–2004 in Amsterdam (Hoek et al. 2011).

5.5 COMPARISON OF MODELED EXPOSURES WITH MEASUREMENTS AT NEW EXTERNAL VALIDATION SITES

Weeklong measurements in outdoor air have been made between September 2021 and April 2023 at 90 locations across the Netherlands. Because of the COVID-19 pandemic, the start of the monitoring has been delayed by about 9 months, as researchers were not allowed to conduct the required home visits. Furthermore, there were concerns about the representativeness of measurements during the COVID-19 pandemic and associated lockdowns. Because of the delays, we were unable to complete the planned two separate weeklong measurements for all 100 homes. We agreed with HEI to limit monitoring to 100 homes, of which 20 homes would be measured twice. We finally measured 90 homes, and in 31 of these homes, measurements were made in two different weeks. Because of the long delay, several people who agreed to participate at the end of 2020 or early 2021 no longer wanted to participate in 2022. In each measurement week, measurements were made at three homes and a continuous reference site near Utrecht, in the middle of the country. Additional Materials 1 Figure S4 shows the distribution over the country.

5.5.1 Quality Assessment and Control

Additional Materials 2 Table S5 lists the number of successful measurements at the homes and reference site. Additional Materials 2 Table S6 presents the limit of detection based on field blanks and the precision based on field duplicates of the $PM_{2.5}$, $PM_{2.5}$-absorbance with the Harvard Impactor, and NO_2 measurements with passive samplers. Limits of detection were low. All samples were above the limit of detection for $PM_{2.5}$ and $PM_{2.5}$ absorbance. For NO_2, all but two samples were above the limit of detection. The precision for all measurements was good and comparable with previous projects (Cyrys et al. 2012; Eeftens et al. 2012).

A fairly large number of real-time measurements of UFPs and BC were missing at the reference site, because of malfunctioning equipment. In contrast, the passive sampler NO_2 and integrated $PM_{2.5}$ sampler data were nearly complete. If there were insufficient properly functioning BC and UFP real-time monitors, we prioritized measurements at the home locations and dropped the reference site measurements. The rationale was that based on previous projects (e.g., ESCAPE), we anticipated that we could impute missing values at the reference site using data from the routine network. Additional Materials 2 Table S7 presents the imputation of missing data for BC and UFPs, using network data (one regional background site) for BC and the $PM_{2.5}$ and $PM_{2.5}$ absorbance measurements at the reference site for UFPs. The BC measurements at our reference site were well explained by the BC measured with a standard monitor (AE33) at a regional background site (Cabauw) about 30 km from our site: R^2 was 0.75. No external UFP measurements were available for the full period to impute data. We imputed missing UFP values by regressing UFP data at our site on $PM_{2.5}$ and $PM_{2.5}$ absorbance from integrated measurements (after filling in the two missing values for $PM_{2.5}$ and $PM_{2.5}$ absorbance with $PM_{2.5}$ and BC data from Cabauw, R^2 = 0.91 and 0.89 respectively), resulting in a moderate R^2 of 0.55. The RMSE was small (906 p/cm^3), suggesting UFPs can be assessed with good precision. The imputation equations for BC, $PM_{2.5}$, and $PM_{2.5}$ absorbance showed an intercept close to zero and a slope close to 1.

Because of malfunctioning of instruments, requiring factory repair, we could not conduct the full campaign with a fixed set of BC and UFP monitors. In total, we used nine different UFPs and seven different BC monitors. Additional Materials 2 Table S8 shows the results of the regular intercomparisons of all real-time BC and UFP measurement instruments we performed. Most intercomparisons showed fairly close agreement in measured levels of the different instruments, with a few exceptions. We used the average ratio of the geometric means of the current and previous intercomparison between instruments to scale the readings of the different instruments. We used the ratio of two previous or two following intercomparisons if no current or previous intercomparison was available (in case of a malfunctioning instrument preventing testing).

Applying the quality check procedures of our standard operating procedure, missing data in measurements, and deletion of measurement weeks with less than 75% of valid data, left us with 75% and 85% of usable data points for UFPs and BC respectively (i.e., usable second and minute measurements) of the total number of expected data points (i.e., number of seconds and minutes in a week times the number of weekly measurements for which equipment was set out [162 for UFPs, 157 for BC, including reference site measurements]). We conclude that the real-time UFP and BC monitors were less robust than we anticipated, based on previous experiences. The passive and integrated samplers used to measure $PM_{2.5}$, $PM_{2.5}$ absorbance, and NO_2 were robust.

Intercomparison with routine monitoring on 9 days showed that our NO_2 measurements corresponded very well with reference measurements (R^2 = 0.96 for both the regional background and traffic location with a mean difference of –0.60 and –0.75 µg/m³ respectively). Our $PM_{2.5}$ measurements on 8 days corresponded moderately well (R^2 = 0.54 at the regional background location with a mean difference of 1.86 µg/m³) (Additional Materials 2 Figure S30). The small number of measurements may have affected this comparison, as the R^2 when comparing with a reference $PM_{2.5}$ monitor at a site about 20 km from our reference site was 0.91 (Additional Materials 2 Table S7, n = 39 days).

5.5.2 Concentration Distributions

Weekly average concentrations at the reference site in the full study period are shown in Additional Materials 2 Figure S31, before and after imputation. Large temporal variation was found for all four measured pollutants. The largest variation was observed for BC (from about 0.25 to 2 µg/m³), NO_2 (from about 4 to 30 µg/m³), and $PM_{2.5}$ (from about 5 to 25 µg/m³). Variation was more moderate for UFPs (from ~3,000 to 8,000 p/cm³). This temporal variation complicates the comparison between sites measured in different periods. The measurements at the reference site were used to correct the measurements at the homes for temporal variation.

Figure 11 shows the distribution of the weekly average measurements, stratified by site type, after correction for

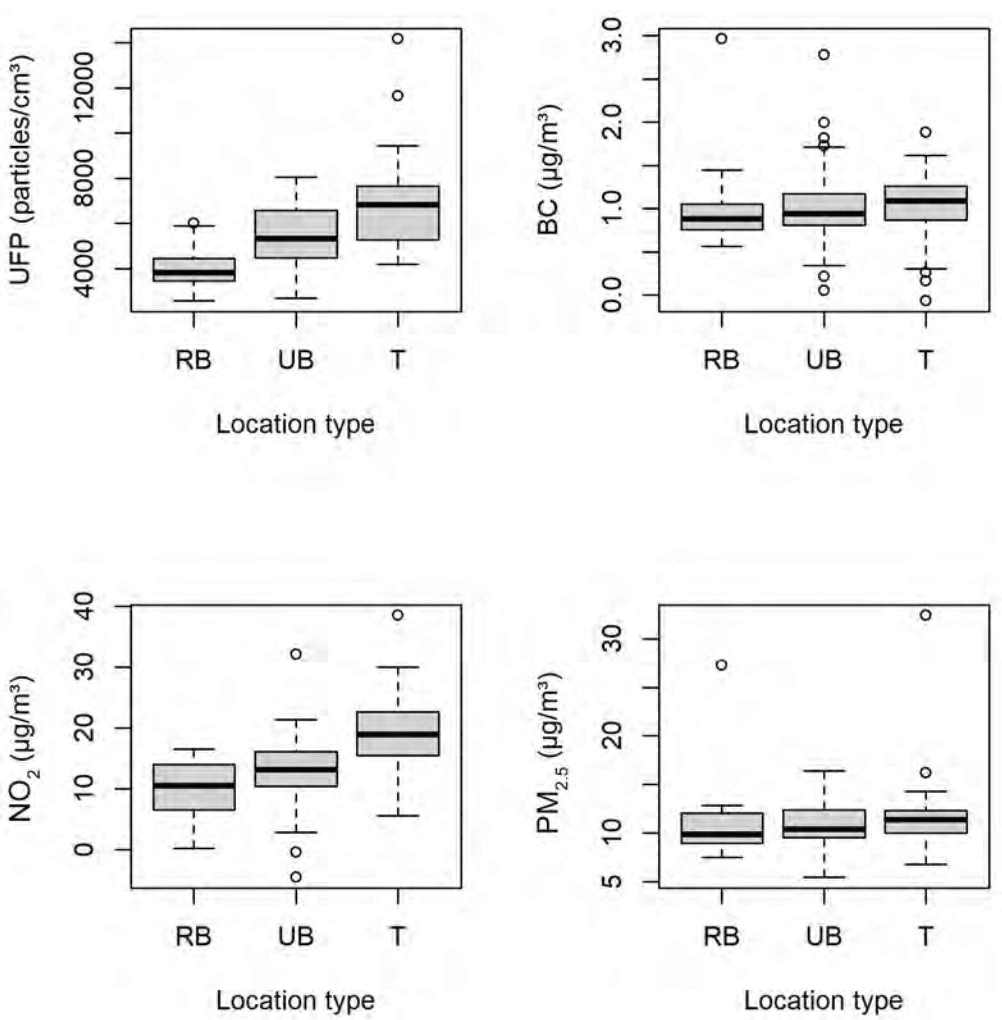


Figure 11. Distribution of average measurements at 90 new external validation sites, adjusted for temporal variation, stratified by site type. RB = regional background, T = traffic, UB = urban background.

temporal variation. Additional Materials 2 Table S9 and Figure S32 show the statistics and the distribution before temporal correction. For NO_2 and UFPs, we observed the expected increase in concentrations from regional background to urban background and traffic sites. For $PM_{2.5}$, the small contrast across site types was also anticipated. For BC, we found less contrast than we anticipated. Additional Materials 2 Figure S33 documents that the concentration patterns were similar for the independent $PM_{2.5}$ absorbance measurements. Furthermore, the two BC measurements were highly correlated (R = 0.78). Because we furthermore observed a very high correlation between our BC measurements at our reference site and BC measurements using a more established BC monitoring instrument (AE33 aethalometer) at a routine regional background site, we conclude that the BC measurements were of sufficient quality to further use as validation data.

Concentration levels of BC were mostly consistent with monitoring from more established monitors at other locations in the Netherlands, see sections 5.3 and 5.4. UFP concentrations were substantially lower than measured in other mostly mobile and short-term monitoring campaigns, however, there is considerable uncertainty on UFP concentration levels, related to differences in monitoring domains (on-road or off-road), instrumentation (lower particle size monitored), and time period.

5.5.3 Stability of the Concentration Estimate

As most sites (59 of 90) only have a single measurement week available, we tested the stability of a single estimate by correlating the first and second measurements for the 31 sites with two measurement rounds (Additional Materials 2 Figure S34). We found a low ($PM_{2.5}$, BC) to moderate (UFPs, NO_2) correlation between the first and second measurements. We found a slight or moderate improvement (NO_2) after temporal correction. The low correlation raises concern about the stability of the spatial contrast across locations to represent a contrast in average concentration. We, therefore, compare spatial model predictions with single measurements at all 90 sites and the average of two measurements at 31 sites. We only interpret the UFP and NO_2 results, the two pollutants with a moderate correlation between the first and second measurements. The data for BC and $PM_{2.5}$ are not informative (comparison data are shown in the Additional Materials section 5.5 for transparency).

The variability between sites was smaller when we assessed the average of two measurements compared to a single time-adjusted concentration (Additional Materials Figure S35). The variability was larger than expected based on spatial contrasts in monitoring campaigns with more temporally stable estimates including ESCAPE (Eeftens 2012) and the NAQMN. This further supports that the measured contrast is affected by both spatial and (insufficiently adjusted for) temporal variability. Finally, we anticipate that the remaining temporal variability in the validation data will substantially lower the explained variance when comparing modeled and measured concentrations because the models are spatial models. We expect the impact on bias would be less, as we included different site types in each monitoring session, suggesting that the average spatial difference is not confounded by temporal variation. We therefore start with discussing bias and then the explained variance.

5.5.4 Comparison of Model Predictions with Measurements

Figure 12 and **Table 7** show the comparison of model predictions with the new external validation measurements for all 90 sites and the 31 sites with two measurements for UFPs and NO_2 (SLR models only). Additional Materials 2 Figure S35 shows scatterplots for UFPs and NO_2 for all models and algorithms. Additional Materials 2 Table S10 shows the comparison for BC and $PM_{2.5}$, for which we only cautiously interpret the bias. We do not further interpret explained variance because of large random errors per site.

For UFPs and NO_2, all models predicted substantial variation of concentrations between sites, typically a two- to three-fold range. We found generally a low to moderate explained variance of measured concentration patterns for all models (R^2 between 16% and 47%), with mostly modest differences between models and algorithms within models. Given the modest number of sites, we only interpreted differences in R^2 of more than 10% for both the single and the two measurement sites as potentially meaningful.

In general, the explained variance was higher for the 31 sites for which the average of two measurements was available. The moderate explained variance is at least partly related to the limited temporal stability of the measured average. An additional post-hoc analysis where we grouped observations of Table 7 in sets of five measurements supports this interpretation. Specifically, we sorted the observations at the 90 sites based on the modeled exposure and then calculated the average modeled and measured exposure in groups of five ordered observations. We then calculated the agreement between average modeled and average measured exposures. The explained variance for UFPs for mobile models was between 50% and 70%, much higher than for the comparison of individual site average measurements and modeling reported in Table 7. For BC, we also found an increase in R^2 from essentially null to about 10%, indicating that averaging was not sufficient to address the larger temporal variability.

We found substantial differences in the level of predicted concentrations between modeled and measured concentrations, which for NO_2 were primarily related to the year of monitoring. The variability was, however, substantially less than for the measurements, especially for the sites with only a single measurement.

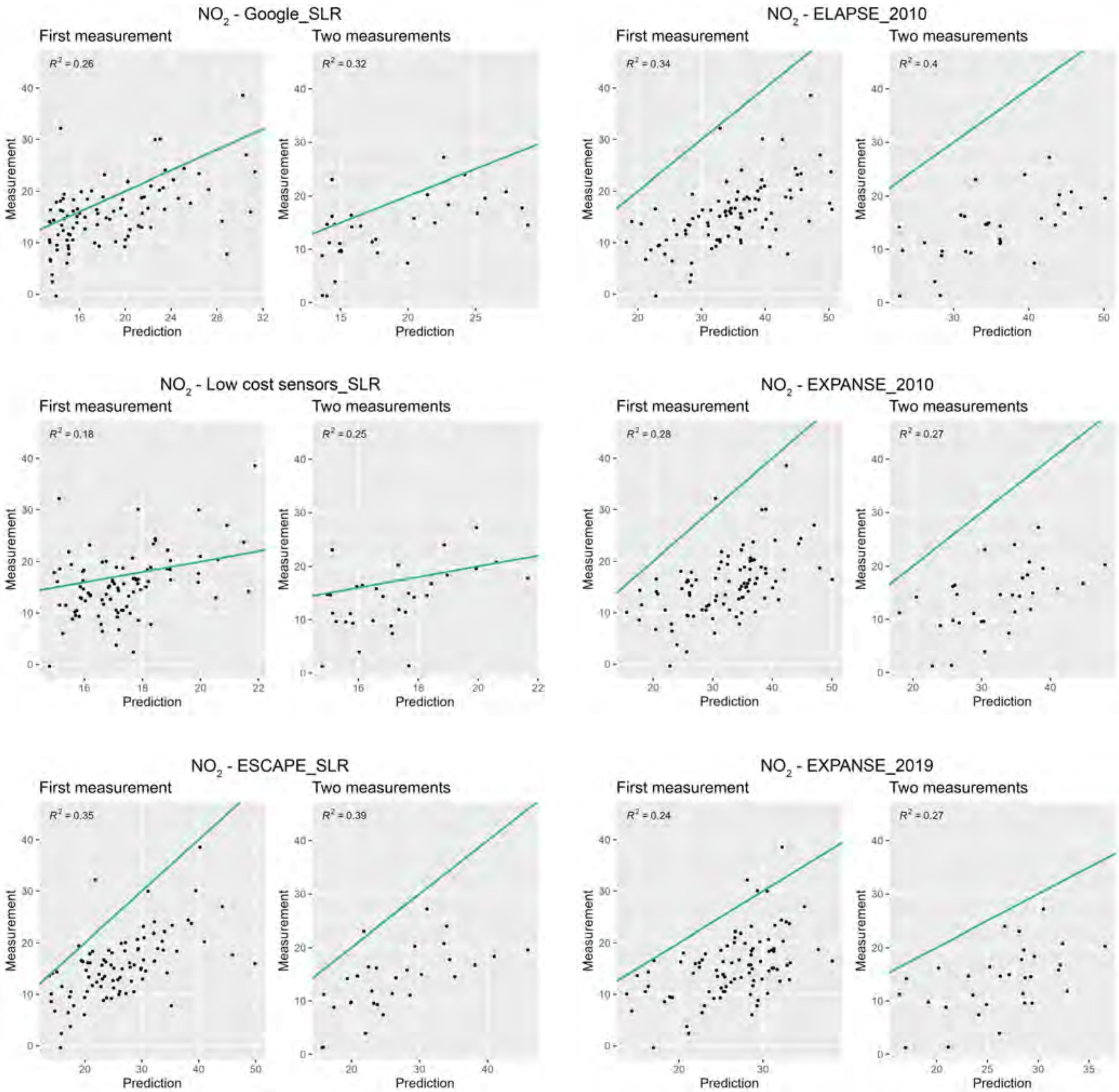


Figure 12. Comparison of exposure predictions from different models with weekly average measurements adjusted for temporal variation at the new external validation sites for NO_2 and UFPs. SLR models only, selected years from dispersion and European models. NO_2 in µg/m³, UFPs in p/cm³. The green line is the 1:1 line.

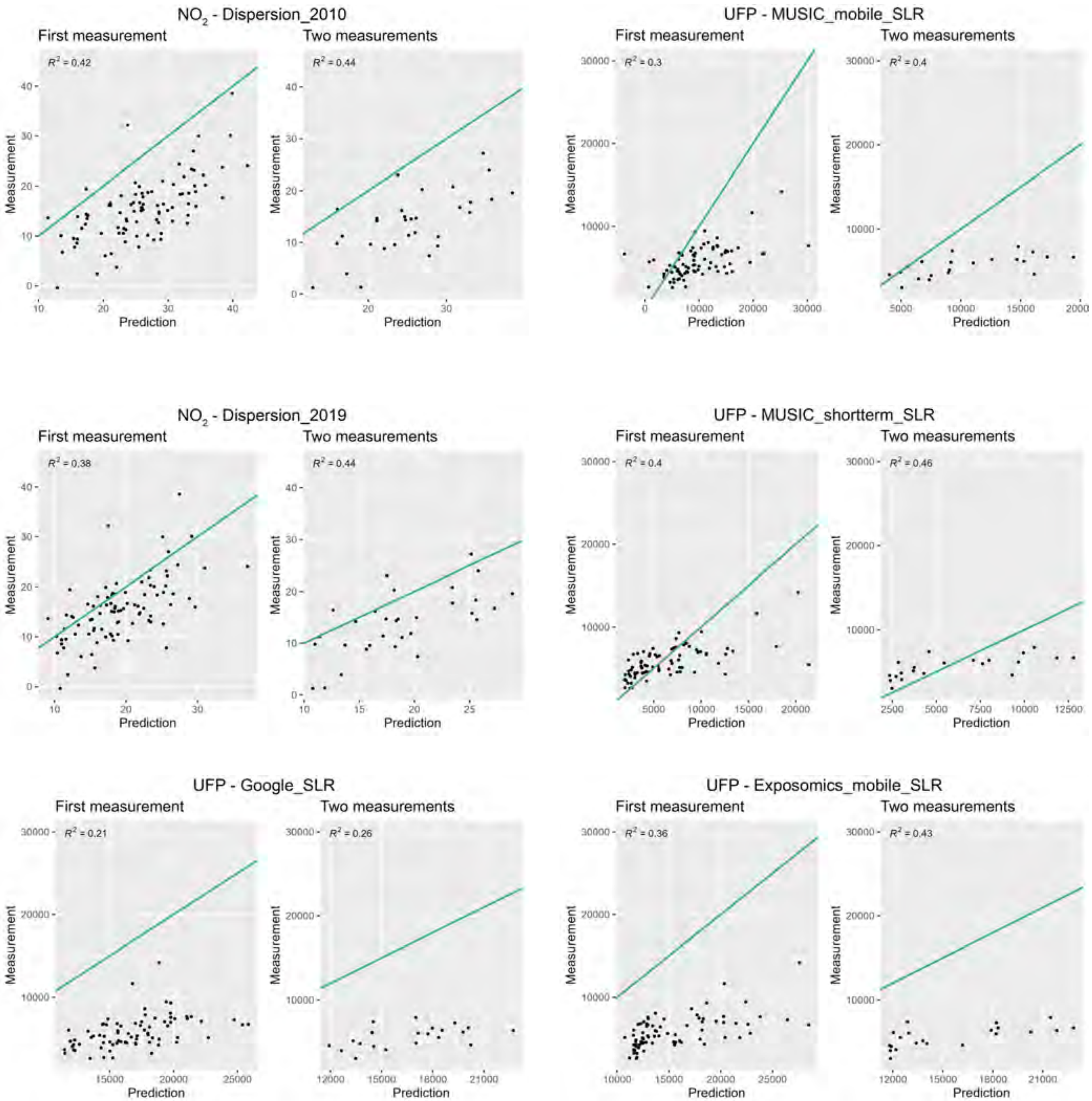


Figure 12. (*continued*)

continued on next page

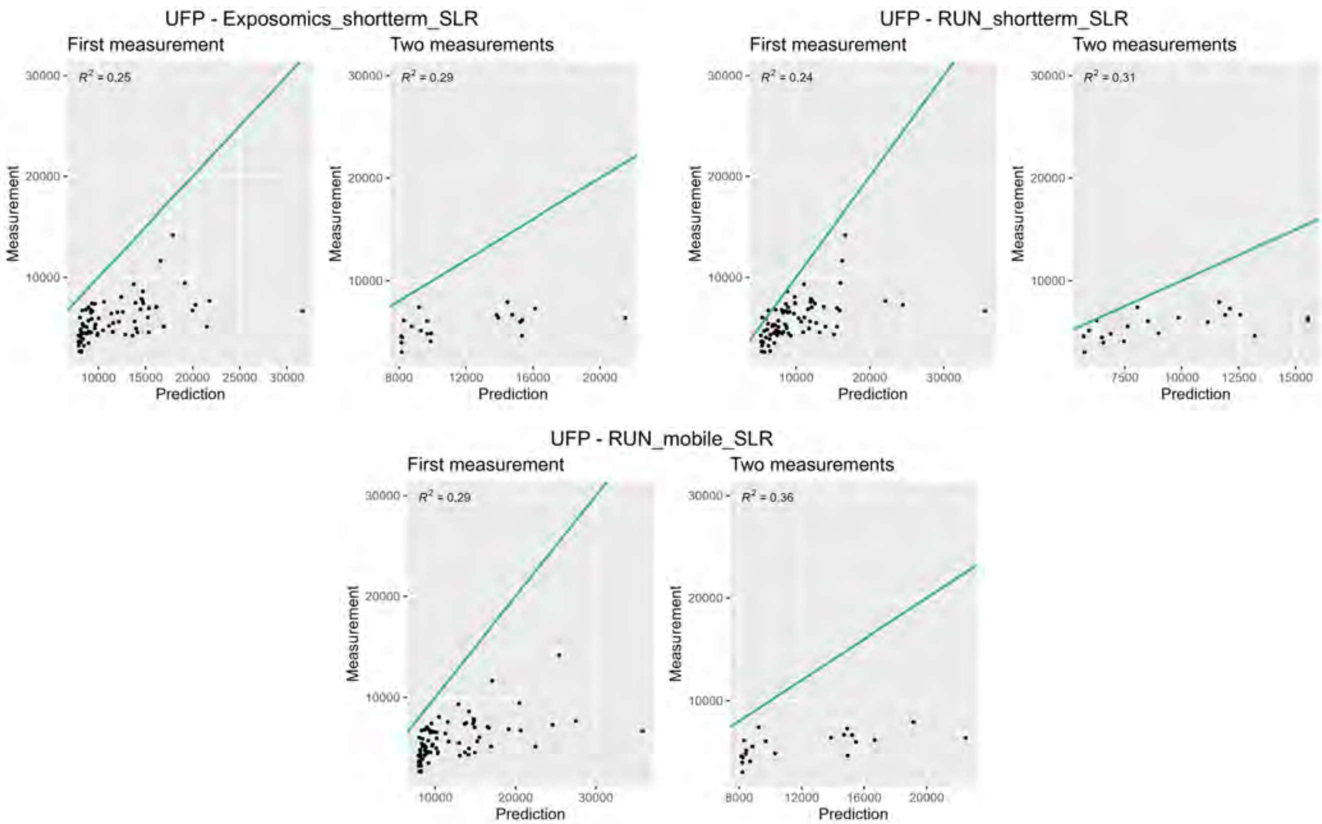


Figure 12. (*continued*)

For UFPs, all models had a substantial positive bias. All models predicted substantially higher UFP concentrations than measured, both for the recent (2019) Google Airview model and the oldest model (MUSIC 2013). The models based on short-term roadside monitoring overpredicted less than the models based on mobile (on-road) monitoring. The bias of the mobile models is likely explained by the on-road measurements that were used to train the model. Differences in instrumentation may have played a role as well. The measurements of the two repeats varied between 3,000 and 8,000 p/cm^3. On an absolute scale, all models predicted larger variability. On a relative scale, the predicted contrast was modestly larger than measured. The MUSIC mobile model had the highest R^2 and the Google Airview model had the lowest R^2. We did not find evidence that the nationwide RUN mobile monitoring campaign performed substantially better than the major city-based MUSIC, Exposomics, and Google Airview models. While for individual models, there were modest differences in performance between algorithms, there was no consistent pattern of one algorithm performing better. For MUSIC, the short-term model had a similar performance to the mobile model. For Exposomics the mobile model had better performance than the short-term model, but for RUN the opposite was observed. The R^2 of the Google Airview model was similar to the R^2 for the Google Airview NO_2 model. The explained variance (R^2) obtained for the average of five ordered modeled and measured observations was 0.69, 0.58, 0.63, and 0.61 for the mobile models MUSIC, Google Airview, Exposomics, and RUN developed with the SLR algorithm, respectively. Explained variance was much higher than for individual observations (Table 7).

For NO_2, all models had a positive bias. For the most recent models (Google Airview, low-cost sensors, and dispersion 2019), the bias was relatively small. For the European EXPANSE and dispersion models, the bias decreased from 2010 to 2019, reflecting the decreasing concentration. The measurements of two repeats varied between 2 and 28 μg m^3. Most models predicted a smaller or similar range, especially the low-cost sensor model. The highest R^2 was found for dispersion models, but the differences were modest. Differences in performance between algorithms were small and not consistent. We found very similar explained variance between exposure models from different years for both the dispersion model estimates and the Europewide models. The more recent models (Google Airview, low-cost sensors LUR, and dispersion model 2019) predicted modestly higher concentrations compared to the measurements. The other models, including the dispersion model for 2010, predicted substantially higher concentrations than measured.

For BC, low bias was found for the more recent models including the Google Airview and RUN mobile models and the dispersion model for 2019. For $PM_{2.5}$, low bias was found for the dispersion model and the European EXPANSE model of recent years (Additional Materials 2 Table S10).

Table 7. Root Mean Squared Error (RMSE), Bias, and Explained Variance (R^2) of Different Exposure Models of Annual Average Concentrations Measured at the New Validation Sites (N = 90 with at Least One Measurement; N = 31 with Two Measurements) for NO_2 in µg/m³ and UFPs in p/cm³

		NO_2						UFPs					
		First Measurement			Two Measurements			First Measurement			Two Measurements		
Model	Algorithm/ Year	RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2
Google	SLR	6.37	3.04	0.26	7.01	4.74	0.32	11,445	11,073	0.21	11,318	11,044	0.26
	RF	9.81	7.82	0.22	11.07	9.51	0.24	12,799	12,630	0.28	13,059	12,922	0.31
	Lasso	6.31	1.25	0.19	6.33	3.51	0.31	12,114	11,896	0.17	12,055	11,930	0.20
MUSIC (mobile)	SLR							6,405	4,209	0.30	6,278	4,932	0.40
	RF							9,801	8,769	0.27	9,854	9,059	0.23
	Lasso							9,363	9,017	0.41	9,511	9,208	0.47
MUSIC (short-term)	SLR							3,385	661	0.40	2,748	460	0.46
	RF							5,734	5,072	0.32	6,323	5,699	0.33
Exposomics (mobile)	SLR							10,155	9,570	0.36	10,391	9,934	0.43
	RF							11,462	10,300	0.22	11,627	10,680	0.35
	Lasso							10,505	10,208	0.38	10,572	10,328	0.45
Exposomics (short-term)	SLR							7,058	6,000	0.25	6,952	6,196	0.29
	RF							6,911	6,208	0.26	7,110	6,529	0.28
	Lasso							7,098	6,711	0.20	7,132	6,905	0.24
RUN (mobile)	SLR							7,782	6,342	0.29	7,332	6,403	0.36
	RF							8,008	7,381	0.30	7,712	7,326	0.24
	Lasso							8,284	7,708	0.25	8,280	7,892	0.28
RUN (short-term)	SLR							5,936	4,181	0.24	4,629	3,806	0.31
	RF							5,209	4,379	0.30	5,152	4,396	0.16
	Lasso							5,217	4,889	0.35	5,164	5,038	0.44
Low-cost sensors LUR	SLR	6.06	1.80	0.18	6.30	3.28	0.25						
	RF	5.74	0.46	0.19	5.85	2.13	0.26						
	Lasso	5.96	0.61	0.23	6.10	2.17	0.26						
ESCAPE	SLR	12.48	10.73	0.35	13.74	12.35	0.39						
	RF	13.98	12.25	0.31	15.42	13.92	0.32						
	Lasso	11.70	10.01	0.31	12.76	11.43	0.33						
ELAPSE	2010	20.10	19.06	0.34	22.35	21.55	0.40						
EXPANSE	SLR_2010	17.84	16.59	0.28	19.24	18.17	0.27						
	SLR_2013	16.41	15.14	0.22	17.75	16.59	0.17						
	SLR_2016	14.74	13.38	0.23	16.20	14.94	0.18						
	SLR_2019	12.48	11.00	0.24	13.57	12.41	0.27						
	RF_2010	16.16	14.41	0.29	17.83	16.21	0.34						
	RF_2013	15.98	14.14	0.28	17.73	15.94	0.30						
	RF_2016	13.10	10.93	0.27	14.80	12.85	0.30						
	RF_2019	9.93	7.46	0.27	11.41	9.24	0.31						
Dispersion	2010	11.87	10.54	0.42	12.74	11.66	0.44						
	2013	8.77	6.75	0.38	9.67	8.15	0.42						
	2016	8.40	6.40	0.38	9.27	7.79	0.43						
	2019	6.38	3.67	0.38	6.76	4.90	0.44						

RF = random forest; SLR = supervised linear regression.

As mentioned, the BC and $PM_{2.5}$ data need to be cautiously interpreted because of the high temporal variation in the measurements, especially affecting R^2.

Figure 7 (section 5.2) shows that the data-fusion model predicted the corresponding weekly average measured concentration well. For the other models, we only had annual average predictions available, the primary aim of the study.

5.6 COMPARISON OF MODEL PREDICTIONS WITH MEASUREMENTS AT EXISTING EXTERNAL VALIDATION SITES

The comparison of model predictions with existing external validation data is presented in **Table 8**, **Figure 13**, and Additional Materials 2 Figure S36. **Box 1** presents a summary of the performance. For BC (in 2019), NO_2 (in 2010 and 2019), and $PM_{2.5}$ (in 2010 and 2019), existing external monitoring data were available based on year-round monitoring (NAQMN and additionally passive samplers for NO_2). For NO_2, passive sampler data were available in 2007 based on four simultaneous weekly measurements across the country. For UFPs, monitoring data were available for Amsterdam in 2002 (RUPIOH), and for Amsterdam–Utrecht in 2014 (Exposomics) from research studies with limited temporal coverage. Section 3.3 provides more detail.

We found generally moderate to high external validation performance for BC, NO_2, and UFPs, with substantial variation between models. Most models had substantially higher model performance at the existing nationwide monitoring sites than at the new validation sites (section 5.5), especially for BC. This is explained by the much better temporal stability of the average concentration measured at the routine monitoring sites. Most models predicted higher concentrations than measured. For NO_2, model predictions were close to measured concentrations. Additional Materials 2 Figure S36 illustrates that only a few comparisons are affected by single outliers, specifically some mobile model predictions for BC. We note that for the dispersion models, the comparison with measurements at the NAQMN network and routine Palmes tube sites represents an underestimation, as dispersion model predictions were assigned based on the nearest residential address. This does not apply to the other validation datasets, which were performed at residential addresses (section 4.1.4).

For BC, exposure models overestimated the measured concentrations at the 2019 NAQMN sites except the dispersion model (Figure 13, Table 8). Bias from short-term roadside models was substantially lower than for the models based on mobile (on-road) monitoring. The recent Google Airview model showed little bias, at about 0.3 µg/m³ (Table 8). Models underestimated the 2002–2004 measurements in Amsterdam. Models explained the variation of measured concentrations at the NAQMN sites in 2019 well, with the highest performance observed for the Europewide ELAPSE model. Differences in performance with the dispersion model, nationwide ESCAPE model, and the best-performing algorithm of each of the mobile monitoring models were, however, small. Especially the dispersion model, Google Airview, ESCAPE, RUN, and Exposomics models explained a large fraction of the variation in BC in the city of Amsterdam in 2002 using RUPIOH data, supporting the temporal stability of the exposure models. All models explained the Exposomics data obtained in 2014 less well, possibly because of the lower temporal stability. Differences in performance between algorithms were occasionally substantial for individual models (partly explained by a single or a few outliers), however none of the three algorithms performed consistently better. The two outliers for the RUN model are a highway site outside the cities and a residential site close to the Amsterdam harbor.

The generally high explained variance was accompanied by large differences in predicted concentrations and contrast (Figure 13; Additional Materials 2 Figure S36). Models from mobile monitoring predicted substantially higher concentrations measured at NAQMN sites in 2019, which cannot be explained by temporal (decreasing) trends. Models predicted substantially lower concentrations than measured in 2002 in the city of Amsterdam, explained by temporal trends.

For NO_2, exposure models explained the level of concentrations and the variation of measured concentrations at the NAQMN sites in 2010 and 2019 well, with the highest performance observed for the dispersion models. Differences in performance with the Europewide ELAPSE model, the Amsterdam-based Google Airview model, and the nationwide ESCAPE model were, however, small. All models explained variation in 2010 modestly better than in 2019, both for models based on 2019 monitoring (Google Airview, EXPANSE 2019) and 2009 monitoring (ESCAPE, EXPANSE 2010). This is likely related to the larger variation in 2010 routine monitoring data. All models explained Palmes tube passive sampler concentration data in Amsterdam and Utrecht moderately well, with the highest performance observed for the Google Airview model (especially in 2011). Model performance for all models was better when all sites across the Netherlands were included. All models explained the variation measured in 2007 with passive samplers across the Netherlands well, with higher performance for the dispersion model, Google Airview model, and the nationwide ESCAPE model compared to the Europewide models.

The models predicted NO_2 levels much better than observed for BC (Figure 13, Table 8). The differences in level between model prediction and measurements were probably mostly due to temporal trends; for example, the ESCAPE models based on 2009 monitoring align well with 2010 NAQMN routine data and overestimate the 2019 NAQMN data. Bias was generally small when comparing models of years close to the monitoring data.

Table 8. Root Mean Squared Error (RMSE), Bias, and Explained Variance (R^2) of Annual Average Concentrations Measured in Existing Validation Datasets[a]

			NAQMN (N = 6) 2011			NAQMN (N = 31) 2019			Exposomics (N = 41) 2014			RUPIOH (N = 46) 2002–2004		
BC			**RMSE**	**Bias**	**R^2**	**RMSE**	**Bias**	**R^2**	**RMSE**	**Bias**	**R^2**	**RMSE**	**Bias**	**R^2**
MUSIC	M	SLR	1.18	0.13	0.33	1.93	0.62	0.66	1.70	1.08	0.17	0.90	0.04	0.37
		RF	1.38	1.11	0.80	2.47	2.18	0.24	2.20	1.82	0.20	1.12	0.25	0.20
		Lasso	0.66	0.45	0.72	1.67	1.46	0.58	1.42	1.30	0.18	0.62	0.08	0.34
	ST	SLR	1.22	–0.91	0.26	1.09	0.52	0.13	0.56	0.22	0.19	1.18	–1.02	0.42
		RF	0.95	–0.73	0.72	0.56	0.51	0.69	0.45	0.16	0.26	1.31	–1.18	0.49
Exposomics	M	SLR	1.63	1.26	0.34	2.59	2.25	0.64	2.48	2.29	0.21	1.41	1.03	0.35
		RF	1.45	1.42	0.91	2.79	2.52	0.44	3.12	2.82	0.15	2.08	1.65	0.24
		Lasso	1.21	0.91	0.25	2.35	2.32	0.52	2.17	2.13	0.17	1.08	0.88	0.33
	ST	SLR	0.46	0.29	0.94	2.08	1.51	0.43	0.99	0.81	0.27	0.83	–0.55	0.40
		RF	0.34	–0.16	0.91	1.22	1.08	0.62	1.01	0.82	0.26	0.75	–0.51	0.52
		Lasso	0.30	–0.17	0.94	1.56	1.24	0.32	0.74	0.60	0.19	1.02	–0.77	0.26
RUN	M	SLR	0.52	–0.01	0.82	1.55	1.05	0.29	0.76	0.50	0.18	0.81	–0.51	0.32
		RF	0.60	–0.23	0.65	1.32	1.04	0.47	0.83	0.61	0.20	0.76	–0.54	0.52
		Lasso	0.79	–0.31	0.66	1.21	0.95	0.27	0.69	0.52	0.17	0.90	–0.63	0.36
	ST	SLR	0.79	–0.77	0.97	1.22	0.58	0.35	0.45	0.01	0.20	1.48	–1.35	0.54
		RF	1.18	–1.05	0.79	0.33	0.17	0.78	0.41	–0.05	0.29	1.53	–1.42	0.58
		Lasso	1.14	–1.04	0.98	0.65	0.34	0.35	0.43	–0.12	0.22	1.63	–1.49	0.56
Google		SLR	1.03	–0.85	0.93	0.44	0.32	0.65	0.42	0.06	0.22	1.50	–1.35	0.48
		RF	1.03	–0.90	0.92	0.55	0.48	0.62	0.42	0.07	0.23	1.47	–1.35	0.56
		Lasso	1.06	–0.89	0.93	0.48	0.39	0.58	0.43	0.11	0.19	1.50	–1.35	0.47
ESCAPE		SLR	0.59	–0.30	0.76	0.89	0.79	0.69	0.65	0.51	0.28	1.09	–0.90	0.41
		RF	0.68	–0.39	0.78	0.83	0.79	0.74	0.66	0.51	0.24	1.04	–0.85	0.46
		Lasso	0.59	–0.31	0.84	0.89	0.81	0.65	0.63	0.48	0.27	1.10	–0.89	0.36
ELAPSE		2010	0.68	–0.30	0.69	0.99	0.92	0.74	0.86	0.73	0.17	0.93	–0.67	0.30
		2013	0.73	–0.41	0.70	0.89	0.82	0.74	0.77	0.63	0.17	1.01	–0.77	0.30
		2016	0.75	–0.44	0.70	0.87	0.80	0.74	0.76	0.61	0.17	1.01	–0.77	0.30
		2019	0.77	–0.47	0.70	0.85	0.77	0.74	0.75	0.59	0.17	1.02	–0.78	0.30
Dispersion		2013	1.46	–1.23	0.27	0.26	0.11	0.64	0.39	–0.01	0.29	1.56	–1.42	0.53
		2016	1.65	–1.37	0.00	0.21	0.04	0.61	0.41	–0.07	0.27	1.90	–1.79	0.55
		2019	1.86	–1.64	0.13	0.31	–0.24	0.59	0.60	–0.43	0.27	0.90	0.04	0.37

M = mobile (on-road); N = number of sites; NAQMN = national air quality monitoring network; RF = random forest; SLR = supervised linear regression; ST = short-term roadside (applicable only to mobile monitoring studies).

[a]Units µg/m³ for $PM_{2.5}$, NO_2, and BC; p/cm³ for UFPs. The first three columns define the model, and the remaining columns refer to the performance of the validation datasets.

continued on next page

Table 8. (*continued*)

		NAQMN (N = 64) 2010			NAQMN (N = 81) 2019			TRACHEA (N = 144) 2007		
NO_2		RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2
Google	SLR	9.99	−8.22	0.80	4.51	0.05	0.69	9.49	−7.33	0.73
	RF	8.88	−5.26	0.67	6.79	3.53	0.50	8.26	−3.92	0.52
	Lasso	11.43	−9.12	0.66	5.56	−0.65	0.58	10.83	−8.78	0.61
ESCAPE	SLR	6.61	−1.40	0.75	11.06	7.21	0.61	5.08	−0.38	0.75
	RF	5.77	−1.16	0.77	9.25	7.08	0.63	5.48	0.82	0.71
	Lasso	6.80	−0.95	0.74	11.70	7.40	0.58	5.38	−0.88	0.72
ELAPSE	2010	6.91	3.95	0.80	14.14	12.30	0.69	9.63	8.02	0.72
EXPANSE	SLR_2010	5.81	0.22	0.78	11.32	8.88	0.64	7.75	4.55	0.59
	SLR_2013	6.38	−2.24	0.75	9.29	6.58	0.60	7.56	3.12	0.51
	SLR_2016	6.87	−3.67	0.76	8.02	5.07	0.62	6.80	1.61	0.56
	SLR_2019	9.15	−6.80	0.74	5.60	2.12	0.64	6.75	−0.31	0.58
	RF_2010	6.96	−2.40	0.71	8.91	5.99	0.63	7.24	3.10	0.56
	RF_2013	7.96	−4.18	0.68	8.11	4.40	0.60	7.21	2.74	0.54
	RF_2016	8.23	−5.38	0.73	6.84	3.14	0.66	6.39	−0.01	0.59
	RF_2019	10.80	−8.71	0.71	5.16	−0.14	0.65	7.34	−3.32	0.59
Dispersion	2010	6.01	−2.27	0.82	7.10	5.94	0.77	5.24	−0.04	0.73
	2013	8.70	−6.44	0.81	4.56	2.39	0.74	6.69	−4.40	0.79
	2016	9.73	−7.56	0.80	3.80	0.86	0.74	6.72	−4.40	0.79
	2019	11.84	−9.86	0.78	3.96	−1.37	0.74	9.21	−7.34	0.78

M = mobile (on-road); N = number of sites; NAQMN = national air quality monitoring network; RF = random forest; SLR = supervised linear regression; ST = short-term roadside (applicable only to mobile monitoring studies).

[a]Units µg/m³ for $PM_{2.5}$, NO_2, and BC; p/cm³ for UFPs. The first three columns define the model, and the remaining columns refer to the performance of the validation datasets.

Table 8. (*continued*)

NO_2		Palmes Tubes Amsterdam–Utrecht and Surroundings (N = 191) 2011			Palmes Tubes Amsterdam–Utrecht and Surroundings (N = 336) 2019			Palmes Tubes, Netherlands (N = 404) 2019		
		RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2
Google	SLR	11.19	–9.02	0.53	5.03	0.12	0.42	4.93	0.51	0.61
	RF	11.25	–8.52	0.39	5.55	1.59	0.39	6.16	2.49	0.49
	Lasso	10.80	–8.44	0.50	5.65	0.14	0.39	6.46	–0.30	0.54
ESCAPE	SLR	8.25	1.29	0.32	12.26	10.05	0.32	11.71	9.42	0.57
	RF	7.25	–0.56	0.41	10.80	9.21	0.37	10.46	8.95	0.61
	Lasso	7.96	–0.43	0.38	11.11	8.57	0.36	10.71	8.27	0.56
ELAPSE	2010	10.21	5.96	0.27	16.75	15.21	0.22	16.00	14.29	0.53
EXPANSE	SLR_2010	7.95	–0.21	0.34	11.55	9.25	0.27	11.20	8.99	0.53
	SLR_2013	8.66	–3.22	0.31	9.30	6.30	0.23	9.09	6.29	0.47
	SLR_2016	8.96	–4.29	0.33	8.32	5.20	0.25	8.07	5.12	0.51
	SLR_2019	9.77	–6.21	0.36	6.75	3.15	0.25	6.45	2.99	0.53
	RF_2010	7.95	–0.61	0.33	10.37	8.29	0.34	9.94	7.82	0.58
	RF_2013	8.08	–2.35	0.35	9.31	6.72	0.27	8.92	6.27	0.53
	RF_2016	8.53	–3.70	0.34	7.59	5.28	0.35	7.28	4.79	0.60
	RF_2019	10.71	–7.29	0.31	5.60	1.50	0.33	5.37	1.25	0.59
Dispersion	2010	7.52	–3.03	0.49	7.82	5.61	0.39	7.68	5.85	0.61
	2013	9.72	–7.13	0.53	5.88	1.64	0.34	5.70	1.94	0.56
	2016	9.65	–7.22	0.59	5.15	1.57	0.43	4.92	1.68	0.65
	2019	12.66	–10.64	0.52	4.86	–1.72	0.47	4.61	–1.35	0.67

M = mobile (on-road); N = number of sites; NAQMN = national air quality monitoring network; RF = random forest; SLR = supervised linear regression; ST = short-term roadside (applicable only to mobile monitoring studies).

[a]Units μg/m^3 for $PM_{2.5}$, NO_2, and BC; p/cm^3 for UFPs. The first three columns define the model, and the remaining columns refer to the performance of the validation datasets.

Table 8. (*continued*)

			Exposomics (N = 41) 2014			RUPIOH (N = 46) 2002–2004		
UFPs			**RMSE**	**Bias**	R^2	**RMSE**	**Bias**	R^2
MUSIC	M	SLR	7,412.11	5,028.41	0.53	16,671.25	–9,276.96	0.25
		RF	6,691.01	5,181.66	0.66	17,849.62	–9,304.56	0.09
		Lasso	5,564.40	4,510.10	0.57	18,119.06	–11,279.80	0.32
	ST	SLR	4,595.36	–1,408.91	0.49	21,587.66	–16,154.89	0.25
		RF	3,405.47	–1,197.13	0.58	21,657.33	–16,805.49	0.35
Exposomics	M	SLR	5,554.95	4,534.48	0.60	17,179.45	–10,485.76	0.38
		RF	9,026.06	6,998.06	0.55	16,112.78	–7,553.95	0.19
		Lasso	5,280.28	4,265.62	0.59	18,183.59	–11,090.42	0.28
	ST	SLR	4,040.11	818.85	0.46	20,761.04	–15,281.46	0.25
		RF	3,359.40	169.43	0.52	21,181.97	–15,389.83	0.22
		Lasso	3,831.27	–730.63	0.40	22,470.83	–16,929.86	0.19
RUN	M	SLR	5,214.18	3,140.90	0.56	17,427.82	–10,866.92	0.31
		RF	6,102.83	3,727.54	0.45	18,016.91	–10,552.01	0.15
		Lasso	4,050.32	2,395.26	0.57	19,233.25	–13,078.09	0.31
	ST	SLR	3,558.39	–380.25	0.50	20,905.29	–14,533.13	0.10
		RF	3,833.46	–200.53	0.54	20,751.34	–15,245.00	0.27
		Lasso	5,985.80	–3,823.55	0.44	25,443.88	–20,010.12	0.11
Google		SLR	6,618.36	4,243.12	0.06	19,960.17	–13,023.24	0.27
		RF	6,390.71	4,282.97	0.06	19,837.58	–13,010.60	0.35
		Lasso	6,139.70	4,094.06	0.11	19,980.46	–12,911.65	0.19

M = mobile (on-road); N = number of sites; NAQMN = national air quality monitoring network; RF = random forest; SLR = supervised linear regression; ST = short-term roadside (applicable only to mobile monitoring studies).

[a]Units µg/m^3 for $PM_{2.5}$, NO_2, and BC; p/cm^3 for UFPs. The first three columns define the model, and the remaining columns refer to the performance of the validation datasets.

Table 8. (*continued*)

$PM_{2.5}$			NAQMN (*N* = 14) 2010 RMSE	Bias	R^2	NAQMN (*N* = 49) 2019 RMSE	Bias	R^2	Exposomics (N = 41) 2014 RMSE	Bias	R^2	RUPIOH (*N* = 46) 2002–2004 RMSE	Bias	R^2
MUSIC	M	SLR	36.21	35.41	0.00	37.03	36.10	0.04	36.86	36.30	0.06	24.42	24.07	0.01
		RF	32.14	30.86	0.00	39.45	38.73	0.02	37.34	36.19	0.02	20.75	20.08	0.01
	ST	SLR	99.56	95.36	0.04	96.88	94.88	0.00	84.86	84.34	0.14	70.82	70.57	0.13
		RF	72.02	71.65	0.03	79.35	78.90	0.00	74.35	74.23	0.01	65.12	64.91	0.06
Exposomics	M	SLR	32.22	30.53	0.20	39.52	37.63	0.01	39.31	38.60	0.09	30.63	30.26	0.15
		RF	46.68	40.96	0.19	47.46	43.34	0.01	66.95	58.05	0.06	71.45	62.16	0.03
		Lasso	59.31	55.21	0.08	63.61	60.23	0.00	59.80	56.17	0.00	53.97	51.61	0.07
	ST	SLR	30.47	29.37	0.04	42.49	40.11	0.03	39.33	38.95	0.12	30.23	29.75	0.07
		RF	46.01	43.57	0.23	56.47	52.03	0.06	49.35	47.54	0.08	46.96	44.49	0.03
RUN	M	SLR	30.92	30.61	0.09	36.67	36.37	0.00	36.53	36.27	0.07	26.91	26.69	0.17
		RF	20.13	19.40	0.19	25.97	25.59	0.00	26.49	25.94	0.04	19.70	19.15	0.18
	ST	SLR	18.69	16.00	0.20	33.92	27.60	0.00	21.53	21.21	0.14	13.74	12.69	0.07
		RF	10.13	8.39	0.06	18.18	17.20	0.09	20.68	20.52	0.09	11.66	11.44	0.18
ESCAPE		SLR	2.22	−1.41	0.12	6.63	6.34	0.02	4.76	4.41	0.06	6.30	−6.02	0.13
		RF	2.12	−0.89	0.04	6.81	6.53	0.03	5.22	4.89	0.06	5.48	−5.22	0.31
		Lasso	1.67	−0.61	0.18	6.97	6.78	0.01	5.04	4.75	0.12	5.64	−5.33	0.22
ELAPSE		2010	2.76	−1.28	0.05	6.82	6.39	0.01	4.99	4.60	0.02	5.95	−5.64	0.11
EXPANSE		SLR_2010	2.46	−1.67	0.03	6.42	6.15	0.08	4.63	4.25	0.01	6.36	−6.02	0.02
		SLR_2013	4.08	−3.74	0.10	3.82	3.44	0.12	2.29	1.57	0.10	8.70	−8.46	0.00
		SLR_2016	6.87	−6.71	0.27	1.82	0.84	0.06	2.21	−1.08	0.00	11.65	−11.47	0.01
		SLR_2019	7.35	−7.18	0.19	1.79	0.56	0.02	2.34	−1.49	0.01	12.00	−11.83	0.02
		RF_2010	2.62	−1.17	0.19	6.92	6.65	0.01	5.46	5.17	0.03	5.26	−4.79	0.08
		RF_2013	4.16	−3.61	0.10	4.37	3.92	0.00	3.11	2.60	0.05	7.72	−7.47	0.03
		RF_2016	5.85	−5.54	0.09	2.43	1.77	0.02	1.81	0.30	0.00	10.05	−9.85	0.00
		RF_2019	6.61	−6.35	0.00	1.86	1.01	0.06	1.90	−0.51	0.00	10.66	−10.47	0.00
Dispersion		2010	3.43	1.80	0.43	7.69	7.30	0.28	6.26	6.00	0.12	4.33	−4.01	0.32
		2013	4.30	−4.09	0.45	3.67	3.20	0.15	2.64	2.05	0.12	8.42	−8.25	0.33
		2016	6.37	−6.23	0.44	1.86	0.47	0.15	1.70	−0.23	0.12	10.43	−10.30	0.32
		2019	7.47	−7.36	0.48	1.69	−0.48	0.15	2.29	−1.53	0.08	11.83	−11.72	0.33

M = mobile (on-road); N = number of sites; NAQMN = national air quality monitoring network; RF = random forest; SLR = supervised linear regression; ST = short-term roadside (applicable only to mobile monitoring studies).

[a] Units μg/m^3 for $PM_{2.5}$, NO_2, and BC; p/cm^3 for UFPs. The first three columns define the model, and the remaining columns refer to the performance of the validation datasets.

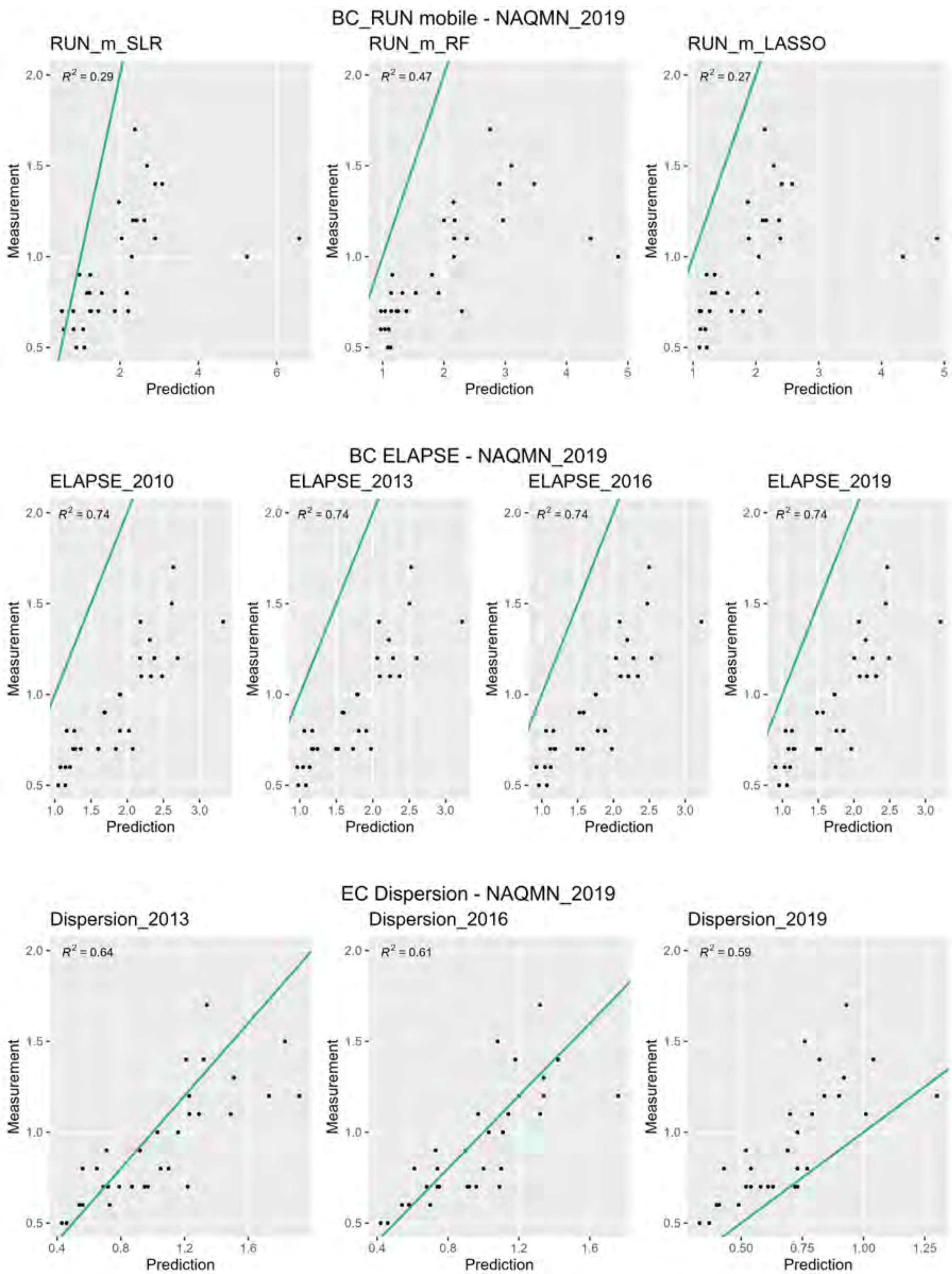


Figure 13. Scatterplots of comparison of exposure predictions with existing external validation data for BC, NO_2, UFPs, and $PM_{2.5}$. The green line is the 1:1 line. Pollutant concentrations are in p/cm³ for UFPs and in µg/m³ for all others. The first element of the title indicates the exposure prediction model (e.g., Google SLR model); the second is the validation data (see section 3.3.1).

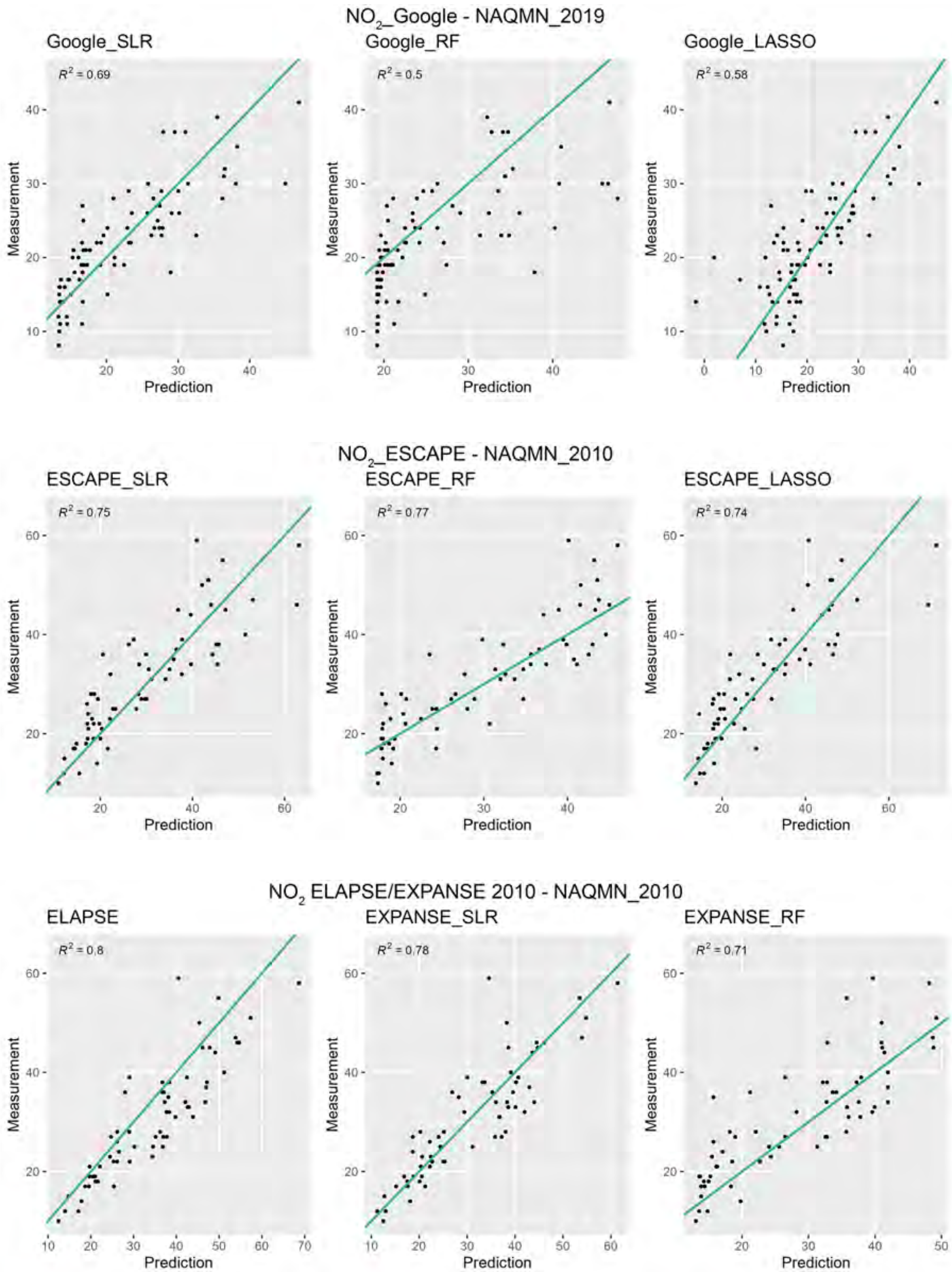


Figure 13. (*continued*)

continued on next page

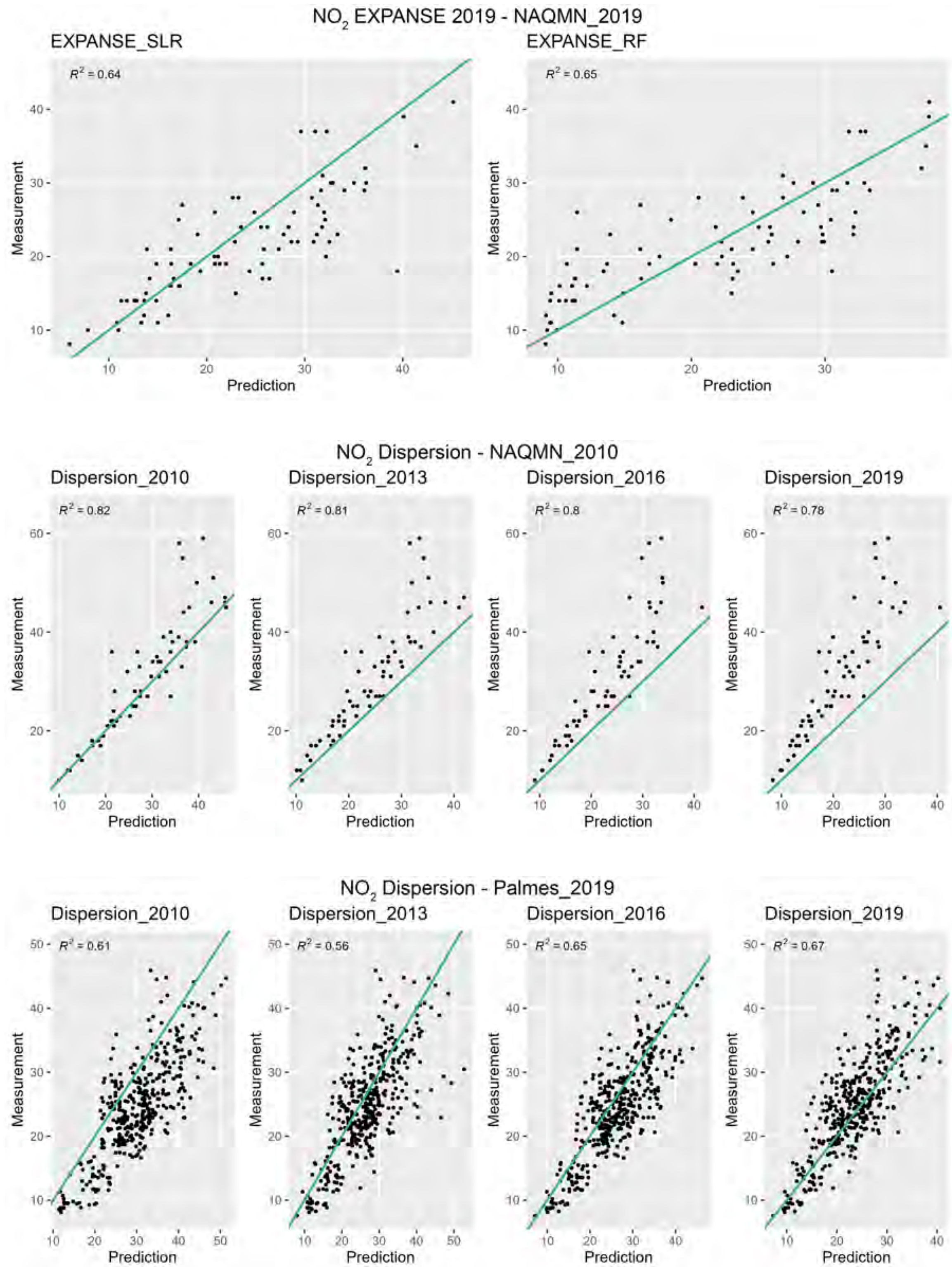


Figure 13. (*continued*)

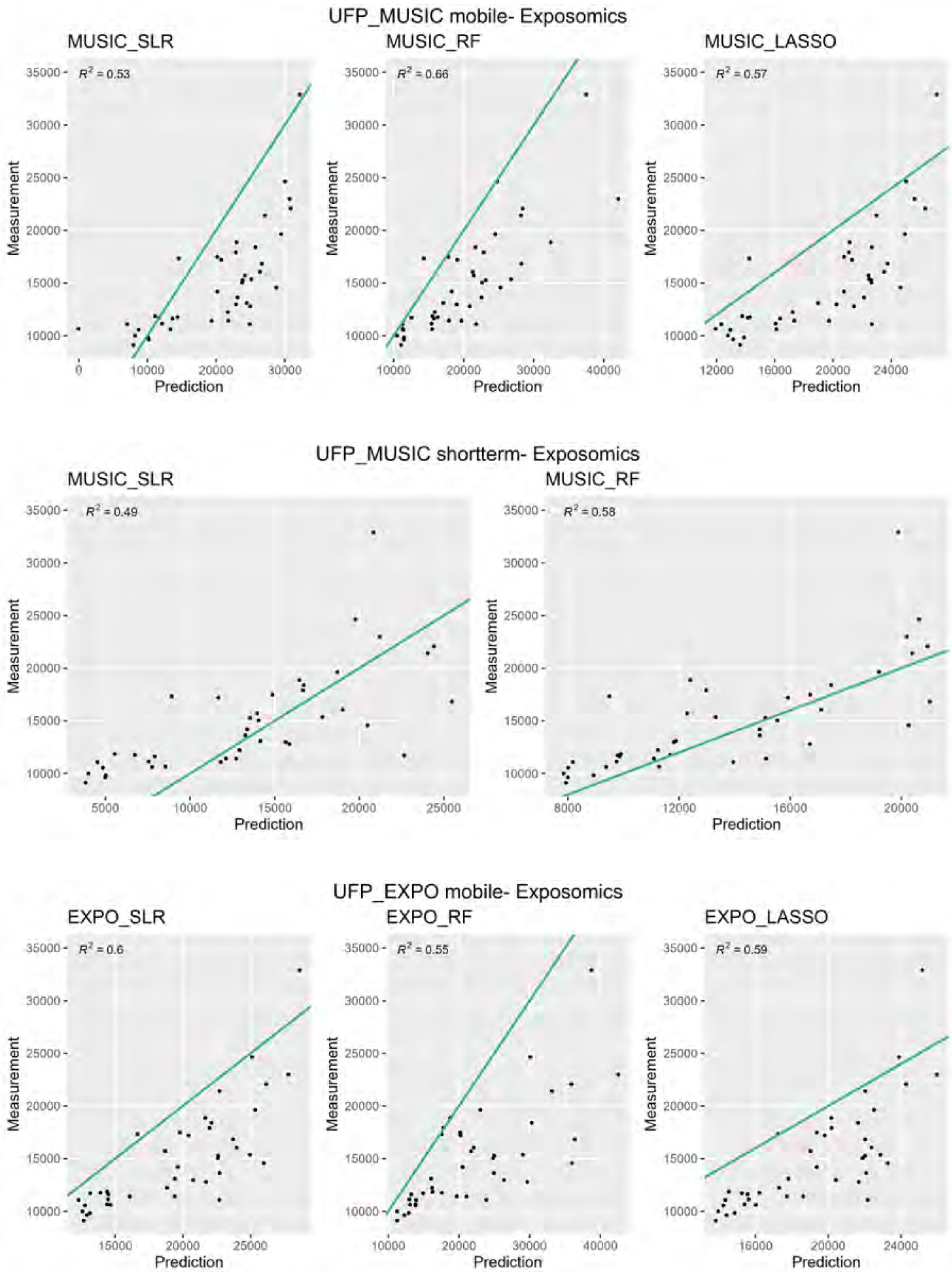


Figure 13. (*continued*)

continued on next page

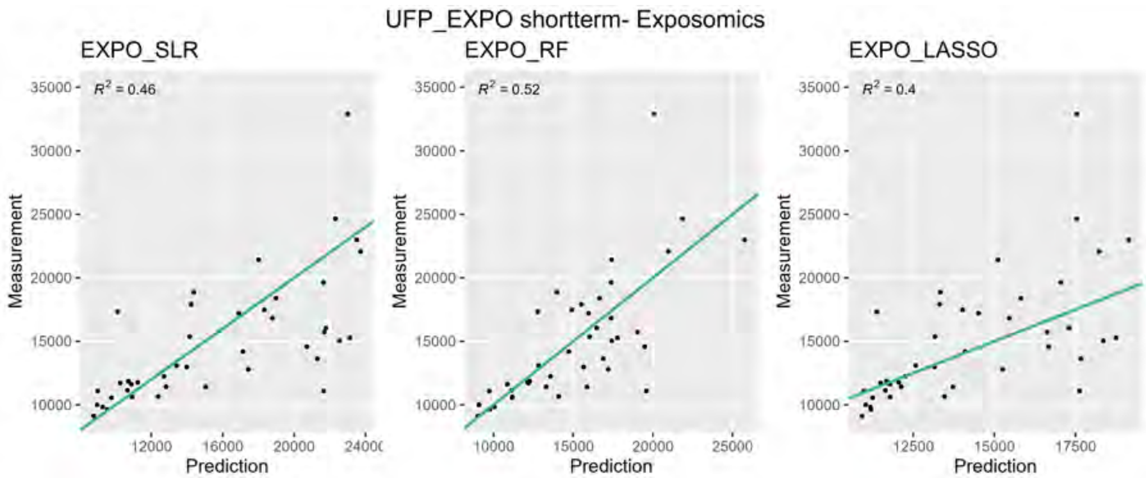


Figure 13. (*continued*)

For UFPs, all mobile monitoring models explained a modest fraction of the variation in UFPs in the city of Amsterdam in 2002 (RUPIOH), supporting the moderate temporal stability of the exposure models. R^2 values were, however, substantially lower than for BC for the same RUPIOH sites. Differences in performance between algorithms were substantial for individual models for the RUPIOH data and less so for the more recent Exposomics data, however none of the three algorithms performed consistently better. All models explained the Exposomics data obtained in 2014 better, except the Google Airview model based on 2019 data.

Large differences in predicted concentrations were found (Figure 13, Table 8, Additional Materials 2 Figure S36). Models predicted substantially lower concentrations than measured in 2002 in the city of Amsterdam using RUPIOH data, probably explained by temporal trends. Differences in instrumentation between monitoring campaigns and validation data have affected this comparison. The large differences observed, for example, in predicted concentrations from the Google Airview 2019 model and the measured concentrations in 2002, both in Amsterdam, cannot be explained by instrumental differences. The comparison between the Exposomics model and the Exposomics validation is not affected by temporal differences: the comparison shows a moderate overestimation of the mobile model and virtually no overestimation of the short-term roadside exposure model. Overestimation of the mobile (on-road) models was substantially larger than for the short-term roadside monitoring-based models. Bias was about 3,000–5,000 p/cm^3 in all four mobile (on-road) campaigns with no obvious trend over time from the 2013 MUSIC to the 2019–2020 Google Airview campaign.

Mobile models performed poorly at predicting external long-term $PM_{2.5}$ concentrations at the NAQMN sites, the Exposomics sites, and with the Amsterdam 2002 data from RUPIOH. The ESCAPE, ELAPSE, and EXPANSE models explained variation at the NAQMN sites poorly, but the level was predicted with little bias. Dispersion models explained variance moderately well for NAQMN 2010 and less so in 2019. The level was explained well. The low explained variance is possibly due to the small variability in measured concentrations at the network sites, especially in recent years. We previously reported that the ELAPSE model did not predict the variability at the ESCAPE sites in the Netherlands well (R^2 = 0.12), whereas it did well across Europe (de Hoogh 2018). The dispersion model explained the variation in 2002 within Amsterdam moderately well. All models explained only a low fraction of the variability of $PM_{2.5}$ in the 2002 data in Amsterdam and Amsterdam–Utrecht, illustrating the dependence of conclusions on the scale of the domain or the year of monitoring. We hypothesize that the BC models explain intraurban variability better than the $PM_{2.5}$ models focusing on between-city variability. Very large differences in predicted concentrations were found (Table 8, Additional Materials 2 Figure S36) for models from mobile monitoring. Bias for the dispersion model and the three fixed-site LUR models when compared to the corresponding year were small.

Model Averaging The median, mean, and weighted mean of the BC models predicted 64%, 69%, and 67% of the variability of BC concentrations at the 31 sites of the NAQMN in 2019. In this analysis, we analyzed the 2019 models for the dispersion model, the European hybrid model, and the SLR model of the empirical models. The ensemble model estimate does not improve upon the best individual models, which explains 74% of the concentration variability (Table 8).

The median, mean, and weighted mean of the NO_2 models predicted 75%, 76%, and 76% of the variability of NO_2 concentrations at the 81 sites of the NAQMN in 2019. In this analysis, we analyzed the 2019 models for the dispersion model, the European hybrid model, and the SLR model of the ESCAPE and Google Airview models. The ensemble model estimate improved marginally upon the best individual model, which also explained 74% of the concentration variability (Table 8).

Box 1. Summary of Performance at External Validation Sites

- The mean explained variance for BC based on the national monitoring network 2019 was 0.48, 0.71, and 0.61 for mobile monitoring, fixed-site LUR, and dispersion models, with large variability within each category. The mean explained variance for BC was 0.48, 0.58, and 0.47 for SLR, Random Forest, and LASSO algorithms.
- The mean explained variance for NO_2 based on the national monitoring network 2019 was 0.59, 0.63, and 0.75 for mobile monitoring, fixed-site LUR, and dispersion models, with large variability within each category. The mean explained variance for NO_2 was 0.63, 0.61, and 0.58 for SLR, Random Forest, and LASSO algorithms.
- The mean explained variance for UFPs based on the new external validation campaign was 0.36, 0.27, and 0.35 for SLR, Random Forest, and LASSO algorithms.

5.7 COMPARISON OF EPIDEMIOLOGICAL FINDINGS: DUELS COHORT

Population characteristics are presented in Additional Materials 1 Table S3. During follow-up, 945,615 natural cause deaths occurred, of which 244,977 were due to cardiovascular disease, 84,734 due to respiratory disease, and 71,622 due to lung cancer. Exposure distributions and the correlations between the different models are not presented separately for the full cohort, as the distributions and correlations presented in section 5.4 for the 20,000 randomly selected residential addresses represent the national cohort. Note that the model-specific IQRs can be found in Appendix A Table S1.

Figure 14 shows the HR for the associations for the four pollutants with natural-cause mortality in the most-adjusted confounder model for fixed increments and model-specific IQRs. For the empirical models, the SLR algorithm is shown. Additional Materials 2 Tables S11–S14 show the HRs for natural, cardiovascular, respiratory, and lung cancer mortality of all models and algorithms. Additional Materials 2 Table S15 shows the HRs for confounder models 1–3 with different level of control for covariates for natural-cause mortality.

Natural-Cause Mortality With all exposure models, positive associations that were generally statistically significant with natural-cause mortality were found in analyses covering the entire country. The strength of the association varied substantially across models, both when expressed for a fixed increment and as model-specific IQRs. The variability of HRs between models was generally smaller when expressed per IQR (Additional Materials 2, Tables S11–S14). Confidence intervals were sizably smaller for some of the mobile monitoring models.

For BC, there was no consistent difference in HRs between the mobile monitoring, LUR models based on fixed-site monitoring, and dispersion models (Additional Materials 2 Table S11). Sizable differences in HRs were found between algorithms within individual models, especially when expressed for fixed increments (e.g., within Google Airview, MUSIC, and Exposomics). For the ESCAPE model, the differences between algorithms were small. There was no consistent pattern of higher HRs for a specific algorithm. HRs and confidence intervals for dispersion models of 2013, 2016, and 2019 were very similar.

Positive generally significant associations were found for NO_2, with only small differences between the Google Airview mobile monitoring model, nationwide ESCAPE, Europewide EXPANSE, and dispersion modeling approaches (Additional Materials 2 Table S12). HRs and confidence intervals for Europewide LUR and dispersion models of 2010, 2013, 2016, and 2019 were very similar. A null association was found for the new low-cost monitor network based on SLR and the Google Airview LASSO algorithm. For other algorithms in these two models, significant associations were found. For the ESCAPE and EXPANSE models, differences in HRs were small between algorithms.

Positive generally significant associations were found for UFPs, with substantial differences between the different mobile monitoring models when expressed per fixed increments (Additional Materials 2 Table S13). HRs for mobile (on-road) and short-term roadside models were not consistently different. We found no consistent difference across algorithms, although in individual models, differences in HRs between algorithms could be large.

Positive generally significant associations were found for $PM_{2.5}$, with only small differences between the nationwide ESCAPE, Europewide EXPANSE, and dispersion modeling approaches (Additional Materials 2 Table S14). HRs and confidence intervals for Europewide LUR and dispersion models of 2010, 2013, 2016, and 2019 were very similar. A null association was found for the EXPANSE models based on Random Forest. For the ESCAPE models, the differences in HRs were small between algorithms. Despite the poor exposure models based on mobile monitoring and low-cost sensor measurements for $PM_{2.5}$, significant associations with natural mortality were found for several of these models.

For most exposures, HRs decreased upon adjustment for individual-level covariates (confounder model 1 to model 2, Additional Materials 2 Table S15) and area-level covariates (confounder model 2 to model 3). For some exposures, the difference between confounder models 2 and 3 was small. There was no consistent pattern of larger changes in HR between models or algorithms. For BC and $PM_{2.5}$, the decrease in HR upon adjustment for covariates was similar for dispersion models and empirical models, whereas, for NO_2, the decrease in HR was substantially smaller for the dispersion model

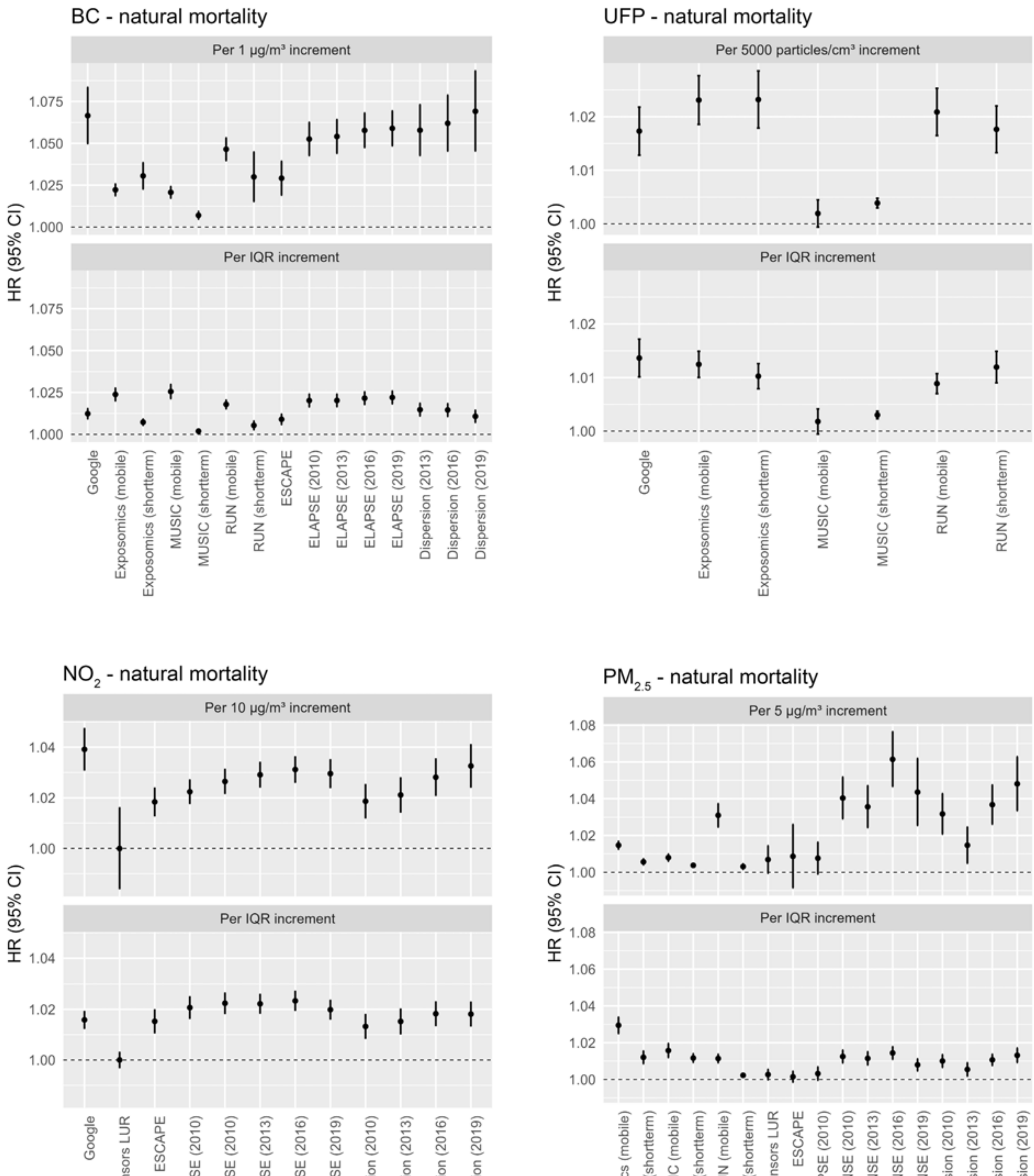


Figure 14. HR for the association of the four pollutants with natural mortality for different exposure models in the nationwide DUELS cohort, using the most adjusted model 3. HR expressed per 1-µg/m³ for BC, 10-µg/m³ for NO_2, 5,000-p/cm³ for UFPs, and 5-µg/m³ for $PM_{2.5}$. Covariates included were at the individual level age (time axis), sex (strata), standardized household income, region of origin (Dutch, Western, non-Western, Morocco, Turkey, Suriname, Antilles Netherlands), and marital status (married, widowed, divorced, single), and at area-level socioeconomic status variables at both the neighborhood and the regional (COROP) scale: mean income per income recipient, social assistance per 1,000 inhabitants, % low education, % non-Western immigrants, and unemployment per 1,000 inhabitants. IQR is shown in Appendix A table S1. Note scales differ per pollutant.

Variability of HR in Relation to Model Performance at External Validation Sites When we combined the external validation statistics (Table 8, NAQMN 2019) with the observed HR for natural-cause mortality, we found a weak pattern of higher HRs for exposure models with higher validation R^2 (Additional Materials 2 Table S16). For both BC and NO_2, there was a weak tendency of increasing HRs with increasing validation R^2 (**Figure 15**, Additional Materials 2 Table S16). The pattern for BC was somewhat clearer when HRs were expressed per IQR. Within models, we found some evidence of higher HRs associated with higher validation R^2. For NO_2, the range in validation R^2 was small.

Amsterdam Analyses When the analyses were restricted to the city of Amsterdam only, we still observed positive associations for natural-cause mortality with BC, NO_2, and UFPs (Additional Materials 2 Table S17). Associations with $PM_{2.5}$ were less consistent. Differences between models and algorithms were somewhat larger than for the nationwide analysis, partly related to wider confidence intervals. Especially for NO_2, consistent associations were found with all models. Associations with UFPs were more consistent with the MUSIC, Exposomics, and RUN campaigns than with the Google Airview mobile monitoring campaign. For $PM_{2.5}$, associations were mostly found with the dispersion model and

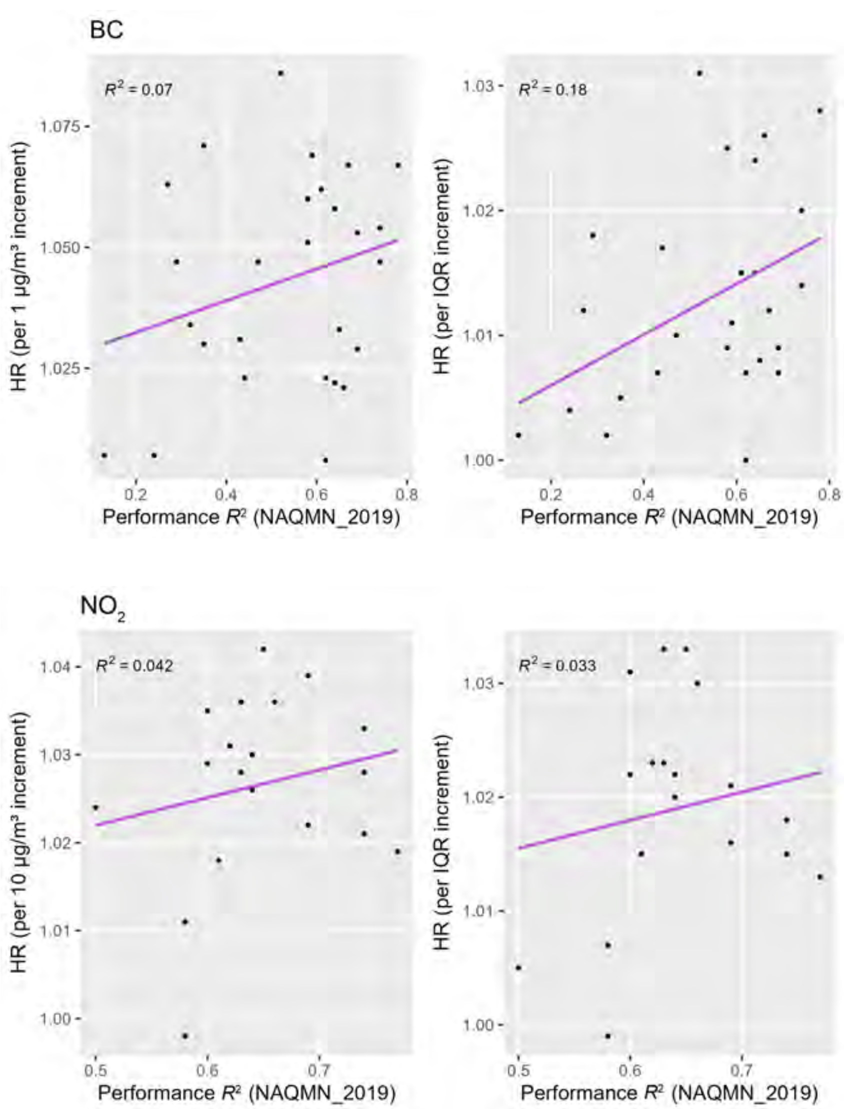


Figure 15. Associations between BC and NO_2 with natural mortality of different exposure models in relation to the explained variance at NAQMN external validation sites in 2019. HR per 1- and 10-µg/m³ for BC and NO_2, and model-specific IQR. Models using the most adjusted model 3 (see Figure 14). IQR is shown in Appendix A Table S1. The purple line is the regression line. Note that scales differ per pollutant.

less so with the ESCAPE, ELAPSE, and EXPANSE models, possibly because of the higher correlation with traffic-related pollutants in the dispersion model. Differences in HRs between algorithms were smaller for the models based on temporally stable monitoring (ESCAPE, ELAPSE, EXPANSE) compared to the mobile monitoring campaigns. We did not find evidence that models developed primarily based on Amsterdam monitoring (Google Airview and MUSIC) showed more consistent associations with mortality than the other models.

Two-Pollutant Models Two-pollutant analyses for natural-cause mortality with either dispersion models or Europewide empirical models (ELAPSE or EXPANSE) showed some differences between the two models. With the dispersion model, in most analyses, only one of the two pollutants remained significantly associated with mortality, whereas with the empirical model, in general, both pollutants remained associated (Additional Materials 2, Table S18). This finding is consistent with the higher correlation between pollutants for dispersion models.

Heterogeneity in Effect Estimates Related to Exposure To illustrate how large the heterogeneity in effect estimates is, we performed random-effect meta-analyses for natural-cause mortality (**Table 9**). We found I^2 values of larger than 80% for the four pollutants, documenting that differences in exposure assessment may contribute substantially to the observed heterogeneity in reviews of epidemiological studies (Boogaard et al. 2023a; Chen et al. 2020; HEI 2022). In the WHO systematic review, the I^2 value for all-cause mortality for $PM_{2.5}$ was 89% based on 25 studies (Chen et al. 2020). In the HEI review of traffic-related air pollutants, I^2 values of 83%, 84%, and 51% were found for NO_2, EC, and $PM_{2.5}$ studies on natural-cause mortality (Boogaard et al. 2023a). Note that the comparison is limited because a different body of evidence is summarized, and most studies differ in many other methods and design features beyond exposure assessment; hence, caution is warranted.

Table 9. Meta-Analysis of Effect Estimates for Associations of Natural-Cause Mortality in DUELS to Illustrate Degree of Heterogeneity Introduced by Differences in Exposure Assessment[a]

Pollutant	*N* Effect Estimates	I^2
BC	10	97.0%
NO_2	5	87.6%
UFPs	7	97.3%
$PM_{2.5}$	10	96.5%

[a]For multiyear models (dispersion and European models), only the model for 2016 was used in the meta-analysis. Only the SLR algorithm is included.

Cause-Specific Mortality For cardiovascular mortality, mostly positive associations were found for BC, UFPs, and $PM_{2.5}$ (Additional Materials 2 Tables S11–S14). However, the significance and strength of the association varied more between models than for natural-cause mortality, likely related to the overall smaller effect estimates for this outcome. For some exposure models, no associations were found, for example in the EXPANSE Random Forest models. For NO_2, HRs were generally close to unity, with some HRs still indicating an association. With the Europewide LUR and Google Airview model, fairly consistent associations were found, whereas with the dispersion model associations were null for all years.

For respiratory mortality, mostly positive associations were found for BC, NO_2, UFPs, and $PM_{2.5}$. The strength of the association varied between models as observed for natural-cause mortality, with generally larger effect estimates for this outcome. HRs for different years were very similar, with the exception of the EXPANSE 2019 SLR model for $PM_{2.5}$.

For lung cancer mortality, consistent positive associations were found for BC, NO_2, UFPs, and $PM_{2.5}$. For EXPANSE Random Forest models of $PM_{2.5}$, no association was found. The strength of the association varied between models as observed for natural-cause mortality, with generally larger effect estimates for this outcome. HRs for different years were very similar, with the exception of the EXPANSE 2019 SLR model for $PM_{2.5}$.

5.8 COMPARISON OF EPIDEMIOLOGICAL FINDINGS: PIAMA BIRTH COHORT

Population characteristics are presented in Additional Materials 1, Table S4. Exposure distributions are shown in Additional Materials 2, Table S19 for the population included in the asthma analysis, including the IQR. Pollutant distributions were similar for the smaller lung function population compared to the asthma analysis population. In general, moderately lower contrasts were found compared to the contrast at the 20,000 randomly selected sites across the Netherlands, consistent with the smaller recruitment area (Additional Materials 2, Table S4). Correlations between exposure models are shown in Additional Materials 2, Tables S20–S23 for the population with asthma data and Additional Materials 2, Tables S24–S27 for the smaller population with lung function data. In general, correlation patterns were similar to those described for the randomly selected 20,000 residential addresses in section 5.4. Briefly, while most exposure estimates were moderately to highly correlated, some exposure estimates only had a low correlation ($R < 0.5$), such as the MUSIC mobile model for BC and the Google Airview UFP model.

Associations with Asthma Associations between air pollution exposure and asthma at age 1–20 years at birth and current address are shown in **Figure 16** for fixed increments, and Additional Materials 2 Figure S37 for model-specific IQRs, including all models with the SLR algorithm. Additional

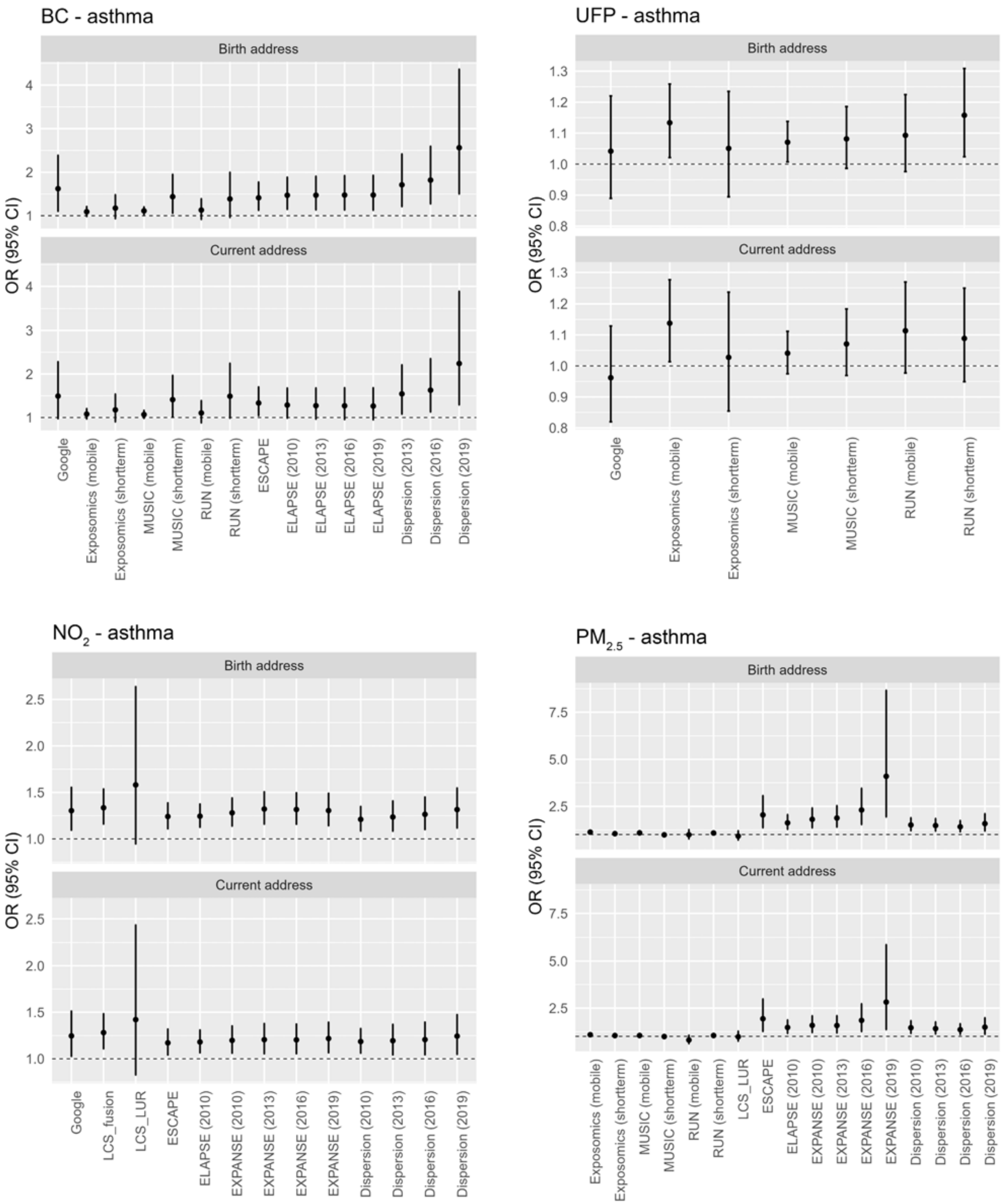


Figure 16. Associations between air pollution exposure and asthma at age 1 to 20 years at the birth and current address, most adjusted model. OR expressed per 1-µg/m³ for BC, 10-µg/m³ for NO_2, 5,000-p/cm³ for UFPs, and 5-µg/m³ for $PM_{2.5}$. Adjusted for sex, maternal and paternal asthma, hay fever, native nationality, parental education, breastfeeding, older siblings, day-care attendance, maternal smoking during pregnancy, parental smoking at home, mold or dampness at home, pets, use of gas for cooking, and active smoking (from age 14). Note scales differ per pollutant.

Materials 2 Table S28 includes all models and algorithms. Additional Materials 2 Figure S38 shows odds ratios (OR) for NO_2 low-cost models using different scenarios for data fusion. For most models, significant positive associations with asthma incidence were found, although the strength of the association varied. Associations were found both with the exposure at the birth address and the current address.

For BC, consistent associations were found with all exposure models. The largest ORs were found for the dispersion model, but the difference between the nationwide and European LUR models and most of the mobile models was small.

For NO_2, consistent associations were found with all exposure models for both the birth and current address. ORs were comparable for the dispersion model, the nationwide ESCAPE, and the European LUR models. The confidence intervals from the low-cost sensor LUR model were particularly wide, albeit still indicative of an association. The ORs from the data fusion model were similar to the dispersion model, consistent with the small contribution of the low-cost sensor measurement to the modeling.

For UFPs, associations were mostly positive with different models, but less consistent than for BC and NO_2. Associations with the Google Airview model and the Exposomics short-term roadside model were generally null.

For $PM_{2.5}$, consistent associations were found with the dispersion model, the nationwide ESCAPE, and the European LUR models, with similar ORs. ORs from the mobile models and the low-cost sensor LUR model were inconsistent and often indicative of a null association.

Figure 17 shows the associations between air pollution exposure and lung function (i.e., forced vital capacity [FVC] and forced expiratory volume in 1 second [FEV_1]) at age 16 at the current address for fixed increments. Additional Materials 2 Figure S39 shows the associations for model-specific IQRs. Additional Materials 2 Table S29 includes all models and algorithms. For most exposure models, negative associations with FEV_1 were found that were most consistent for BC, NO_2, and $PM_{2.5}$. Associations with FVC were less consistent.

For BC, the negative associations with FEV_1 were similar in magnitude for the dispersion model, the ESCAPE model, and the ELAPSE model. Most mobile monitoring models had smaller associations and sometimes wide confidence intervals. Associations with FVC were mostly null.

For NO_2, the negative associations with FEV_1 were similar in magnitude for the dispersion model, the ESCAPE model, the ELAPSE and EXPANSE models, and the Google Airview mobile monitoring model. A null association with a wide confidence interval was found for the low-cost sensor model. Associations with FVC were indicative of a negative association with most models, except for the low-cost sensor LUR model, which also had wide confidence intervals. Associations were very similar for the fusion model based on low-cost sensor measurements and the dispersion model.

For UFPs, the weak negative associations with FEV_1 were similar in magnitude for different mobile monitoring models. Associations with FVC were mostly null.

For $PM_{2.5}$, the negative associations with FEV_1 were similar in magnitude for the dispersion, ESCAPE, ELAPSE, and EXPANSE (except 2019) models. Associations were inconsistent for the mobile monitoring models. Associations with FVC were mostly null.

Meta-analysis of Heterogeneity **Table 10** presents an assessment of the heterogeneity in effect estimates as an illustration of the potential contribution of exposure assessment differences to previously reported heterogeneity. Effect estimates for asthma incidence showed substantial heterogeneity for BC (I^2 = 64% and 37% for exposure at birth address and current address, respectively), whereas no or little heterogeneity was found for NO_2, UFPs, and $PM_{2.5}$, with I^2's close to 0%.

In the meta-analyses for lung function, some heterogeneity was found for associations of BC and $PM_{2.5}$ with FEV_1 (I^2 = 42% and 33% respectively), whereas no heterogeneity was found for the other pollutants or FVC. Including the models developed with the Random Forest and LASSO algorithms in the meta-analyses increased the I^2 for some associations. Heterogeneity in previous meta-analyses was low to moderate (Additional Materials 2 Table S30).

Relation Between Effect Estimates and Predictive Performance of the Exposure Model When comparing effect estimates for asthma per fixed increment with the external validation R^2 of the model, we found a weak positive relation for BC and UFPs (Additional Materials 2 Figure S40), indicating that models with a higher external validation R^2 corresponded to a higher OR. The relation was strongest for UFPs (R^2 = 0.39 at the birth address), slightly positive for BC (R^2 = 0.14 at the birth address), and close to null or negative for NO_2 and $PM_{2.5}$. When we regressed the OR of a model on the model performance and the IQR of the predicted exposure together, we found that the relations with model performance for BC and UFPs remained positive and mostly significant (P value < 0.10) (**Table 11**). Associations for NO_2 and $PM_{2.5}$ remained highly nonsignificant. For BC and $PM_{2.5}$, higher predicted exposure contrast was associated with a lower effect estimate (Table 11).

For the lung function estimates, a higher R^2 corresponded to a stronger negative percentage difference in lung function estimate for BC, NO_2, and UFPs with FEV_1 (**Figure 18**). The relation was fairly strong for NO_2 with an R^2 of 0.44, indicating that 44% of the variability of the percentage difference in lung function associated with NO_2 was explained by exposure model performance. For $PM_{2.5}$ no clear correlation was observed, in a setting where all models had low validation R^2. For FVC, for which no main effects were found for any models, only UFPs showed a clear correlation between the OR and model performance (R^2 = 0.32).

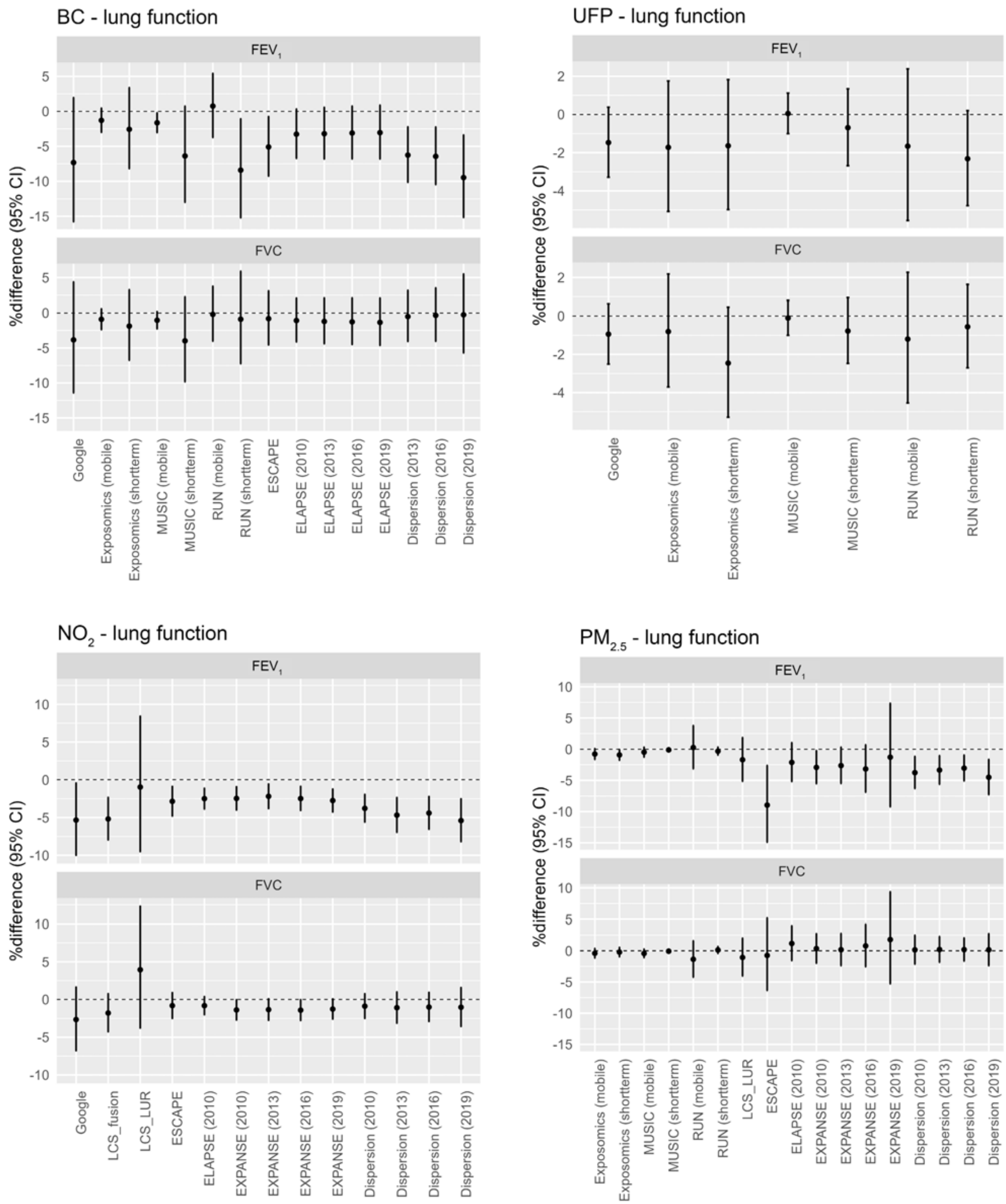


Figure 17. Associations between air pollution exposure and lung function (FVC and FEV1) at age 16 at the current address, most adjusted model. Expressed as percentage difference for fixed increments: 1-µg/m^3 for BC, 10-µg/m^3 for NO_2, 5,000-p/cm^3 for UFPs, and 5-µg/m^3 for $PM_{2.5}$. Adjusted for sex, log-transformation of age, weight, and height, parental education, maternal atopy, paternal atopy, breastfeeding, respiratory infections in the last 3 weeks (before the medical examination), Dutch nationality, indoor tobacco smoke exposure in the home, maternal smoking in pregnancy, furry pets in the home, mold in the home, gas cooking, and average air pollution concentrations for the 7 days preceding the lung function measurement. Note that scales differ per pollutant.

Table 10. Meta-analysis of Effect Estimates of Asthma Incidence Until Age 20 and Lung Function at Age 16 to Assess Level of Heterogeneity

		Asthma Incidence		Lung Function	
		Birth Address	Current Address	FEV_1	FVC
Pollutant	Number of Models[a]	I^2	I^2	I^2	I^2
Only SLR models for empirical models					
BC	7	64.2%	37.0%	42.3%	0.0%
NO_2	5	0.0%	0.0%	0.0%	0.0%
UFPs	4	0.0%	18.8%	0.0%	0.0%
$PM_{2.5}$	3	1.6%	0.0%	33.3%	0.0%
Including available SLR, RF, and Lasso models for empirical models					
BC	17	62.3%	39.5%	38.8%	0.0%
NO_2	12	17.8%	0.0%	28.9%	0.0%
UFPs	12	0.0%	33.3%	0.7%	0.0%
$PM_{2.5}$	6	19.6%	8.6%	0.0%	0.0%

[a]For the multiyear models, only the earliest available year is included (2013 for the BC dispersion model, 2010 for the rest). If both the ELAPSE and EXPANSE models for 2010 were available, only the EXPANSE model was included.

Table 11. Associations Between Model Performance R^2 at External Validation Sites and the Predicted Exposure Contrast (IQR) of Different Exposure Models and the OR for Asthma Incidence (Dependent Variable)[a]

			Performance R^2		IQR	
Pollutant	External Validation Set	Health Model	Beta (SE)	*P* Value	Beta (SE)	*P* Value
BC	NAQMN 2019	Performance R^2	0.91 (0.55)	0.12		
		Performance R^2 + IQR	1.06 (0.53)	0.06	–0.54 (0.30)	0.09
NO_2	NAQMN 2019	Performance R^2	–0.15 (0.15)	0.34		
		Performance R^2 + IQR	–0.07 (0.18)	0.70	–0.00 (0.01)	0.47
NO_2	Palmes 2019	Performance R^2	–0.11 (0.20)	0.59		
		Performance R^2 + IQR	–0.09 (0.19)	0.66	–0.00 (0.00)	0.28
NO_2	TRACHEA 2007	Performance R^2	–0.19 (0.09)	0.06		
		Performance R^2 + IQR	–0.17 (0.10)	0.11	–0.00 (0.00)	0.46
UFPs	New external validation 2021–2023	Performance R^2	0.81 (0.32)	0.03		
		Performance R^2 + IQR	0.82 (0.31)	0.03	0.00 (0.00)	0.24
$PM_{2.5}$	NAQMN 2019	Performance R^2	–6.11 (2.28)	0.02		
		Performance R^2 + IQR	1.40 (2.80)	0.63	–0.60 (0.18)	0.00

NAQMN = National Air Quality Monitoring Network.
[a]Fully adjusted models.

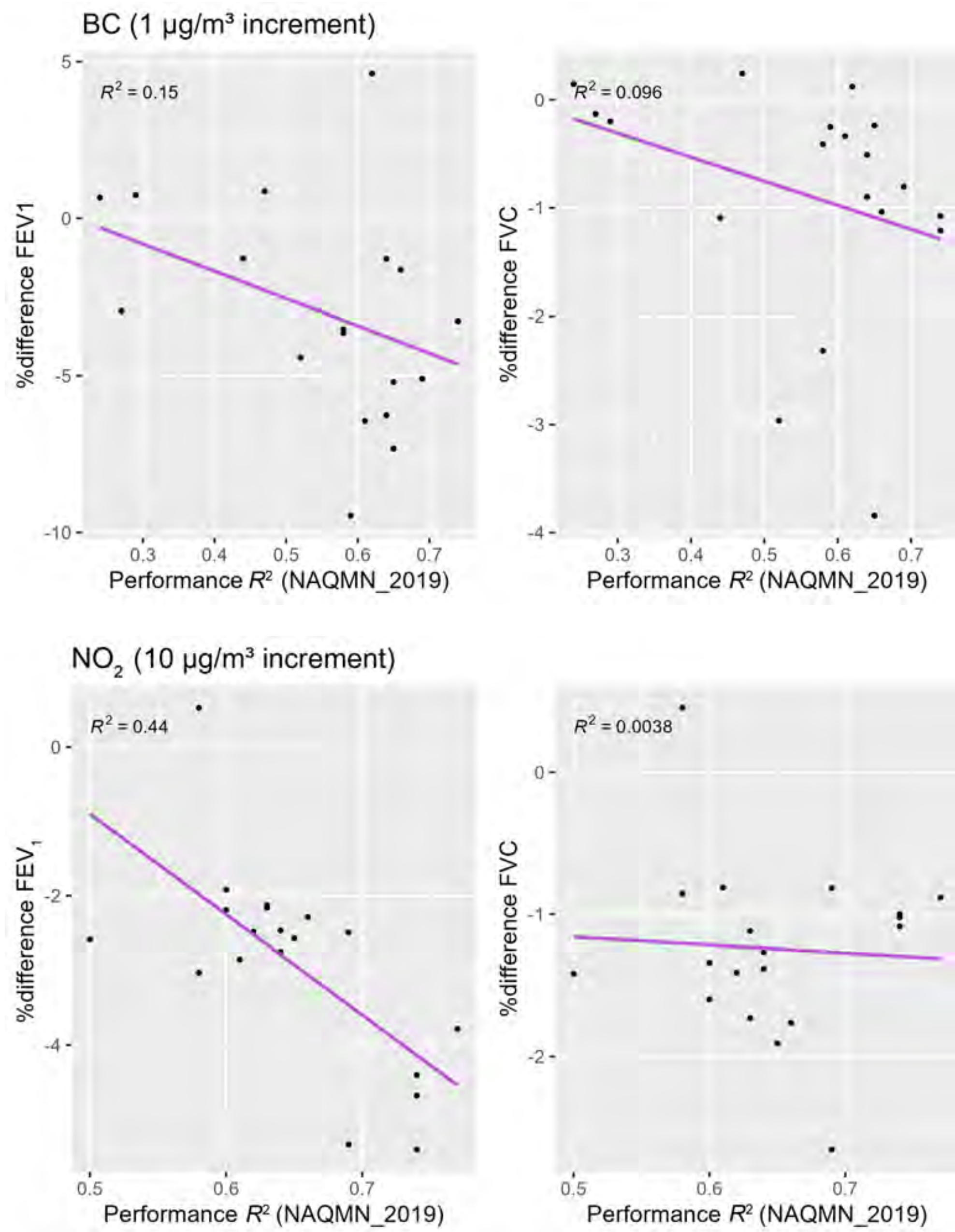


Figure 18. Associations between BC, NO_2, UFPs, and $PM_{2.5}$ with lung function of different exposure models in relation to the explained variance at NAQMN external validation sites in 2019. Most adjusted model. The purple line is the regression line.

continued on next page

5.9 COMPARISON OF EPIDEMIOLOGICAL FINDINGS: EPIC-NL COHORT

Population characteristics are presented in Additional Materials 1 Table S5. Exposure distributions are shown in Additional Materials 2 Table S31. The majority of the cohort was female, as one of the two subcohorts was female only. The average age was 50. Slightly more than 50% of the population was overweight or obese. About 40% of the population had never smoked.

Associations between air pollution exposure and natural-cause mortality are shown in **Figure 19** and Additional Materials 2 Table S32. Associations for BC, NO_2, and UFPs with natural-cause mortality were generally slightly below unity — consistently for all models, although the vast

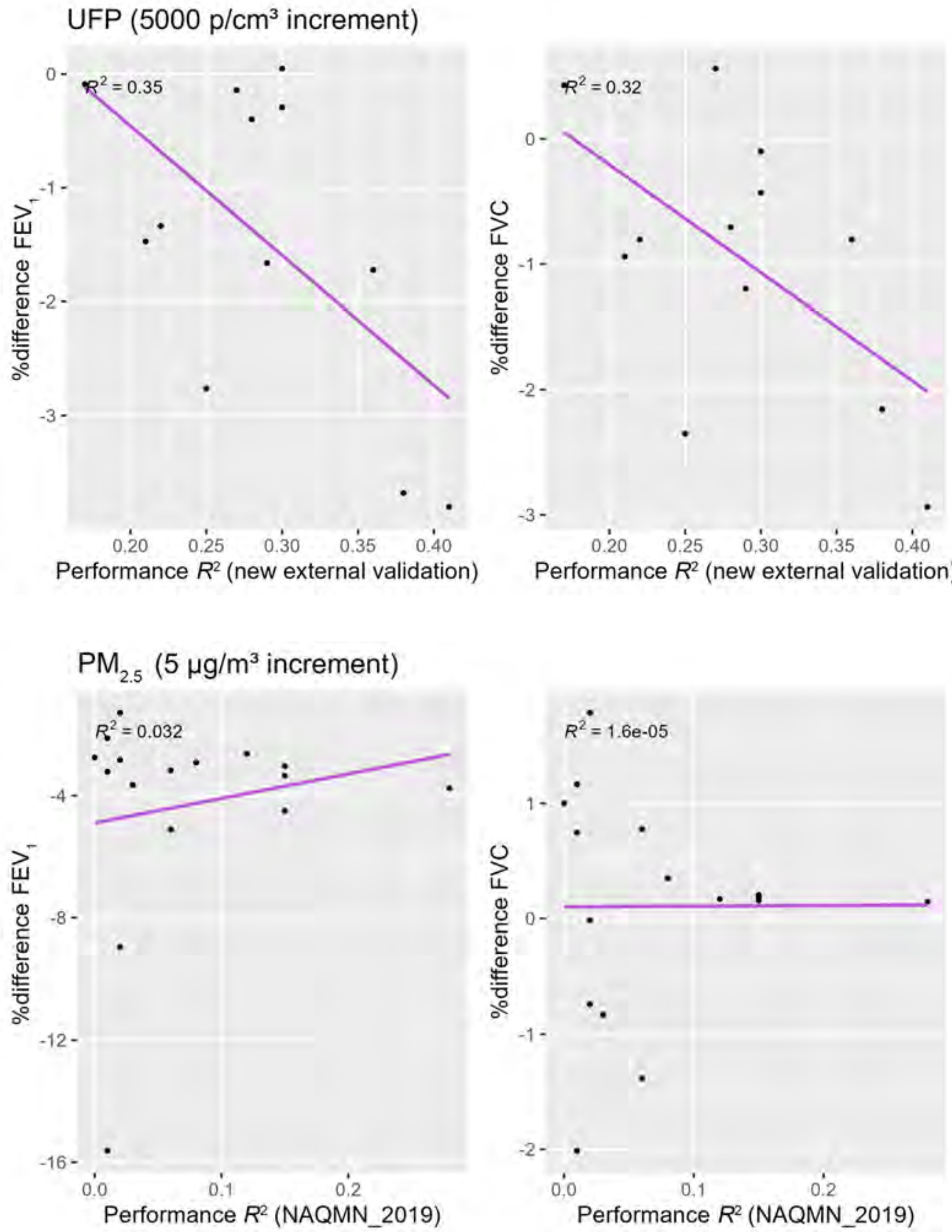


Figure 18. (*continued*)

majority of confidence intervals include unity. Confidence intervals were wider than for the large DUELS cohort.

For BC, HRs were all nonsignificantly below unity, with small differences between models and algorithms. For most models, the HR for Random Forest was closer to unity, but the confidence intervals overlapped widely with those of other models. For $PM_{2.5}$, differences in HRs were larger between models, along with wide confidence intervals. Consistent positive associations were found only with the EXPANSE models for the different years and the LUR model based on the new low-cost sensor measurements. Those significant associations with $PM_{2.5}$ are not consistent with the low explained variance of the $PM_{2.5}$ models. Associations with dispersion modeling, the ESCAPE models, the mobile monitoring models, and the 2010 ELAPSE model were null. Associations were similar for the SLR and Random Forest algorithms (ESCAPE and EXPANSE).

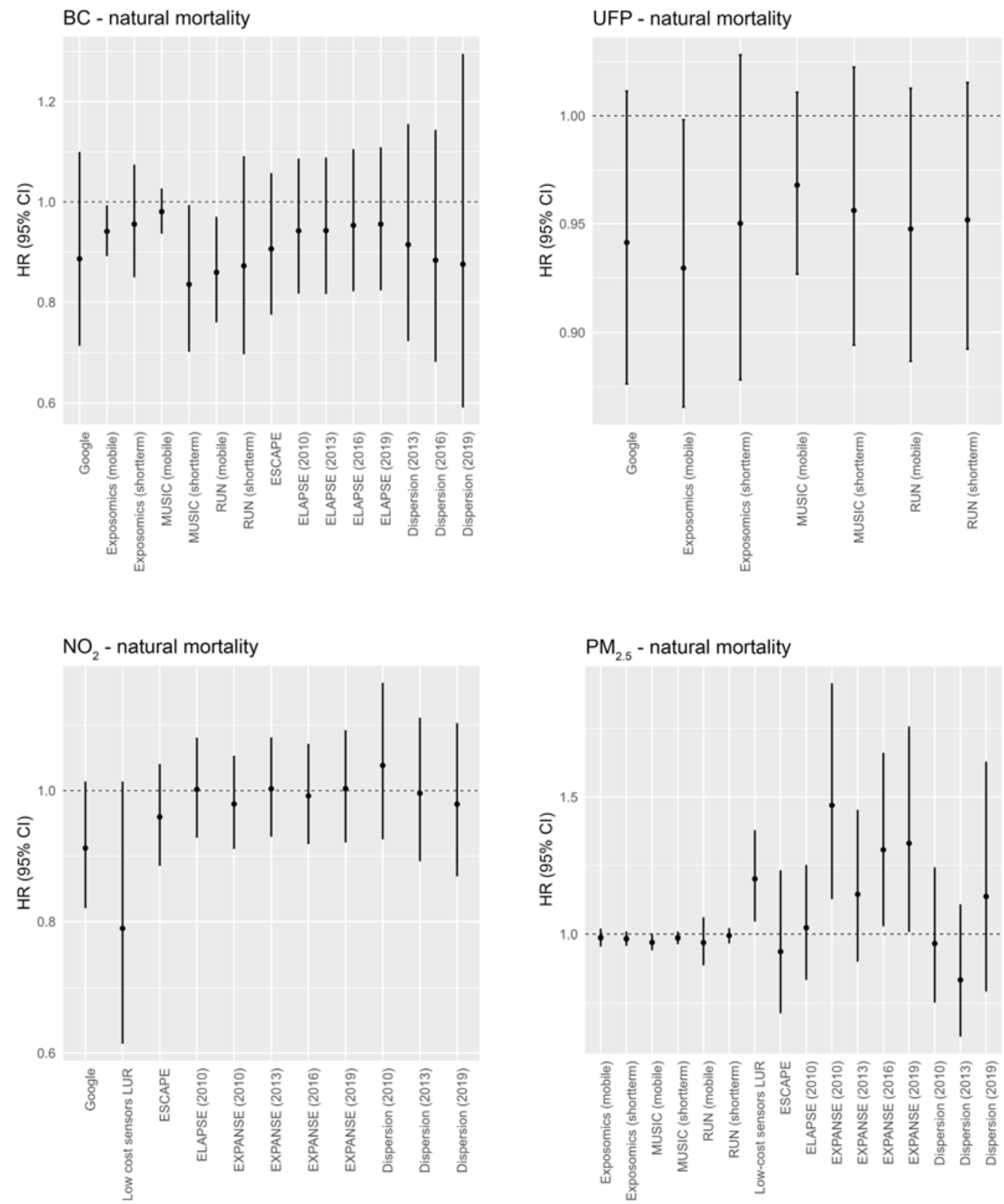


Figure 19. Associations between air pollution exposure and natural-cause mortality in EPIC-NL, most-adjusted model (*N* = 33,269). HR expressed per 1-μg/m³ for BC, 10-μg/m³ for NO_2, 5,000-p/cm³ for UFPs and 5-μg/m³ for $PM_{2.5}$. Covariates adjusted for include age (time axis), sex (strata), smoking status (never/former/current), smoking intensity (number of cigarettes/day), smoking duration (years of smoking), fruit and vegetable intake, alcohol consumption (low, medium, high), body mass index (BMI) (categories defined by WHO), educational level (low, medium, high), employment and occupational status (employed, unemployed, housekeeping, retired), marital status (single, married/with a partner, divorced, widowed) at the individual level, and mean income, ethnicity, and unemployment rate at the neighborhood level.

Associations between air pollution exposure and the incidence of coronary events and stroke are shown in Additional Materials 2 Table S33. Similar patterns as observed for natural-cause mortality were found, though with even wider confidence intervals. Associations for BC, NO_2, and UFPs were generally close to unity, consistently for all models, with confidence intervals that included unity in all but two occasions.

5.10 COMPARISON OF EFFECT ESTIMATES ACROSS THE THREE STUDIES

In this section, we evaluate the consistency of differences in effect estimates across the three cohorts for the different models, to get more insight in potential factors behind the differences. We evaluated the effect estimates in the DUELS

Table 12. Comparison of Models with Relatively High and Relatively Low Effect Estimates in the Large DUELS Cohort in the Two Other Cohort Studies

	Low Estimate					High Estimate				
Pollutant	DUELS Mortality	PIAMA Asthma	PIAMA FEV_1	EPIC-NL Mortality	Possible Reason (IQR, R^2)	DUELS Mortality	PIAMA Asthma	PIAMA FEV_1	EPIC-NL Mortality	Possible Reason
BC	Google-RF	L	L	0	IQR low					
	MUSIC_m_RF	L	L	0	R^2-val					
	MUSIC-ST_SLR	H	0	L	R^2-val					
NO_2	Google Lasso	0	L	0		Low-cost Lasso	H	H	L	IQR low
	Low-cost SLR	0	L	L	IQR low	Low-cost RF	H	0	0	IQR low
UFPs	MUSIC-m_SLR	0	L	0	IQR high	RUN-st-Lasso	H	H	L	IQR very low
	MUSIC-st-SLR	0	L	0	IQR high					
$PM_{2.5}$	ESCAPE-SLR	0	0	L						
	Low-cost SLR	L	L	H	R^2-val					

0 = average estimate; H = high estimate; L = low estimate; mortality = natural-cause mortality; val = validation.

cohort with natural-cause mortality, as the cohort with the most precise effect estimates. We identified models for which the HR was remarkably lower or higher than the majority of HRs, acknowledging that there is ambiguity in this analysis. We used Additional Materials 2 Table S11 to assess these models. All exposure models were evaluated, but not all models (e.g., ELAPSE and EXPANSE) showed remarkably high or low effect estimates. We then evaluated whether these exposure models resulted in low or high estimates in the other cohorts as well. Comparing the different cohorts and endpoints, we found some evidence of consistency of low and high effect estimates across studies (**Table 12**).

The low-cost sensor LUR models for NO_2 fairly consistently show either low or high effect estimates across the cohorts. The model predicts a much smaller IQR than the other models. Two of the MUSIC mobile (on-road) monitoring models showed deviating effect estimates, possibly related to the low validation R^2 of these models. There was no clear pattern with respect to algorithm. Moreover, we did not find clear differences in the model structure of the SLR models identified in Additional Materials 2 Table S1. Perhaps of note is that the MUSIC short-term roadside model for BC did not contain any buffers, and the low-cost sensor model of NO_2 included traffic intensity for heavy-duty vehicles only as a predictor.

6. DISCUSSION AND CONCLUSIONS

We compared exposure models that are commonly applied in epidemiological studies. The models differed widely in terms of modeling (deterministic dispersion models and empirical LUR models) and monitoring approaches (mobile, fixed-site). Empirical models differed additionally in the monitoring year. Empirical models were developed with three algorithms: SLR, Random Forest, and LASSO.

6.1 COMPARISON OF EXPOSURE BETWEEN DIFFERENT MODELS

Box 2 presents the observations for exposure comparisons. Our findings of fairly high correlations for diverse exposure models agree with the previous comparison of dispersion models and LUR models based on ESCAPE (de Hoogh et al. 2014; Klompmaker et al. 2021; Wang et al. 2015). Correlations were generally lower for $PM_{2.5}$ compared to more traffic-related pollutants in our current and earlier studies. An additional factor for the lower correlations for $PM_{2.5}$ may be due to the smaller contrast in exposure for $PM_{2.5}$ in the Netherlands. We did find sizable differences in level and exposure contrast, with mobile monitoring generally resulting in higher levels, related to the on-road monitoring. Bland-Altman plots showed that the difference in predicted exposures was not constant across the range of exposures for most comparisons. The higher exposure levels from models based on mobile monitoring compared to short-term roadside models have been documented before for the MUSIC and Exposomics mobile monitoring studies (Kerckhoffs et al. 2016, 2017). In these studies, the difference was about 30% and could not be explained by specific predictors such as traffic intensity.

Our study extends these observations to more diverse models, including between dispersion, fixed-site LUR, and mobile monitoring. We furthermore extended the comparison to include comparison at external validation sites, finding only modest differences between empirical models based on mobile monitoring, fixed-site monitoring, and dispersion modeling (mostly based on BC). A strength of our study was the use of external validation data to compare the performance of very different exposure models. Internal cross-validation statistics can be difficult to interpret and certainly impossible to compare across models based on datasets with different temporal resolutions (Klompmaker et al. 2015). Because of limited routine monitoring, validation of long-term exposure models has been challenging for unregulated pollutants such as BC and UFPs. This applies to both empirical models and deterministic dispersion models. The practice of using data at a small number of sites to validate dispersion models mixes temporal and spatial variation and cannot be interpreted as validation for spatial contrasts. We used existing data from different periods (2002–2019), allowing us to assess how well recent models predicted past and recent exposure contrasts. Remarkably, we found that models developed based on monitoring between 2014 and 2019 predicted the contrast in exposure in 2002 well for BC and moderately for UFPs.

Box 2. Exposure Observations and Conclusions

- Exposure predictions of BC, NO_2, and $PM_{2.5}$ for different years from 2010 to 2019 with the same method (dispersion model, European LUR model) were highly correlated, documenting stable spatial contrast patterns. Exposure levels of BC, NO_2, and $PM_{2.5}$ declined in this period.
- Different exposure models generally resulted in highly to moderately correlated exposure predictions. The correlation between methods was generally higher for the nationwide scale than for the major city of Amsterdam. The predicted level of exposure and exposure contrast differed substantially between methods and algorithms within methods.
- Small differences were found between the different exposure models and algorithms in explaining spatial variation at new and existing validation sites.
- Mobile monitoring studies generally resulted in modestly higher BC concentrations and exposure contrasts compared to other exposure methods. Mobile monitoring resulted in poorly performing models for $PM_{2.5}$.
- Models explained historical exposure patterns at external sites covering more than 10 years moderately well, especially for BC.
- Validation of long-term exposure models for the nonroutinely monitored pollutants BC and especially UFPs was challenging. Despite generally successful monitoring, the new external validation monitoring campaign with 1-week monitoring resulted in temporally unstable estimates of the spatial contrast across sites, especially for BC and $PM_{2.5}$. Grouping observations to reduce random error resulted in moderately high explained variance for UFP mobile monitoring models (R^2 = 58–69%)
- There was little benefit in using low-cost sensors for NO_2 and $PM_{2.5}$. The addition of low-cost sensor data did not improve NO_2 estimates in models that combined dispersion model estimates and data from the national monitoring network data. The NO_2 sensors were able to distinguish diurnal concentration patterns.

We generally found moderate to low agreement between exposure model predictions and measurements of UFPs and NO_2 at the new validation sites. The relatively low agreement is at least partly related to the modest temporal stability related to performing a single 1-week measurement per site to represent the annual average at all sites (for 31 sites two measurement weeks were available). The explained variance (R^2) obtained for the average of five ordered modeled and measured observations was 0.69, 0.58, 0.63, and 0.61 for the MUSIC, Google Airview, Exposomics, and RUN UFP mobile models, respectively. The explained variance was much higher than for individual observations. By averaging five observations (ordered based on the modeled exposure), random error in the measurements is reduced and a better estimate of the average exposure is obtained. The much higher explained variance supports that mobile monitoring of UFPs effectively models the variability of residential exposures. By calculating the average of five observations, we do lose the individual comparison. There was little difference in modeled exposure in the groups of five ordered observations.

We previously documented that an imprecisely estimated annual average concentration results in lower R^2 values (Kerckhoffs et al. 2016, 2017; Klompmaker et al. 2015; Montagne et al. 2015). Consistently, the R^2 for comparing model predictions of NO_2 and BC at existing validation sites with more temporally stable estimates was considerably higher. However, with both the new and existing validation data, comparisons between exposure models are still valid. For UFPs we did not have previous nationwide data available. Models explained intraurban spatial contrasts in 2014 and 2002 moderately well.

In the comparison of the explained variance of measurements at NAQMN sites between dispersion and empirical models, the performance of the dispersion models is underestimated because we did not perform specific modeling for the exact site but assigned the nearest residential address. This limitation does not apply to specific research studies, where monitoring was performed at specific residential addresses. As monitoring data have been used in the past to calibrate the models, a comparison at these sites between dispersion and empirical models is difficult to interpret. For NO_2, the TRACHEA study provides the cleanest comparison set, typically with small differences in performance between models.

The generally comparable results for the dispersion and the empirical models have been observed in a setting with a fine-resolution dispersion model that uses a detailed emission database and has been calibrated in previous years with a comprehensive national monitoring network for NO_2 and $PM_{2.5}$. These findings likely will not extend to coarse-scale models, or to study areas or pollutants with less reliable emission data such as UFPs.

The comparison between BC exposure models is moderately affected by the different metrics used to represent BC. In the ESCAPE and ELAPSE European models and the RUPIOH validation data, $PM_{2.5}$ absorbance was measured. In the mobile monitoring campaigns and NAQMN validation data, actual BC was measured with optical models. Dispersion models are expressed in EC. Correlations between the different metrics are generally very high, but differences in level of about 30% between the various metrics have been documented (HEI 2022). For UFPs, the Google Airview campaign used a different instrument (EPC3783) than the other mobile monitoring campaigns (CPC3007), albeit both devices captured particles 10 nm or larger. Differences in absolute levels are more difficult to quantify and are dependent on the particle size distribution.

Comparison of Exposure Predictions Between Different Algorithms Within Models We found no consistent differences in model performance and estimated health effects. The correlation of predictions between the three evaluated algorithms within the empirical models was generally moderate to high, suggesting that these algorithms can be used interchangeably. We note that these comparisons are not affected by temporal differences as the same monitoring data is used. Our study further extends the analysis presented by Kerckhoffs et al. (2019) for mobile and short-term monitoring studies of UFPs and Chen et al. (2020) for the Europewide hybrid models from ELAPSE for NO_2 and $PM_{2.5}$. Our conclusions agree with these studies in showing generally small differences in the prediction of exposure contrasts. In addition, we documented no consistent differences in validation statistics or health effect estimates. Findings may be different in different settings and may depend on whether internal or external validation data are used to compare model performance. In a study at a small number of fixed monitoring sites, Random Forest outperformed SLR in predicting particle concentrations (Brokamp et al. 2018). In a Canadian study, no difference in performance between a machine-learning model and linear regression was found for UFP mobile monitoring (Weichenthal et al. 2016). A review concluded that machine learning methods explained more variance in ambient air pollution than traditional linear models (Vachon et al. 2024).

We note that our study involved comparisons of predicting solely spatial contrasts. It has been speculated that in spatiotemporal settings, due to stronger nonlinear associations, the benefit of machine-learning models compared to SLR may be larger (see Discussion in Chen et al. 2019). In a study of wildfire exposure, machine-learning models consistently performed better than linear regression approaches, likely because these models can handle nonlinear relationships and interaction between predictors better (Reid et al. 2015). We further note we used default values in the Random Forest and LASSO models. Further tuning may increase the performance of these models, though, for the Europewide models, differences in performance between the untuned and tuned Random Forest models were minimal (Chen et al. 2019).

We consistently compared three algorithms in the different exposure models. However, more algorithmic variations in developing exposure models have been applied. In the Google Airview mobile monitoring study with measurements on all streets, a mixed-effect model was shown to modestly outperform other algorithms for NO_2 based on 82 sites with long-term passive sampler data (Kerckhoffs et al. 2022). In the nationwide mobile monitoring campaign RUN, the model that separated regional, urban, and local scale (deconvolution model) was the preferred model, although the differences in performance with SLR, Random Forest, and LASSO were small (Kerckhoffs et al. 2021).

6.2 COMPARISON OF HEALTH EFFECTS BETWEEN DIFFERENT MODELS

Box 3 presents the observations for comparisons of health effect estimates from different models. Our findings of fairly consistent associations for diverse exposure models with mortality in a large nationwide cohort agree with the findings of the US-based American Cancer Society study (Jerrett et al. 2017). In our study, we evaluated more diverse models including deterministic dispersion models, mobile monitoring models, and hybrid models based on fixed-site routine monitoring, the latter comparable to the models in the American Cancer Society study. Consistent with the American Cancer Society study, we also found large differences in HRs between different models, especially when expressed for fixed exposure contrasts. The smaller difference in HRs when expressed for model-specific IQRs documents that part of this difference is due to differences in the predicted exposure contrast. There are merits in comparing HRs for fixed increments (as is common in meta-analysis) and for model-specific increments (to investigate differences beyond the exposure contrast). In the case of highly correlated exposures, a smaller predicted exposure contrast will lead to a higher observed HR when expressed per fixed increment. In a previous analysis within the DUELS cohort, we also found consistent associations between the same dispersion model evaluated in this study, the national ESCAPE LUR model, and the Europewide ELAPSE hybrid LUR model for $PM_{2.5}$, BC, and NO_2 (Klompmaker et al. 2021). In the ELAPSE study, associations with mortality were generally consistent for the local ESCAPE models and the Europewide ELAPSE model (Brunekreef et al. 2021). In ELAPSE, overall associations between $PM_{2.5}$ and mortality were consistent between the satellite-based exposure model from the Canadian MAPLE study — a model with global coverage including Europe — and the ELAPSE model, though in local areas differences were found (Brunekreef et al. 2021; Chen et al. 2023).

Box 3. Health Effects Observations and Conclusions

- The year of the exposure model did not affect the presence of an association with mortality and morbidity, and modestly affected effect estimates related to the smaller exposure contrast.
- Consistent with the high correlation of the different exposure models, the application of these models generally resulted in similar conclusions on the presence of associations with natural-cause, respiratory, and lung cancer mortality in a large nationwide cohort as well as a large birth cohort for asthma and lung function. Effect estimates differed substantially, especially if expressed per fixed increments of exposure.
- There were no consistent differences in effect estimates between mobile monitoring, fixed-site LUR, and dispersion models as grouped methods.
- For cardiovascular mortality in the nationwide cohort and in particular for the health outcomes investigated in the smaller EPIC-NL cohort, differences between models were more variable, suggesting that for small effect estimates and small studies, differences in conclusions related to the exposure assessment method can be more distinct.
- Differences in exposure assessment may, therefore, contribute substantially to the observed heterogeneity of effect estimates in systematic reviews of epidemiological studies.
- Factors that explained some of the heterogeneity of effect estimates included the performance of the model at external validation sites, the predicted exposure contrast, and the choice of algorithm (for fixed increments).
- Application of SLR, Random Forest, and LASSO resulted in similar associations with health, especially if expressed per IQR.

Our findings for the lung function analysis in the PIAMA birth cohort agree with a previous analysis within the PIAMA birth cohort, showing comparable effect estimates for another dispersion model and the ESCAPE model of NO_2, $PM_{2.5}$, and BC for lung function at age 8 (Wang et al. 2015). High correlations (R = 0.86–0.90) between LUR- and dispersion-modeled concentrations for NO_2, $PM_{2.5}$, and BC were found.

The generally robust associations across very different exposure models further support the confidence in an association between outdoor air pollution and mortality and morbidity.

In the analyses of the large administrative cohort, we did not have data on lifestyle available. In an analysis of UFPs and mortality based on the nationwide RUN model performed in the framework of the current study, we documented only a small decrease in HRs using an indirect adjustment approach for smoking and body mass index, using the indirect adjustment models applied in the ELAPSE study (Bouma et al. 2023a).

In the EPIC-NL cohort, associations were mostly null or (nonsignificantly) negative. A lack of positive associations was found for most exposure models. Considering the relatively wide confidence intervals, our EPIC-NL effect estimates were mostly in line with previous papers in this cohort, albeit with lower point estimates (Brunekreef et al. 2021; Downward et al. 2018; Wolf et al. 2021). Additional Materials 2 Table S36 presents effect estimates in the current and previous studies.

6.2.1 Comparison of Health Effects Between Different Algorithms Within Models

The sometimes substantial differences in effect estimates for algorithms in individual models suggest the merit of using more than one algorithm in empirical models to evaluate the robustness of epidemiological findings. Our findings are consistent with a study within the ELAPSE project, in which fairly larger differences were found for some elemental components in association with natural-cause mortality in a large, pooled cohort (Chen et al. 2021). Despite differences in associations between individual models, we found no consistent difference in performance and estimated health effects, and there were generally moderate to high correlations of predictions between the three evaluated algorithms within the empirical models, suggesting these models can be used interchangeably.

In an analysis of the association between UFPs and allergic sensitization, we consistently found a null association by applying six different exposure estimates (the three algorithms evaluated with or without deconvolution) (Bouma 2023b).

6.2.2 Comparison of Health Effect Estimates Between Models from Different Years

In a setting of declining concentrations of BC, NO_2, and $PM_{2.5}$ in the period of 2010–2019, we documented very high correlations between exposure predictions of multiple years for BC, NO_2, and $PM_{2.5}$, consistent with previous observations for NO_2 (Cesaroni et al. 2012; Eeftens et al. 2011; Gulliver et al. 2013; Wang et al. 2013). We also documented very similar health effect estimates for exposure models from different years, both for dispersion models and Europewide hybrid LUR models. The high correlations imply that the year of monitoring of empirical models is likely not an important factor when comparing associations based on different models, particularly on the direction and presence of an association between exposure and health; however, the predicted concentration level and the exposure contrast may differ. We found only small differences in the magnitude of health effect estimates when applying models of different years, mostly related to a decreased exposure contrast in more recent years. However, in settings of declining concentrations, the level at which health effects occur will be affected by the choice of a more or less recent year. This will affect air quality standard setting.

Our study thus provides further support for the application of exposure models based on monitoring for years that are not fully temporally aligned with the follow-up period of epidemiological studies. It is quite common that exposure models are based on more recent years than the follow-up period of the epidemiological study, for example, because mobile monitoring of UFPs has been performed only in more recent years (Downward et al. 2018; Weichenthal et al. 2017) or because $PM_{2.5}$ monitoring on a sufficiently large scale only started recently in Europe (Stafoggia et al. 2022; Strak et al. 2021).

The conclusion of stable spatial contrasts applies to a setting with a largely developed urban structure. In rapidly developing areas or industrial settings where a major source closes down, this conclusion likely will not apply.

6.2.3 Heterogeneity of Effect Estimates Related to Exposure Assessment Model

The heterogeneity found between effect estimates for different exposure models for natural-cause mortality in the nationwide cohort was comparable to the reported overall heterogeneity in effect estimates in systematic reviews of outdoor air pollutants (Chen et al. 2020; HEI 2022; Huangfu and Atkinson 2020). In the PIAMA birth cohort, we found modest heterogeneity based on the I^2 statistic, consistent with the modest heterogeneity reported in previous meta-analyses. We cannot compare the I^2 statistics between DUELS and PIAMA, as the confidence intervals for DUELS were smaller, leading to larger I^2 values for the same difference in effect estimates between models. The large heterogeneity for natural-cause mortality in DUELS shows that differences in exposure assessment between studies may contribute substantially to observed heterogeneity in effect estimates across studies. This has been speculated previously based on between-studies comparisons of effect estimates and exposure methods but not documented systematically as studies differ in many aspects beyond the exposure assessment model. This further suggests that in evidence synthesis, one should be careful to downgrade confidence in the presence of an association based on heterogeneity statistics such as the I^2. For a further discussion on this issue, see (Boogaard et al. 2023b).

We note that there are more differences in the representation of exposure than in the models we have addressed. The averaging time of the exposure period and potential lags may affect the observed effect estimates, as has been documented in analyses of the MAPLE study (Brauer et al. 2022) and the ELAPSE study (Stafoggia et al. 2022; Strak et al. 2021). A few cohort studies on mortality that typically report high HRs have been affected by an analytical error that assigned exposure to events, with the average exposure from baseline to the date of the event for cases, and the average of the full follow-up period to the participants without events (noncases), see the explanation in (Ostro et al. 2010, including erratum). This error continues to occur in recent papers as well (Wang et al. 2023).

6.2.4 Two-Pollutant Models

Our finding of higher correlations between pollutants for dispersion models versus empirical LUR models is consistent with an earlier analysis in the DUELS administrative cohort (Klompmaker et al. 2021) and an earlier analysis in the PIAMA birth cohort (Wang et al. 2015). The implication is that it is more difficult to disentangle associations with specific pollutants when exposure has been assessed with dispersion models. One explanation for the higher correlation between pollutants is shared methodological processes of dispersion (e.g., the assumption that fine particles behave as inert gases at the local scale) and emissions in dispersion models. An alternative explanation is that correlations may be too low in empirical models related to partly arbitrary selection of variables, or too high because of shared GIS predictor variables.

6.2.5 Additional Findings

While not the main aim of our study, we provided evidence that long-term UFP exposure is associated with natural-cause and cause-specific mortality, both in single- and two-pollutant models adjusting for NO_2, $PM_{2.5}$, and BC in the large DUELS cohort (Bouma et al. 2023a). The evidence on the health effects of long-term exposure remains scarce, and our study strengthens the evidence base of the health effects of this currently unregulated pollutant. We did not find associations between UFPs and mortality from neurological diseases in DUELS, whereas we did find associations between $PM_{2.5}$ and NO_2 with dementia mortality (Peters et al. 2024).

6.3 FACTORS EXPLAINING VARIABILITY OF EFFECT ESTIMATES BETWEEN EXPOSURE MODELS AND ALGORITHMS

We evaluated whether the variability in effect estimates across models and algorithms could be explained by the performance of the exposure model at external validation sites, the predicted exposure contrast, the year of the model (temporal misalignment), the algorithm, part of the 24-hour covered, and the structure of the models. For the large DUELS cohort, we also calculated mean HRs for groups of models and algorithms based on natural-cause mortality (**Table 13**). Additional Materials 2 Table S35 shows the analyses conducted exploiting Bland-Altman plots in relation to differences in HRs between models from natural-cause mortality in DUELS. **Table 14** provides a summary of our findings.

Differences between the type of model (dispersion, mobile [both on-road and roadside], and fixed-site LUR) were small for BC and NO_2 (Table 13). For UFPs, only mobile monitoring models were assessed. For the three pollutants, HRs were similar for SLR and Random Forest algorithms and somewhat larger for the LASSO algorithm when expressed per fixed increments. The larger HR for LASSO was mostly related to a smaller IQR, as HRs did not differ much when expressed per IQR.

The slope from Bland-Altman plots especially explained the variability of the differences in HRs between models for BC (Additional Materials 2 Table S35). For NO_2, none of the statistics from the Bland-Altman plot was associated with differences in HRs. The magnitude of the relationship was small; a slope of 1 (a large slope) was associated with a 0.02 lower HR.

In the large administrative cohort, we found a weak tendency of higher HRs for models with a higher explained variance of NO_2 and BC concentrations at external validation sites. The weak relationship is in agreement with a simulation study by Szpiro and coworkers (Szpiro et al. 2011). For BC, we observed that the two models with low validation R^2 resulted in HRs close to unity. For $PM_{2.5}$, we found significant associations with natural-cause mortality for the low-cost sensor and mobile monitoring $PM_{2.5}$ models, which explained variation in validation measurements poorly. These LUR models include traffic and population predictors that represent traffic-related pollution, but not total $PM_{2.5}$. Traffic-related pollution is related to mortality as indicated by the findings for BC and NO_2. Hence, an exposure model with poor explanatory performance may still result in significant associations if applied in epidemiological studies, though with potentially severe under- or overestimation of the effect. The poorly performing In the PIAMA birth cohort, we found stronger evidence that, especially for lung function, models with higher performance at external validation sites had larger effect estimates. The higher effect estimates are consistent with the impact of lower measurement error, assuming the error has a large classical component.

In general, we found that the predicted exposure contrast was negatively associated with the magnitude of the effect

Table 13. Mean HRs for Natural-Cause Mortality in the DUELS Cohort Across Grouped Models and Algorithms

	BC		NO_2		UFPs	
	Per 1-µg/m³	Per IQR	Per 10-µg/m³	Per IQR	Per 5,000-p/cm³	Per IQR
Mobile monitoring LUR	1.041	1.013	1.020	1.007		
Fixed-site monitoring LUR	1.048	1.016	1.029	1.024		
Dispersion	1.063	1.014	1.025	1.016		
Supervised Linear Regression	1.032	1.013	1.027	1.019	1.015	1.009
Random Forest	1.034	1.011	1.039	1.021	1.016	1.009
Lasso	1.057	1.014	1.101	1.013	1.105	1.013

[a]HRs per fixed increments and IQR, most-adjusted models. See Figure 14 for models.

Table 14. Summary of Factors Affecting Magnitude of Effect Estimates

Factor	Hypothesis For Effect Estimates	Evidence From This Study	Comment
Exposure model performance	Higher, because of less measurement error	Moderate evidence in both Duels and PIAMA cohorts that higher validity results in higher effect estimates Strongest for lung function in PIAMA	Relationship is weak, consistent with Szpiro et al. (2011) Imprecision in effect estimates, type of measurement error (Berkson, classical), and confounding affect relationship
Predicted exposure contrast	Lower for higher contrasts of highly correlated models	Clear evidence in Duels and PIAMA by comparing HRs and ORs with fixed increments and model-specific IQR Table 11 documents significant differences in PIAMA for $PM_{2.5}$ and BC	Differences were occasionally substantial (e.g., in the case of some Lasso models and BC mobile monitoring models)
Deterministic (yes / no)	Both directions possible	No evidence was found for systematic differences	With dispersion models, more difficult to separate multiple pollutants because of higher correlation Dispersion models allow source-specific analyses
Algorithm (SLR/RF/ Lasso)	Both directions possible	Small differences, related to exposure contrast RF and Lasso smaller IQR for BC and UFPs HR for Lasso higher when expressed per fixed increment, identical when expressed per IQR	
Year of model (temporal misalignment)	Both directions	No impact on HR, as evidenced from applying the Europewide model and the dispersion model for multiple years	
Time of day covered (mobile monitoring)	Lower if only daytime covered (Blanco et al. 2023a, 2023b)	No differences in effect estimates between the three mobile monitoring studies excluding rush hour and the Google Airview studies Mobile models resulted in similar HRs compared to dispersion and fixed-site LUR	Limited ability to assess differences, as no mobile study covered 24-hour and weekend days A previous study suggested timing was not important, as spatial contrasts were stable across time of the day (Downward et al. 2018)
Model structure	Both directions	No clear differences	All models included both fine-scale and large-scale predictors

estimates expressed for fixed increments. Fixed increments are used in meta-analyses. The explanation for this finding is that a smaller contrast in the independent variable linked to the same contrast in health results in a steeper slope. A deviating predicted exposure contrast (IQR) from an exposure model may be a tool to identify potentially different effect estimates.

The year of the exposure model (section 6.2.2) did not affect the effect estimates.

We did not find consistent differences in effect estimates between deterministic and empirical models of exposure assessment. We argue that both models have merit, depending on, for example, the availability of solid fine-resolution emission data, understanding of physical-chemical processes, and the availability of major determinants of concentration contrasts in GIS. For a more detailed generic discussion, see (Hoek 2017; Jerrett et al. 2005;).

Differences in effect estimates can further be related to factors, which are hard to evaluate quantitatively and difficult to separate from other factors. For example, the domain of monitoring may affect how well the model reflects exposure contrasts. The number of sites may be important as well, but it is difficult to compare the number of street segments in mobile monitoring with sites in fixed-site monitoring, given the larger difference in temporal coverage.

6.4 LESSONS LEARNED ON VALIDATION DATA FOR NONREGULATED POLLUTANTS

We made a large effort to collect new validation data at 100 locations across the Netherlands, primarily for the less monitored pollutants UFPs and BC. We set out to measure each home for 2 weeks in different seasons. Mostly related to delays due to the COVID-19 pandemic, we were unable to achieve this goal. For 31 homes we obtained data for 2 weeks, while 59 homes recorded data for only 1 week. We found poor stability of the spatial contrast between sites based on comparing the two repeated measurements, especially for $PM_{2.5}$ and BC. For UFPs and NO_2 we found at least moderate stability. We found only limited improvement after correction for temporal variation at a continuous reference site in the center of the country. We had expected better stability after correction for temporal variation, based on previous studies in Amsterdam in 2002–2004 with a single 1-week measurement (Hoek et al. 2011), and Amsterdam–Utrecht with three 24-hour measurements in 2014 across the year (van Nunen et al. 2017). We speculate that a reduced spatial contrast in exposure in recent years may have contributed to this finding. Furthermore, the fact that measurements were conducted across the entire country may have made the correction for temporal variation less effective than in the two previous studies we based our design on. Hence, the continuous monitoring instruments for BC and UFPs were less robust for weeklong measurements than we anticipated based on previous studies (Kerckhoffs et al. 2021). However, this does not fully explain the findings, as integrated sampler measurements were robust for $PM_{2.5}$, and the temporal correlation across locations across the Netherlands for $PM_{2.5}$ was very high. One reason for the poor stability was that measurements could not be performed simultaneously for all sites, because of limitations in the availability of equipment. Given the cost of instruments, simultaneous measurements at a large number of locations (e.g., 40 or more) will not be feasible in the near future.

The limited stability of the new external validation data especially affected the explained variance R^2, and the evaluation of bias to a lesser extent. For UFPs and NO_2, we still used the data to compare different models. The absolute value of the R^2 is likely underestimated, but differences across exposure models are less affected.

Lessons Learned from the Current Validation Effort When monitoring data is used for validation purposes, the requirement for temporal completeness is more crucial than for model development. The approach with nonsimultaneous weekly data and correction with a reference site in a nationwide setting did not result in sufficiently stable estimates per site to assess how well models explain the variability of concentrations. For assessing bias in predicted concentrations, the approach worked sufficiently well. Furthermore, when we calculated the correlation between the average of five ordered observations for modeled and measured exposures in the full set of 90 locations, we found much higher explained variances for UFPs (up to R^2 = 69%) compared to the individual observations.

There are no easy solutions to this problem for future studies, depending on the resources available. Covering a larger fraction of time and doing more simultaneous measurements is an obvious solution, but one that requires a larger number of instruments than we had available, despite a sizable instrument budget and instruments available from previous projects. In a previous study, we calculated intraclass correlation coefficients and ratios of between and intrasite variability (Klompmaker et al. 2015). Intraclass correlation coefficients and ratios increased sizably when a larger amount of time was covered per site, comparing short-term roadside campaigns (3 × 30 minutes per site as in MUSIC), 1 week as in RUPIOH, six times a week in a personal monitoring project, and three times in 14 days in ESCAPE (Klompmaker et al. 2015). Temporal correction with a reference site improved intraclass correlation coefficients and variance ratios for the longer-duration campaigns, but not the short-term campaign (Klompmaker et al. 2015). In our study, the temporal correction for NO_2 improved the stability. The rationale for using a background site is the generally high temporal correlation of 24-hour average concentrations across metropolitan and larger areas, forming the basis of epidemiological time series studies of short-term exposure health effects. For shorter-duration exposures, the correlation is lower. In a study in Amsterdam and three other European cities, high temporal correlations were found across the city for $PM_{2.5}$, $PM_{2.5}$ absorbance, and UFPs: 0.98, 0.94, and 0.76 of 24-hour average exposures (Puustinen et al. 2007). In a large study area such as the Netherlands, correlations may be lower, particularly for local pollutants such as UFPs. For regional background sites across the country, the correlation of 1-week average UFP concentrations at sites across the country with the reference site was 0.58 (Kerckhoffs et al. 2021).

Simultaneous measurements have been performed in the TRACHEA study, with four times 1-week samples with passive samplers at 144 sites (Eeftens et al. 2011). The dataset resulted in high explained variances of the different models in the current study. This illustrates that validation does not need to cover a full year. Simultaneous sampling at all sites was feasible because of the availability of passive samplers for NO_2.

Considerations for Validation of Long-Term Exposure Predictions It is crucial that validation involves spatial comparison between modeled and measured concentrations. Temporal comparisons using a single or a few routine monitoring sites do not inform how well a model performs spatially.

1. The need for spatial validation creates a major issue for the validation of exposure predictions of nonregulated pollutants because of limited routine monitoring. Until now, no good solutions for this problem have been developed.

2. The balance between number of repeats and number of sites may be shifted more toward number of repeats

than number of sites. We spent 40 1-week measurement sessions (with three sites and a reference site simultaneously) to measure 90 sites once, and at 31 of those sites, an additional 1-week measurement was performed in another season. With the same (large) effort, we could have collected 1-week measurements six times at 20 sites, four times at 30 sites, and three times at 40 sites.

3. The number of sites should be sufficient to calculate a meaningful correlation. A minimum of 20 sites may be a rule of thumb, but preferably more.
4. The distribution of site types is important. The selected sites should cover the variability of sources, including motorized road traffic. Overrepresentation of traffic sites as done in our study generally improves exposure contrast.
5. Simultaneous measurements, at as many sites as feasible, are preferable.
6. A reference site with continuous measurements is helpful, though this may not be universally true, especially if the study area is large and includes multiple airsheds.
7. More routine monitoring of BC and UFPs at multiple sites is needed, which at the minimum would reduce the need for a reference site.
8. More robust instruments than the DiskMini and MA200 at a reasonable cost may improve precision for long-term monitoring.
9. Development of lower-cost instruments for UFPs and BC would allow more simultaneous deployment.
10. One current option is to perform repeated mobile monitoring on a sizable number of roads differing in traffic intensity and other important source influences. This recommendation is based on the observation that after about 20–25 repeats (Apte et al. 2017; Kerckhoffs 2022), the overall average is reliably measured. This approach only validates the on-road concentration domain; it does not address differences between on-road and residential locations.

6.5 MOBILE MONITORING AND LOW-COST SENSORS

Mobile Monitoring Studies For both BC and NO_2, we found that exposure predictions at residential addresses from mobile (both on-road and short-term roadside) monitoring were highly correlated with predictions from dispersion models and hybrid LUR models based on fixed-site monitoring. Mobile monitoring studies did result in higher concentration predictions and larger contrast in exposure. We did find higher positive bias (overestimation) for the mobile (on-road) monitoring-based models compared to other approaches, and compared to the short-term roadside monitoring models. Using external validation, we found no consistent differences in explained variance between mobile monitoring (both on-road and short-term roadside) and the other approaches. We found no consistent differences in the size of the health effects estimated from mobile monitoring and other models. This provides strong support for applying mobile monitoring approaches, especially for pollutants with limited routine monitoring.

In the first mobile monitoring studies (MUSIC, Exposomics, and RUN), we purposefully did not include rush-hour monitoring. The rationale for that was to increase comparability between monitoring at different locations (Montagne et al. 2015; van Nunen et al. 2017). An analysis made on week-long continuous UFP monitoring data at 46 sites in Amsterdam in 2002–2004 found that intraurban exposure contrasts for specific hours were highly correlated (Downward et al. 2018). Moreover, absolute levels of daytime concentration excluding rush hour deviated less than 20% from overall averages (Downward et al. 2018). In contrast, a series of studies from the US based on short-term roadside monitoring reported that short-term roadside monitoring designs were more biased when they did not cover the full day (in this case from the hours of 5:00 to 23:00) and all days of the week (weekdays and weekends) (Blanco et al. 2023a; Blanco et al. 2023b). In the Google Airview mobile monitoring campaign, a longer time period was included (8–22 hours). We did not find a higher performance of the Google Airview model compared to the previous models. The other exposure models and the validation data covered the full 24-hour period.

Mobile monitoring resulted in poorly performing models for $PM_{2.5}$, suggesting that mobile monitoring is not an attractive approach for this pollutant. Few previous studies have attempted to develop models based on mobile monitoring for $PM_{2.5}$. The poor performance is consistent with $PM_{2.5}$ being mostly a pollutant that varies on a regional scale with a large background and generally small local source contribution (Eeftens et al. 2012; HEI 2022). For $PM_{2.5}$, routine monitoring, satellite monitoring, and models based on them, and dispersion models have been well established.

We found evidence that empirical models can be applied outside the domain they were developed in. The Google Airview BC and NO_2 models developed based on monitoring for the city of Amsterdam performed well in explaining variation across the Netherlands. This agrees with previous findings reported for the MUSIC study, in which BC short-term roadside monitoring models explained nationwide variations of BC measured in ESCAPE well (Montagne et al. 2015). The explanation for this finding is that models included large-scale predictors, such as a 5,000-meter buffer for population density, and land use variables such as distance to a port or industrial area, to explain differences between large urban areas and smaller towns. The high correlation between predictions of our national UFP model (RUN) with models developed in the major cities only (MUSIC, Exposomics), is in agreement with this interpretation.

Low-Cost Sensors There was little benefit in using low-cost sensors for NO_2 and $PM_{2.5}$. While the investment cost for sensors was low, a large personnel effort was needed to obtain interpretable data (communication, calibration). Furthermore, the NO_2 sensor required individual calibration with a large dynamic range of concentrations for a period of more than a month. The calibration formula for NO_2 was stable only for a few months, after which a night calibration with routine monitoring was applied. Finally, a data fusion model documented that the low-cost sensor network hardly improved concentration fields compared to the national dispersion model estimates and NAQMN monitoring alone. In settings with less dense routine monitoring and less developed dispersion models, low-cost sensor models may be more informative.

For NO_2, well-established passive samplers such as the Palmes tube and Ogawa badge may be the model of choice, if the interest is primarily in long-term exposure. Low-cost sensors do provide diurnal variation, which can be useful, for example, when linked to time–activity data.

6.6 IMPACT OF COVID-19 PANDEMIC AND ASSOCIATED LOCKDOWN

The COVID-19 pandemic affected the study primarily by causing delays in the start of the fieldwork for the new validation campaign and the low-cost sensor network. Because of the lockdown and the associated unusual traffic conditions, the fieldwork of both studies was delayed until September 2021. This avoided the most stringent lockdowns. A second impact was that due to the delays, we were unable to complete the two repeats at all monitoring sites of the new validation campaign. During a short period between the middle of December 2021 and February 2022, a stricter lockdown was in place, possibly affecting the measurements. The absolute concentrations measured between March 2022 and April 2023 may have been affected more moderately as traffic levels were not fully back to the pre-COVID-19 period. An analysis by RIVM estimated that the lockdown in early 2020 reduced observed NO_2 concentrations by 30% (95% CI: 25–35%), 26% (21–32%), and 18% (10–25%) for traffic, urban, and rural background locations, respectively (Velders et al. 2021). Slightly smaller reductions of 8–28% were found with the EMEP4NL simulations for urban and regional background locations. Reductions in observed $PM_{2.5}$ concentrations of about 20% (10–25%) were found for all locations, which is somewhat larger than the estimates of 5–16% based on the model simulations. The contrast in exposure between locations was thus less affected than the absolute levels. This is further supported by a separate analysis we performed that showed that the contrast in exposure between traffic and background sites of the NAQMN did not differ between holidays (characterized by lower traffic intensity) and nonholidays. None of the models and none of the existing validation data were affected by the COVID-19 pandemic, as they were developed before 2020. Overall, our conclusions are not materially affected by the lockdowns related to COVID-19.

6.7 STRENGTHS AND LIMITATIONS

We systematically compared exposure models that are widely applied in air pollution epidemiology. Models ranged from deterministic to a variety of empirical models based on mobile and fixed-site monitoring data. We compared exposure predictions and supplemented that with evaluations at new and existing validation sites, covering a 20-year period. We further compared health effect estimates in adult and child cohorts.

Comparisons have been made of exposure models within the Netherlands, with models being developed for a period of 10 years (2009–2019). In other study areas and periods, comparisons between models may have been different, an issue discussed more specifically before; for example, with respect to year.

Exposure models differed in principle but also in monitoring years. Differences between models could be attributed to model or time in most cases, as for BC, NO_2, and $PM_{2.5}$, we had data from multiple years. For UFPs, this was more challenging because of the lack of data for the same model over multiple years.

We developed SLR models for all empirical models and Random Forest for all except the ELAPSE European BC model. LASSO models were attempted for all mobile monitoring campaigns and ESCAPE. LASSO models were not available for multiple years of the EXPANSE project, and we did not have the resources to develop these models. We consider the comparison between the easy-to-interpret SLR and the more sophisticated Random Forest machine-learning model as the most important algorithm comparison. We compared models and algorithms differing widely in principle, but acknowledge that more differences in assigned exposure may occur; for example, related to the temporal representation of exposure (lag and averaging period) (Brauer et al. 2022).

We assigned exposures to the residential address at baseline for the two adult cohorts. For the nationwide cohort, this was a recent address (2013). For the EPIC-NL study, the baseline address was the mid-1990s. Because our main objective was to compare effect estimates between models, only baseline exposure was considered reasonable. In the PIAMA birth cohort study, exposures were assigned to the birth address and the current address during the specific survey.

We furthermore focused on differences in outdoor exposure only. We did not address indoor–outdoor relationships and the inclusion of time–activity data. Particularly, the latter may introduce differences if the predicted spatial contrast differs between models. The topic will be addressed in an upcoming HEI report by Kees de Hoogh (Swiss Tropical and Public Health Institute, Allschwil, Switzerland) funded under the same Request for Application as the current study. A review conducted in the framework of that study concluded that there is little bias in assessing outdoor exposure at the residential address only (Hoek et al. 2024).

We compared four major air pollutants, but traffic-related air pollution includes more pollutants including nontailpipe elements, carbon monoxide, and polycyclic aromatic hydrocarbons. Comparisons may be different for these pollutants, which have fewer related epidemiological studies.

7. IMPLICATIONS OF FINDINGS

We discuss implications for exposure assessment in epidemiological studies, interpretation of epidemiological studies, evidence synthesis of epidemiological studies, health impact assessment, and policy. The implications follow from the previous discussion and are, therefore, brief.

Exposure Assessment Models based on mobile monitoring data are useful to assess residential exposure in epidemiological studies, although models are based on on-road measurements. Exposure levels and contrasts may be larger than for empirical models based on fixed-site measurements. Mobile models may potentially be applied in a wider area than the monitoring domain, depending on the model structure. By driving in surrounding towns of a major city, the model structure can more generally be applicable compared to monitoring within the city only.

Exposure assessment based on current models can be applied to assess exposures back in time, for at least up to a decade before. These findings apply to settings of established infrastructure and urban areas, comparable to the Netherlands.

In studies of long-term exposure, there was little benefit in using low-cost real-time sensors for NO_2 and $PM_{2.5}$. For NO_2, well-established passive samplers such as the Palmes tube and Ogawa badge may be the model of choice to measure at large numbers of locations.

Additional routine monitoring of UFPs is recommended to provide much-needed data on temporal trends of this scarcely measured pollutant.

Epidemiological Studies Based on validation performance, no clear preference for a specific exposure assessment or model development algorithm can be stated. Health effect estimates of studies using different exposure models can be integrated. The application of multiple models in individual studies such as deterministic and empirical models is useful to inform on the robustness of epidemiological findings.

The application of multiple algorithms to develop models based on monitoring data is also useful for evaluating the robustness of epidemiological findings. Based on validation performance, no clear preference for a specific algorithm can be stated when spatial contrasts at residential locations are evaluated.

Two-pollutant models may be more difficult to interpret when dispersion models are used compared to empirical models, because of the generally higher correlations across pollutants.

Evidence Synthesis and Health Impact Assessment The finding of associations between air pollution and mortality and morbidity with very different exposure models adds to the confidence in these associations.

We have shown that large heterogeneity in effect estimates occurs due to differences in exposure models in the same epidemiological study. This was previously speculated, but difficult to document, as between-study comparison of effect estimates is affected by multiple differences between studies. In most cases, heterogeneity was due to the magnitude of the effect estimate, much less due to the presence of an association. This further suggests that in evidence synthesis, one should be careful to downgrade confidence in the presence of an association based on heterogeneity statistics such as the I^2. These statistics do not differentiate between the direction and magnitude of the association.

The observed heterogeneity has implications for the debate on whether to use local effect estimates from a single or few local studies (with a specific exposure assessment model) versus those from the wider body of evidence in health impact assessment. In this debate, heterogeneity of effect estimates across locations and potential unadjusted biases due to unmeasured covariates in single studies should additionally be considered.

The difference in predicted exposure levels between models raises concern when population exposure for the health impact assessment is assessed with different models than those used in the epidemiological studies used to derive the exposure–response function.

Policy If exposure models are used to check compliance with air quality limit values or standards, the large differences in level between models may result in large differences in conclusions regarding meeting the standard across the country. In practice, most empirical models are not used for compliance testing.

The finding of independent associations between long-term UFP exposure and mortality suggests that to protect public health, additional policies may be needed beyond those addressing $PM_{2.5}$ and NO_2.

DATA AVAILABILITY STATEMENT

Because of the European privacy regulations (GDPR), individual cohort data cannot be shared. R code and anonymized exposure data are available from the authors.

ACKNOWLEDGMENTS

This work additionally was supported by an ASPASIA grant from the Dutch Research Council (NWO) to Dr. Ulrike Gehring (project number 015.010.044), the Environmental Defense Fund, EXPOSOME-NL (NWO grant number 024.004.017), EXPANSE (EU-H2020 Grant number 874627) and the Dutch Ministry of Infrastructure and Water Management (grant M/240121 and part of E/122521 supported by the Innovation Program for Environmental Monitoring). We thank Kees Meliefste and Matthijs Dijk for their technical support of the fieldwork, and Daan Aarts for contributing to the modeling of low-cost data and analysis of the EPIC-NL cohort study.

REFERENCES

Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, et al. 2017. High-resolution air pollution mapping with Google Street View cars: exploiting big data. Environ Sci Technol 51:6999–7008, *https://doi.org/10.1021/acs.est.7b00891*.

van de Beek E, Kerckhoffs J, Hoek G, Sterk G, Meliefste K, Gehring U, et al. 2021. Spatial and spatiotemporal variability of regional background ultrafine particle concentrations in the Netherlands. Environ Sci Technol 55:1067–1075, *https://doi.org/10.1021/acs.est.0c06806*.

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli, X, et al. 2013. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe: the ESCAPE project. Atmos Environ 72:10–23, *http://dx.doi.org/10.1016/j.atmosenv.2013.02.037*.

Blanco MN, Doubleday A, Austin E, Marshall JD, Seto E, Larson TV, et al. 2023a. Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. J Expo Sci Environ Epidemiol 33:465–473, *https://doi.org/10.1038/s41370-022-00470-5*.

Blanco MN, Bi J, Austin E, Larson TV, Marshall JD, Sheppard L. 2023b. Impact of mobile monitoring network design on air pollution exposure assessment models. Environ Sci Technol 57:440–450, *https://doi.org/10.1021/acs.est.2c05338*.

Boogaard H, Samoli E, Patton AP, Atkinson RW, Brook JR, Chang HH, et al. 2023a. Long-term exposure to traffic-related air pollution and non-accidental mortality: a systematic review and meta-analysis. Environ Int 176:107916, *https://doi.org/10.1016/j.envint.2023.107916*.

Boogaard H, Atkinson RW, Brook JR, Chang HH, Hoek G, Hoffmann B, et al. 2023b. Evidence synthesis of observational studies in environmental health: lessons learned from a systematic review on traffic-related air pollution. Environ Health Perspect 131:115002, *https://doi.org/10.1289/ehp11532*.

Bouma F, Janssen NA, Wesseling J, van Ratingen S, Strak M, Kerckhoffs J, et al. 2023a. Long-term exposure to ultrafine particles and natural and cause-specific mortality. Environ Int 175:107960, *https://doi.org/10.1016/j.envint.2023.107960*.

Bouma F, Hoek G, Koppelman GH, Vonk JM, Kerckhoffs J, Vermeulen R, et al. 2023b. Exposure to ambient ultrafine particles and allergic sensitization in children up to 16 years. Environ Res 219:115102, *https://doi.org/10.1016/j.envres.2022.115102*.

Brauer M, Brook JR, Christidis T, Chu Y, Crouse DL, Erickson A, et al. 2022. Mortality–Air Pollution Associations in Low Exposure Environments (MAPLE): Phase 2. Research Report 212. Boston, MA: Health Effects Institute.

Brokamp C, Jandarov R, Hossain M, Ryan P. 2018. Predicting daily urban fine particulate matter concentrations using a Random Forest model. Environ Sci Technol 52:4173–4179, *http://dx.doi.org/10.1021/acs.est.7b05381*.

Brunekreef B, Strak M, Chen J, Andersen ZJ, Atkinson R, Bauwelinck M, et al. 2021. Mortality and Morbidity Effects of Long-Term Exposure to Low-Level $PM_{2.5}$, BC, NO_2, and O_3: An Analysis of European Cohorts in the ELAPSE Project. Research Report 208. Boston, MA: Health Effects Institute.

Carey IM, Atkinson RW, Kent AJ, van Staa T, Cook DG, Anderson HR. 2013. Mortality associations with long-term exposure to outdoor air pollution in a national English cohort. Am J Respir Crit Care Med 187:1226–1233, *https://doi.org/10.1164/rccm.201210-1758oc*.

Cesaroni G, Porta D, Badaloni C, Stafoggia M, Eeftens M, Meliefste K, Forastiere F. 2012. Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. Environ Health 11:48, *https://doi.org/10.1186/1476-069x-11-48*.

Chen J, Braun D, Christidis T, Cork M, Rodopoulou S, Samoli E et al. 2023. Long-term exposure to low-level $PM_{2.5}$ and mortality: investigation of heterogeneity by harmonizing analyses in large cohort studies in Canada, United States, and Europe. Environ Health Perspect 131:127003, *https://doi.org/10.1289/ehp12141*.

Chen J, Hoek G. 2020. Long-term exposure to PM and all-cause and cause-specific mortality: a systematic review and meta-analysis. Environ Int 143:105974, *https://doi.org/10.1016/j.envint.2020.105974*.

Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ Int 130:104934, *https://doi.org/10.1016/j.envint.2019.104934*.

Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. 2020. Development of Europe-wide models for particle elemental composition using supervised linear regression and random forest. Environ Sci Technol 54:15698–15709, *https://doi.org/10.1021/acs.est.0c06595*.

Chen J, Rodopoulou S, de Hoogh K, Strak M, Andersen ZJ, Atkinson R, et al. 2021. Long-term exposure to fine particle elemental components and natural and cause-specific mortality: a pooled analysis of eight European cohorts within the ELAPSE project. Environ Health Perspect 129:47009, *https://doi.org/10.1289/ehp8368*.

Cyrys J, Eeftens M, Heinrich J, Ampe C, Armengaud A, Beelen R, et al. 2012. Variation of NO_2 and NO_x concentrations between and within 36 European study areas: results from the ESCAPE study. Atmos Environ 62:374–390, *http://dx.doi.org/10.1016/j.atmosenv.2012.07.080*.

Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. 2016. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environ Sci Technol 50:4712–4721, *https://doi.org/10.1021/acs.est.5b06121*.

Downward GS, van Nunen EJHM, Kerckhoffs J, Vineis P, Brunekreef B, Boer JMA, et al. 2018. Long-term exposure to ultrafine particles and incidence of cardiovascular and cerebrovascular disease in a prospective study of a Dutch cohort. Environ Health Perspect 126:127007, *https://doi.org/10.1289/ehp3047*.

Eeftens M, Beelen R, Fischer P, Brunekreef B, Meliefste K, Hoek G. 2011. Stability of measured and modelled spatial contrasts in NO_2 over time. Occup Environ Med 68:765–770, *https://doi.org/10.1136/oem.2010.061135*.

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. 2012. Development of land use regression models for PM(2.5), PM(2.5) absorbance, PM(10), and PM(coarse) in 20 European study areas; results of the ESCAPE project. Environ Sci Technol 46:11195–11205, *https://doi.org/10.1021/es301948*.

Fischer PH, Marra M, Ameling CB, Hoek G, Beelen R, de Hoogh K, et al. 2015. Air pollution and mortality in seven million adults: the Dutch Environmental Longitudinal Study (DUELS). Environ Health Perspect 123:697–704, *https://doi.org/10.1289/ehp.1408254*.

Gehring U, Beelen R, Eeftens M, Hoek G, de Hoogh K, de Jongste JC, et al. 2015. Particulate matter composition and respiratory health: the PIAMA birth cohort study. Epidemiology 26:300–309, *https://doi.org/10.1097/ede.0000000000000264*.

Gehring U, Gruzieva O, Agius R, Beelen R, Custovic A, Cyrys J, et al. 2013. Air pollution exposure and lung function in children: the ESCAPE project. Environ Health Perspect 121:1357–1364, *https://doi.org/10.1289/ehp.1306770*.

Gehring U, Wijga AH, Koppelman GH, Vonk JM, Smit HA, Brunekreef B. 2020. Air pollution and the development of asthma from birth until young adulthood. Eur Respir J 56:2000147, *https://doi.org/10.1183/13993003.00147-2020*.

Gulliver J, de Hoogh K, Hansell A, Vienneau D. 2013. Development and back-extrapolation of NO_2 land use regression models for historic exposure assessment in Great Britain. Environ Sci Technol 47:7804–7811, *https://doi.org/10.1021/es4008849*.

Health Effects Institute (HEI). 2022. Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. A Special Report of the HEI Panel on the Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute.

Hoek G. 2017. Methods for assessing long-term exposures to outdoor air pollutants. Curr Environ Health Rep 4:450–462, *https://doi.org/10.1007/s40572-017-0169-5*.

Hoek G, Beelen R, Kos G, Dijkema M, van der Zee SC, Fischer PH, et al. 2011. Land use regression model for ultrafine particles in Amsterdam. Environ Sci Technol 45:622–628, *https://doi.org/10.1021/es1023042*.

Hoek G, Meliefste K, Cyrys J, Lewné M, Bellander T, Brauer M, et al. 2002. Spatial variability of fine particle concentrations in three European areas. Atmos Environ 36:4077–4088, *https://doi.org/10.1016/S1352-2310(02)00297-2*.

Hoek G, Vienneau D, de Hoogh K. 2024. Does residential address-based exposure assessment for outdoor air pollution lead to bias in epidemiological studies? Environ Health 23:75, *http://dx.doi.org/10.1186/s12940-024-01111-0*.

de Hoogh K, Chen J, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. 2018. Spatial $PM_{2.5}$, NO_2, O_3 and BC models for Western Europe: evaluation of spatiotemporal stability. Environ Int 120:81–92, *https://doi.org/10.1016/j.envint.2018.07.036*.

de Hoogh K, Gulliver J, Donkelaar AV, Martin RV, Marshall JD, Bechle MJ, et al. 2016. Development of West-European $PM_{2.5}$ and NO_2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ Res 151:1–10, *https://doi.org/10.1016/j.envres.2016.07.005*.

de Hoogh K, Korek M, Vienneau D, Keuken M, Kukkonen J, Nieuwenhuijsen MJ, et al. 2014. Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. Environ Int 73:382–392, *https://doi.org/10.1016/j.envint.2014.08.011*.

Huangfu P, Atkinson R. 2020. Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: a systematic review and meta-analysis. Environ Int 144:105998, *https://doi.org/10.1016/j.envint.2020.105998*.

Hvidtfeldt UA, Sørensen M, Geels C, Ketzel M, Khan J, Tjønneland A, et al. 2019. Long-term residential exposure to $PM_{2.5}$, PM_{10}, black carbon, NO_2, and ozone and mortality in a Danish cohort. Environ Int 123:265–272, *https://doi.org/10.1016/j.envint.2018.12.010*.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. 2005. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 15:185–204, *https://doi.org/10.1038/sj.jea.7500388*.

Jerrett M, Turner MC, Beckerman BS, Pope CA, van Donkelaar A, Martin RV, et al. 2017. Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates. Environ Health Perspect 125:552–559, *https://doi.org/10.1289/ehp575*.

Kerckhoffs J, Hoek G, Gehring U, Vermeulen R. 2021. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ Int 154:1065–1069, *https://doi.org/10.1016/j.envint.2021.106569*.

Kerckhoffs J, Hoek G, Messier KP, Brunekreef B, Meliefste K, Klompmaker JO, et al. 2016. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ Sci Technol 50:12894–12902, *https://doi.org/10.1021/acs.est.6b03476*.

Kerckhoffs J, Hoek G, Portengen L, Brunekreef B, Vermeulen RCH. 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ Sci Technol 53:1413–1421, *https://doi.org/10.1021/acs.est.8b06038*.

Kerckhoffs J, Hoek G, Vlaanderen J, van Nunen E, Messier K, Brunekreef B, et al. 2017. Robustness of intra-urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ Res 159:500–508, *https://doi.org/10.1016/j.envres.2017.08.040*.

Kerckhoffs J, Khan J, Hoek G, Yuan Z, Ellermann T, Hertel O, et al. 2022a. Mixed-effects modeling framework for Amsterdam and Copenhagen for outdoor NO_2 concentrations using measurements sampled with Google Street View cars. Environ Sci Technol 56:7174–7184, *https://doi.org/10.1021/acs.est.1c05806*.

Kerckhoffs J, Khan J, Hoek G, Yuan Z, Hertel O, Ketzel M, et al. 2022b. Hyperlocal variation of nitrogen dioxide, black carbon, and ultrafine particles measured with Google Street View cars in Amsterdam and Copenhagen. Environ Int 170:107575, *https://doi.org/10.1016/j.envint.2022.107575*.

Klompmaker JO, Janssen N, Andersen ZJ, Atkinson R, Bauwelinck M, Chen J, et al. 2021. Comparison of associations between mortality and air pollution exposure estimated with a hybrid, a land-use regression and a dispersion model. Environ Int 146:106306, *https://doi.org/10.1016/j.envint.2020.106306*.

Klompmaker JO, Montagne DR, Meliefste K, Hoek G, Brunekreef B. 2015. Spatial variation of ultrafine particles and black carbon in two cities: Results from a short-term measurement campaign. Sci Total Environ 508:266–275, *https://doi.org/10.1016/j.scitotenv.2014.11.088*.

Messier KP, Chambliss SE, Gani S, Alvarez R, Brauer M, Choi JJ. 2018. Mapping air pollution with Google Street View Cars: efficient approaches with mobile monitoring and land use regression. Environ Sci Technol 52:12563–12572, *https://doi.org/10.1021/acs.est.8b03395*.

Milanzi EB, Koppelman GH, Smit HA, Wijga AH, Oldenwening M, Vonk JM, et al. 2018. Air pollution exposure and lung function until age 16 years: the PIAMA birth cohort study. Eur Respir J 52:1800218, *https://doi.org/10.1183/13993003.00218-2018*.

Montagne DR, Hoek G, Klompmaker JO, Wang M, Meliefste K, Brunekreef B. 2015. Land use regression models for ultrafine particles and black carbon based on short-term monitoring predict past spatial variation. Environ Sci Technol 49:8712–8720, https://doi.org/10.1021/es505791g.

van Nunen E, Vermeulen R, Tsai MY, Probst-Hensch N, Ineichen A, Davey M, et al. 2017. Land use regression models for ultrafine particles in six European areas. Environ Sci Technol 51:3336–3345, https://doi.org/10.1021/acs.est.6b05920.

Ostro B, Hu J, Goldberg D, Reynolds P, Hertz A, Bernstein L, Kleeman MJ. 2015. Associations of mortality with long-term exposures to fine and ultrafine particles, species and sources: results from the California Teachers Study Cohort. Environ Health Perspect 123:549–556, https://doi.org/10.1289/ehp.1408565.

Ostro B, Lipsett M, Reynolds P, Goldberg D, Hertz A, Garcia C, et al. 2010. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the California Teachers Study. Environ Health Perspect 118:363–369, https://doi.org/10.1289/ehp.0901181.

Peters S, Bouma F, Hoek G, Janssen N, Vermeulen R. 2024. Air pollution exposure and mortality from neurodegenerative diseases in the Netherlands: A population-based cohort study. Environ Res 259:119552, https://doi.org/10.1016/j.envres.2024.119552.

Puustinen A, Hämeri K, Pekkanen J, Kulmala M, Hartog J de, Meliefste K, et al. 2007. Spatial variation of particle number and mass over four European cities. Atmos Environ 41:6622–6636, https://doi.org/10.1016/j.atmosenv.2007.04.020.

Ratingen S van, Vonk J, Blokhuis C, Wesseling J, Tielemans E, Weijers E. 2021. Seasonal influence on the performance of low-cost NO_2 sensor calibrations. Sensors 21:7919, https://doi.org/10.3390/s21237919.

Reid CE, Jerrett M, Petersen ML, Pfister GG, Morefield PE, Tager IB, et al. 2015. Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. Environ Sci Technol 49:3887–3896, https://doi.org/10.1021/es505846r.

Shen Y, de Hoogh K, Schmitz O, Clinton N, Tuxen-Bettman K, Brandt J, et al. 2022. Europe-wide air pollution modeling from 2000 to 2019 using geographically weighted regression. Environ Int 168:107485, https://doi.org/10.1016/j.envint.2022.107485.

Stafoggia M, Oftedal B, Chen J, Rodopoulou S, Renzi M, Atkinson RW, et al. 2022. Long-term exposure to low ambient air pollution concentrations and mortality among 28 million people: results from seven large European cohorts within the ELAPSE project. Lancet Planet Health 6:e9–e18, https://doi.org/10.1016/s2542-5196(21)00277-1.

Strak M, Weinmayr G, Rodopoulou S, Chen J, de Hoogh K, Andersen ZJ, et al. 2021. Long-term exposure to low level air pollution and mortality in eight European cohorts within the ELAPSE project: pooled analysis. BMJ 374:n1904, https://doi.org/10.1136/bmj.n1904.

Szpiro AA, Paciorek CJ, Sheppard, L. 2011. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology 22:680–685, https://doi.org/10.1097/EDE.0b013e3182254cc6.

Vachon J, Kerckhoffs J, Buteau S, Smargiassi A. 2024. Do machine learning methods improve prediction of ambient air pollutants with high spatial contrast? A systematic review. Environ Res 262(Pt 2):119751, https://doi.org/10.1016/j.envres.2024.119751.

Velders G, Aben J, Geilenkirchen G, den Hollander H, Megens L, van der Swaluw E, et al. 2016. New maps of concentrations and depositions for NSL and PAS: 2015 (Grootschalige concentratie- en depositiekaarten Nederland). RIVM report 2016. The Netherlands: Dutch National Institute for Public Health and the Environment.

Velders GJM, Willers SM, Wesseling J, van den Elshout S, van der Swaluw E, Mooibroek D, et al. 2021. Improvements in air quality in the Netherlands during the corona lockdown based on observations and model simulations. Atmos Environ 247:118158, https://doi.org/10.1016/j.atmosenv.2020.118158.

Wang M, Beelen R, Eeftens M, Meliefste K, Hoek G, Brunekreef B. 2012. Systematic evaluation of land use regression models for NO_2. Environ Sci Technol 46:4481–4489, https://doi.org/10.1021/es204183v.

Wang M, Gehring U, Hoek G, Keuken M, Jonkers S, Beelen R, et al. 2015. Air pollution and lung function in Dutch children: a comparison of exposure estimates and associations based on land use regression and dispersion exposure modeling approaches. Environ Health Perspect 123:847–851, https://doi.org/10.1289/ehp.1408541.

Wang R, Henderson SB, Sbihi H, Allen RW, Brauer M. 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmos Environ 64:312–319, https://doi.org/10.1016/j.atmosenv.2012.09.056.

Wang X, Ran S, Xia H, Shi H, Wu G, Zhang Z, et al. 2023. Ambient air pollution associated with incident asthma, subsequent cardiovascular disease and death: a trajectory analysis of a national cohort. J Hazard Mater 460:132372, https://doi.org/10.1016/j.jhazmat.2023.132372.

Weichenthal S, Bai L, Hatzopoulou M, Van Ryswyk K, Kwong JC, Jerrett M, et al. 2017. Long-term exposure to ambient ultrafine particles and respiratory disease incidence in Toronto, Canada: a cohort study. Environ Health 16:64, https://doi.org/10.1186/s12940-017-0276-7.

Weichenthal S, Ryswyk KV, Goldstein A, Bagg S, Shekkarizfard M, Hatzopoulou M. 2016. A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach. Environ Res 146:65–72, https://doi.org/10.1016/j.envres.2015.12.016.

Wesseling J, Nguyen L, Hoogerbrugge R. 2016. Measured and calculated concentrations of nitrogen oxides and particulates in the period 2010–2015 (update). RIVM Rapport 2016–0106. The Netherlands: Dutch National Institute for Public Health and the Environment.

Wesseling J, de Ruiter H, Blokhuis C, Drukker D, Weijers E, Volten H, et al. 2019. Development and implementation of a platform for public information on air quality, sensor measurements, and citizen science. Atmosphere 10:445, http://dx.doi.org/10.3390/atmos10080445.

Wesseling J, van Velze K. 2014. Technical description of Standard Calculation Method 2 (SRM-2). RIVM Briefrapport 2014-0109. The Netherlands: Dutch National Institute for Public Health and the Environment.

Wijga AH, Kerkhof M, Gehring U, de Jongste JC, Postma DS, Aalberse RC, et al. 2014. Cohort profile: the prevention and incidence of asthma and mite allergy (PIAMA) birth cohort. Int J Epidemiol 43:527–535, https://doi.org/10.1093/ije/dys231.

Wolf K, Hoffmann B, Andersen ZJ, Atkinson RW, Bauwelinck M, Bellander T, et al. 2021. Long-term exposure to low-level ambient air pollution and incidence of stroke and coronary heart disease: a pooled analysis of six European cohorts within the ELAPSE project. Lancet Planet Health 5:e620–e632, https://doi.org/10.1016/s2542-5196(21)00195-9.

Yu Z, Koppelman GH, Boer JMA, Hoek G, Kerckhoffs J, Vonk JM, et al. 2022. Ambient ultrafine particles and asthma onset until age 20: the PIAMA birth cohort. Environ Res 214:113770, https://doi.org/10.1016/j.envres.2022.113770.

Yu Z, Koppelman GH, Hoek G, Kerckhoffs J, Vonk JM, Vermeulen R, et al. 2021. Ultrafine particles, particle components and lung function at age 16 years: the PIAMA birth cohort study. Environ Int 157:106792, https://doi.org/10.1016/j.envint.2021.106792.

HEI QUALITY ASSURANCE STATEMENT

The conduct of this study was subjected to an independent audit by Mr. David Bush and Scott Adamson of Trinity Consultants, Inc. Mr. Bush and Mr. Adamson are experts in quality assurance for air quality monitoring studies and data management. The audit included a review of data quality for conformance to the study protocol as detailed in the final report and the study's quality assurance plan. The date of the audit is listed below along with the phase of the study examined.

QUALITY ASSURANCE AUDITS

Date: December 2024

Phase of Study: The final report was reviewed, including verification of data quality for each of the study components. An off-site audit was conducted via a teleconferencing platform with primary study personnel. The audit concentrated on the study's quality assurance and data management activities and included a review of the overall process utilized to collect new data and to manage and combine the exposure, air quality, epidemiological, and modeling data. Also evaluated were the procedures and measures undertaken to ensure quality and consistency in the processed databases and modeling results. Examples of data sets for the different types of data and modeling files were displayed by study personnel and reviewed for consistency, clarity, and completeness.

Written reports of the audit were provided to the HEI project manager, who transmitted the findings to the Principal Investigator. The quality assurance audit demonstrated that the study was conducted by an experienced team with a high concern for data quality. Study personnel were responsive to audit questions and recommendations, providing formal responses that adequately addressed all issues. The report appears to be an accurate representation of the study.

David H. Bush, Quality Assurance Officer

ADDITIONAL MATERIALS ON THE HEI WEBSITE

Additional Materials 1 and 2 contain material not included in the main report. They are available on the HEI website at *www.healtheffects.org/publications.*

Additional Materials 1. Detailed Methods

Additional Materials 2. Detailed Results

ABOUT THE AUTHORS

Gerard Hoek has a PhD from Wageningen University, the Netherlands. He is an associate professor at the Institute for Risk Assessment Sciences, Utrecht University, the Netherlands. His research is focused on air pollution exposure assessment and air pollution epidemiology.

Femke Bouma has an MSc from Goteborg University, Sweden. She studied global health after receiving her Medicine (BSc) degree at Leiden University. She is a PhD student at the Institute for Risk Assessment Sciences, Utrecht University, the Netherlands.

Nicole Janssen has a PhD from Wageningen University. She is a senior scientist at the National Institute for Public Health and the Environment. Her research interests include the health effects of air pollution, combined environmental exposures, and interventions. She has coordinated a large research program on the health effects of ultrafine particles around Schiphol Airport.

Joost Wesseling has a PhD from the National Institute of Nuclear Physics and High Energy Physics. As a senior scientist at the National Institute for Public Health and the Environment, he has co-developed the present air quality monitoring system in the Netherlands. For almost 20 years, he has been involved in all aspects of calculating air quality and comparing model results with measurements. In recent years, he has spent a lot of time testing air quality sensors and developing measurement strategies for them.

Sjoerd van Ratingen has an MSc in theoretical physics at Utrecht University. As a senior scientist at the National Institute for Public Health and the Environment, he has experience in the modeling of local-scale dispersion processes. In past projects, he has applied a broad range of tools and methods to model dispersion within the context of air quality assessment. He has been investigating the use of low-cost air quality sensors, satellite data (TROPOMI), and, for example, real-time traffic data as a means to improve air quality assessment.

Jules Kerckhoffs has a PhD from Utrecht University. He is an assistant professor at the Institute for Risk Assessment Sciences, Utrecht University, the Netherlands. His research interests include developing methods for environmental exposure assessment with a focus on mobile monitoring of air pollution.

Ulrike Gehring has a PhD from the Ludwig Maximilian University of Munich, Germany. She is an associate professor at the Institute for Risk Assessment Sciences, Utrecht University, the Netherlands. Her research includes environmental epidemiology and biostatistics with a focus on assessing the long-term health effects of ambient air pollution and the determinants of asthma and allergies in children. She is the principal investigator of the PIAMA birth cohort study.

Wouter Hendricx has an MSc in Earth and Environment from Wageningen University. As a scientist at the National Institute for Public Health and the Environment, he has experience in the use of low-cost air quality sensors as a means to improve air quality assessment.

Roel Vermeulen has a PhD from Wageningen University, the Netherlands. He is a distinguished university professor at Utrecht University, the Netherlands, with a focus on developing new methods to quantify exposures and their biological impacts. He is a professor at the Institute for Risk Assessment Sciences, Utrecht University, the Netherlands. He is the chair of the planetary health program at Utrecht University. He leads the Dutch program on the Exposome and the International Human Exposome Network. He coordinates the EU-funded multicenter study EXPANSE, a 5-year project on the impact of the Urban Exposome on cardiometabolic disease.

Kees de Hoogh has a PhD from the University of Leicester. He currently is a group leader at the Swiss Tropical and Public Health Institute, Allschwil, Switzerland. He has coordinated exposure assessments to air pollution and other environmental stressors in national and international projects.

OTHER PUBLICATIONS RESULTING FROM THIS RESEARCH

Bouma F, Janssen NA, Wesseling J, van Ratingen S, Strak M, Kerckhoffs J, et al. 2023a. Long-term exposure to ultrafine particles and natural and cause-specific mortality. Environ Int 175:107960, *https://doi.org/10.1016/j.envint.2023.107960*.

Bouma F, Hoek G, Koppelman GH, Vonk JM, Kerckhoffs J, Vermeulen R, et al. 2023b. Exposure to ambient ultrafine particles and allergic sensitization in children up to 16 years. Environ Res 219:115102, *https://doi.org/10.1016/j.envres.2022.115102*.

Klompmaker JO, Janssen N, Andersen ZJ, Atkinson R, Bauwelinck M, Chen J, et al. 2021. Comparison of associations between mortality and air pollution exposure estimated with a hybrid, a land-use regression and a dispersion model. Environ Int 146:106306, *https://doi.org/10.1016/j.envint.2020.106306*.

Peters S, Bouma F, Hoek G, Janssen N, Vermeulen R. 2024. Air pollution exposure and mortality from neurodegenerative diseases in the Netherlands: a population-based cohort study. Environ Res 259:119552, *https://doi.org/10.1016/j.envres.2024.119552*.

APPENDIX A: TABLE S1

Table S1. Predicted annual average exposure concentrations at 20,000 residential addresses in the Netherlands. BC, NO_2 and $PM_{2.5}$ in μg/m³. UFPs in p/cm³. Median and interquartile range (IQR) presented. Additional Materials Table S4 contains the full distribution. SLR = supervised linear regression; RF = Random Forest.

Pollutant (BC)	Median	IQR
Google_SLR	1.00	0.32
Google_RF	1.06	0.08
Google_Lasso	1.07	0.22
MUSIC_mobile_SLR	1.08	2.18
MUSIC_mobile_RF	1.80	1.13
MUSIC_mobile_Lasso	2.10	0.80
MUSIC_shortterm_SLR	0.99	0.68
MUSIC_shortterm_RF	1.10	0.24
EXPO_mobile_SLR	2.68	1.71
EXPO_mobile_RF	2.83	1.18
EXPO_mobile_Lasso	3.09	0.52
EXPO_shortterm_SLR	1.47	0.49
EXPO_shortterm_RF	1.45	0.45
EXPO_shortterm_Lasso	1.55	0.12
RUN_mobile_SLR	1.25	0.90
RUN_mobile_RF	1.34	0.76
RUN_mobile_Lasso	1.43	0.59
RUN_shortterm_SLR	0.93	0.32
RUN_shortterm_RF	0.90	0.20
RUN_shortterm_Lasso	0.92	0.14
ESCAPE_SLR	1.45	0.43
ESCAPE_RF	1.49	0.36
ESCAPE_Lasso	1.47	0.36
ELAPSE_2010	1.72	0.53
Dispersion _13	0.96	0.29
Dispersion_16	0.96	0.30
Dispersion_19	0.67	0.16

Pollutant (NO_2)	Median	IQR
Google_SLR	19.5	8.6
Google_RF	21.3	3.5
Google_Lasso	17.9	9.3
ESCAPE_SLR	28.3	13.0
ESCAPE_RF	28.6	11.0
ESCAPE_Lasso	27.0	11.3
ELAPSE_2010	36.8	10.9
EXPANSE_2010	31.7	9.8
EXPANSE_2013	30.0	8.2
EXPANSE_2016	28.5	8.2
EXPANSE_2019	26.6	7.3
Dispersion_10	28.4	9.4
Dispersion_13	24.4	7.8
Dispersion_16	23.9	8.3
Dispersion_19	20.8	7.3
EXPANSE_RF_2010	31.3	12.2
EXPANSE_RF_2013	30.3	11.0
EXPANSE_RF_2016	28.4	10.4
EXPANSE_RF_2019	24.8	10.7
LCS_SLR	18.3	3.4
LCS_RF	17.3	2.7
LCS_Lasso	16.6	1.7
LCS fusion_All_0	21.2	7.6
LCS fusion_All_0.5	20.5	7.2
LCSfusion_Region_NS_0.5	20.7	7.3
LCSfusion Region_WE0.5	20.5	6.9

Pollutant (UFPs)	Median	IQR
Google_SLR	17414.2	3451.6
Google_RF	17386.3	1384.5
Google_Lasso	17386.2	2026.6
MUSIC_mobile_SLR	11958.4	12508.7
MUSIC_mobile_RF	14777.4	7957.4
MUSIC_mobile_Lasso	15107.6	6511.7
MUSIC_shortterm_SLR	7806.5	9600.9
MUSIC_shortterm_RF	10097.4	4033.5
EXPO_mobile_SLR	15100.0	7029.6
EXPO_mobile_RF	16217.3	7475.1
EXPO_mobile_Lasso	16240.1	5420.2
EXPO_shortterm_SLR	10966.5	4657.2
EXPO_shortterm_RF	12393.6	4489.7
EXPO_shortterm_Lasso	11791.2	1875.6
RUN_mobile_SLR	11538.0	7832.0
RUN_mobile_RF	12626.9	6906.3
RUN_mobile_Lasso	12855.8	4384.4
RUN_shortterm_SLR	10785.7	7902.9
RUN_shortterm_RF	10323.8	6652.1
RUN_shortterm_Lasso	10809.8	635.6

Pollutant ($PM_{2.5}$)	Median	IQR
MUSIC_mobile_SLR	44.7	8.3
MUSIC_mobile_RF	50.2	8.7
MUSIC_shortterm_SLR	98.0	15.1
MUSIC_shortterm_RF	86.5	7.4
EXPO_mobile_SLR	48.0	13.7
EXPO_mobile_RF	47.5	18.5
EXPO_mobile_Lasso	61.7	35.1
EXPO_shortterm_SLR	49.0	14.4
EXPO_shortterm_RF	51.8	30.3
RUN_mobile_SLR	45.7	3.2
RUN_mobile_RF	37.5	8.6
RUN_shortterm_SLR	31.7	6.3
RUN_shortterm_RF	32.1	4.1
ESCAPE_SLR	16.7	0.9
ESCAPE_RF	17.0	1.0
ESCAPE_Lasso	17.3	0.3
ELAPSE_2010	17.6	2.0
EXPANSE_SLR_2010	17.2	1.2
EXPANSE_SLR_2013	14.4	1.1
EXPANSE_SLR_2016	11.8	1.0
EXPANSE_SLR_2019	11.5	0.9
DISPERSION_10	17.7	1.7
DISPERSION_13	14.3	1.4
DISPERSION_16	11.7	1.4
DISPERSION_19	10.6	1.3
EXPANSE_RF_2010	18.3	1.1
EXPANSE_RF_2013	15.6	0.8
EXPANSE_RF_2016	13.3	0.5
EXPANSE_RF_2019	12.4	0.5
LCS_SLR	12.8	2.3

Research Report 226, *Comparison of Long-Term Air Pollution Exposure from Mobile and Routine Monitoring, Low-Cost Sensors, and Dispersion Models,* by G. Hoek et al.

INTRODUCTION

Outdoor air pollution is a major global public health risk factor. There is now broad expert consensus that exposure to ambient air pollution causes an array of adverse health effects based on evidence from a large body of scientific literature that has grown exponentially since the mid-1990s (IARC 2016; US EPA 2016, 2019, 2022; WHO 2021).

The assessment of long-term exposure to ambient air pollution for epidemiological studies, however, remains challenging. Early cohort studies characterized exposure by assigning the average concentration measured at one or a few central sites within a city to each participant within that city (Dockery et al. 1993; Pope et al. 2002). Fixed-site networks — even those in North America and Western Europe — continue to have relatively limited spatial coverage in many areas, particularly in suburban and rural locations, and insufficient density to capture small-scale (within-city) variations of air pollution.

Recent developments in measurement technologies and modeling approaches for long-term exposure to air pollution have increasingly been used to estimate air pollution at finer spatial scales for epidemiological studies of large populations. Advances include novel air pollution sensors, mobile monitoring, satellite data, hybrid models, and machine learning approaches (Hoek 2017). Even with those advances, important limitations and challenges remain when predicting long-term air pollution exposure, particularly for pollutants that vary widely in space and time.

In 2019, HEI issued the Request for Applications 19-1: Applying Novel Approaches to Improve Long-Term Exposure Assessment of Outdoor Air Pollution for Health Studies (see Preface). The goal of the RFA* was to develop and apply novel, scalable approaches to improve assessments of long-term exposures to outdoor air pollutants that vary widely in space and time, such as ultrafine particles (UFPs), black carbon (BC), and nitrogen dioxide (NO_2). Studies were intended to evaluate exposure measurement error quantitatively and to determine how exposure assessment approaches might ultimately influence the estimated health effects.

Dr. Hoek and colleagues proposed to compare the performance of a suite of air pollution exposure assessment methods in the Netherlands, including a comparison of health effects estimates among different methods. The HEI Research Committee recommended the study for funding because it would compare different exposure modeling approaches, including models based on data from low-cost sensors. They also appreciated the focus on UFPs and the leveraging of a wealth of data including three cohort studies.

This Commentary provides the HEI Improved Exposure Assessment Studies Review Panel's evaluation of the study. It aims to aid the HEI's sponsors and the public by highlighting the study's strengths and limitations and placing the results presented in the Investigators' Report within a broader scientific and regulatory context.

SCIENTIFIC AND REGULATORY BACKGROUND

Traffic-related air pollution continues to be an important risk factor for poor health worldwide, with the highest exposures in urban settings and at residences near busy roadways (HEI 2022). It is a complex mixture of gases and particles resulting from using motor vehicles. Motor vehicles emit various pollutants, including NO_2, BC, and UFPs (HEI 2022). Exposure assessment of those pollutants is challenging because they are characterized by high spatial and temporal variability.

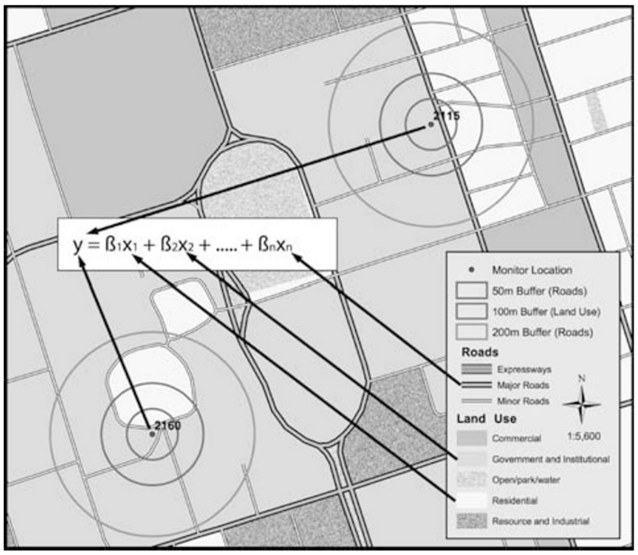
Epidemiological studies have used different approaches to address these challenges, including land use regression (LUR) models based on fixed-site routine monitoring, low-cost sensor networks, mobile monitoring, and dispersion models.

LUR models are empirically derived by regressing observations of air pollution at a number of locations on land use variables derived via geographic information systems (GIS). Typically, LUR models are developed using a variety of algorithms, such as supervised stepwise procedures or machine-learning algorithms, to identify which land use

Dr. Gerard Hoek's 3-year study, "Comparison of Long-term Air Pollution Exposure Assessment Based on Mobile Monitoring, Low-Cost Sensors, Dispersion Modelling, and Routine Monitoring-Based Models (CLAIRE)," began in June 2020. Total expenditures were $800,000. The draft Investigators' Report from Hoek and colleagues was received for review in February 2024. A revised report, received in July 2024, was accepted for publication in September 2024. During the review process, the HEI Improved Exposure Assessment Studies Review Panel and the investigators had the opportunity to exchange comments and clarify issues in the Investigators' Report and the Panel's Commentary. As a co-investigator of the Hoek Investigators' Report, Dr. Ulrike Gehring was not involved in its evaluation by the Panel.

* A list of abbreviations and other terms appears at the end of this volume.

variables are the best predictors of local concentrations of air pollution (Hoek 2017). Predictor variables often include indicators of land use, proximity to or density of nearby roads, traffic intensity, and population density. After fitting, the LUR model is used to estimate pollutant concentrations where no measurements are available in the study area, and ultimately, to predict exposures at locations important for epidemiological studies (**Commentary Figure** 1).



Commentary Figure 1. Illustration of elements of a LUR model (Jerrett et al. 2005).

LUR models can be based on fixed-site measurements, mobile measurements, or a combination of both. Mobile monitoring strategies can involve on-road measurements made while driving predefined routes, or repeated short-term measurements made while in a parked vehicle at various locations. LUR models can also be based on existing routine monitoring networks and specifically designed monitoring campaigns, either using research-grade measurement devices or low-cost sensors.

LUR models are empirical models in contrast to dispersion or chemical transport models, which are often deterministic models that are based on physical principles and estimated emissions. A wide variety of dispersion models exist that differ in their spatial scales (e.g., street, urban, regional, continental, or even global) and the processes they include (e.g., dispersion only or dispersion plus chemical transformations). The quality of the input data is a key determinant of the performance of any air pollution model (Hoek 2017). Recognizing the limitations of any single approach, researchers have also developed hybrid models that combine monitoring data, land use data, satellite observations, and dispersion models.

Little information is available about the relative performance of the various approaches to assess long-term exposure to traffic-related air pollution, and in turn, how those different approaches might influence health effects estimates. As noted earlier, important limitations and challenges remain when predicting long-term air pollution exposure for pollutants that vary highly in space and time.

Exposure models are applied in epidemiological studies that underpin the air quality standards and guidelines. Governments in the United States and Europe have recently moved toward more stringent $PM_{2.5}$ annual standards — 9 and 10 μg/m³, respectively — which align more closely with the 2021 WHO Air Quality Guidelines of 5 μg/m³. A more stringent NO_2 annual standard was also set in Europe (**Commentary Table 1**). There are no specific ambient air quality standards or guidelines for UFPs and BC, and regulatory agencies do not commonly measure them. Hence, international or national standard methods to characterize them have not been established (HEI 2010; HEI Review Panel on Ultrafine Particles 2013). Although no air quality guidelines have been developed for UFPs and BC, the WHO has provided "good practice statements" for these pollutants geared toward additional monitoring, mitigation, and epidemiological research (WHO 2021).

Commentary Table 1. Annual NO_2 and $PM_{2.5}$ Standards in the US, EU, and WHO Guidelines

Organization	Annual $PM_{2.5}$ (μg/m³)	Annual NO_2 (μg/m³)	Notes
US EPA (2024)	9	100	NAAQS
US EPA (Previous)	12	100	Previous NAAQS
EU (2024)	10	20	Limit value for 2030
EU (Previous)	25	40	Previous limit value
WHO (2021)	5	10	Air Quality Guidelines
WHO (Previous)	10	40	Previous Air Quality Guidelines

NAAQS = National Ambient Air Quality Standards.

SPECIFIC AIMS

The specific aims of Dr. Hoek's study were to accomplish the following:

1. Develop long-term ambient air pollution exposure estimates for selected epidemiological studies based on low-cost sensors, mobile and fixed-site monitoring, and dispersion modeling
2. Compare different exposure assessment methods in terms of their ability to predict the spatial variation of long-term average concentrations using independent validation data
3. Compare different exposure assessment methods in terms of air pollution effect estimates in selected epidemiological studies

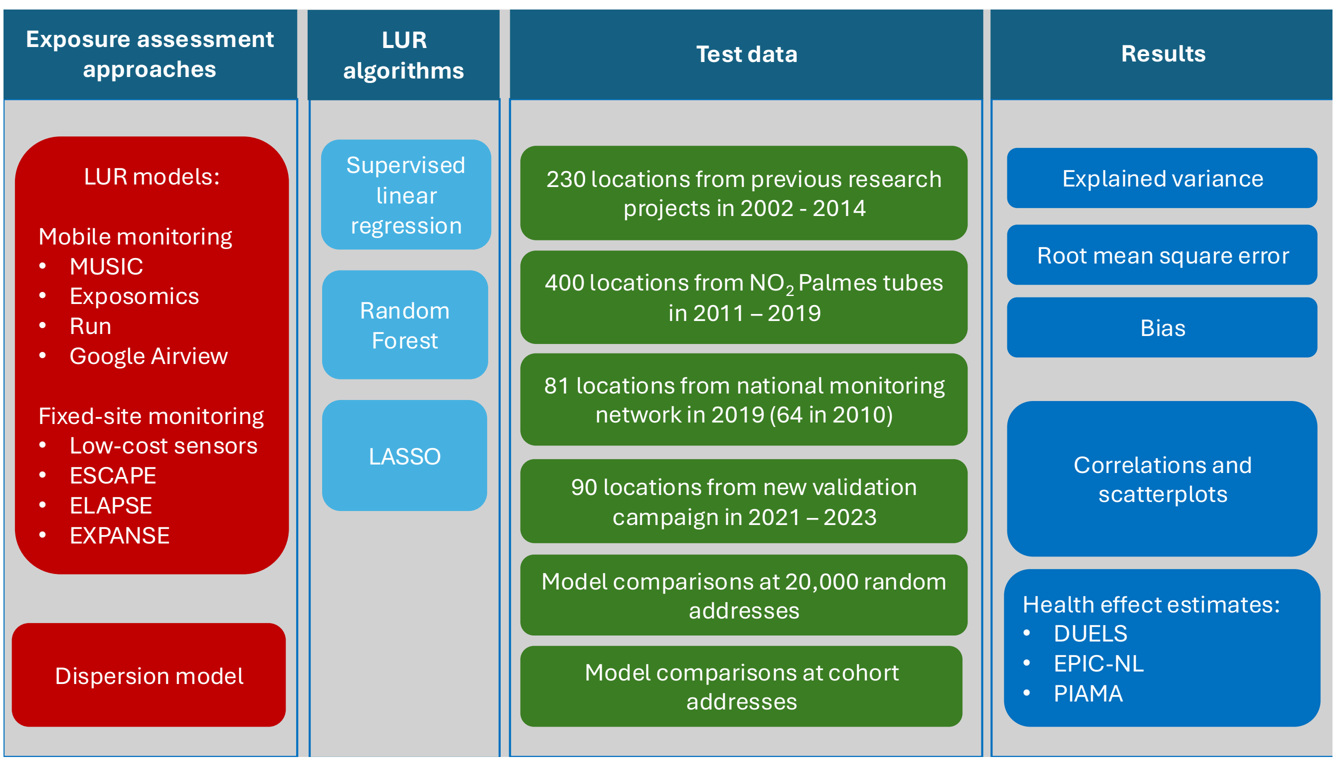
SUMMARY OF APPROACH AND METHODS

Dr. Hoek and colleagues compared the performance of a suite of air pollution exposure assessment methods in the Netherlands. The predictions of the exposure models were compared at 20,000 random residential addresses in the Netherlands and tested on existing and new validation data at residential locations where appropriate (**Commentary Figure 2**). The investigators also conducted epidemiological analyses in three cohort studies to compare health effect estimates of the different exposure assessment methods. They assessed four pollutants (UFPs, BC, $PM_{2.5}$, NO_2) and used existing data and models from previous collaborative projects where available.

EXPOSURE ASSESSMENT APPROACHES

Characteristics of the various exposure assessment approaches, including pollutants, monitoring approach, modeling approach, monitoring area, and year are provided in **Commentary Table 2**. The investigators evaluated annual average outdoor air pollution concentrations using a suite of exposure assessment methods that differ in their monitoring data (low-cost sensors, mobile monitoring, and fixed-site routine or fixed-site study-specific monitoring using research-grade instruments) and modeling approaches (LUR models or air pollution dispersion models). In addition, they tested various model development algorithms in the LUR-based models beyond linear regression, including machine learning methods.

The exposure assessment approaches also differed in the pollutants covered, the monitoring area, the time of day, and the study period. The greatest diversity of methods was available for BC and $PM_{2.5}$, while results for UFPs were only available from mobile monitoring models from different monitoring areas and years. For two approaches (the dispersion model and the EXPANSE model), the investigators were able to produce exposure predictions for multiple years to



Commentary Figure 2. Schematic overview of the study design.

Commentary Table 2. Characteristics of the Various Exposure Assessment Approaches

Model Name	Pollutants	Monitoring Approach	Monitoring Area	Modeling Approach	Year	Reference
MUSIC	UFPs, BC, $PM_{2.5}$	Mobile monitoring While parked: 3 × 30 minutes per year in 161 locations On-road: 2,964 road segments	Two cities: Amsterdam, Rotterdam	LUR	2013	Kerckhoffs et al. 2016
Exposomics	UFPs, BC, $PM_{2.5}$	Mobile monitoring While parked: 3 × 30 minutes per year in 240 locations On-road: 5,236 road segments	Three cities: Amsterdam, Maastricht, Rotterdam	LUR	2014–2015	Kerckhoffs et al. 2017
RUN	UFPs, BC, $PM_{2.5}$	Mobile monitoring While parked: 3 × 30 minutes per year in 400 locations On-road: 14,392 road segments	Netherlands	LUR	2016–2017	Kerckhoffs et al. 2021
Google Air View	UFPs, BC, $PM_{2.5}$, NO_2	Mobile monitoring On-road: 46,664 road segments, five to 10 repeats	One city: Amsterdam	LUR	2019–2020	Kerckhoffs et al. 2022
Low-cost sensors	$PM_{2.5}$, NO_2	Fixed-site monitoring 1 × 6 months in 84 locations	Netherlands	LUR and data-fusion model	2021–2023	New for the current study
ESCAPE	BC, $PM_{2.5}$, NO_2	Fixed-site monitoring 3 × 14-day average per year on 40–80 locations	Netherlands	LUR	2009–2010	Beelen et al. 2013; Eeftens et al. 2012
ELAPSE	BC, $PM_{2.5}$, NO_2	Fixed-site monitoring Annual average from 543–2,399 routine monitoring locations 436 sites from ESCAPE for BC	Western Europe	Hybrid LUR	2010	De Hoogh et al. 2018
EXPANSE	$PM_{2.5}$, NO_2	Fixed-site monitoring Annual average from 699–3,176 routine monitoring locations	Europe	Hybrid LUR	2010–2019	Shen et al. 2022
Dispersion model	EC, $PM_{2.5}$, NO_2	Not applicable	Netherlands	Deterministic	2010–2019	Velders et al. 2021

assess changes in spatial patterns between 2010 and 2019. All exposure assessment approaches were developed on data collected before the COVID-19 pandemic in 2020, except for the new low-cost sensor models.

New Low-Cost Sensor Campaign

A new low-cost sensor campaign was conducted at 84 residential outdoor locations across the Netherlands in 2021–2023 as part of a larger citizen-science project coordinated by the National Institute for Public Health and the Environment (Wesseling et al. 2019). Locations were either close to busy roads (~75%) or (urban) background locations (25%). Continuous measurements of NO_2 and $PM_{2.5}$ were collected using the Alphasense B43F sensor and the Sensirion SPS30 particle counter in two consecutive half-year measurement periods. Data from the national monitoring network were used to account for temporal differences between the two measurement periods. For NO_2, sensor calibration was performed before the first and second 6-month periods using six co-located sensors alongside reference measurements. Night-time NO_2 measurements were used to adjust differences between sensors on a daily basis to account for potential drift of the sensor within the 6-month period. For this night calibration, measurements at nearby reference stations were used to recalibrate the sensors at the participants' home addresses. The $PM_{2.5}$ sensors were not calibrated.

Low-cost sensor data were used primarily in data-fusion models, where these data were combined with dispersion model estimates and data from the routine national monitoring network to explore the added value of low-cost sensor data for NO_2. In addition, LUR models were developed using low-cost sensor data for both NO_2 and $PM_{2.5}$.

Algorithms Tested

For the various LUR-based models, the investigators tested three model development algorithms: supervised linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest. Supervised linear regression is widely applied, which starts with univariate linear regression models for each potential predictor, and chooses the model with the highest adjusted R^2 as the starting point. Additional predictor variables were included in the model if they improved the adjusted R^2 of the previous model step, and only if the variable coefficient was in the expected direction of effect (e.g., positive for traffic intensity).

Other algorithms have increasingly been applied to alleviate some of the concerns related to standard linear regression. LASSO is a shrinkage algorithm that is used to weight the predictor coefficients more reliably when the predictors are highly correlated. Random Forest is a machine learning algorithm that can model nonlinearity and the potentially complex interactions among predictors.

The predictors available for inclusion in the algorithms tested in this study were based primarily on the ESCAPE model and included land use, traffic, and population density variables. For the European hybrid models, satellite observations and air pollution chemical transport model estimates were also available.

Existing and New Validation Data

The predictions of the different exposure approaches and algorithms were tested using existing and new validation data that were not used in model development.

Existing validation data were used from various research projects at more than 230 residential outdoor locations across the Netherlands covering more than 10 years. In addition, extensive monitoring was conducted using Palmes tubes for NO_2 across the Netherlands at 400 locations from 2011 to 2019. Validation data for the various pollutants were collected using a variety of monitoring approaches that also used different monitoring areas and study periods. NO_2 was the pollutant with the most validation data. Data were available for UFPs and BC at only 87 locations in Utrecht and Amsterdam, with relatively short sampling times (up to a week) per site from 2002 to 2004 and in 2014.

Data from the routine national monitoring network were also used. The number of locations differed per pollutant and varied across years. For 2019, the most data were available for NO_2 (81 sites), fewer for $PM_{2.5}$ (49 sites), and even fewer for BC (31 sites). No data were available for UFPs because they are not measured routinely in the Netherlands.

To increase the number of UFPs and BC measurements for validation purposes, Hoek and colleagues conducted a new campaign to acquire independent validation data from 2021 to 2023 — after most COVID-19 restrictions were lifted. Measurements of $PM_{2.5}$ and NO_2 were also collected for completeness. In total, measurements were conducted at 90 residential outdoor locations across the Netherlands. Similar to the new low-cost sensor campaign, traffic sites were oversampled. 1-week measurements were collected at the 90 locations using various research-grade instruments, including the DISCmini for UFPs; at 31 of those sites, an additional 1-week measurement was performed in a different season. Measurements were not taken simultaneously at all sites because of equipment limitations; only four locations were measured simultaneously. Data from a reference site in the middle of the country were used to account for temporal differences between measurement weeks.

To represent long-term exposure, validation data were aggregated to an average concentration per location, which also accounted for temporal differences between measurement weeks where applicable.

Assignment of Exposure Estimates to Locations

The investigators assigned the annual air pollution model predictions to 20,000 random residential addresses, addresses of cohort participants, and validation locations using a standard Dutch database of geocoded addresses from 2019. For the assignment to the routine network sites, they used the nearest residential address; for all other validation data, the exact residential address was used because monitoring took place at residential home locations.

For all cohort participants, exposure data were assigned to the residential address at the time of recruitment to the cohort. For PIAMA, exposure was also assigned at the current address.

HEALTH STUDIES

Epidemiological analyses in three cohort studies were conducted to compare the health effect estimates of the different exposure assessment approaches and algorithms.

Study Population and Health Outcomes

For the health analyses, Hoek and colleagues selected three population-based cohorts in the Netherlands that differ in size, population, location, health outcomes, and study period (**Commentary Table 3**). DUELS is a very large administrative cohort that includes all Dutch adult citizens aged 30 years or older (10.8 million), starting from 2013 and following until 2019. DUELS was formed by linking census data, population registries, and death registries, but contains less detailed covariate data than the other selected cohorts. In addition, the investigators selected two smaller, "conventional" epidemiological cohorts with detailed information available on lifestyle factors (EPIC-NL and PIAMA). EPIC-NL is an adult cohort of mostly women (76%), whereas PIAMA is a birth cohort that has 20 or more years of follow-up data. Both DUELS and EPIC-NL were also included in the HEI-funded ELAPSE study that investigated the health effects of low levels of air pollution (Brunekreef et al. 2021).

DATA ANALYSES

Data Analysis Comparing Exposure Predictions and Performance

Hoek and colleagues compared annual average predictions of the different exposure assessment methods by preparing scatterplots and calculating Pearson correlations between predictions from the different models. They also prepared Bland-Altman plots to identify systematic differences between the exposure predictions. They calculated the correlations across years where available and across pollutants within methods. They assessed model performance in terms of explained variance (R^2), root-mean-square error (RMSE), and bias (model minus measurement) by comparing model predictions with validation data. Comparisons were made for three spatial domains: country-wide, four major cities (Amsterdam, Rotterdam, Utrecht, and The Hague), and Amsterdam.

Data Analysis in the Health Studies

The investigators applied Cox proportional hazards models to assess the association between air pollution exposure from the different exposure assessment approaches and various health outcomes in DUELS and EPIC-NL. For PIAMA, they applied linear regression and discrete-time hazard models to investigate the associations between the different air pollution exposure estimates and lung function and asthma incidence, respectively.

Commentary Table 3. Key Study Characteristics of the Three Health Studies in the Netherlands

Study Name	Study Population	Location	Health Outcomes	Study Period	Sample Size (rounded)	Age	Reference
DUELS	Adult administrative cohort	Nationwide	Natural-cause and cause-specific mortality	2013–2019	10.8 million	30+	Brunekreef et al. 2021; Klompmaker et al. 2021
EPIC-NL	Adult cohort	Four cities: Amsterdam, Doetinchem, Utrecht and Maastricht	Natural-cause mortality Coronary events Stroke events	1993–2013	34,000	20+	Brunekreef et al. 2021
PIAMA	Birth cohort	3 regions: northern, middle, and southwestern	Asthma incidence	1996–2016	3,700	0–20	Yu et al. 2022
		3 regions: north, middle, and southwest	Lung function (FEV_1, FVC)	1996–2012	700	0–16	Yu et al. 2021

DUELS = Dutch Environmental Longitudinal Study; EPIC-NL = European Prospective Investigation into Cancer and Nutrition: Netherlands cohort; FEV_1 = forced expiratory volume in 1 second; FVC = forced vital capacity; PIAMA = Prevention and Incidence of Asthma and Mite Allergy.

The main results for DUELS were adjusted for age, sex, and individual and area-level socioeconomic status. In addition, results were adjusted for various lifestyle factors, such as smoking and diet for EPIC-NL and PIAMA.

The main analyses focused on single-pollutant models, although limited two-pollutant models were applied for DUELS for natural-cause mortality and a subset of exposure approaches. The investigators expressed the effect estimates using fixed increments across methods (5 µg/m³ for $PM_{2.5}$, 10 µg/m³ for NO_2, 1 µg/m³ for BC, and 5,000 p/cm³ for UFPs) and method-specific interquartile ranges of exposure.

To further explore if and why effect estimates differed among exposure approaches, the investigators conducted some additional analyses for DUELS and PIAMA. For example, they used scatterplots and linear regression analysis to evaluate whether the performance of the different exposure models and the predicted exposure contrast were associated with the effect estimate. Furthermore, they conducted a random-effects meta-analysis with the different exposure assessment methods to illustrate the degree of heterogeneity potentially introduced in meta-analyses, because in meta-analysis it is common practice to combine the results of studies using different exposure assessment methods.

SUMMARY OF RESULTS

COMPARISON OF EXPOSURES ACROSS DIFFERENT METHODS

In most comparisons, the investigators observed only small differences in performance (<0.1 difference in R^2) among the three LUR algorithms (supervised linear regression, Random Forest, or LASSO). Hence, this summary focuses on the results of the supervised linear regression models.

In some cases, the predicted exposure levels and exposure contrasts varied widely between methods. For example, the interquartile range at the 20,000 residential locations differed by a factor of up to 14 for BC across methods and was even higher for $PM_{2.5}$ (a factor of 17). Exposure contrast factors of 4 and 8 were reported for UFPs and NO_2, respectively. In general, the mobile monitoring models predicted higher BC and $PM_{2.5}$ levels and contrasts at residential sites than other models, probably related to the on-road monitoring. Exposure contrasts for $PM_{2.5}$ were comparatively low for all models except the mobile monitoring models (the interquartile range varied from 0.9 to 2.3 µg/m³ in all nonmobile monitoring models; a factor of 3).

The various exposure models generally resulted in moderately to highly correlated exposure predictions for BC, NO_2, and UFPs at residential sites across the Netherlands (correlation coefficient $R > 0.7$). In contrast, the correlations for $PM_{2.5}$ between the different models were lower ($R < 0.4$), particularly for the poorly performing mobile monitoring models.

Exposure predictions for different years in the period of 2010–2019 (dispersion model and EXPANSE model) were highly correlated for BC, NO_2, and $PM_{2.5}$ ($R > 0.9$), and indicated stable spatial contrasts over a 10-year period. Exposure levels of BC, NO_2, and $PM_{2.5}$ declined during this period.

Exposure models explained a low to moderate amount of spatial variance in UFPs and NO_2 at the new validation sites (R^2 between 16% and 47%), and a positive bias was reported for all models (models overestimated measurements up to 13,000 p/cm³ for UFPs and 17 µg/m³ for NO_2).

Most models explained a moderate amount of spatial variance in past exposures at existing validation sites for BC and NO_2 ($R^2 > 0.5$) and explained less variance for UFPs (R^2 >0.25). Most models predicted past exposures of $PM_{2.5}$ poorly (**Commentary Table 4**).

Commentary Table 4. The Explained Variance (in %) of the Mobile, Fixed-Site, and Air Pollution Dispersion Exposure Assessment Approaches at Existing Validation Sites

	2019 National Monitoring Network			2019 Palmes Tubes, National	2007 TRACHEA, National	2002–2004 RUPIOH, Amsterdam	2014 Exposomics, Amsterdam and Utrecht
Exposure Assessment Approaches	**BC (N = 31)**	**$PM_{2.5}$ (N = 49)**	**NO_2 (N = 81)**	**NO_2 (N = 404)**	**NO_2 (N = 144)**	**UFPs (N = 46)**	**UFPs (N = 41)**
Mobile monitoring[a,b]	45	1	69	61	73	24	46
Fixed-site monitoring[a]	72	5	63	52	62	NA	NA
Air pollution dispersion	61	18	75	62	77	NA	NA

NA = not available; RUPIOH = Relationship between Ultrafine and fine Particulate matter in Indoor and Outdoor air and respiratory Health; TRACHEA = Traffic-Related Air pollution and Children's respiratory HEalth and Allergies.
[a] Based on LUR models using supervised linear regression.
[b] Includes on-road and short-term parked monitoring.

The addition of low-cost sensor data did not improve NO_2 estimates in models that combined dispersion model estimates and data from the routine national monitoring network.

COMPARISON OF HEALTH EFFECTS FOR THE DIFFERENT METHODS

In most cases, the application of the various exposure assessment approaches in the health studies led to similar findings in terms of direction (positive, null, or negative associations) between air pollutant exposure and various mortality and morbidity outcomes (**Commentary Table 5**). There were no consistent differences in effect estimates between exposure assessment approaches based on mobile monitoring, fixed-site monitoring, or dispersion models. These results were consistent with the typically moderate-to-high correlations of the different exposure assessment approaches. Positive (adverse) associations were observed most clearly in DUELS and to a lesser extent in PIAMA. Null or sometimes even negative associations were reported in the EPIC-NL cohort.

The strength (magnitude) of the positive associations, however, differed based on the exposure assessment method. For example, the association for natural-cause mortality in DUELS for a 1 μg/m^3 increase in BC ranged from a hazard ratio of 1.007 (95% confidence interval: 1.005–1.009) to 1.07 (1.05–1.09). Typically, differences in the magnitude of the associations across methods were smaller when expressed per method-specific interquartile range. Heterogeneity in the effect estimates was high (I^2 > 88%) for natural-cause mortality in DUELS, but low to moderate for the effect estimates reported in PIAMA.

Factors that explained some of the heterogeneity of effect estimates included the performance of the model at validation sites and the predicted exposure contrast. The year of the exposure model did not explain the heterogeneity.

Commentary Table 5. Summary of Positive, Null (Nonstatistically Significant), or Negative Findings for Selected Health Outcomes in the Three Cohort Studies

	DUELS				EPIC–NL				PIAMA			
	Natural-Cause Mortality								Asthma Incidence			
Model Name[a]	**UFPs**	**BC**	**NO_2**	**$PM_{2.5}$**	**UFPs**	**BC**	**NO_2**	**$PM_{2.5}$**	**UFPs**	**BC**	**NO_2**	**$PM_{2.5}$**
MUSIC (mobile)	0	+	NA	+	0	0	NA	–	+	+	NA	+
MUSIC (short-term parked)	+	+	NA	+	0	–	NA	0	0	+	NA	0
Exposomics (mobile)	+	+	NA	+	–	–	NA	0	+	0	NA	+
Exposomics (short–term parked)	+	+	NA	+	0	0	NA	0	0	0	NA	+
RUN (mobile)	+	+	NA	+	0	–	NA	0	0	0	NA	0
RUN (short-term parked)	+	+	NA	+	0	0	NA	0	+	0	NA	+
Google Air View	+	+	+	NA	0	0	+	NA	0	+	+	NA
Low-cost sensors	NA	NA	0	+	NA	NA	0	+	NA	NA	0	0
ESCAPE	NA	+	+	0	NA	0	0	0	NA	+	+	+
ELAPSE (2010)	NA	+	+	0	NA	0	0	0	NA	+	+	+
EXPANSE (2010)	NA	NA	+	+	NA	NA	0	+	NA	NA	+	+
EXPANSE (2013)	NA	NA	+	+	NA	NA	0	0	NA	NA	+	+
EXPANSE (2016, 2019)	NA	NA	+	+	NA	NA	0	+	NA	NA	+	+
Dispersion model (2010, 2013, 2016, 2019)	NA	NA	+	+	NA	NA	0	0	NA	NA	+	+

NA = not available.
[a] Models from supervised linear regression were chosen for the LUR models.

HEI IMPROVED EXPOSURE ASSESSMENT STUDIES REVIEW PANEL'S EVALUATION

In its independent review of the study, the Panel thought the study was well-motivated and effectively leveraged a wealth of air pollution and health data. They thought the study was comprehensive with thorough analysis and findings that will be of broad interest and value to a wide audience. The study documented that the various exposure models generally resulted in moderately to highly correlated exposure predictions for all pollutants except $PM_{2.5}$. Findings on the presence of an association with various mortality and morbidity outcomes were similar, albeit with sometimes notable differences in the magnitude of the associations.

STRENGTHS OF THE STUDY

The Panel noted several strengths of the research. First, the comparison of a large suite of exposure models commonly used in epidemiological studies was notable and made the results relevant and widely applicable. The study included a variety of models such as LUR models based on fixed-site routine monitoring, mobile monitoring, and dispersion models. The study also investigated state-of-the-art hybrid models that combined monitoring data, land-use data, satellite observations, and estimates from chemical transport models. Moreover, the study assessed four pollutants, including UFPs and BC for which the evidence base is still limited, partly due to the lack of comprehensive routine monitoring.

Second, the Panel thought the specific assessment of the potential role of low-cost sensors in developing exposure models was particularly useful because low-cost sensors are increasingly used. The investigators performed state-of-the-art calibrations of the NO_2 low-cost sensors, yet they still did not find them helpful in predicting long-term exposures once other more common sources of exposure were included in prediction models.

Third, the extensive validation efforts at both the nationwide and city-specific domains and the reporting of multiple measures to test the performance of the exposure models are strengths. The investigators made use of new and existing validation data spanning a 20-year period. The comparisons of predictions at 20,000 randomly selected addresses and the cohort participants' addresses were useful for the evaluation. Moreover, the investigators reported multiple measures of performance — not only correlations and explained variance but also possible bias and systematic differences using, for example, Bland-Altman plots — thereby providing an in-depth performance assessment.

Fourth, the application of the various exposure models in relation to health outcomes was another strength. The investigators applied the various exposure estimates to three population-based cohorts in the Netherlands that differ in size, population, location, health outcomes, and study period. In particular, the health analysis for a very large population (10.8 million) that included all Dutch adult citizens aged 30 years or older was considered informative. The inclusion of both mortality and morbidity outcomes and the availability of lifestyle factors (e.g., smoking and diet) in the two smaller cohorts were also noted as strengths.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results, as explained next.

LIMITATIONS IN THE VALIDATION DATA

The Panel was impressed by the extensive validation efforts and the leverage of many datasets but noted some limitations, particularly in the utility of the new validation data. The short duration of the measurements, the nonsimultaneous sampling, and the small number of repeated measurements for the different seasons hampered the utility of these new data for validating the different exposure assessment approaches. The monitoring campaign had to be postponed because of the COVID-19 pandemic, and only limited equipment (4 instruments) was available. Eventually, in 31 locations, 1-week measurements were available in two different seasons, but this monitoring approach fell short of the original plan to conduct measurements in 100 locations in two different seasons. Those issues limited the ability to extrapolate the measurements reliably to annual mean exposures and then compare them with the various predictions from annual average air pollution models.

Indeed, large temporal variation was evident for all four pollutants in the new measurement data, even after accounting for temporal differences between measurement weeks using a reference site. Moreover, the investigators did not interpret the data for BC and $PM_{2.5}$ because of the low correlation between the first and second week of measurements. The investigators added a post-hoc analysis where instead of comparing individual site average measurements and modeling, they ranked observations and averaged over 5 observations to address the temporality issue better, which indeed increased model performance. Regardless, the Panel concurred with the investigators that the temporal coverage of the new validation data was a limitation, but thought there was value for other researchers in the lessons learned from this effort, as nicely described in the report's discussion section.

The comparisons using existing RUPIOH and Exposomics validation data were limited by similar temporal coverage issues. The cleanest comparison set was from TRACHEA, where simultaneous 1-week measurements of NO_2 were available in four seasons in 2007 at 144 locations. The comparison using fixed-site national monitor data with complete temporal coverage was comprehensive, but problematic for a few models (i.e., dispersion models or low-cost sensor models) because these models were developed using those same data. Moreover, to compare the dispersion model results using fixed-site national monitor data, the investigators used the nearest residential address instead of the exact monitoring site location, which could have influenced the results.

Due to limitations in the available validation data, and because model results were generally highly correlated, no clear preference for a specific exposure assessment approach or model development algorithm was stated. Hence, more work is needed to evaluate alternative exposure models and assessment approaches. Despite those limitations, the validation efforts were extensive and thorough and should interest many readers.

FOCUS ON ESTIMATING OUTDOOR RESIDENTIAL EXPOSURE

Like many other outdoor air pollution and health studies, the current study only focused on estimating outdoor concentrations at residential locations. There is also interest in how this outdoor pollution contributes to personal exposure, although there is no reason to expect this to vary by the exposure assessment methods evaluated in the current study. Steps toward assessing personal exposure to outdoor pollution include accounting for individual time–activity patterns, including commuting and time spent at nonresidential locations, rather than only residential addresses. Outdoor and indoor concentration measurements can also be paired to estimate infiltration factors. These steps are challenging, costly, and demanding to perform for large cohorts, and thus were understandably beyond the scope of this study.

Lack of consideration of infiltration rates and time–activity data adds exposure measurement error, which is often assumed to bias the estimated outdoor air pollution and health estimates toward the null, although the nature of the potential bias cannot be fully known (Kioumourtzoglou et al. 2014; Sheppard et al. 2012). A recent review, which was part of the upcoming HEI report of de Hoogh funded under the same RFA as the current study, reported similar health effects among studies using residential vs. time-integrated exposure assessment of air pollution exposure, documenting that the overall bias in epidemiological studies might likely be small (Hoek et al. 2024).

Methodological complexity and lack of data prevented the investigators from propagating exposure measurement error into the health effects estimations. How to propagate exposure measurement error into health effects estimation in long-term air pollution and health studies remains an area of future research (Samoli and Butland 2017).

HETEROGENEITY IN EFFECT ESTIMATES

Substantial heterogeneity was found in the magnitude — but not direction — of the air pollution associations with the different exposure assessment methods within and across studies. The Panel appreciated the investigators' attempts to explore heterogeneity, even though it was not fully explained by the various factors considered in the study. Several factors were considered, such as the performance of the model, the predicted exposure contrast, and the year of the exposure model. Some of the factors, however, such as spatial resolution, temporal coverage, number of sites, and modeling domain, are difficult to separate from other factors because many of them are related, as noted by the authors. The Panel commented that the heterogeneity could be informative and warrants further examination. Relatedly, the Panel thought the discussion regarding the null or sometimes even negative associations in the EPIC-NL cohort could have been deepened.

The Panel found the implications from an illustrative meta-analysis of the different exposure assessment methods intriguing for evidence synthesis. The investigators showed elegantly that differences in exposure assessment can already lead to high heterogeneity in the same epidemiological study. Hence, it is unsurprising that high heterogeneity is typically reported in systematic reviews and meta-analyses of the health effects of air pollution (e.g., Chen and Hoek 2020; Huangfu and Atkinson 2020; Kasdagli et al. 2024; Orellano et al. 2024) because these analyses compare and integrate results from studies that differ in many methods and design features beyond the exposure assessment. Similarly, in HEI's low-exposure epidemiology initiative, substantial heterogeneity was found in the magnitudes of the positive associations within and across studies (Boogaard et al. 2024), with heterogeneity only slightly reduced in a harmonized analysis using the same exposure model, outcome definition, population age, covariates, and statistical models (Chen 2023). The meta-analysis in the current study implies that for evidence synthesis, caution is warranted in downgrading a body of literature based on heterogeneity statistics — an important point also made in a paper that summarizes the lessons learned from HEI's systematic review on traffic-related air pollution (Boogaard et al. 2023; HEI 2022). This finding is important because heterogeneity in study results is often misinterpreted as a sign of weak or inconsistent evidence, potentially undermining confidence in a body of literature. Recognizing this nuance allows for a more accurate and fair synthesis of evidence, supporting well-informed public health and policymaking decisions.

Of note in the cohort applications is that the main analyses focused on single-pollutant models. This choice is understandable given the volume of results already generated by the project. However, it is important to note that conclusions drawn from two-pollutant models might differ importantly from those drawn from single-pollutant models. This was demonstrated by the limited two-pollutant models that were run using DUELS. In most analyses using the air pollution dispersion model, only one of the two pollutants remained statistically significantly associated with natural-cause mortality. In contrast, in analyses with the ELAPSE and EXPANSE models, both pollutants remained associated with that outcome. Future research should prioritize the inclusion of multipollutant models to enhance the robustness and accuracy of findings regarding the health effects of air pollution.

SUMMARY AND CONCLUSION

The study by Dr. Hoek and colleagues compared the performance of a suite of air pollution exposure assessment methods in the Netherlands for four pollutants (UFPs, BC, $PM_{2.5}$, NO_2). It included a comparison of health effects estimates derived from these methods. The predictions of the exposure models were compared at random residential addresses in the Netherlands and tested on existing and new validation data. Epidemiological analyses in three cohort studies were conducted to compare health effect estimates of the various exposure assessment methods. For all cohort participants, the various exposure estimates were assigned to the residential address at recruitment to the cohort.

The study leveraged a wealth of air pollution and health data from previous collaborative projects and included a large variety of models commonly applied in epidemiological studies, making the results relevant and widely applicable. The extensive validation efforts at both the nationwide and city-specific domains and the reporting of multiple measures to test the performance of the exposure models were considered additional strengths. The investigators used new and existing validation data spanning a 20-year period. Applying those models in relation to various health outcomes in three different cohorts was another strength. In particular, the health analysis for a very large population (10.8 million) that included all Dutch adult citizens aged 30 years or older was considered informative.

The study documented that the various exposure models generally resulted in moderately to highly correlated exposure predictions for all pollutants except $PM_{2.5}$. Findings on the presence of an association with various mortality and morbidity outcomes were similar, albeit with sometimes notable differences in the magnitude of the associations.

Although the Panel broadly agreed with the investigators' conclusions, some limitations should be considered when interpreting the results. The comparison of the different exposure approaches using new validation data was hampered by the short duration of the measurements, the nonsimultaneous sampling, and the small number of repeated measurements for the different seasons. Those issues limited the ability to extrapolate the measurements reliably to annual mean exposures and then compare them with the various annual average air pollution models. Some comparisons using existing validation data were limited by similar temporal coverage issues. Substantial heterogeneity was found in the magnitude — but not direction — of the air pollution associations of the different exposure assessment methods within and across studies. The Panel thought the heterogeneity was not fully explained by the various factors considered in the study and warrants further examination.

The comprehensive report includes many findings that will be of broad interest and value to a wide audience.

ACKNOWLEDGMENTS

The HEI Review Committee is grateful to the Improved Exposure Assessment Studies Review Panel for their thorough review of the study. The Committee is also grateful to Dan Crouse for his oversight of the study, to Hanna Boogaard for assistance with reviewing the report and in preparing its Commentary, to Tom Zaczkiewicz for editing this Report and its Commentary, and to Kristin Eckles and Hope Green for their roles in preparing this Research Report for publication.

REFERENCES

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X. 2013. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe: the ESCAPE project. Atmos Environ 72:10–23, *https://doi.org/10.1016/j.atmosenv.2013.02.037*.

Boogaard H, Atkinson RW, Brook JR, Chang HH, Hoek G, Hoffmann B, et al. 2023. Evidence synthesis of observational studies in environmental health: lessons learned from a systematic review on traffic-related air pollution. Environ Health Perspect 131:115002, *https://doi.org/10.1289/ehp11532*.

Boogaard H, Crouse DL, Tanner E, Mantus E, van Erp AM, Vedal S, et al. 2024. Assessing adverse health effects of long-term exposure to low levels of ambient air pollution: The HEI experience and what's next? Environ Sci Technol 58:12767–12783, *https://doi.org/10.1021/acs.est.3c09745*.

Brunekreef B, Strak M, Chen J, Andersen ZJ, Atkinson R, Bauwelinck M, et al. 2021. Mortality and Morbidity Effects of Long-Term Exposure to Low-Level $PM_{2.5}$, BC, NO_2, and O_3: An Analysis of European Cohorts in the ELAPSE project. Research Report 208. Boston, MA: Health Effects Institute.

Chen J, Braun D, Christidis T, Cork M, Rodopoulou S, Samoli E, et al. 2023. Long-term exposure to low-level $PM_{2.5}$ and mortality: investigation of heterogeneity by harmonizing analyses in large cohort studies in Canada, United States, and Europe. Environ Health Perspect 131:127003, *https://doi.org/10.1289/ehp12141*.

Chen J, Hoek G. 2020. Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis. Environ Int 143:105974, *https://doi.org/10.1016/j.envint.2020.105974*.

de Hoogh K, Chen J, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. 2018. Spatial $PM_{2.5}$, NO_2, O_3 and BC models for Western Europe: evaluation of spatiotemporal stability. Environ Int 120:81–92, *https://doi.org/10.1016/j.envint.2018.07.036*.

Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, et al. 1993. An association between air pollution and mortality in six US cities. N Engl J Med 329:1753–1759, *https://doi.org/10.1056/nejm199312093292401*.

EC (European Commission). 2022. Proposal for a Directive of the European Parliament and the Council on ambient air quality and cleaner air for Europe. Brussels, 26.10.2022. COM (2022) 542 final.

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. 2012. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas: results of the ESCAPE project. Environ Sci Technol 46:11195–11205, *https://doi.org/10.1021/es301948k*.

HEI Panel on the Health Effects of Traffic-Related Air Pollution. 2010. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Special Report 17. Boston, MA: Health Effects Institute.

Health Effects Institue (HEI) 2022. Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute.

HEI Review Panel on Ultrafine Particles. 2013. Understanding the Health Effects of Ambient Ultrafine Particles. Perspectives 3. Boston MA: Health Effects Institute.

Hoek G. 2017. Methods for assessing long-term exposures to outdoor air pollutants. Curr Environ Health Rep 4:450–462, *https://doi.org/10.1007/s40572-017-0169-5*.

Hoek G, Vienneau D, de Hoogh K. 2024. Does residential address-based exposure assessment for outdoor air pollution lead to bias in epidemiological studies? Environ Health 23:75, *https://doi.org/10.1186/s12940-024-01111-0*.

Huangfu P, Atkinson R. 2020. Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: a systematic review and meta-analysis. Environ Int 144:105998, *https://doi.org/10.1016/j.envint.2020.105998*.

IARC (International Agency for Research on Cancer) Working Group on the Evaluation of Carcinogenic Risks to Humans. 2016. Outdoor air pollution. IARC Monogr Eval Carcinog Risks Hum 109:9–444.

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. 2005. A review and evaluation of intraurban air pollution exposure models. J Exposure Anal Environ Epidemiol 15:185–204, *https://doi.org/10.1038/sj.jea.7500388*.

Kasdagli MI, Orellano P, Pérez Velasco R, Samoli E. 2024. Long-term exposure to nitrogen dioxide and ozone and mortality: update of the WHO air quality guidelines systematic review and meta-analysis. Int J Public Health 69:1607676, *https://doi.org/10.3389/ijph.2024.1607676*.

Kioumourtzoglou MA, Spiegelman D, Szpiro AA, Sheppard L, Kaufman JD, Yanosky JD, et al. 2014. Exposure measurement error in $PM_{2.5}$ health effects studies: a pooled analysis of eight personal exposure validation studies. Environ Health 13:2, *https://doi.org/10.1186/1476-069x-13-2*.

Kerckhoffs J, Hoek G, Messier KP, Brunekreef B, Meliefste K, Klompmaker JO, Vermeulen R. 2016. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ Sci Technol 50:12894–12902, *https://doi.org/10.1021/acs.est.6b03476*.

Kerckhoffs J, Hoek G, Vlaanderen J, van Nunen E, Messier K, Brunekreef B, et al. 2017. Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ Res 159:500–508, *https://doi.org/10.1016/j.envres.2017.08.040*.

Kerckhoffs J, Hoek G, Gehring U, Vermeulen R. 2021. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ Int 154:106569, *https://doi.org/10.1016/j.envint.2021.106569*.

Kerckhoffs J, Khan J, Hoek G, Yuan Z, Hertel O, Ketzel M, et al. 2022. Hyperlocal variation of nitrogen dioxide, black carbon, and ultrafine particles measured with Google Street View cars in Amsterdam and Copenhagen. Environ Int 170:107575, *https://doi.org/10.1016/j.envint.2022.107575*.

Klompmaker JO, Janssen N, Andersen ZJ, Atkinson R, Bauwelinck M, Chen J, et al. 2021. Comparison of associations between mortality and air pollution exposure estimated with a hybrid, a land-use regression and a dispersion model. Environ Int 146:106306, *https://doi.org/10.1016/j.envint.2020.106306*.

Orellano P, Kasdagli MI, Pérez Velasco R, Samoli E. 2024. Long-term exposure to particulate matter and mortality: an update of the WHO global air quality guidelines systematic review and meta-analysis. Int J Public Health 69:1607683, *https://doi.org/10.3389/ijph.2024.1607683*.

Pope CA 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287:1132–1141, *https://doi.org/10.1001/jama.287.9.1132*.

Samoli E, Butland BK. 2017. Incorporating measurement error from modeled air pollution exposures into epidemiological analyses. Curr Environ Health Rep 4:472–480, *https://doi.org/10.1007/s40572-017-0160-1*.

Sheppard L, Burnett RT, Szpiro AA, Kim SY, Jerrett M, Pope CA 3rd, et al. 2012. Confounding and exposure measurement error in air pollution epidemiology. Air Qual Atmos Health 5:203–216, *https://doi.org/10.1007/s11869-011-0140-9*.

Shen Y, de Hoogh K, Schmitz O, Clinton N, Tuxen-Bettman K, Brandt J, et al. 2022. Europe-wide air pollution modeling from 2000 to 2019 using geographically weighted regression. Environ Int 168:107485, *https://doi.org/10.1016/j.envint.2022.107485*.

US EPA (United States Environmental Protection Agency). 2016. Integrated science assessment for oxides of nitrogen–health Criteria. EPA/600/R-15/068. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency), 2019. Integrated science assessment for particulate matter. EPA/600/R-19/188. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency). 2022. Supplement to the 2019 integrated science assessment for particulate matter. EPA/600/R-22/028. Washington, DC: US EPA.

US EPA (United States Environmental Protection Agency). 2024. Final rule: Reconsideration of the national ambient air quality standards for particulate matter. Available: *https://www.epa.gov/pm-pollution/final-reconsideration-national-ambient-air-quality-standards-particulate-matter-pm* [accessed November 13, 2024].

Velders GJM, Willers SM, Wesseling J, den Elshout SV, van der Swaluw E, Mooibroek D, et al. 2021. Improvements in air quality in the Netherlands during the corona lockdown based on observations and model simulations. Atmos Environ 247:118158, *https://doi.org/10.1016/j.atmosenv.2020.118158*.

Wesseling J, de Ruiter H, Blokhuis C, Drukker D, Weijers E, Volten H, et al. 2019. Development and implementation of a platform for public information on air quality, sensor measurements, and citizen science. Atmosphere 10:445, *http://dx.doi.org/10.3390/atmos10080445*.

WHO (World Health Organization). 2021. WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. Geneva: WHO.

Yu Z, Koppelman GH, Boer JMA, Hoek G, Kerckhoffs J, Vonk JM, et al. 2022. Ambient ultrafine particles and asthma onset until age 20: the PIAMA birth cohort. Environ Res 214:113770, *https://doi.org/10.1016/j.envres.2022.113770*.

Yu Z, Koppelman GH, Hoek G, Kerckhoffs J, Vonk JM, Vermeulen R, et al. 2021. Ultrafine particles, particle components and lung function at age 16 years: the PIAMA birth cohort study. Environ Int 157:106792, *https://doi.org/10.1016/j.envint.2021.106792*.

ABBREVIATIONS AND OTHER TERMS

BAG	Dutch address database
BC	black carbon
CI	confidence interval
DUELS	Dutch Environmental Longitudinal Study
EC	elemental carbon
ESCAPE	European Study of Cohorts for Air Pollution Effects
ELAPSE	Effects of Low-Level Air Pollution: A Study in Europe
EPIC-NL	European Prospective Investigation into Cancer and Nutrition: Netherlands cohort
EXPANSE	EXposome Powered tools for healthy living in urbAN Settings
FEV_1	forced expiratory volume in 1 second
FVC	forced vital capacity
GIS	geographic information system
HR	hazard ratio
IQR	interquartile range
LASSO	Least Absolute Shrinkage and Selection Operator
LUR	land use regression
MAPLE	Mortality–Air Pollution Associations in Low-Exposure Environments
MORGEN	Monitoring Project on Risk Factors for Chronic Diseases
MUSIC	Measurements of Ultrafine particles and Soot In Cities
NAQMN	National Air Quality Monitoring Network
NO_2	nitrogen dioxide
O_3	ozone
OR	odds ratio
PIAMA	Prevention and Incidence of Asthma and Mite Allergy
$PM_{2.5}$	particulate matter ≤ 2.5 µm in aerodynamic diameter
PM_{10}	particulate matter ≤ 10 µm in aerodynamic diameter
R	correlation coefficient
R^2	coefficient of determination or explained variance
RIVM	National Institute for Public Health and the Environment
RMSE	root mean squared error
RR	relative risk
RUN	nationwide mobile monitoring project in the Netherlands
RUPIOH	Relationship between Ultrafine and fine Particulate matter in Indoor and Outdoor air and respiratory Health
SLR	supervised linear regression
TRACHEA	Traffic-Related Air pollution and Children's respiratory HEalth and Allergies
UFP	ultrafine particles
WHO	World Health Organization

RELATED HEI PUBLICATIONS

Number	Title	Principal Investigator	Date
Research Reports			
217	Long-Term Exposure to Outdoor Ultrafine Particles and Black Carbon and Effects on Mortality in Montreal and Toronto, Canada	S. Weichenthal	2024
216	Scalable Multipollutant Exposure Assessment Using Routine Mobile Monitoring Platforms	J. Apte	2024
212	Mortality–Air Pollution Associations in Low-Exposure Environments (MAPLE): Phase 2	M. Brauer	2022
211	Assessing Adverse Health Effects of Long-Term Exposure to Low Levels of Ambient Air Pollution: Implementation of Causal Inference Methods	F. Dominici	2022
208	Mortality and Morbidity Effects of Long-Term Exposure to Low-Level $PM_{2.5}$, BC, NO_2, and O_3: An Analysis of European Cohorts in the ELAPSE Project	B. Brunekreef	2021
203	Mortality–Air Pollution Associations in Low-Exposure Environments (MAPLE): Phase 1	M. Brauer	2019
202	Enhancing Models and Measurements of Traffic-Related Air Pollutants for Health Studies Using Dispersion Modeling and Bayesian Data Fusion	S. Batterman	2020
200	Assessing Adverse Health Effects of Long-Term Exposure to Low Levels of Ambient Air Pollution: Phase 1	F. Dominici	2019
196	Developing Multipollutant Exposure Indicators of Traffic Pollution: The Dorm Room Inhalation to Vehicle Emissions (DRIVE) Study	J. Sarnat	2018
194	A Dynamic Three-Dimensional Air Pollution Exposure Model for Hong Kong	B. Barratt	2018
177	National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components	M. Lippman	2013
139	Effects of Long-Term Exposure to Traffic-Related Air Pollution on Respiratory and Cardiovascular Mortality in the Netherlands: The NLCS-AIR Study	B. Brunekreef	2009
Perspectives			
3	Understanding the Health Effects of Ambient Ultrafine Particles	HEI	2013
Special Reports			
23	Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution	HEI	2022
17	Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects	HEI	2010

PDFs are available for free downloading at www.healtheffects.org/publications.

HEI BOARD, COMMITTEES, and STAFF

BOARD OF DIRECTORS

Richard A. Meserve, Chair *Senior of Counsel, Covington & Burling LLP; President Emeritus, Carnegie Institution for Science; former Chair, US Nuclear Regulatory Commission, USA*

Jo Ivey Boufford *Clinical Professor of Global Health and Pediatrics, New York University, USA*

Homer A. Boushey *Emeritus Professor of Medicine, University of California San Francisco, USA*

Stephen Corman *President, Corman Enterprises, USA*

Martha J. Crawford *Operating Partner, Macquarie Asset Management, USA*

Ana V. Diez Roux *Dana and David Dornsife Dean and Distinguished University Professor of Epidemiology, Dornsife School of Public Health, Drexel University; Director, Drexel Urban Health Collaborative, USA*

Michael J. Klag *Dean Emeritus and Second Century Distinguished Professor, Johns Hopkins Bloomberg School of Public Health, USA*

Alan I. Leshner *CEO Emeritus, American Association for the Advancement of Science, USA*

Catherine L. Ross *Regents' Professor Emerita, City and Regional Planning and Civil and Environmental Engineering, Georgia Institute of Technology; Chairman of the Board of Directors of the Auto Club Group, American Automobile Association, USA*

Martha E. Rudolph *Environmental Attorney, Former Director of Environmental Programs, Colorado Department of Public Health and Environment, USA*

Karen C. Seto *Frederick C. Hixon Professor of Geography and Urbanization Science, Yale School of the Environment, Yale University, USA*

Jared L. Cohon *President Emeritus and Professor, Civil and Environmental Engineering and Engineering and Public Policy, Carnegie Mellon University, USA, In Memoriam 1947–2024*

RESEARCH COMMITTEE

David A. Savitz, Chair *Professor of Epidemiology, School of Public Health, and Professor of Obstetrics and Gynecology and Pediatrics, Alpert Medical School, Brown University, USA*

Benjamin Barratt *Professor, Environmental Research Group, School of Public Health, Imperial College London, United Kingdom*

David C. Dorman *Professor, Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, USA*

Christina H. Fuller *Associate Professor, School of Environmental, Civil, Agricultural and Mechanical Engineering, University of Georgia College of Engineering, USA*

Marianne Hatzopoulou *Professor, Civil and Mineral Engineering, University of Toronto, Research Chair in Transport Decarbonization and Air Quality, Canada*

Heather A. Holmes *Associate Professor, Department of Chemical Engineering, University of Utah, USA*

Neil Pearce *Professor of Epidemiology and Biostatistics, London School of Hygiene and Tropical Medicine, United Kingdom*

Evangelia (Evi) Samoli *Professor of Epidemiology and Medical Statistics, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece*

Alexandra M. Schmidt *Professor of Biostatistics, School of Population and Global Health, McGill University, Canada*

Neeta Thakur *Associate Professor of Medicine, University of California San Francisco, USA*

Gregory Wellenius *Professor, Department of Environmental Health, Boston University School of Public Health and Director, BUSPH Center for Climate and Health, USA*

continued on next page

HEI BOARD, COMMITTEES, and STAFF

REVIEW COMMITTEE

Melissa J. Perry, Chair *Dean, College of Public Health, George Mason University, USA*

Sara D. Adar *Professor of Epidemiology and Global Public Health, Department of Epidemiology, University of Michigan School of Public Health, USA*

Kiros T. Berhane *Cynthia and Robert Citrone-Roslyn and Leslie Goldstein Professor and Chair, Department of Biostatistics, Mailman School of Public Health, Columbia University, USA*

Katherine B. Ensor *Noah G. Harding Professor of Statistics, Rice University, USA*

Ulrike Gehring *Associate Professor, Institute for Risk Assessment Sciences, Utrecht University, Netherlands*

Michael Jerrett *Professor, Department of Environmental Health Sciences, Fielding School of Public Health, University of California Los Angeles, USA*

Frank Kelly *Humphrey Battcock Chair in Community Health and Policy and Director of the Environmental Research Group, Imperial College London School of Public Health, United Kingdom*

Jennifer L. Peel *Professor of Epidemiology, Department of Environmental and Radiological Health Sciences, Colorado State University, and the Colorado School of Public Health, USA*

Eric J. Tchetgen Tchetgen *University Professor and Professor of Biostatistics and Epidemiology, Perelman School of Medicine, and Professor of Statistics and Data Science, The Wharton School, University of Pennsylvania, USA*

John Volckens *Professor, Department of Mechanical Engineering, Walter Scott Jr. College of Engineering, Colorado State University, USA*

Scott Weichenthal *Professor, Department of Epidemiology, Biostatistics, and Occupational Health, School of Population and Global Health, McGill University, Canada*

STAFF AND CONSULTING SCIENTISTS

Elena Craft *President and CEO*

Ellen K. Mantus *Director of Science*

Donna J. Vorhees *Director of HEI Energy*

Thomas J. Champoux *Director of Science Communications*

Jacqueline C. Rutledge *Director of Finance and Administration*

Jason Desmond *Deputy Director of Finance and Administration*

Nikola Sobot *Director of Development*

Emily Alden *Corporate Secretary*

Amy Andreini *Science Communications Specialist*

Hanna Boogaard *Consulting Principal Scientist*

Aaron J. Cohen *Consulting Principal Scientist*

Jacki Collins *Senior Staff Accountant*

Dan Crouse *Senior Scientist*

Cloelle Danforth *Senior Scientist*

Gabriela Daza *Research Assistant*

Philip J. DeMarco *Compliance Manager*

Continues next page

(*Staff and Consulting Scientists, continued*)

Kristin C. Eckles *Senior Editorial Manager*
Elise G. Elliott *Staff Scientist*
Hope Green *Editorial Project Manager*
Hlina Kiros *Research Assistant*
Lissa McBurney *Senior Science Administrator*
Janet McGovern *Executive Assistant*
Samantha Miller *Research Assistant*
Victor Nthusi *Consulting Research Fellow*
Pallavi Pant *Head of Global Initiatives*
Allison P. Patton *Senior Scientist*
Yasmin Romitti *Staff Scientist*
Anna Rosofsky *Senior Scientist and Community Health and Environmental Research Initiatives Lead*
Abinaya Sekar *Consulting Research Fellow*
Robert Shavers *Operations Manager*
Eva Tanner *Staff Scientist*
Alexis Vaskas *Digital Communications Manager*

Daniel S. Greenbaum *President Emeritus, In Memoriam 1952–2024*
Robert M. O'Keefe *Vice President Emeritus*
Annemoon M. van Erp *Deputy Director of Science Emerita*

RESEARCH REPORT

NUMBER 226
MARCH 2025

Health Effects Institute

75 Federal Street
Suite 1400
Boston, Massachusetts, 02110, USA
+1-617-488-2300

www.healtheffects.org