HEI

ADDITIONAL MATERIALS

Research Report 222

Cardiometabolic Health Effects of Air Pollution, Noise, Green Space, and Socioeconomic Status: The HERMES Study

Ole Raaschou-Nielsen et al.

Appendix Tables A1–A4, Appendix Figures A1–A4, and Method for Multiexposure Analyses

The Additional Materials were reviewed for spelling, grammar, and cross-references to the main report. It has not been formatted or fully edited by HEI. This document was reviewed by the HEI Review Committee.

Correspondence may be addressed to Dr. Ole Raaschou-Nielsen at the Danish Cancer Institute, Strandboulevarden 49,2100 Copenhagen, Denmark; email: *ole@cancer.dk*.

Although this document 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, therefore, may not necessarily reflect the views of the Agency; no official endorsement by it should be inferred. The contents of this document have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties; no endorsement by them should be inferred.

HEI Raaschou-Nielsen Research Report 222

Additional Materials

Contents

Appendix Table A1. Associations between air pollution and type-2 diabetes in the Danish population, 2005–2017, in three adjustment models.

		HR (95% CI) per IQR		
Air Pollutant	**IQR**	**Model 1**[1]	**Model 2**[2]	**Model 3**[3]
$PM_{2.5}$ ($\mu g/m^3$)				
Total	1.85	1.013 (1.003; 1.024)	1.043 (1.032; 1.054)	1.043 (1.031; 1.056)
Other sources	1.63	0.988 (0.976; 0.999)	1.031 (1.019; 1.043)	1.020 (1.007; 1.032)
Local traffic	0.37	1.020 (1.016; 1.025)	1.019 (1.015; 1.024)	1.026 (1.020; 1.031)
Ultrafine particles (#/cm^3)				
Total	4,248	1.006 (0.998; 1.014)	1.043 (1.035; 1.052)	1.052 (1.042; 1.063)
Other sources	2,769	0.993 (0.985; 1.000)	1.030 (1.023; 1.037)	1.027 (1.019; 1.036)
Local traffic	1,698	1.021 (1.014; 1.028)	1.036 (1.030; 1.043)	1.049 (1.040; 1.058)
Elemental carbon ($\mu g/m^3$)				
Total	0.28	1.007 (1.002; 1.013)	1.021 (1.016; 1.026)	1.022 (1.016; 1.027)
Other sources	0.12	0.989 (0.984; 0.995)	1.005 (1.002; 1.008)	1.003 (0.999; 1.007)
Local traffic	0.17	1.023 (1.017; 1.028)	1.028 (1.022; 1.033)	1.037 (1.030; 1.043)
NO_2 ($\mu g/m^3$)				
Total	7.15	1.023 (1.016; 1.030)	1.041 (1.034; 1.048)	1.056 (1.046; 1.065)
Other sources	2.68	1.006 (0.998; 1.014)	1.046 (1.037; 1.054)	1.043 (1.034; 1.053)
Local traffic	5.17	1.023 (1.016; 1.029)	1.030 (1.024; 1.036)	1.039 (1.031; 1.047)

[1] Model 1: adjusted for age, sex, and calendar year.
[2] Model 2: Model 1 plus adjustment for marital status, individual and family income, country of origin, occupational status, and education.
[3] Model 3: Model 2 plus an adjustment for area-level percentage of the population with low income, with only basic education, who are unemployed, with manual labor, with a non-Western background, with a criminal record, who are sole-provider, and who live in social housing.

Appendix Table A2. Associations between air pollution and myocardial infarction in the Danish population, 2005–2017, in three adjustment models.

		HR (95% CI) per IQR		
Air Pollutant	**IQR**	**Model 1[1]**	**Model 2[2]**	**Model 3[3]**
$PM_{2.5}$ (µg/m^3)				
Total	1.85	0.983 (0.969; 0.998)	1.019 (1.004; 1.034)	1.053 (1.035; 1.071)
Other sources	1.63	0.989 (0.973; 1.006)	1.032 (1.015; 1.049)	1.051 (1.032; 1.069)
Local traffic	0.37	0.991 (0.985; 0.998)	0.996 (0.990; 1.003)	1.011 (1.003; 1.018)
Ultrafine particles (#/cm^3)				
Total	4,248	0.961 (0.950; 0.972)	1.001 (0.990; 1.013)	1.040 (1.025; 1.055)
Other sources	2,769	0.974 (0.964; 0.985)	1.014 (1.003; 1.024)	1.034 (1.022; 1.046)
Local traffic	1,698	0.963 (0.953; 0.972)	0.984 (0.974; 0.994)	1.011 (0.999; 1.024)
Elemental carbon (µg/m^3)				
Total	0.28	0.966 (0.957; 0.975)	0.989 (0.981; 0.998)	1.009 (1.000; 1.019)
Other sources	0.12	0.969 (0.960; 0.977)	0.994 (0.987; 1.001)	1.001 (0.996; 1.007)
Local traffic	0.17	0.981 (0.973; 0.989)	0.992 (0.984; 1.000)	1.013 (1.003; 1.023)
NO_2 (µg/m^3)				
Total	7.15	0.969 (0.960; 0.979)	0.994 (0.984; 1.004)	1.027 (1.013; 1.040)
Other sources	2.68	0.984 (0.973; 0.996)	1.025 (1.014; 1.037)	1.048 (1.034; 1.062)
Local traffic	5.17	0.972 (0.963; 0.981)	0.986 (0.977; 0.995)	1.009 (0.998; 1.020)

[1] Model 1: adjusted for age, sex, and calendar year.

[2] Model 2: Model 1 plus adjustment for marital status, individual and family income, country of origin, occupational status, and education.

[3] Model 3: Model 2 plus an adjustment for area-level percentage of the population with low income, with only basic education, who are unemployed, with manual labor, with a non-Western background, with a criminal record, who are sole-provider, and who live in social housing.

Appendix Table A3. Associations between air pollution and stroke in the Danish population, 2005–2017, in three adjustment models.

		HR (95% CI) per IQR		
Air Pollutant	**IQR**	**Model 1[1]**	**Model 2[2]**	**Model 3[3]**
PM$_{2.5}$ (μg/m^3)				
Total	1.85	1.067 (1.054; 1.081)	1.083 (1.069; 1.097)	1.077 (1.061; 1.094)
Other sources	1.63	1.083 (1.068; 1.099)	1.108 (1.092; 1.124)	1.091 (1.074; 1.108)
Local traffic	0.37	1.008 (1.002; 1.014)	1.006 (1.001; 1.012)	1.004 (0.998; 1.011)
Ultrafine particles (#/cm^3)				
Total	4,248	1.021 (1.011; 1.031)	1.041 (1.030; 1.051)	1.039 (1.026; 1.052)
Other sources	2,769	1.025 (1.016; 1.034)	1.045 (1.035; 1.054)	1.038 (1.028; 1.049)
Local traffic	1,698	1.004 (0.995; 1.012)	1.010 (1.002; 1.019)	1.003 (0.992; 1.014)
Elemental carbon (μg/m^3)				
Total	0.28	1.007 (1.000; 1.014)	1.014 (1.007; 1.021)	1.009 (1.001; 1.018)
Other sources	0.12	1.002 (0.998; 1.007)	1.008 (1.004; 1.011)	1.005 (1.000; 1.009)
Local traffic	0.17	1.007 (1.000; 1.014)	1.008 (1.002; 1.015)	1.005 (0.996; 1.013)
NO$_2$ (μg/m^3)				
Total	7.15	1.020 (1.012; 1.029)	1.029 (1.020; 1.038)	1.028 (1.017; 1.040)
Other sources	2.68	1.058 (1.047; 1.069)	1.082 (1.071; 1.093)	1.077 (1.065; 1.089)
Local traffic	5.17	1.005 (0.997; 1.013)	1.007 (0.999; 1.015)	1.001 (0.991; 1.010)

[1] Model 1: adjusted for age, sex, and calendar year.

[2] Model 2: Model 1 plus adjustment for marital status, individual and family income, country of origin, occupational status, and education.

[3] Model 3: Model 2 plus an adjustment for area-level percentage of the population with low income, with only basic education, who are unemployed, with manual labor, with a non-Western background, with a criminal record, who are sole-provider, and who live in social housing.

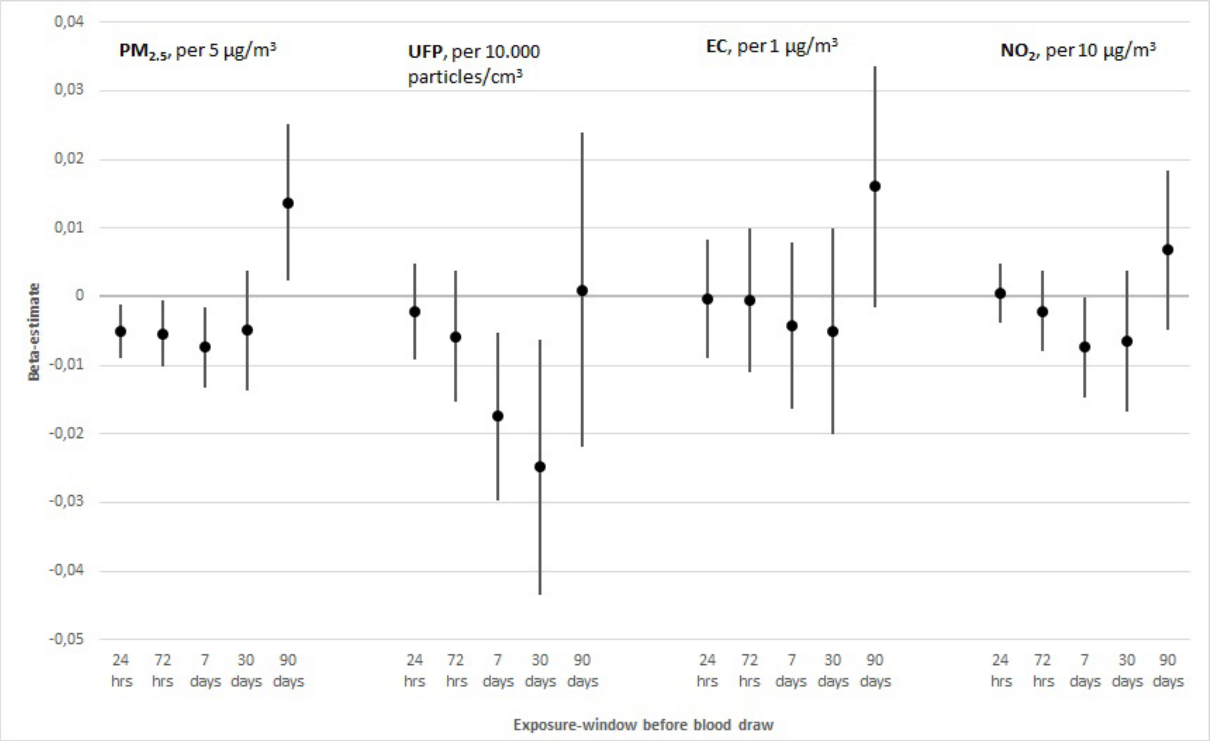
Appendix Table A4. Seasonal variation in lipid levels and blood pressure (mean ± SD).

Reproduced from Roswall et al. 2023 by permission of Elsevier. © 2023 Environmental Research.

	HDL	Non-HDL	Systolic Blood Pressure	Diastolic Blood Pressure
Spring, March–May	1.59 ± 0.43	3.38 ± 1.00	116.08 ± 15.58	80.49 ± 10.79
Summer, June–August	1.59 ± 0.44	3.39 ± 1.01	114.44 ± 15.63	79.47 ± 10.89
Fall, September–November	1.57 ± 0.44	3.44 ± 1.01	116.20 ± 15.85	80.60 ± 10.90
Winter, December–February	1.58 ± 0.44	3.50 ± 1.05	117.18 ± 16.06	80.97 ± 10.90

Appendix Figure A1. Associations (beta-estimates with 95% CI)1 between air pollution means of five time windows, and HDL (“good cholesterol”).

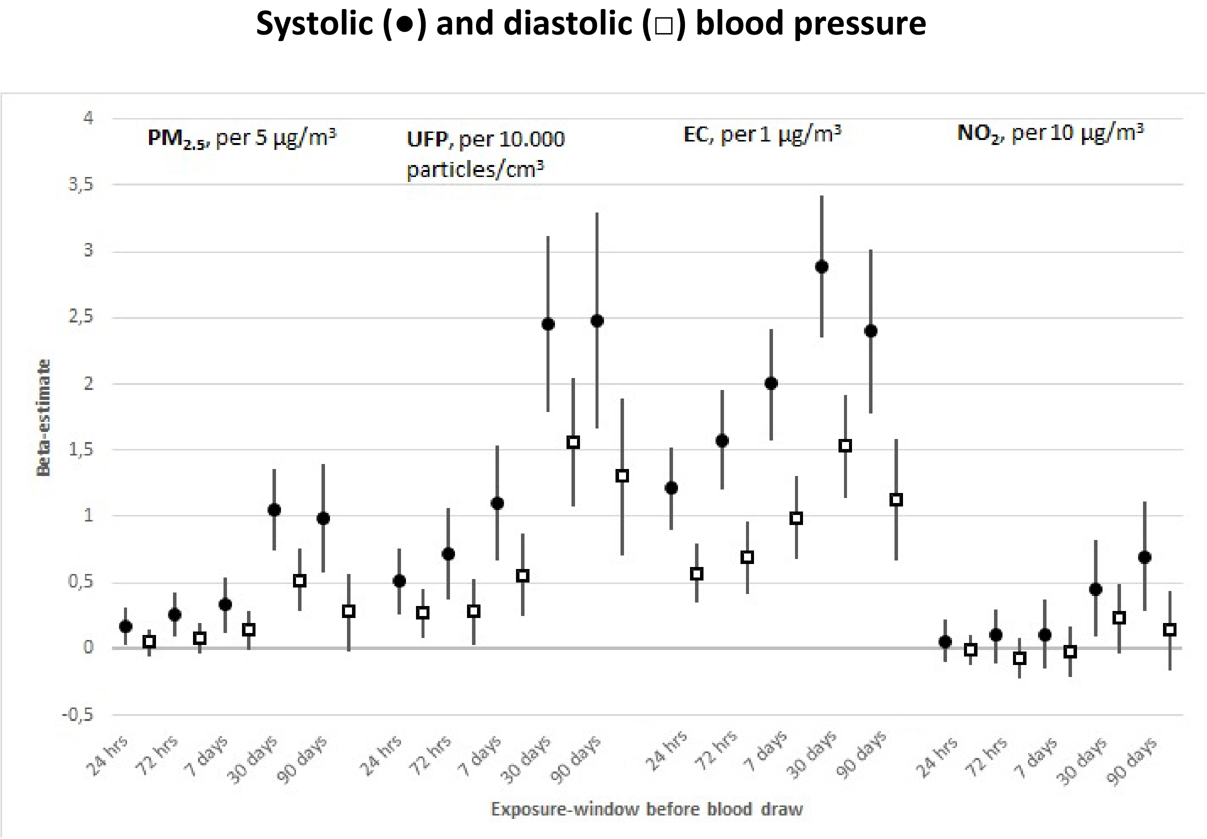
Reproduced from Roswall et al. 2023 by permission of Elsevier. © 2023 Environmental Research.



[1] Adjusted for age, age-squared, sex, marital status, education, income, smoking before blood draw (yes/no), hours since last smoke, environmental tobacco smoke, alcohol before blood draw, physical activity (yes/no), hours of physical activity/week, body mass index, percentage of parish population having low income, having only basic education, living in social housing, and green space at the residence.

Appendix Figure A2. Associations (beta-estimates with 95% CI)1 between air pollution means of five time windows, and systolic and diastolic blood pressure.

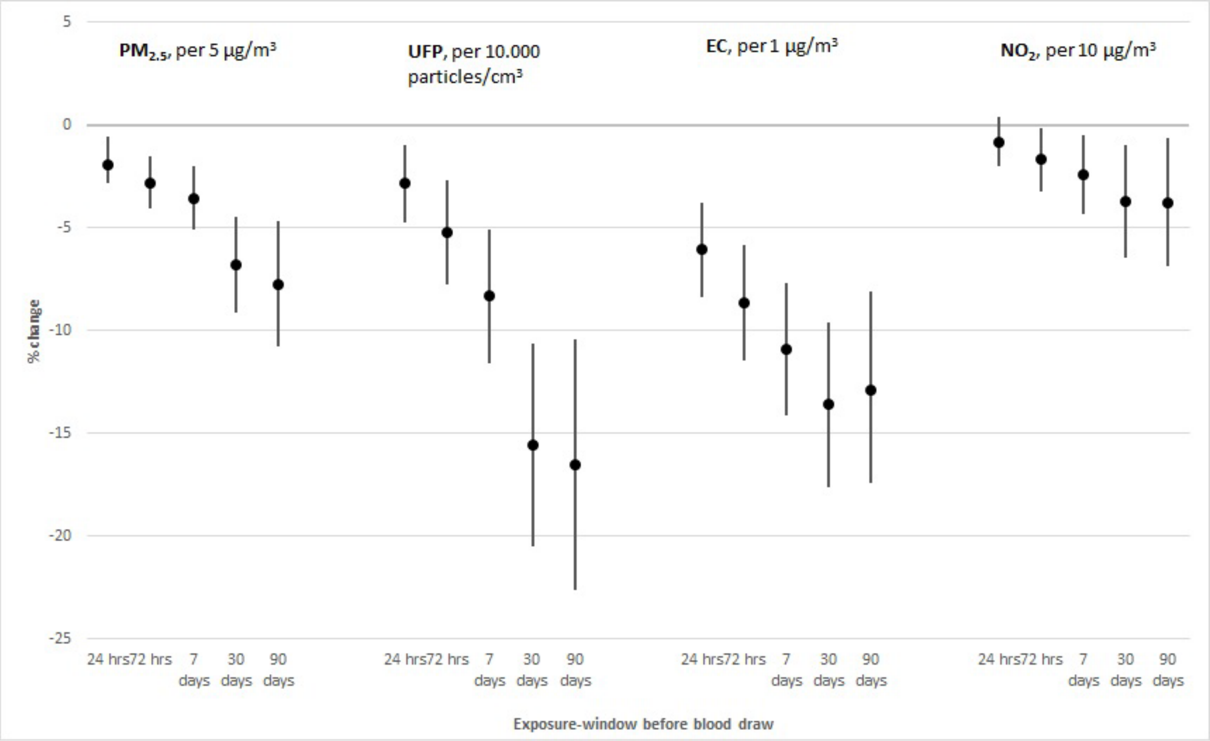
Reproduced from Roswall et al. 2023 by permission of Elsevier. © 2023 Environmental Research.



[1] Adjusted for age, age-squared, sex, marital status, education, income, smoking before blood draw (yes/no), hours since last smoke, environmental tobacco smoke, alcohol before blood draw, physical activity (yes/no), hours of physical activity/week, body mass index, percentage of parish population having low income, having only basic education, living in social housing, and green space at the residence.

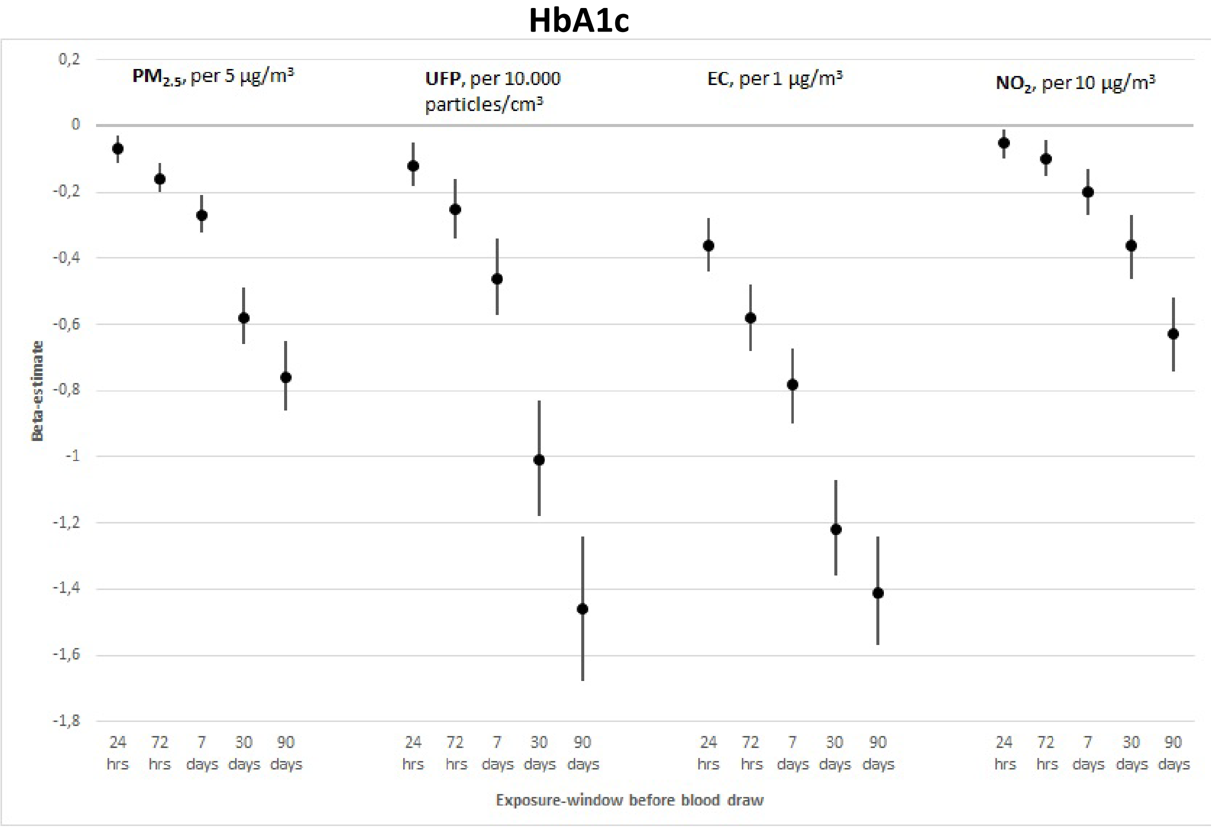
Appendix Figure A3. Associations (percentage with 95% CI)1 between air pollution means of five time windows and CRP.

CRP



[1] Adjusted for age, age-squared, sex, marital status, education, income, smoking before blood draw (yes/no), hours since last smoke, environmental tobacco smoke, alcohol before blood draw, physical activity (yes/no), hours of physical activity/week, body mass index, percentage of parish population having low income, having only basic education, living in social housing, and green space at the residence.

Appendix Figure A4. Associations (beta-estimates with 95% CI)1 between air pollution means of five time windows and HbA1c.



[1] Adjusted for age, age-squared, sex, marital status, education, income, smoking before blood draw (yes/no), hours since last smoke, environmental tobacco smoke, alcohol before blood draw, physical activity (yes/no), hours of physical activity/week, body mass index, percentage of parish population having low income, having only basic education, living in social housing, and green space at the residence.

Method for Multiexposure Analysis

Introduction. It was a secondary aim of the HERMES study to develop a new statistical method to separate the health effects of correlated exposures (multiexposure analyses).

The traditional approach for multiexposure models has been to include only a limited number of pollutants with limited correlation using standard regression methodology. The reason is that the inclusion of highly correlated pollutants in a regression analysis may make parameter estimates unstable and hard to interpret. Reduction in included pollutants has been achieved by different means such as pre-screening to ensure that the correlation between two pollutants is below a threshold, pre-regression analyses where one pollutant is regressed on another and subsequently only including the first pollutant along with the residuals from the pre-regression in the final regression model, or penalized regression as in LASSO, where the model-fitting procedure is tuned to find the smallest subset of pollutants, which still provides an acceptable fit of the observed outcomes. While these approaches each have their merits, they share the same underlying weakness: they depend on a number of parametric assumptions. In particular, it is assumed *a priori* that the pollutants have a linear effect on the outcome and that this effect is not modified by levels of other pollutants or covariates. The assumptions can be relaxed by building more complex regression models. This, however, introduces more arbitrary and hard-to-justify choices in the statistical analysis, which reduces the reproducibility and generalizability of the results. If the underlying parametric assumptions are violated for the pollutant in focus or even for covariates, the conclusions can be invalid.

The idea of a new method. We intended to solve this problem by a fundamentally different modeling approach inspired by the principles guiding the rapidly growing scientific field of causal inference (Pearl 2009; VanderWeele 2015) in combination with the tools of machine learning.

From a statistical perspective, the traditional regression-based approaches for multipollutant data as discussed above aim to describe the expected outcomes for any combination of values for all pollutants, co-pollutants, and covariates. This is a very demanding aim, which can only be achieved by introducing hard-to-justify parametric assumptions. Moreover, that ambition is more than what is needed to address our true scientific question: "How does a change in one of the pollutants affect the outcome, if everything else is kept fixed?" We would address that scientific question without relying on a battery of parametric assumptions.
Specifically, we would approach the problem in two distinct phases. In phase 1, we would employ a random forest, which is a computer-intensive approach. The method works by constructing a large number of decision trees to determine if a given observation is a case; to reduce the correlation between trees, each tree is built on a random subset of the data and includes only a random subset of the predictor variables. The final probability is obtained as the average probability across thousands of such trees. The random forest provides a likelihood of being a case for a given combination of exposures and covariates. (i.e., "if covariate A is below X and covariate B is above Y, but below Z, etc."). The advantage is that the method does not require parametric assumptions or limited dependency between the predictor variables. The random forest methodology has been developed to include survival outcomes and competing risks (Mogensen et al. 2012). While random forests are good at capturing interdependencies and nonlinearities of multipollutant data, they do not provide interpretable parameter estimates, which would be achieved in phase 2 of our approach. In

phase 2, we would initially introduce the pollution scenarios, which we would like to compare to pinpoint the effect of the single pollutants. For example, to shed light on the effect of pollutant A, we could consider scenarios such as the following:

- Scenario 0: Assume all pollutants and covariates were as observed.
- Scenario 1: Assume pollutant A was increased by one unit, while all other pollutants and covariates were as observed.

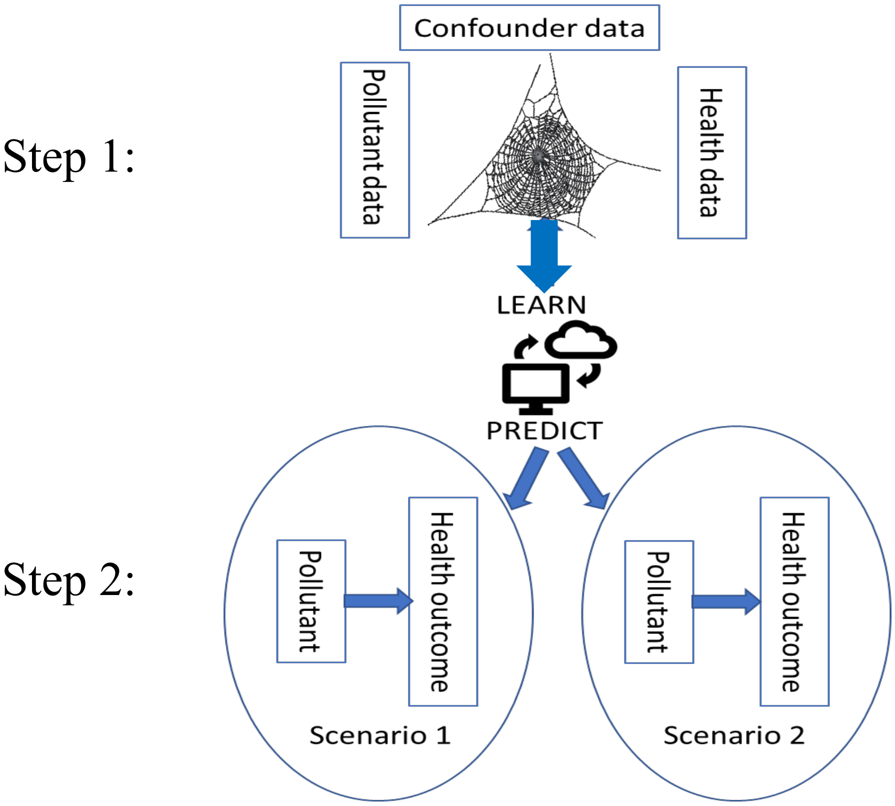
For each of these scenarios, the random forest from Phase 1 would be employed to simulate the likelihood of outcomes. These simulated data sets could be compared using whatever scale is most informative (e.g., hazard ratios to facilitate comparison with traditional Cox models). Standard errors are obtained by multiple iterations of phase 2 for a given pair of scenarios. The use of specified scenarios to analyze the effects of components follows the ideas introduced in the causal inference field. The proposed method can be viewed as a version of G-computation (Robins 1986). While the random forest — and specifically the individual prediction trees — will not in themselves be biologically interpretable, this is not a shortcoming because we only require that the predictions based on averages are accurate.

The script. We have developed a script intended for analyses of multipollutant data following the principles outlined above and shown in Figure 1. The script has two steps: afforestation and then prediction and analysis.

Step 1: Afforestation
Using the R-function *randomForest*, we created a random forest-based fit on the same variables as included in our Cox analyses. All parameters had the same resolution as in the Cox analyses. For each pollutant, the variables included were: the pollutant, age, sex, calendar year, education level, personal and household income, marital status, and ethnicity as well as area-level information on the proportion of inhabitants living in social housing, being sole providers, being unemployed, having non-Western background, having only basic education, having manual labor, or having a criminal record. Due to limitations of the server capacity, a separate random forest was generated for each year of age and combined in the "predict" stage (described below). Random forests were generated for $PM_{2.5}$, UFP, EC, and NO_2. An annotated sample script is included below. A forest was also created including simultaneously all four pollutants and further forests were generated later in the exploration process, but coding-wise they were similar to the one described below. In each age stratum, we looked at one-year events. This made censoring negligible.

Step 2: Prediction and analysis
We defined scenarios of interest, e.g., comparing all exposed to $PM_{2.5}$ as observed with all exposed to $PM_{2.5}$ as observed + IQR/10, with IQR being the interquartile range of $PM_{2.5}$ exposure in the data set. All other factors were observed in both scenarios.

Figure 1. The principle of the proposed new method for multiexposure analyses.



Based on the age-specific random forest generated above, the probability of each record (a specific person, at a specific age) being a case in each scenario, was determined (R-function *predict*) and according to this likelihood they were randomly classified as cases or non-cases. Data for all ages and the two scenarios were combined, forming a simulation data set containing one record, per person, per year, per scenario. Finally, the relative risk of being a case in the two scenarios was calculated by a traditional Cox model. An annotated sample script is included below. Note that the Cox model is not used to understand the pollutant as such. It is merely used to quantify/simplify the predictions created by the *predict* step.
The estimates of uncertainty/SD were determined from at least 50 simulation runs. However, as these very computation-intensive and precise variance estimates were irrelevant to model development, this step was not included in the scenarios investigated. The SD estimates in this document are, therefore, crude estimates and likely to be underestimated.

Table 1. Comparison of linear estimates in traditional Cox model and new random forest-based model. Both models were based on the same national data set (HR relates to risk of type 2 diabetes)

	$PM_{2.5}$		UFP		EC		NO_2	
Interquartile range (IQR)	1.84 µg/m^3		4.068 particles/cm^3		0.27 µg/m^3		6.44 µg/m^3	
Cox-estimates per IQR	1.04 (1.03-1.06)		1.05 (1.04-1.06)		1.02 (1.02-1.03)		1.06 (1.05-1.07)	
Scenario[a]	**HR**	**SD**[b]	**HR**	**SD**	**HR**	**SD**	**HR**	**SD**
Random forest-based estimates, forest including all four pollutants								
+IQR/10	1.058	0.004	1.054	0.004	1.069	0.004	1.041	0.004
-IQR/10	1.066	0.004	1.086	0.004	1.074	0.004	1.098	0.004
Random forest-based estimates, forest including only one pollutant								
+IQR/10	1.054	0.004	1.098	0.004	1.022	0.004	1.014	0.004
-IQR/10	1.072	0.004	1.062	0.004	1.071	0.004	1.090	0.004
+IQR/100	0.985	0.004	0.985	0.004	0.985	0.004	0.981	0.004
-IQR/100	0.988	0.004	0.986	0.004	0.994	0.004	0.992	0.004
+IQR/50	0.989	0.004	0.984	0.004	0.986	0.004	0.974	0.004
-IQR/50	0.995	0.004	0.994	0.004	0.997	0.004	0.998	0.004
+IQR/20	1.011	0.004	0.999	0.004	0.992	0.004	0.987	0.004
-IQR/20	1.022	0.004	1.016	0.004	1.022	0.004	1.029	0.004

a: Comparing a scenario with air pollutants as observed and a scenario where all have changed by a fraction of the interquartile range (IQR).
b: SDs are crude estimates.

Model validation. The aim of the method was to enable the disentangling of effects of multiple correlated exposures. To investigate the basic validity of the approach, we initially conducted analyses focusing on single pollutants using type 2 diabetes as the health endpoint. We evaluated scenarios contrasting exposure as observed with exposure as observed plus or minus IQR/10. The HRs were substantially larger than those observed in a traditional Cox model based on an identical data set and identical covariates. Even more disconcerting, regardless of whether exposure increased or decreased, the diabetes HR increased (Table 1).

We subsequently followed different paths to try to find the explanation for these unrealistic results:

1) One possible explanation might be imprecise exposure estimates due to too many persons being assigned exposures rarely observed in the input data. We, therefore, explored changing exposure by smaller fractions of IQR. Both increasing and decreasing exposure by IQR/100 and IQR/50 produced HRs close to the null, likely due to the increment being too small to produce discernable effects. For IQR/20, both increasing and decreasing $PM_{2.5}$ exposure produced elevated point estimates, whereas, for UFP, EC, and NO_2, risk increased with decreasing exposure and was close to 1 for increasing exposure (Table 1).
2) These inexplicable associations led us to explore another issue: as attested by plots produced when predicting the likelihood of being a case, the air pollutant was by far the most influential factor in determining the likelihood of being a case (Figure 2). As

all variables in the random forest prediction approach are essentially treated as categorical (with the cut-points determined as part of the fitting procedure), we speculated that the high resolution of air pollution allowed it to reflect other risk factors. For NO_2 and $PM_{2.5}$, we therefore created random forests based on the air pollutants being categorized into seven categories as in previous Cox analyses. This reduced the relative influence of the air pollutant substantially (Figure 3).

3) In this dataset, we then investigated scenarios comparing all individuals set to the lowest/reference category and all individuals set to each of the six other categories. However, this did not produce HRs resembling the results of the Cox analysis (Table 2).
4) We also evaluated scenarios "all as observed" versus all increased one category level, except for those in the highest category who remained at that level thus preventing exposures not observed in input data. We also tried to reduce all exposures to one categorical level, except the lowest. This produced increasing risk with both increasing and decreasing exposure levels (Table 2).

Figure 2. Influence plot of all covariates in the random forest-based prediction of diabetes risk, with $PM_{2.5}$ as a continuous variable. The plot depicts results for age=79. Other ages were similar (data not shown).

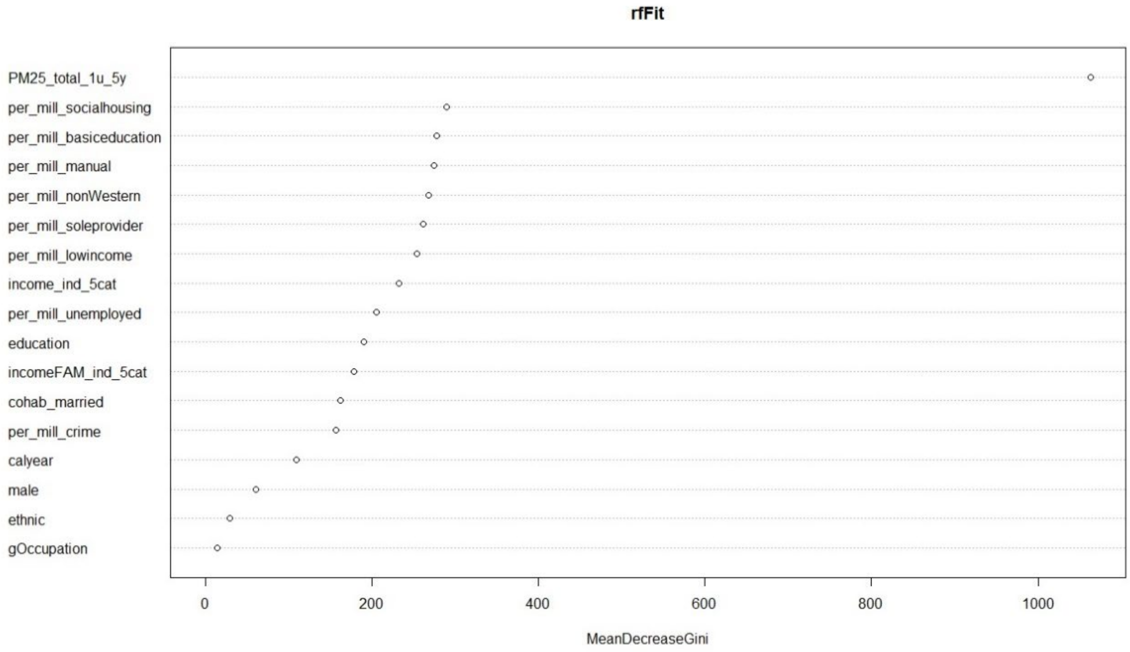


Figure 3. Influence plot of all covariates in random forest-based prediction of diabetes risk, with $PM_{2.5}$ in seven categories. The plot depicts results for age=79. Other ages were similar (data not shown).

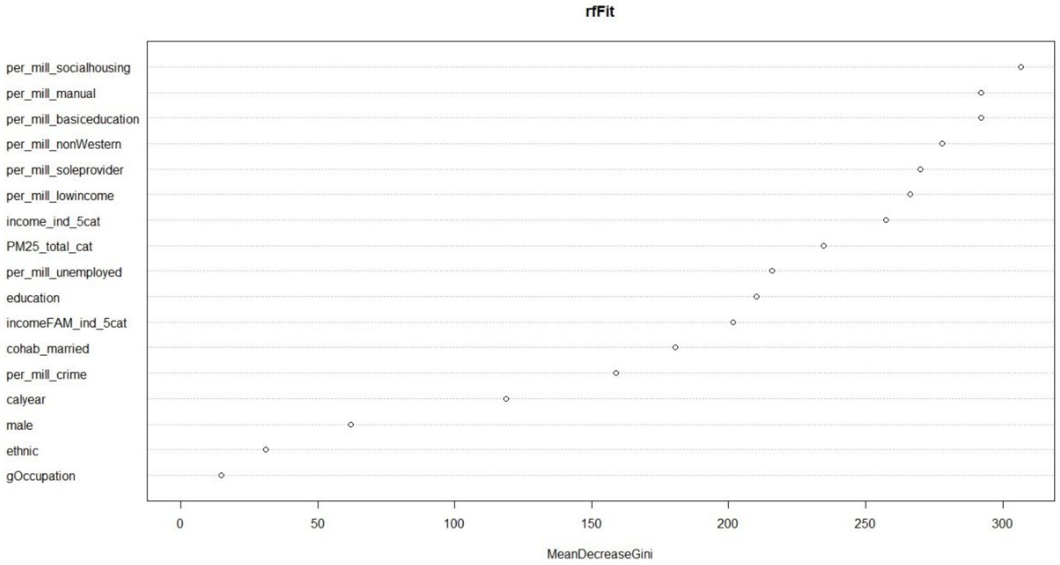
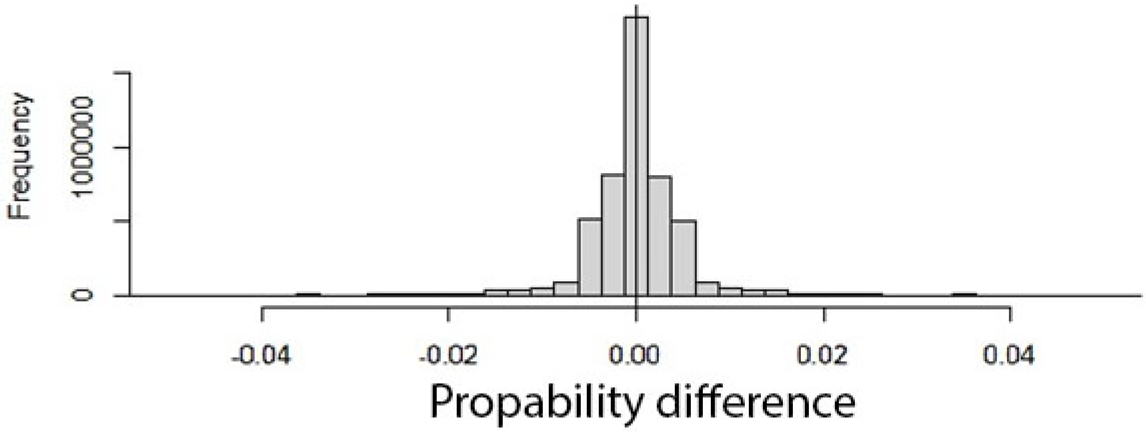


Table 2. Comparison of categorical estimates in traditional Cox model and new random forest-based model

$PM_{2.5}$	Cox Cases	Cox HR	RF HR		NO_2	Cox Cases	Cox HR	RF HR	
<8.69	19.722	1.00	ref.		<9.88	18.969	1.00	ref	
8.69-9.40	25.779	0.99 (0.97-1.00)	0.76	(0,75-0,76)	9.88-11.8	25.697	1.03 (1.01-1.05)	0.79	(0.78-0.79)
9.40-10.3	39.069	1.01 (0.99-1.03)	0.67	(0.66-0.67)	11.8-14.7	38.999	1.06 (1.04-1.08)	0.68	(0.67-0.68)
10.3-11.3	35.343	1.06 (1.03-1.08)	0.74	(0.73-0.74)	14.7-19.0	33.831	1.06 (1.04-1.08)	0.71	(0.71-0.72)
11.3-11.9	17.980	1.09 (1.06-1.12)	0.92	(0.91-0.92)	19.0-23.4	19.060	1.13 (1.11-1.16)	0.91	(0.90-0.91)
11.9-12.3	5.279	1.06 (1.02-1.10)	1.16	(1.15-1.17)	23.4-26.8	5.808	1.16 (1.13-1.20)	1.18	(0.85-1.17)
>12.3	4.860	1.10 (1.06-1.14)	1.52	(1.51-1.53)	>26.8	5.668	1.15 (1.11-1.19)	1.57	(0.64-1.56)
All increased 1 categorical level[a]			1.14	(1.13-1.15)				1.21	(0.83-1.20)
All decreased 1 categorical level[b]			1.23	(1.21-1.24)				1.13	(0.89-1.12)

a: Except highest category
b: Except lowest category

Figure 4. Frequency of difference in predicted probability of being a diabetes case, between the two scenarios: "all observations having $PM_{2.5}$ exposure as observed plus IQR/10" and "all observations having $PM_{2.5}$ exposure as observed minus IQR/10."



5) Air pollution in Denmark shows geographical gradients due to long-distance transport from neighboring countries and urban areas generate their own peaks. Therefore, as an alternative way of addressing the potential issue of the finely detailed air pollution data perhaps capturing other risk factors, we generated a random forest for $PM_{2.5}$ including population density and geographical region of Denmark as well as living in a single-family home and proportion of green area within 1000 meters of residence. Both when increasing and decreasing all exposure by IQR/10 we found increased HRs (1.04, SD 0.004 and 1.05, SD: 0.005, respectively).
6) Finally, to determine if the problems occurred already in the machine learning stage, we calculated the difference in probability of being a case for each person under the +IQR/10 and –IQR/10 scenarios. Figure 4 shows that these differences were normally distributed around zero, indicating that the issue creating the strange results occurred already in the random forest step.

Altogether, we are uncertain about the reason why the method creates unrealistic results and how it might be fixed. Since the method does not work with one pollutant, we saw no reason to try applying it to multiexposure problems.

We are still confident that relaxing the assumptions inherent in traditional Cox models will be an advantage for multiexposure models. However, based on our experience, we suggest that the underlying methods – whether that be Random Forest or other machine learning methods – need to be redesigned to accommodate survival-type data in a more efficient way and handle the situation of extremely rare outcomes. In our setting, the outcomes are rare because we have a high temporal resolution, which implies that for any specific person in a specific time window, the probability of diabetes (for example) is minimal. The usual "tricks" of rebalancing would likely not work as we also need the absolute risk estimates to be right. It could be that headway could be made by exploring different loss functions for the fit; again, the loss functions should reflect the survival nature of the problem.

Conclusions

With just 8 months left of the HERMES study, we felt forced to give up on making the new method work. Instead, we changed course and applied traditional Cox models for two-exposure and multiexposure analyses to try to identify the most important exposure(s).

Annotated sample code for random forest generation (explanatory notes are marked in bold)

```
library(foreign)
library(survival)
library(timereg) #Aalen
library(descr)
library(ltmle)
library(randomForest)
setwd("g:/data/workdata/707239/Aslak/THEIS method")

OUTCOME<-"DIABETES"
EXPOSET<-"PM25TOTAL"
```

#input and restrict data

```
HERMES100 <- read.csv
('G:\\Data\\workdata\\707239\\HERMES_analyses\\MetteSorensen\\Diabetes\\Analyses\\Additive_models\\
R\\Data\\pro100tilR_age35plus_1yr.csv')
HERMES100 <- subset(HERMES100, (age_start >= 50) & (age_start < 80))
HERMES100$nowEvent <- HERMES100$case_diabetes
```

#AFFORESTATION, (by year of age due to size,
#each age-specific part of the forest saved separately
#Forrest populated by the same parameters as used in Cox-models

```
ageInts <- 50:79
ageDelta <- 1

for(ageTemp in ageInts)
{
  rowNumbers <- which((ageTemp<= HERMES100$age_start) & (HERMES100$age_start <
(ageTemp+ageDelta)))
  workData <- HERMES100[rowNumbers, ]

  workData$nowEvent <- factor(workData$nowEvent)
  rfFit <- randomForest(nowEvent ~  PM25_total_1u_5y +
male + calyear + education + gOccupation + incomeFAM_ind_5cat + income_ind_5cat + cohab_married +
ethnic + per_mill_socialhousing + per_mill_soleprovider + per_mill_unemployed + per_mill_manual +
 per_mill_nonWestern + per_mill_crime + per_mill_lowincome + per_mill_basiceducation, data =
workData)

 fname<-paste("G:\\Data\\workdata\\707239\\Aslak\\THEIS
method\\data\\",OUTCOME,"_",EXPOSET,"_skov_age",ageTemp,".Rdata",sep="")

  save(rfFit, file=fname)
}
```

#similar forests created including all four pollutants as well as separately for each of the other exposures: NO2_total_1u_5y,
UFP_total_1u_5y, EC_total_1u_5y : Code not shown

Annotated sample code for prediction and analysis (explanatory notes are marked in bold)

```
library(foreign)
library(survival)
library(timereg) #Aalen
library(descr)
library(ltmle)
library(randomForest)
library(tibble)
setwd("g:/data/workdata/707239/Aslak/THEIS method")

#Import and restrict data
HERMES100 <- read.csv
('G:\\Data\\workdata\\707239\\HERMES_analyses\\MetteSorensen\\Diabetes\\Analyses\\Additive_models\\
R\\Data\\pro100tilR_age35plus_1yr.csv')
# do reductions
HERMES100 <- subset(HERMES100, (age_start >= 50) & (age_start < 80))
HERMES100$nowEvent <- HERMES100$case_diabetes


OUTCOME<-"DIABETES"
EXPO<-"PM25_total_1u_5y"

# Start of scenario def
# creates 3 scenarios:
#ScenRef, exposure as observed PropRef=likelihood of being a case
#ScenP1, exposure as observed + IQR/10 PropP1=likelihood of being a case
#ScenM1, exposure as observed - IQR/10 PropPM=likelihood of being a case

 HERMES100[paste0(EXPO,"PropRef")]<-NA
 HERMES100[paste0(EXPO,"PropP1")]<-NA
 HERMES100[paste0(EXPO,"PropM1")]<-NA

 HERMES100[paste0(EXPO,"ScenRef")] <- HERMES100[,EXPO]
 HERMES100[paste0(EXPO,"ScenP1")] <- HERMES100[,EXPO]+(IQR(HERMES100[,EXPO]))/10
 HERMES100[paste0(EXPO,"ScenM1")] <- HERMES100[,EXPO]-(IQR(HERMES100[,EXPO]))/10
# end of scenario def

# For each year of age, the appropriate random forests is imported and the likelihood of being a case
is predicted with the predict function

ageInts <- 50:79
ageDelta <- 1

for(ageTemp in ageInts)
{
 fname<-paste("G:\\Data\\workdata\\707239\\Aslak\\THEIS
method\\data\\",OUTCOME,"_",EXPOSET,"_skov_age",ageTemp,".Rdata",sep="")

 load(file=fname)

 rowNumbers <- which((ageTemp<= HERMES100$age_start) & (HERMES100$age_start <
(ageTemp+ageDelta)))
 workData <- HERMES100[rowNumbers, ]

#  summary(rfFit)
#  varImpPlot(rfFit)
#  importance(rfFit)


# then we use predict

 predData <- HERMES100[rowNumbers, ]

 predData[EXPO] <- predData[paste0(EXPO,"ScenRef")]
 temp <- predict(rfFit, newdata = predData, type = "prob")[,2]
 HERMES100[rowNumbers,paste0(EXPO,"PropRef")]<-temp

 predData[EXPO] <- predData[paste0(EXPO,"ScenM1")]
 temp <- predict(rfFit, newdata = predData, type = "prob")[,2]
 HERMES100[rowNumbers,paste0(EXPO,"PropM1")]<-temp

 predData[EXPO] <- predData[paste0(EXPO,"ScenP1")]
 temp <- predict(rfFit, newdata = predData, type = "prob")[,2]
 HERMES100[rowNumbers,paste0(EXPO,"PropP1")]<-temp

}
#ANALYSIS P1 vs REF

HERMES100$scen1<-HERMES100[,paste0(EXPO,"PropRef")]
HERMES100$scen2<-HERMES100[,paste0(EXPO,"PropP1")]
#randomly assign case status based on case probability in the two scenarios

simData <- data.frame(simEvent = c(
            rbinom(nrow(HERMES100), size = 1, prob = HERMES100$scen1),
            rbinom(nrow(HERMES100), size = 1, prob = HERMES100$scen2)),
            age_start = rep(HERMES100$age_start,2),
            age_end = rep(HERMES100$age_end,2))

#for cases, event time is assigned to a random time within age
simData$age_end_sim <- ifelse(simData$simEvent==1, runif(nrow(simData), min=simData$age_start,
max = simData$age_end), simData$age_end)

# appending all records and cases status in the two scenarios
simData$grp <- c(rep(0, nrow(HERMES100)), rep(1, nrow(HERMES100)))

# calculate relative risk of being case under scen2 vs scen1
fitCox <- coxph(Surv(age_start, age_end, simEvent) ~ factor(grp), data = simData)
summary(fitCox)

#As the sample size is doubled above, new SD is calculated here based on 50 simulations
#(This time-consuming step was omitted from development runs reported in the present report)

simHRs <- rep(NA, 50)
for(ii in 1:length(simHRs))
{simData$simEvent <- c(rbinom(nrow(HERMES100), size = 1, prob = HERMES100$prop1),
rbinom(nrow(HERMES100), size = 1, prob = HERMES100$prop2))

 simData$age_end_sim <- ifelse(simData$simEvent==1, runif(nrow(simData), min=simData$age_start,
max = simData$age_end), simData$age_end)
 fitCoxTemp <- coxph(Surv(age_start, age_end, simEvent) ~ factor(grp), data = simData)
 simHRs[ii] <- coef(fitCoxTemp)[1]
}
```

```
sd(simHRs)
```

#The above steps were repeated for comparison of other scenarios and for Air pollutants – code not included

REFERENCES

Mogensen UB, Ishwaran H, Gerds TA. 2012. Evaluating random forests for survival analysis using prediction error curves. Journal of Statistical Software 50:1-23.

Pearl J. 2009. Causality: Models, reasoning and inference, 2nd edition: Cambridge University Press.

Robins J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 7:1393–1512.

VanderWeele TJ. 2015. Explanation in causal inference: Methods for mediation and interaction, 1st edition: Oxford University Press.