Associations between lifestyle and air pollution exposure: Potential for confounding in large administrative data cohorts

CrossMark

Maciej Strak[a,*], Nicole Janssen[b], Rob Beelen[b], Oliver Schmitz[c], Derek Karssenberg[c], Danny Houthuijs[b], Carolien van den Brink[b], Martin Dijst[d], Bert Brunekreef[a,e], Gerard Hoek[a]

[a] Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, Netherlands
[b] National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands
[c] Department of Physical Geography, Faculty of Geosciences, Utrecht University, Netherlands
[d] Department of Human Geography and Spatial Planning, Faculty of Geosciences, Utrecht University, Netherlands
[e] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands

A B S T R A C T

Background: Cohorts based on administrative data have size advantages over individual cohorts in investigating air pollution risks, but often lack in-depth information on individual risk factors related to lifestyle. If there is a correlation between lifestyle and air pollution, omitted lifestyle variables may result in biased air pollution risk estimates. Correlations between lifestyle and air pollution can be induced by socio-economic status affecting both lifestyle and air pollution exposure.

Objectives: Our overall aim was to assess potential confounding by missing lifestyle factors on air pollution mortality risk estimates. The first aim was to assess associations between long-term exposure to several air pollutants and lifestyle factors. The second aim was to assess whether these associations were sensitive to adjustment for individual and area-level socioeconomic status (SES), and whether they differed between subgroups of the population. Using the obtained air pollution–lifestyle associations and indirect adjustment methods, our third aim was to investigate the potential bias due to missing lifestyle information on air pollution mortality risk estimates in administrative cohorts.

Methods: We used a recent Dutch national health survey of 387,195 adults to investigate the associations of PM_{10}, $PM_{2.5}$, $PM_{2.5-10}$, $PM_{2.5}$ absorbance, OP^{DTT}, OP^{ESR} and NO_2 annual average concentrations at the residential address from land use regression models with individual smoking habits, alcohol consumption, physical activity and body mass index. We assessed the associations with and without adjustment for neighborhood and individual SES characteristics typically available in administrative data cohorts. We illustrated the effect of including lifestyle information on the air pollution mortality risk estimates in administrative cohort studies using a published indirect adjustment method.

Results: Current smoking and alcohol consumption were generally positively associated with air pollution. Physical activity and overweight were negatively associated with air pollution. The effect estimates were small (mostly < 5% of the air pollutant standard deviations). Direction and magnitude of the associations depended on the pollutant, use of continuous *vs.* categorical scale of the lifestyle variable, and level of adjustment for individual and area-level SES. Associations further differed between subgroups (age, sex) in the population. Despite the small associations between air pollution and smoking intensity, indirect adjustment resulted in considerable changes of air pollution risk estimates for cardiovascular and especially lung cancer mortality.

Conclusions: Individual lifestyle-related risk factors were weakly associated with long-term exposure to air pollution in the Netherlands. Indirect adjustment for missing lifestyle factors in administrative data cohort studies may substantially affect air pollution mortality risk estimates.

1. Introduction

Long-term air pollution exposure has been linked to a range of health outcomes in several individual cohort studies (Beelen et al., 2014; Pope III et al., 2002). These studies have included information on individual risk factors related to lifestyle, such as smoking, alcohol

consumption, diet, body mass index (BMI) or physical activity as potential confounders. In recent years several studies have employed national mortality registries and combined them with existing databases of air pollution, residential address and neighborhood characteristics (Cesaroni et al., 2013; Crouse et al., 2012; Fischer et al., 2015; Zeger et al., 2008). Such administrative data cohorts have size advantages over individual cohorts, but often lack in-depth information on individual lifestyle factors. If there is a correlation between lifestyle and air pollution then omitted lifestyle variables result in biased risk estimates. As administrative data cohort studies do typically have information on basic individual socioeconomic variables (*e.g.,* household income from tax records) and area/neighborhood socioeconomic factors, the key issue is whether lifestyle is still associated with air pollution after adjusting for these potential confounders.

Some administrative cohort studies include proxies of lifestyle factors to address the potential for residual confounding by unavailable lifestyle information, *e.g.,* preexisting comorbidities or observed lung cancer or COPD rates assumed to represent smoking habits, typically as area-level variables (Cesaroni et al., 2013; Fischer et al., 2015; Zeger et al., 2008). Shin and co-workers recently proposed a method to indirectly adjust hazard ratios (HR) for missing individual risk factors (Shin et al., 2014). The method uses ancillary information on (1) the omitted risk factors representative of the subjects in the cohort (*e.g.,* national health surveys) and their association with air pollution, and (2) the relationship between the omitted risk factors and the health endpoints from the literature. Such indirect adjustment may have substantial effects – the indirect adjustment for smoking habits and obesity in the Shin et al. (2014) study increased the association between particulate matter and ischemic heart disease previously observed in a Canadian administrative cohort (Crouse et al., 2012) by 3–123%, with the smallest of these increases (3%) observed for the associations between lifestyle and air pollution, adjusted for the socioeconomic factors which were also available in the administrative mortality cohort.

The overall aim of the current manuscript was to assess potential confounding bias in air pollution mortality risk estimates due to missing lifestyle factors. The first aim was to investigate the associations between residential address concentrations of a range of air pollutants and several individual lifestyle factors, including smoking, alcohol consumption, physical activity and BMI available in a large Dutch national health survey. The second aim was to assess whether the associations between air pollution and lifestyle were sensitive to adjustment for individual and area-level SES, and whether the association differed between subgroups of the population. The third aim was to investigate how much air pollution mortality effect estimates may change when including lifestyle-related confounders, using the associations observed for smoking and air pollution and the indirect adjustment method by Shin et al. (2014).

2. Methods

2.1. Design

We used a large national health survey conducted in the Netherlands in 2012 to provide data on individual lifestyle factors. The survey included self-reported lifestyle and health, but did not include a mortality follow-up. Long-term air pollution concentrations at the 2012 residential addresses were calculated by land use regression (LUR) models. Fig. 1 presents the outline of the paper. To assess potential confounding bias in air pollution mortality risk due to missing lifestyle data, we first investigated the association between the concentration of seven major air pollutants and individual smoking habits, alcohol consumption, physical activity and BMI using linear regression. We assessed the crude association between air pollution and lifestyle and then adjusted for age, sex and individual and area-level SES variables often available in administrative cohorts. The adjustment

for SES was performed to test the hypothesis that the association between air pollution and lifestyle is driven by socioeconomic factors. We then assessed whether the air pollution and lifestyle associations differed in subgroups defined by age, sex and ethnicity. These subgroup analyses are essential in judging how important differences in population characteristics between the survey population and a potential cohort may be in applying the Shin et al. (2014) indirect adjustment approach. We then used the Shin indirect adjustment approach to calculate the magnitude of the bias of air pollution mortality risk in a hypothetical cohort when lifestyle data is missing. We focused the bias calculation on smoking and the pollutants $PM_{2.5}$ and NO_2 and cardiovascular and lung cancer mortality. Inputs for the bias calculation were the observed associations between smoking and air pollution, as well as smoking-related relative risks for cardiovascular and lung cancer mortality from the literature (see Fig. 1). Bias calculations were performed for the full population and stratified by age and sex subgroups. The Shin et al. (2014) approach then calculates indirectly adjusted air pollution risk estimates in an administrative cohort by subtracting the bias – calculated using the air pollution lifestyle associations in the ancillary study and literature derived risks for lifestyle – from the air pollution risks observed in the administrative cohort. Section 2.4.2 provides more detail. In the current paper, we calculated the bias (the change in air pollution risks), but because of the lack of an actual administrative cohort we could not calculate updated air pollution risks.

2.2. Population

The Public Health Monitor (*Gezondheidsmonitor Volwassenen GGD-en, CBS en RIVM 2012*) was a 2012 national health survey by 28 Public Health Services (GGD), Statistics Netherlands (CBS) and the National Institute for Public Health and the Environment (RIVM). The population was selected across the Netherlands with both major cities and smaller towns sampled. Subjects older than 65 years were oversampled. The response rate was 45–50%. The Monitor includes information on 387,195 citizens aged ≥ 19 years and covers a range of issues related to physical and mental health status, socioeconomic situation, lifestyle and personal characteristics. Statistics Netherlands have further enriched the Public Health Monitor with information on standardized household income and ethnicity.

We linked the Public Health Monitor dataset with information on socioeconomic status (SES) at a neighborhood level (four-digit postal code). A four-digit postal code included on average 2178 addresses. This indicator represents the educational, occupational and economical status of the neighborhood and is derived every four years by The Netherlands Institute for Social Research (SCP) (Knol, 1998). We used the latest available data, being the 2006 SES score.

To examine the representativeness of the Public Health Monitor population for the general adult Dutch population we used the 2012 data from the StatLine electronic databank of Statistics Netherlands (http://statline.cbs.nl/).

2.3. Air pollution exposure assessment

We used the LUR models developed within the framework of the European Study of Cohorts for Air Pollution Effects (ESCAPE) for the Netherlands to assess long-term average air pollution exposure at the 2012 home address of participants of the Public Health Monitor. Models were developed for PM_{10} and $PM_{2.5}$ (particulate matter < 10 μm and 2.5 μm in aerodynamic diameter, respectively), $PM_{2.5-10}$ (the coarse fraction of PM), absorbance of $PM_{2.5}$ (a measure of black carbon particles), NO_2 (nitrogen dioxide) and NO_X (nitrogen oxides) (Beelen et al., 2013; Eeftens et al., 2012). As NO_2 and NO_X were highly correlated, we present only the former in our analyses. Further, we used the $PM_{2.5}$ oxidative potential (OP) LUR models developed on ESCAPE samples (Yang et al., 2015), which includes two OP metrics –
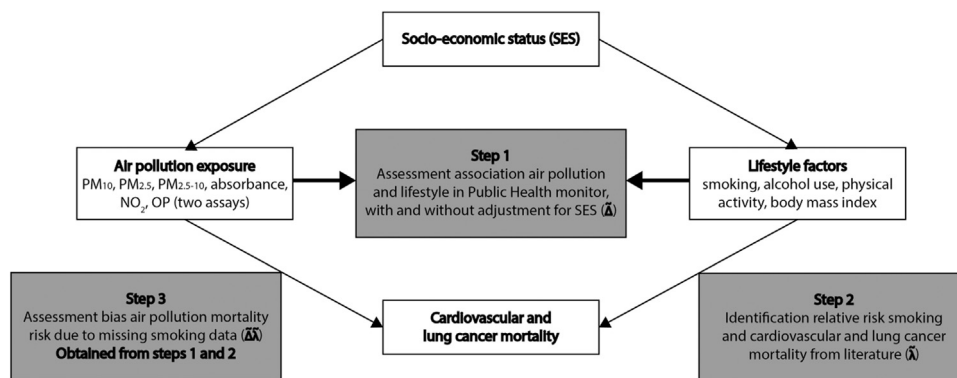
Fig. 1. Overview of calculations assessing potential bias due to missing lifestyle information. Shaded boxes are steps in calculations, transparent boxes are causal framework; $\widetilde{\Delta}$ and $\widetilde{\lambda}$ are parameters of formula (1) (Section 2.4.2).

electron spin resonance (OP^{ESR}) and dithiothreitol (OP^{DTT}). OP may be a metric potentially more health relevant than PM mass (Ayres et al., 2008; Borm et al., 2007). All the LUR models are shown in Table S1. We focus on ESCAPE LUR models because they were developed using a consistent methodology for multiple relevant air pollutants; were applied extensively in epidemiological studies and will be applied in future studies based on administrative data.

Home-address concentrations of the seven selected air pollutants were calculated by applying the ESCAPE LUR models in the PCRaster environment using grids of 5×5 m (Karssenberg et al., 2010).

The air pollution estimates at home addresses were merged with the Public Health Monitor data in a secure environment of Statistics Netherlands. Information on home address (except four-digit postal code) was removed to warrant privacy.

2.4. Data analyses

2.4.1. Lifestyle and air pollution

Analyses were carried out with SAS 9.4 (SAS Institute Inc., Cary, NC, USA) in the secure remote access environment of Statistics Netherlands. We used generalized linear regression models (Proc GLM) to obtain estimates and 95% confidence intervals (CI) of the associations between long-term exposure to PM_{10}, $PM_{2.5}$, $PM_{2.5-10}$, absorbance, OP^{DTT}, OP^{ESR} and NO_2 (as dependent variables) and individual lifestyle factors (as independent variables). Smoking was represented by continuous (number of cigarettes smoked per day) and categorical (current, former, and never smokers) variables. Alcohol consumption was represented by a continuous variable (number of alcohol glasses consumed per week) and a categorical variable (current, former, and never users). Physical activity was characterized by a continuous variable for total moderate-intensity physical activity expressed in minutes per week. Moderate intensity is defined as a score of ≥ 3 MET, where MET stands for Metabolic Equivalent of Task, ranging from one for "inactivity" (*e.g.*, sitting or lying quietly) to 18 for vigorous running (at 3.4 min/km) (Ainsworth et al., 2000). Activities with a MET-score ≥ 3 which we included were: walking or cycling to commute; leisure time walking or cycling; gardening; doing odd jobs in the home; leisure sports (up to four, only those ≥ 3 MET included); vigorous activities in the home, work or school. WHO advises adults at least 150 min of moderate physical activity per week (WHO, 2015b). BMI, derived from reported height and weight, was coded according to WHO classification: underweight (< 18.5 kg/m^2), normal range (18.5–24.9 kg/m^2), overweight (25–29.9 kg/m^2) and obese (≥ 30 kg/m^2) (WHO, 2015a).

We specified five confounder models *a priori*, with an increasing level of adjustment for variables often available in administrative database cohort studies. Model 1 was a crude model (*air pollution = smoking* or *alcohol consumption* or *physical activity* or *BMI*). In Model 2, we added sex and age (five-year categories for 19–64 years and ten-year categories for 65+ years). Model 3a expanded Model 2 with *neighbor-*

hood SES score. Model 3b expanded Model 2 with *individual* SES characteristics, *i.e.,* highest achieved education level (primary or less, lower-secondary, higher-secondary, university), paid occupation (yes/no), household income (in quintiles), marital status (married, unmarried, divorced, widowed), ethnicity (Dutch, Moroccan, Turkish, Surinamese, Dutch Antillean & Aruban, other Western, other non-Western). Model 4 included all the variables from previous models and was chosen as the main confounder model.

Using the main confounder model, we also performed subgroup analyses by sex, age (19–39, 40–64, or 65+ years) and ethnicity (Dutch or non-Dutch) to investigate whether the association between air pollution and lifestyle was modified by those variables. These analyses were performed to gain insight in potential problems of indirect adjustment when applying the associations to an external cohort with different population characteristics than the (national) health survey.

To facilitate comparison between pollutants, we divided the regression slopes (air pollution as dependent and smoking habits, alcohol consumption, physical activity and BMI as independent variables) by the standard deviation of the concentration in the full population and then multiplied by 100 to obtain percentages. Expression relative to the standard deviation instead of the mean concentration, is more informative for the evaluation of potential confounding.

To investigate linearity of the association between lifestyle and air pollution, we calculated associations with deciles of pollutants.

2.4.2. Potential bias in air pollution mortality risk estimates

To illustrate the potential bias in the associations between air pollution and mortality resulting from lack of information on relevant confounders in administrative database cohorts, we used the indirect adjustment method proposed by Shin et al. (2014). We applied this indirect adjustment to the associations between NO_2 and $PM_{2.5}$ and mortality from lung cancer and cardiovascular disease. We used NO_2 and $PM_{2.5}$ because of their extensive application in previous cohort studies. We focused on smoking as it is a very strong risk factor for lung cancer mortality and a moderately strong risk factor for cardiovascular mortality.

Indirect adjustment of mortality HRs uses the following equation (Shin et al., 2014) (see also Fig. 1):

$$\widetilde{\beta} = \hat{\gamma} - \widetilde{\Delta}\widetilde{\lambda} \tag{1}$$

$\widetilde{\beta}$ is the indirectly adjusted air pollution mortality effect estimate;
$\hat{\gamma}$ is the air pollution mortality effect estimate in a cohort with missing lifestyle factors;
$\widetilde{\Delta}$ is the estimate of the association between missing lifestyle factors (as dependent variable) and air pollutants (as independent variable) in the ancillary dataset, here the Public Health Monitor;
$\widetilde{\lambda}$ is the estimate of the association between missing lifestyle factors and mortality, derived from literature.

All these parameters are expressed as slopes of proportional hazard models ($\widetilde{\beta}$, $\hat{\gamma}$, $\overline{\lambda}$) or linear regression models ($\widetilde{\Delta}$). To calculate air pollution mortality risks, these slopes need to be exponentiated after multiplication with a realistic increment in the respective predictors (see Section 3.5).

In the current study, we focus on the term $\widetilde{\Delta}\widetilde{\lambda}$ that represents the calculated bias in the air pollution mortality slope. We calculated the bias term and then exponentiated the bias multiplied by the difference between the 5th and 95th percentile of the air pollution concentration in the ancillary population. This HR can be interpreted as the HR that may be observed due to confounding by missing smoking data, in the absence of a true pollution effect.

We obtained the components of the bias term ($\widetilde{\Delta}\widetilde{\lambda}$) from the associations between air pollution and lifestyle in Section 2.4.1 ($\widetilde{\Delta}$) and the scientific literature ($\widetilde{\lambda}$). The association between NO_2, $PM_{2.5}$ and cigarette smoking ($\widetilde{\Delta}$) needed for the calculations is expressed as the difference in number of cigarettes/day per 1 µg/m³ increase in air pollution as observed in the ancillary data. We obtained $\widetilde{\lambda}$ as the natural logarithm of the relative risks (RR) for lung cancer mortality, cardiovascular disease mortality and smoking from the American Cancer Society – Cancer Prevention Study II (ACS CPS-II) (Pope et al., 2011), following the Shin et al. (2014) paper. The adjusted RR for lung cancer was 11.63 (95% CI: 9.51, 14.24) and for cardiovascular diseases 2.01 (95% CI: 1.84, 2.19), expressed per 10 cigarettes/day.

As we did not have an administrative cohort with air pollution mortality risk estimates without adjustment for smoking ($\hat{\gamma}$), we did not calculate adjusted estimates ($\widetilde{\beta}$).

3. Results

3.1. Population

The characteristics of the included Public Health Monitor population are shown in Table 1. Information on 387,152 adults was available for the analysis. The data is skewed towards the elderly population, with almost 43% being 65 years or older. In the general Dutch population, only 16% is 65 years or older (http://statline.cbs.nl/). The overrepresentation of elderly is part of the Public Health Monitor design. Further, in the Public Health Monitor people of Dutch origin are overrepresented (87% compared with 79% in the general population) and people in the lowest household income quintile (10%) are underrepresented, probably due to differential response rates.

3.2. Air pollution concentrations

Concentrations at home addresses during the 2012 survey of the seven investigated air pollution components are presented in Table 2. Contrasts were larger for NO_2, absorbance and the two OP metrics than for PM_{10} and $PM_{2.5}$. Correlations between the components are shown in Table S2. NO_2 was moderately correlated with OP metrics (0.54–0.60) and highly correlated with PM_{10}, $PM_{2.5-10}$ and absorbance (0.78–0.79).

3.3. Associations between air pollution and lifestyle

3.3.1. Smoking

Nearly 20% of the population were current smokers, smoking on average 10.3 cigarettes/day; almost 40% were former smokers. Compared to never smokers, current smokers lived at addresses with slightly higher concentrations of all investigated air pollutants. The differences were mostly small, after adjustment not exceeding 5% of the standard deviation of the concentrations, and somewhat lower when expressing smoking with a continuous variable (active smoking; per 10.3 cigarettes/day, Fig. 2A1) compared to a categorical variable (Fig. 2A2). Adjustment for neighborhood SES score made these

Table 1
Characteristics of the included Public Health Monitor study population (N=387,152).

Characteristic	Category	Value
Sex	Female	211,264 (54.6)
Age[a]	19–39	75,107 (19.4)
	40–64	146,840 (37.9)
	≥65	165,205 (42.7)
Education (n=373,957)	Primary or less	37,132 (9.9)
	Lower-secondary	131,072 (35.1)
	Higher-secondary	105,853 (28.3)
	University	99,900 (26.7)
Paid occupation (n=360,328)	Yes	172,819 (48.0)
Household income (n=384,802)	< €15,200	39,058 (10.2)
	€15,200–19,400	74,430 (19.3)
	€19,400–24,200	82,149 (21.4)
	€24,200–31,000	91,371 (23.7)
	≥ €31,000	97,794 (25.4)
Marital status[b] (n=381,157)	Married/living together	269,217 (70.6)
	Unmarried/never married/ divorced/widowed	111,940 (29.4)
Ethnicity[c]	Dutch	335,076 (86.6)
	Other western	34,091 (8.8)
	Non-western	17,985 (4.6)
Smoking habits (n=363,411)	Current	71,810 (19.8)
	Former	144,848 (39.9)
	Never	146,753 (40.4)
Alcohol consumption (n=371,773)	Current	305,759 (82.2)
	Former	22,538 (6.1)
	Never	43,476 (11.7)
BMI (n=371,767)	Underweight	5058 (1.4)
	Normal range	172,989 (46.5)
	Overweight	143,088 (38.5)
	Obese	50,632 (13.6)
Neighborhood SES score [mean (SD)] (n=386,026)		37.0 (9.2)
Number of cigarettes/day[d] [mean (SD)] (n=359,945)		10.3 (8.3)
Number of alcohol glasses/week[e] [mean (SD)] (n=353,028)		8.5 (9.6)
Minutes of moderate-intensity (≥3 MET) physical activity/week [mean (SD)] (n=361,907)		1005 (949)

Unless otherwise stated, values are frequency (%) and N=387,152.
[a] Age was analyzed in 12 categories: 1st (19–24 years) to 9th (60–64 years) were five-year categories, whereas 10th (65–74 years) to 12th (≥85 years) were ten-year categories.
[b] Marital status was analyzed in four categories: married/living together, unmarried/never married, divorced, and widowed.
[c] Ethnicity was analyzed in seven categories: Moroccan, Turkish, Surinamese, Dutch Antillean & Aruban, other non-western, other western, and Dutch.
[d] Information for 95% of "current" smoking habit available.
[e] Information for 94% of "current" alcohol consumption available.

differences even smaller. The impact of SES adjustment was large for OP measured by DTT and small for $PM_{2.5}$. Former smokers had lower air pollution concentrations at their home address in Models 1–3a and higher air pollution after adjusting for individual SES characteristics (Models 3b–4) compared to never smokers, but the differences rarely exceeded 2% of the standard deviation of air pollution concentrations (Fig. 2A3).

3.3.2. Alcohol consumption

About 82% of the population were alcohol consumers, with a mean consumption of almost 8.5 alcohol glasses/week; 6% were former consumers. We observed substantial differences in the association

Table 2
Distribution of air pollution concentrations at home addresses (n = 386,831).

Component	Mean	SD	Min	P5	P25	P50	P75	P95	Max
PM_{10} (µg/m^3)	24.76	1.11	23.73	23.76	23.95	24.41	25.16	27.03	34.75
$PM_{2.5}$ (µg/m^3)	16.73	0.69	14.86	15.56	16.34	16.70	17.15	17.71	21.72
$PM_{2.5-10}$ (µg/m^3)	8.30	0.75	7.60	7.63	7.77	8.05	8.55	9.90	14.16
Absorbance (10^{-5}/m)	1.28	0.22	0.85	0.99	1.15	1.24	1.39	1.69	3.14
NO_2 (µg/m^3)[a]	23.83	6.06	9.11	15.46	19.53	22.95	27.32	34.70	138.11
OP^{DTT} (nmol DTT/min/m^3)	1.18	0.20	0.48	0.82	1.05	1.20	1.33	1.49	2.18
OP^{ESR} (A.U./m^3)[b]	0.90	0.15	0.65	0.72	0.80	0.89	0.98	1.15	2.05

[a] There were 145 addresses with NO_2 concentrations > 60 µg/m^3, two of which > 100 µg/m^3.
[b] A.U. is arbitrary unit.

between air pollution and alcohol consumption depending whether the latter was expressed on a continuous or categorical scale (Fig. 2B1 and B2). Adjustment for ethnicity, individual and neighborhood SES reduced associations substantially (Fig. 2B1, B2 and B3). The difference between current and never alcohol consumers with all the air pollution components were negative in Models 1–2, and became marginally positive or disappeared after fully adjusting for the neighborhood and individual SES characteristics (Fig. 2B2).

3.3.3. Physical activity

Average moderate-intensity physical activity in the study population was 1005 min/week (2.4 h/day). Physically more active subjects had lower air pollution concentrations at their home address for all the investigated air pollutants (Fig. 2C). The differences were small, never exceeding 6% of the pollutant standard deviation, highest for NO_2 and OP^{DTT} and lowest for the PM mass metrics. Adjustment for individual SES characteristics reduced the differences in most air pollution concentrations related to physical activity. OP^{DTT} associations were not affected by adjustment.

3.3.4. BMI

Almost half of the population had a BMI in the normal range (46.5%), with 38.5% being overweight, nearly 14% obese and 1.5% underweight. The overweight population (Fig. 2D) had decreased concentrations of nearly all air pollution components compared to the normal weight group (no difference for $PM_{2.5}$, increased OP^{DTT}) with differences in the range of 2–6%. For the obese population (Fig. S1A) the positive or negative direction of the association depended on the adjustment for either the neighborhood or the individual SES characteristics; it was only more robust (5% increase) for OP^{DTT}. The underweight population (Fig. S1B) lived at the addresses with higher concentrations of all included air pollutants, the differences becoming smaller with adjustment.

3.3.5. Sensitivity analysis

Table 1 illustrates that there is a small fraction of missing values in the adjustment variables. A complete case analysis per lifestyle factor (Fig. S2) showed very similar associations to the associations reported in Fig. 2. Associations were also similar when we restricted the analysis to complete cases for all lifestyle and adjustment variables (Fig. S2; Model 4 CC). The fully complete case was based upon 289,703 subjects. In Fig. 2, we reported all available data because of concern with selection bias when using a complete case analysis.

3.4. Subgroup analysis of the association between air pollution and lifestyle

We observed some differences in the association between air pollution and lifestyle between subgroups defined by age, sex and ethnicity, overall somewhat more pronounced for NO_2. For subjects aged 65+ years, smoking intensity was associated with a higher NO_2 concentration contrast compared to the population aged 19–39 years (Fig. 3 and Fig. S3). Among women, associations with smoking intensity

tended to be somewhat stronger than among men, and among Dutch somewhat higher than among the full population.

Associations between alcohol consumption and air pollution differed between men and women and age groups (Fig. S4).

Physical activity had a stronger association with NO_2 exposure in men than in women, with similar although smaller associations observed for almost all included air pollutants (Fig. S5).

Obesity and overweight were negatively associated with NO_2, especially among 19–39 years old subjects (Fig. S6). No association was found for 65+ years olds.

3.5. Bias calculation of air pollution mortality risk estimates for active smoking as an example

Table 3 documents the potential impact of indirect adjustment for smoking for the pollutants NO_2 and $PM_{2.5}$ and mortality from lung cancer and cardiovascular disease. Expressed for the difference between the 5th and 95th percentile, HRs of about 1.105 and 1.073 were calculated for lung cancer mortality and NO_2 and $PM_{2.5}$, respectively, when not adjusting for active smoking. These HRs can be interpreted as the HR potentially observed in a cohort due to confounding related to missing smoking data when the true HR is 1. The potential bias was much smaller for cardiovascular mortality, consistent with the smaller effect of smoking on cardiovascular mortality. HRs for cardiovascular mortality in relation to NO_2 and $PM_{2.5}$ of, respectively, 1.029 and 1.020 can result from missing smoking data.

Motivated by the observation of different associations between air pollution and lifestyle in subgroups of the population, we also performed the indirect adjustment stratified for men and women and in three age categories. Indirect adjustment of air pollution effect estimates stratified by age mostly resulted in the largest bias in the population aged 40–64 years, followed by the population 65+ years (Table 4). The combined age-stratified indirectly-adjusted bias taking into account number of observations, did not differ compared to the bias in the full population (Table 4). We observed similar patterns for stratification by sex: indirect adjustment of HR caused a larger change in women than men (Table 5).

An assumption in these bias calculations is that the association between lifestyle and air pollution is linear. Fig. S7 based on deciles of pollutants, suggests that for none of the pollutants there was strong evidence of a nonlinear association. Especially for NO_2 and OP^{DTT}, associations were essentially linear. For $PM_{2.5}$, the small contrasts between the first deciles limited the linearity evaluation.

4. Discussion

We investigated the associations of individual lifestyle factors, including smoking habits, alcohol consumption, physical activity and BMI, with long-term exposure to a range of air pollutants (PM_{10}, $PM_{2.5}$, $PM_{2.5-10}$, absorbance, NO_2, OP^{DTT}, OP^{ESR}) at the home address among 387,152 adults of the Public Health Monitor in the Netherlands. Individual risk factors related to lifestyle were weakly associated with
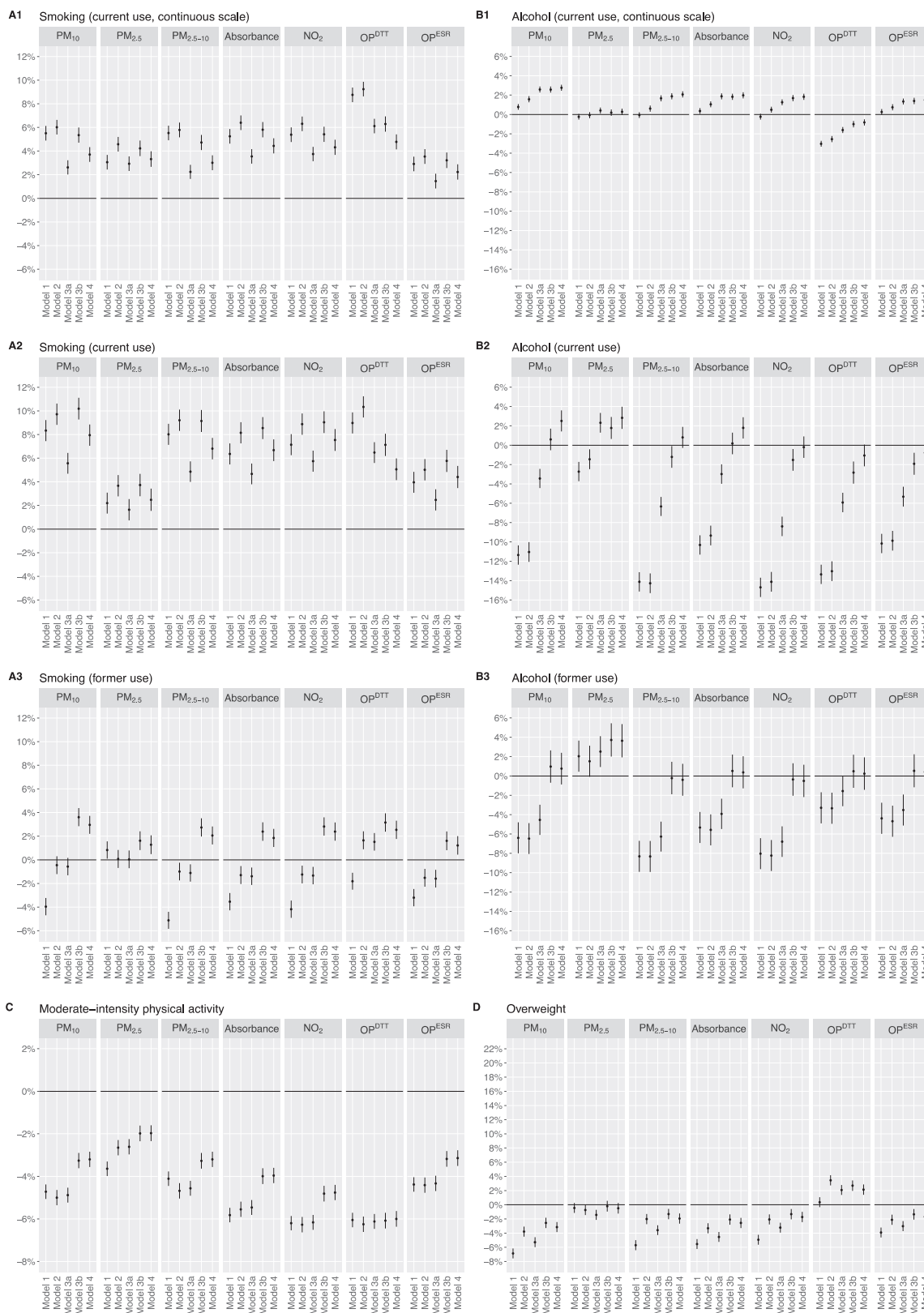
Fig. 2. Percentage differences in air pollution concentrations, relative to the standard deviation of air pollution, depending on the level of adjustment (Models 1–4) for basic subject characteristics: (A1) smoking intensity expressed per 10.3 cigarettes/day, (A2) between current and never smokers, (A3) between former and never smokers, (B1) alcohol consumption expressed per 8.5 alcohol glass/week, (B2) between current and never alcohol consumers, (B3) between former and never alcohol consumers, (C) physical activity expressed per 1005 min/week (2.4 h/day), (D) between overweight and normal-range BMI subjects. Calculated as slope of linear regression (air pollution dependent and lifestyle independent) multiplied by relevant increment of lifestyle (*e.g.*, 10.3 cigarettes) and divided by the standard deviation of air pollution. Note: Model 1: *air pollution = smoking* or *alcohol consumption* or *physical activity* or *BMI*; Model 2 = Model 1 + *sex + age*; Model 3a = Model 2 + *neighborhood SES score*; Model 3b = Model 2 + *individual SES characteristics*; Model 4 = Model 3a + Model 3b.
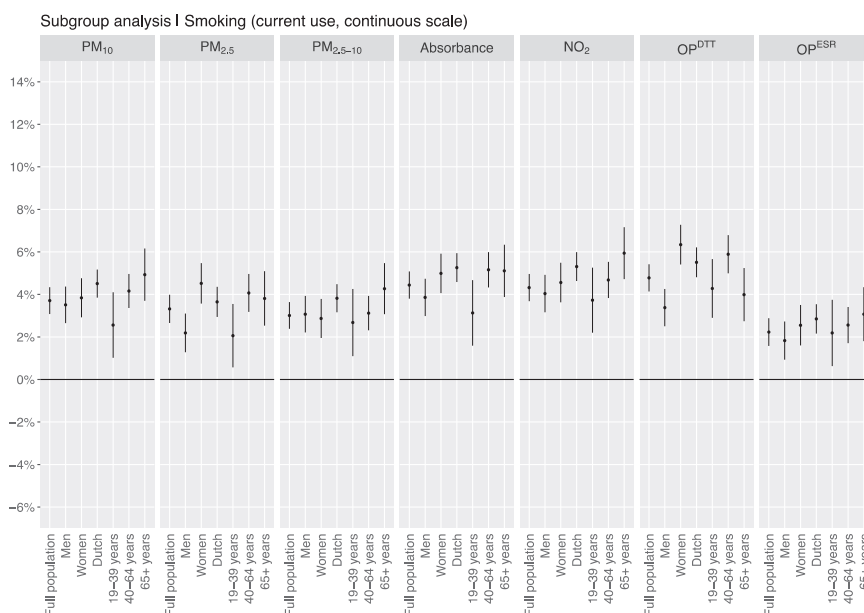
Fig. 3. Subgroup analysis of the differences in air pollution concentrations, comparing with the full population (Model 4), for smoking intensity expressed per 10.3 cigarettes/day. Calculated as slope of linear regression (air pollution dependent and lifestyle independent) multiplied by relevant increment of lifestyle (*e.g.*, 10.3 cigarettes) and divided by the standard deviation of air pollution.

Table 3
Potential bias due to not adjusting for active smoking intensity on air pollution effect estimates for lung cancer and cardiovascular mortality.

Pollutant	Mortality	$\widetilde{\Delta}$[a]	$\widetilde{\lambda}$[b]	Bias[c]	Bias*(P95-P5)[d]	HR[e]
NO$_2$	Lung cancer	0.021	0.245	0.0052	0.100	1.105
PM$_{2.5}$	Lung cancer	0.134	0.245	0.0329	0.071	1.073
NO$_2$	Cardiovascular disease	0.021	0.070	0.0015	0.028	1.029
PM$_{2.5}$	Cardiovascular disease	0.134	0.070	0.0094	0.020	1.020

[a] $\widetilde{\Delta}$ parameter estimate of linear regression with smoking as the dependent and air pollution as the independent variable, in cigarettes per µg/m^3 (see also Table S3).
[b] $\widetilde{\lambda}$ slope of the association between smoking and mortality, per cigarette.
[c] Calculated as $\widetilde{\Delta}\widetilde{\lambda}$ (formula (1) Section 2.4.2), expressed at the scale of slope per unit pollutant.
[d] Bias multiplied by contrast between 5th and 95th percentile (2.15 µg/m^3 for PM$_{2.5}$ and 19.24 µg/m^3 for NO$_2$).
[e] HR per contrast between 5th and 95th percentile (exponential of previous column), can be interpreted as the HR that may be observed due to confounding by missing smoking data, in the absence of a true pollution effect.

long-term exposure to air pollution. These associations generally diminished but were not fully explained by individual and neighborhood SES. Despite the generally weak associations, indirect adjustment for active smoking was found to potentially, considerably change effect estimates for combinations of NO$_2$, PM$_{2.5}$ and mortality from lung cancer and cardiovascular disease.

4.1. Associations between air pollution and lifestyle

Long-term air pollution concentrations estimated at the home address in the Netherlands were associated with individual risk factors related to lifestyle. For most lifestyle-related risk factors, unhealthy lifestyle was associated with higher air pollution exposure in our survey data. However, being overweight was associated with lower air pollution concentrations. The differences in concentrations between "healthy" and "unhealthy" lifestyles, although often statistically significant, were generally small (< 5% of the pollutant standard deviations) and the observed associations differed per pollutant. Different associations by pollutant are plausible as the degree of spatial variability of pollutant concentrations related to region of the country on the one hand, and intra-urban contrasts on the other hand, vary. The magnitude of the observed associations depended on the amount of controlling for individual- and area-level characteristics (Models 1–4). After adjusting for variables often available in epidemiological studies based upon administrative databases, *i.e.*, age, sex, education, occupation, income, marital status, ethnicity and neighborhood SES, the associations between lifestyle factors and air pollution were often reduced but still present. This suggests that either the association between air pollution and lifestyle is not exclusively due to differences in socioeconomic status or that the indicators we had available do not fully characterize SES. At the individual level, we had household income and education but not occupational class available. At the area-level, we had a score comprising income, education and occupa-

Table 4
Bias calculation of air pollution effect estimates stratified by age.

Pollutant	Mortality	19–39 year		40–64 year		65+ year		Combined	
		Bias*(P95-P5)[a]	HR[b]	Bias*(P95-P5)[a]	HR[b]	Bias*(P95-P5)[a]	HR[b]	Bias*(P95-P5)[a]	HR[b]
NO$_2$	Lung cancer	0.073	1.075	0.151	1.163	0.097	1.102	0.113	1.119
PM$_{2.5}$	Lung cancer	0.041	1.042	0.117	1.124	0.055	1.056	0.076	1.079
NO$_2$	Cardiovascular disease	0.021	1.021	0.043	1.044	0.028	1.028	0.032	1.033
PM$_{2.5}$	Cardiovascular disease	0.012	1.012	0.033	1.034	0.016	1.016	0.022	1.022

[a] Calculated as $\widetilde{\Delta}\widetilde{\lambda}$ (formula (1) Section 2.4.2) expressed at the scale of slope per unit pollutant and then multiplied by contrast between 5th and 95th percentile (2.15 for PM2.5 and 19.24 µg/m^3 for NO$_2$);
[b] HR per contrast between 5th and 95th percentile (exponential of previous column), most comparable for NO$_2$ and PM$_{2.5}$.

Table 5
Bias calculation of air pollution effect estimates stratified by sex.

Pollutant	Mortality	Men		Women		Combined	
		Bias*(P95-P5)[a]	HR[b]	Bias*(P95-P5)[a]	HR[b]	Bias*(P95-P5)[a]	HR[b]
NO_2	Lung cancer	0.105	1.111	0.093	1.098	0.099	1.104
$PM_{2.5}$	Lung cancer	0.053	1.054	0.085	1.089	0.070	1.073
NO_2	Cardiovascular disease	0.030	1.030	0.027	1.027	0.028	1.028
$PM_{2.5}$	Cardiovascular disease	0.015	1.015	0.024	1.025	0.020	1.020

[a] Calculated as $\widetilde{\Delta\lambda}$ (formula (1) Section 2.4.2) expressed at the scale of slope per unit pollutant and then multiplied by contrast between 5th and 95th percentile (2.15 for PM2.5 and 19.24 µg/m³ for NO₂);
[b] HR per contrast between 5th and 95th percentile (exponential of previous column), most comparable for NO_2 and $PM_{2.5}$.

tion. We did not have deprivation available. Furthermore, the use of neighborhood data for the year 2006 may have limited the ability to adjust for SES. However, in a six-year period, the ranking of neighborhoods socioeconomic scores has likely not changed in any major way.

In a study by Cesaroni et al. (2013), analysis of a small sample of the study population (0.6%) for whom information on smoking habits was available from an earlier investigation showed no association at all between NO₂ or PM₂.₅ exposure and ever smoking (after adjusting for sex, marital status, place of birth, education, occupation and the small area socioeconomic position indicator). Fischer et al. (2015) investigated the associations between PM₁₀ and NO₂ estimated using a different LUR model, correcting for lifestyle factors (*i.e.*, smoking habits, alcohol consumption, physical activity and BMI) among almost 64,000 participants of health surveys conducted in 2003–2005 in 11 different regions in the Netherlands. In their study, current smokers had 0.5% higher and former smokers had 0.1% higher PM₁₀ at four-digit postal code area level than never smokers. For NO₂ concentrations, current smokers had 2% higher concentration, whereas former smokers had 0.8% lower concentration than never smokers. All these associations were age and sex adjusted and statistically significant. After further adjustment for marital status, ethnicity, education and neighborhood SES score, the concentration differences became smaller and only remained significant for NO₂. In our much larger survey dataset, the fully adjusted associations were higher for PM₁₀ and NO₂ among current smokers, whereas the association for NO₂ among former smokers was positive. One cannot directly compare the findings as we expressed our estimates relative to the standard deviation of air pollution, instead of the mean. Expression relative to variability is more relevant when comparing different pollutants. Shin et al. (2014) found small, negative correlations between PM₂.₅ and cigarette smoking habits or BMI ranging from −0.04 to −0.02 in a Canadian population. In our study, the associations between current smoking and air pollution components were positive (Table S3), as for former smokers in the fully adjusted model. In the Canadian study, associations between lifestyle and air pollution were reduced to virtually null after adjustment for socioeconomic variables. This comparison shows that indirect adjustment for missing lifestyle factors may either increase or decrease observed air pollution effect estimates. This reinforces the notion that such adjustments should only be attempted after very careful evaluation of the comparability of the administrative data base source population and the survey population used to derive associations between air pollution and covariates missing in the administrative data base.

We focus in this paper on the observed associations and not on the likely complex mechanisms behind these associations, as correlation with air pollution exposure is a sufficient condition for a lifestyle variable – such as smoking – to confound air pollution effect estimates – on the condition that the covariate itself is related to the health endpoint. The observed associations are likely driven by differences in lifestyle between populations in urban and more rural areas and populations in neighborhoods of low, medium and high socio-economic position. Neighborhood socioeconomic characteristics were recently

shown to be associated with air pollution concentrations in the Netherlands (Fecht et al., 2015). The highest PM₁₀ and NO₂ mean concentrations were observed in the most deprived neighborhoods and contrast between the most and least deprived quintiles of neighborhoods was higher for NO₂ than for PM₁₀ (6.1 *vs.* 0.3 µg/m³, respectively, representing a 25% and 1% increase for the most deprived over the least deprived quintiles). The association between air pollution and alcohol use was affected by ethnicity. Many subjects of non-western origin do not consume alcohol for religious reasons and predominantly live in the major cities and within these cities in the more deprived neighborhoods with higher air pollution concentrations.

If both air pollution and lifestyle predict participation in the survey, spurious associations between the two may arise. Though, for the survey used in this paper, we do not have information that would help us address this possibility, it seems unlikely that lifestyle factors such as smoking are related to participation, other than through socioeconomic factors, age, and gender, for which we adjusted.

4.2. Potential bias in air pollution effect estimates

Using the Shin et al. (2014) methodology, we illustrated that the air-pollution associated mortality HR could be substantially biased when unadjusted for lifestyle-related risk factors in administrative cohorts. The effect of adjustment was larger for lung cancer than for cardiovascular mortality, because the relative risk for smoking is much higher for lung cancer than for cardiovascular disease. The potential bias is similar for PM₂.₅ and NO₂ when expressed for a comparable contrast in exposure.

The direction and relatively large magnitude of the potential bias is consistent with the ESCAPE analysis of natural-cause mortality, where smoking variables were found to be responsible for a decrease in the combined air pollution effect estimates of about 60% between the crude and fully adjusted models for 19 cohorts across Europe (Beelen et al., 2014). The potential bias is not directly comparable with the ESCAPE analysis, as there was no ESCAPE analysis with adjustment for the individual and area-level SES variables only. Fully adjusted PM₂.₅ h remained positive and significant in the ESCAPE study (Beelen et al., 2014). In a study by Fischer et al. (2015), adjustment for smoking-attributable mortality reduced the HR (especially for PM₁₀), but significant positive associations remained. In contrast, in a study by Cesaroni et al. (2013), inclusion of smoking status in a survival analysis restricted to the small sample of study population with smoking habits information available did not change the observed association between air pollution and non-accidental mortality. This shows that the results of indirect adjustments of air pollution effect estimates in administrative cohorts are study-specific.

Subgroup analysis showed substantial differences of the associations between lifestyle and air pollution by age and sex. Consistently, substantially different bias was found in these subgroups. The observed associations in the full population of the Public Health Monitor, which differs in age characteristics from the general population, are therefore not directly applicable to cohorts that are random samples of the

population. In such cases, the stratified indirect adjustment calculations as performed in this study are needed. Alternatively, the associations between air pollution and lifestyle could be calculated using probability weights to achieve a random sample of the population.

Further, even if the associations between lifestyle and air pollution are small as in our study, high RRs between lifestyle and health outcome (as obtained from the literature) can strongly affect the magnitude of indirect adjustment. The RR for lung cancer mortality related to active smoking from the ACS CPS-II study [11.63 (95% CI: 9.51, 14.24)] had substantially more influence on the indirect adjustment than the much lower RR for cardiovascular diseases mortality observed in the ACS CPS-II study [2.01 (95% CI: 1.84, 2.19)] (Table 3).

Consistent with the difference in direction of association between air pollution and smoking between our study and Shin et al. (2014), no adjustment for smoking resulted in bias away from the null in our study and towards the null in the Canadian study.

Lastly, Shin et al. (2014) assume that the relationship between air pollution and missing lifestyle factors ($\widetilde{\Delta}$) is linear. In our Survey data, the relationships between air pollutants and lifestyle factors do not clearly deviate from linear for all components (Fig. S7). Indirect adjustment based on an assumed linear relationship between air pollution and lifestyle covariate in cases of clear non-linearity may be problematic as it would produce biases in the adjusted air pollution effect estimates which could vary by concentration level. We note that the same assumption is also made in cohort studies that do have smoking intensity data.

In future, an ensemble of methods, including stratified analyses and consideration of non-linear associations, is likely necessary to address the complexity of the indirect adjustment issues. Comparison of the effect of indirect adjustment methods in large cohorts with actual adjustment for individual lifestyle data would be useful as well.

4.3. Strengths and limitations

The Public Health Monitor data was collected in 2012 whereas the LUR models were from 2009. However, the differences in exposure estimation are unlikely to have a demonstrable effect as spatial distribution of air pollution is stable over such a short period of time (Eeftens et al., 2011; Gulliver et al., 2011). Proper adjustment for, *e.g.*, smoking often requires data on a number of dimensions of a confounder. While in our calculations, we focussed only on smoking intensity to illustrate potential bias, the Shin et al. (2014) method allows for simultaneous indirect adjustment of several missing lifestyle variables. For most confounders, higher air pollution was associated with unhealthier lifestyle, except overweight. Bias due to all lifestyle factors would depend on the importance of each for specific health endpoints.

The strength of the study is that we were able to investigate the associations of lifestyle factors and air pollution concentrations among nearly 400,000 individuals with national coverage and a number of air pollution components. Additionally, the ESCAPE LUR models are stable and explain a large fraction of the spatial variance in measured annual averages of PM_{10}, $PM_{2.5}$, $PM_{2.5-10}$, absorbance, NO_2 and both OP metrics (Beelen et al., 2013; Eeftens et al., 2012; Yang et al., 2015). Applying those models allows for a robust estimation of outdoor home address concentrations for a range of air pollutants.

5. Conclusions

Individual risk factors related to lifestyle were weakly associated with long-term exposure to air pollution in the Netherlands. Indirect adjustment for missing lifestyle factors in administrative data cohort studies may substantially affect air pollution mortality risk estimates.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.envres.2017.03.050.

References

Ainsworth, B.E., Haskell, W.L., Whitt, M.C., Irwin, M.L., Swartz, A.M., Strath, S.J., et al., 2000. Compendium of physical activities: an update of activity codes and MET intensities. Med. Sci. Sports Exerc. 32, S498–504.

Ayres, J.G., Borm, P., Cassee, F.R., Castranova, V., Donaldson, K., Ghio, A., et al., 2008. Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential–a workshop report and consensus statement. Inhal. Toxicol. 20, 75–99. http://dx.doi.org/10.1080/08958370701665517.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., et al., 2013. Development of NO2 and NOx 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.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., et al., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet 383, 785–795. http://dx.doi.org/10.1016/S0140-6736(13)62158-3.

Borm, P.J.A., Kelly, F., Kunzli, N., Schins, R.P.F., Donaldson, K., 2007. Oxidant generation by particulate matter: from biologically effective dose to a promising, novel metric. Occup. Environ. Med. 64, 73–74. http://dx.doi.org/10.1136/oem.2006.029090.

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., et al., 2013. Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. Environ. Health Perspect. 121, 324–331. http://dx.doi.org/10.1289/ehp.1205862.

Crouse, D.L., Peters, P.A., van Donkelaar, A., Goldberg, M.S., Villeneuve, P.J., Brion, O., et al., 2012. Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national-level cohort study. Environ. Health Perspect. 120, 708–714. http://dx.doi.org/10.1289/ehp.1104049.

Eeftens, M., Beelen, R., Fischer, P., Brunekreef, B., Meliefste, K., Hoek, G., 2011. Stability of measured and modelled spatial contrasts in NO2 over time. Occup. Environ. Med. 68, 765–770. http://dx.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 PM2.5, PM2.5 absorbance, PM10 and PM coarse in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205. http://dx.doi.org/10.1021/es301948k.

Fecht, D., Fischer, P., Fortunato, L., Hoek, G., de Hoogh, K., Marra, M., et al., 2015. Associations between air pollution and socioeconomic characteristics, ethnicity and age profile of neighbourhoods in England and the Netherlands. Environ. Pollut. 198, 201–210. http://dx.doi.org/10.1016/j.envpol.2014.12.014.

Fischer, P.H., Marra, M., Ameling, C.B., 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. http://dx.doi.org/10.1289/ehp.1408254.

Gulliver, J., Morris, C., Lee, K., Vienneau, D., Briggs, D., Hansell, A., 2011. Land use regression modeling to estimate historic (1962−1991) concentrations of black smoke and sulfur dioxide for Great Britain. Environ. Sci. Technol. 45, 3526–3532. http://dx.doi.org/10.1021/es103821y.

Karssenberg, D., Schmitz, O., Salamon, P., de Jong, K., Bierkens, M.F.P., 2010. A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. Environ. Model. Softw. 25, 489–502. http://dx.doi.org/10.1016/j.envsoft.2009.10.004.

Knol, F., 1998. Van hoog naar laag; van laag naar hoog: de sociaal-ruimtelijke ontwikkeling van wijken tussen 1971–1995 (in Dutch). Sociaal en Cultureel Planbureau, Den Haag.

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA J. Am. Med. Assoc. 287, 1132–1141. http://dx.doi.org/10.1001/jama.287.9.1132.

Pope III, C.A., Burnett, R.T., Turner, M.C., Cohen, A., Krewski, D., Jerrett, M., et al., 2011. Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure–response relationships. Environ. Health Perspect. 119, 1616–1621. http://dx.doi.org/10.1289/ehp.1103639.

Shin, H.H., Cakmak, S., Brion, O., Villeneuve, P., Turner, M.C., Goldberg, M.S., et al., 2014. Indirect adjustment for multiple missing variables applicable to environmental epidemiology. Environ. Res. 134, 482–487. http://dx.doi.org/10.1016/j.envres.2014.05.016.

WHO, 2015a. BMI Classification. Available: ⟨http://apps.who.int/bmi/index.jsp?IntroPage=intro_3.html⟩ (Accessed 16 November 2015).

WHO, 2015b. Physical Activity and Adults. Available: ⟨http://www.who.int/dietphysicalactivity/factsheet_adults/en/⟩ (Accessed 16 November 2015).

Yang, A., Wang, M., Eeftens, M., Beelen, R., Dons, E., Leseman, D.L., et al., 2015. Spatial variation and land use regression modeling of the oxidative potential of fine particles. Environ. Health Perspect. http://dx.doi.org/10.1289/ehp.1408916.

Zeger, S.L., Dominici, F., McDermott, A., Samet, J.M., 2008. Mortality in the medicare population and chronic exposure to fine particulate air pollution in urban centers (2000–2005). Environ. Health Perspect. 116, 1614–1619. http://dx.doi.org/10.1289/ehp.11449.