Supplemental Material

Appendix 1. Differences between full cohort and case-cohort analyses

Essential in the NLCS-AIR study is that complete confounder information is only available for the subcohort and for subjects who died or developed cancer during follow-up. As a result, analyses adjusted for all available confounders from the questionnaire are only possible using the case-cohort approach. These potential confounders were chosen a priori (age, gender, BMI, active smoking, passive smoking, education, occupational exposure, marital status, alcohol use, vegetable intake, fruit intake, energy intake, fatty acids intake, folate intake, fish consumption, and area level socio-economic status variables). For the full cohort analyses only a limited number of confounders are available (age, gender, smoking status and area level socio-economic status variables).

Adjusting for all available confounders in the case-cohort analyses with the large confounder model led to a strong reduction of the number of subjects available for analysis (~60% of the original number) because of missing values in confounder variables. In the full cohort analyses this reduction was much smaller (~90% available for analysis) due to the limited number of confounders. This appendix includes the results of two sets of analyses we conducted to help interpret the results with the case-cohort and the full cohort approach. First, we assessed the impact of different models / populations on the effect estimates. Second, we assessed the role of random variability by repeating the case-cohort analyses after randomly drawing 100 new subcohorts from the full cohort.

Impact of confounder models and populations

We assessed four analysis models that differed in treatment of confounders: 1) adjusted for age and gender; 2) adjusted for all available confounders; 3) adjusted for age and gender, but only including the subjects that had complete information for all possible confounders included in model 2; and 4) adjusted only for the limited number of confounders available in the full cohort, but only including the subjects that had complete information for all possible confounders included in the model 2.

In the case-cohort dataset among the natural cause mortality cases 39.4% had a partner who never smoked, while 24.7% had a partner who was a former smoker and 35.9% had a partner who was a current smoker. For the subcohort members these percentages were 33.2%, 31.2% and 35.6%, respectively. Among the cases 18.5% was low exposed and 7.8% was high exposed to biological dust, with the remaining 73.7% being classified as non-exposed. Among the subcohort members 22.1% was low and 6.3% was high exposed to biological dust. For exposure to mineral dust the percentages low and high exposed were also slightly higher for the cases compared with the subcohort members: 17.0% versus 15.0% for low exposure, and 11.1% and 8.2% for high exposure. Among the cases 28.5% was low exposed to gases and fumes, and 11.8% was high exposed, while for the subcohort members these percentages were 25.9% and 8.9%, respectively. The median fruit consumption among cases was 137.2 g/day (interquartile range: 70.4 – 221.0 g/day); for subcohort members the median fruit consumption was 153.3 g/day (89.0 – 233.5 g/day). The median vegetable consumption was also slightly higher among subcohort members: 178.3 g/day (133.7 – 232.0 g/day), compared with the cases: 168.6 g/day (124.1 – 222.5 g/day).

Results of the confounder analyses are shown in Table 1 of the Supplemental Material for the black smoke (BS) background concentration and the traffic intensity on the nearest road for the various mortality outcomes. Important differences between the effect estimates of the BS background concentrations in the age-gender adjusted model and the age-gender adjusted model with only subjects with complete confounder information were found in the case-cohort analyses

for all mortality outcomes, showing that the occurrence of missing values introduced bias in the effect estimates of the background concentration. This difference was much smaller for the traffic variables in the case-cohort analyses. Similar analyses in the full cohort with the limited set of confounders showed much less evidence of such a selection effect for both background concentrations and traffic variables. The results also showed that in the case-cohort analyses there was little difference between the effect estimates adjusted for all available confounders and the effect estimates adjusted for the limited number of confounders in the full cohort, suggesting that inclusion of the full set of potential confounder variables in fact made little difference in the case-cohort analysis. The biggest difference was for respiratory mortality where the effect estimate of the model adjusted for all available confounders for the background concentration in the case-cohort analysis was higher than for the age-gender adjusted models or the model adjusted for the limited number of confounders available in the full cohort. This does suggest, but not guarantee, that the pattern of adjustment would be the same in the full cohort analysis if data on all confounders had been available for analysis.

We investigated whether information of specific confounders in the case-cohort analysis were primarily responsible for this selection effect. However, not just one or two confounders were responsible for the reduction in the number of subjects available for analysis, and therefore responsible for the selection effect, but the combination of all available confounders in the case-cohort analysis was responsible for the selection effect.

Role of random variability

Table 1 of the Supplemental Material also shows that the results of the age-gender adjusted model for the case-cohort and full cohort analyses produced nearly identical results for background concentrations but not for traffic intensity on the nearest road. The traffic variable was positively

associated with natural cause, cardiovascular, respiratory and lung cancer mortality in the full cohort analyses, but there were no associations in the case-cohort analyses.

We further explored this issue by randomly generating one hundred subcohorts of 5,000 subjects from the complete study population, and then repeating the age-gender adjusted case-cohort analysis using each of these one hundred subcohorts in turn as reference. The results for cardiopulmonary mortality are shown in Table 2 of the Supplemental Material. The average RRs of the 100 case-cohort analyses were, as expected, very close to the RR obtained in the full cohort. According to expectations under normal sampling theory, the RRs of the 100 case-cohort analyses varied, with the effect estimates of the original case-cohort analyses clearly within the range of effect estimates of the 100 new case-cohort analyses. However, the results also indicate that for the variables "traffic intensity on the nearest road" and "living near a major road" the results of the age and gender adjusted case-cohort analysis using the original subcohort are different from what was found for the average of the 100 randomly drawn subcohorts. For the other exposure variables there was no such difference. These results suggest that the effect estimates in the case-cohort analyses can be sensitive to sampling variation, i.e. sensitive to the selection of the subcohort even though it was completely random, probably due to the small fraction of high exposed subjects ("living near a major road") and the skewness of the exposure distribution ("traffic intensity on the nearest road" – see also Figure 1). This sampling variation results in effect estimates that do not reflect the underlying effect estimates in the study population as a whole.

Supplemental Material, Table 1: Relative risks (95% CIs) for the association between background concentration (period 1987-1996) and traffic intensity with cause-specific mortality in case-cohort and full cohort analyses, using different confounder models.

Exposure model	Confounder model [a]	Population [a]	Case-cohort analyses	N [b]	Full cohort analyses	N [b]
Natural cause mortality						
Black smoke background	Age-gender adjusted	All	1.15 (0.97 – 1.35)	21,457	1.14 (1.07 – 1.22)	117,499
Traffic intensity on nearest road			0.99 (0.91 – 1.08)		1.04 (1.00 – 1.08)	
Black smoke background	Fully adjusted	Complete confounder data	0.99 (0.75 – 1.31)	12,720	1.09 (1.00 – 1.19)	105,296
Traffic intensity on nearest road			0.99 (0.88 – 1.11)		1.03 (1.00 – 1.08)	
Black smoke background	Age-gender adjusted	Complete confounder data	1.03 (0.83 – 1.28)	12,720	1.15 (1.07 – 1.24)	105,296
Traffic intensity on nearest road			1.00 (0.90 – 1.12)		1.04 (1.01 – 1.09)	
Black smoke background	Partially adjusted	Complete confounder data	0.99 (0.76 – 1.19)	12,720	-	-
Traffic intensity on nearest road			0.98 (0.88 – 1.09)		-	-
Cardiovascular mortality						
Black smoke background	Age-gender adjusted	All	1.14 (0.94 – 1.38)	10,762	1.14 (1.02 – 1.28)	117,499
Traffic intensity on nearest road			1.00 (0.91 – 1.10)		1.06 (1.00 – 1.13)	
Black smoke background	Fully adjusted	Complete confounder data	1.00 (0.72 – 1.40)	6,510	1.11 (0.96 – 1.28)	105,296
Traffic intensity on nearest road			1.03 (0.90 – 1.17)		1.05 (0.99 – 1.12)	
Black smoke background	Age-gender adjusted	Complete confounder data	1.05 (0.81 – 1.36)	6,510	1.16 (1.03 – 1.31)	105,296
Traffic intensity on nearest road			1.02 (0.91 – 1.16)		1.06 (1.00 – 1.13)	

Black smoke background	Partially adjusted	Complete confounder data	1.01 (0.74 – 1.39)	6,510	-	-
Traffic intensity on nearest road			1.00 (0.88 – 1.13)		-	-
Respiratory mortality						
Black smoke background	Age-gender adjusted	All	1.42 (1.01 – 2.00)	5,847	1.41 (1.06 – 1.88)	117,499
Traffic intensity on nearest road			1.04 (0.91 – 1.19)		1.13 (0.99 – 1.27)	
Black smoke background	Fully adjusted	Complete confounder data	1.52 (0.80 – 2.88)	3,607	1.22 (0.86 – 1.74)	105,296
Traffic intensity on nearest road			0.94 (0.71 – 1.25)		1.10 (0.95 – 1.26)	
Black smoke background	Age-gender adjusted	Complete confounder data	1.31 (0.82 – 2.10)	3,607	1.34 (0.99 – 1.82)	105,296
Traffic intensity on nearest road			1.01 (0.80 – 1.27)		1.11 (0.97 – 1.27)	
Black smoke background	Partially adjusted	Complete confounder data	1.33 (0.77 – 2.31)	3,607	-	-
Traffic intensity on nearest road			0.97 (0.77 – 1.21)		-	-
Lung cancer mortality						
Black smoke background	Age-gender adjusted	All	1.17 (0.89 – 1.53)	6,692	1.15 (0.94 – 1.42)	
Traffic intensity on nearest road			1.00 (0.89 – 1.13)		1.06 (0.95 – 1.18)	
Black smoke background	Fully adjusted	Complete confounder data	1.02 (0.61 – 1.71)	4,075	1.01 (0.78 – 1.32)	
Traffic intensity on nearest road			1.03 (0.87 – 1.22)		1.07 (0.96 – 1.19)	
Black smoke background	Age-gender adjusted	Complete confounder data	1.03 (0.71 – 1.48)	4,075	1.09 (0.87 – 1.37)	
Traffic intensity on nearest road			1.07 (0.92 – 1.24)		1.09 (0.97 – 1.21)	
Black smoke background	Partially adjusted	Complete confounder data	0.93 (0.59 – 1.48)	4,075	-	-
Traffic intensity on nearest road			1.01 (0.86 – 1.17)		-	-

Other mortality						
Black smoke background	Age-gender adjusted	All	1.12 (0.94 – 1.33)	13,098	1.11 (1.01 – 1.22)	
Traffic intensity on nearest road			0.97 (0.89 – 1.05)		1.00 (0.95 – 1.06)	
Black smoke background	Fully adjusted	Complete confounder data	0.95 (0.71 – 1.26)	7,883	1.09 (0.96 – 1.23)	
Traffic intensity on nearest road			0.93 (0.82 – 1.06)		1.00 (0.94 – 1.06)	
Black smoke background	Age-gender adjusted	Complete confounder data	0.98 (0.78 – 1.23)	7,883	1.13 (1.02 – 1.26)	
Traffic intensity on nearest road			0.96 (0.85 – 1.08)		1.00 (0.95 – 1.06)	
Black smoke background	Partially adjusted	Complete confounder data	0.96 (0.73 – 1.25)	7,883	-	-
Traffic intensity on nearest road			0.94 (0.84 – 1.06)		-	-

[a] Used confounder models: *Age-gender adjusted*: adjusted for age and gender; *Fully adjusted*: adjusted for all available potential confounders; and *Partially adjusted*: adjusted only for confounders of the limited full cohort confounder model.

Populations: *All*: All subjects; and *Complete confounder data:* Only including the subjects that had complete information for all possible confounders included in the fully adjusted confounder model.

Used confounders in fully adjusted confounder model:

Case-cohort analysis: age, gender, BMI, active smoking, passive smoking, education, occupational exposure, marital status, alcohol use, vegetable intake, fruit intake, energy intake, fatty acids intake, folate intake, fish consumption, and area level indicators of socio-economic status.

Full cohort analysis: age, gender, smoking status, and area level indicators of socio-economic status.

RRs were calculated for concentration changes from the 5^{th} to the 95^{th} percentile: 10 $\mu g/m^3$ for BS and 10,000 mvh/24 h for traffic intensity on the nearest road.

[b] N is the number of observations available for analysis. The number of observations in case-cohort analyses is the sum of subcohort members and the number of mortality cases of the studied cause.

Supplemental Material, Table 2: Distribution of RR estimates and 95% CIs for cardiopulmonary mortality from case-cohort analyses of 100 randomly drawn subcohorts, and RRs of the case-cohort analyses with original subcohort and RRs of the full cohort analyses (adjusted for age and gender).[a]

Exposure model	RR (95%-CI) for case-cohort with original subcohort	RR (95%-CI) for full cohort	Average RR (min – max) [SD] of 100 case-cohort analyses
Black smoke background	1.17 (0.97 – 1.42)	1.17 (1.05 – 1.30)	1.16 (0.97 – 1.38) [0.09]
Traffic intensity on nearest road	1.01 (0.92 – 1.11)	1.07 (1.02 – 1.13)	1.08 (0.90 – 1.26) [0.05]
Black smoke background	1.13 (0.93 – 1.38)	1.16 (1.04 – 1.29)	1.15 (0.94 – 1.41) [0.09]
Traffic intensity in a 100 m buffer	1.08 (0.95 – 1.22)	1.06 (0.99 – 1.14)	1.07 (0.89 – 1.21) [0.06]
Black smoke background	1.18 (0.97 – 1.42)	1.18 (1.06 – 1.31)	1.17 (0.98 – 1.39) [0.09]
Living near a major road	1.00 (0.83 – 1.21)	1.10 (0.99 – 1.22)	1.10 (0.84 – 1.37) [0.09]

[a] RRs were calculated for concentration changes from the 5th to the 95th percentile: 10 µg/m^3 for BS background/overall estimate; for the traffic intensity on the nearest road: 10,000 mvh/24h, for the sum of traffic intensity in a buffer of 100 m: 335,000 mvh/100m. RRs for living near a major road were calculated with as reference category not living near a major road.