Protocol for Conducting Epidemiological Studies on Residential Environmental Exposures and Health Outcomes

Summary

This protocol provides a description of key steps in conducting epidemiological studies on environmental exposures and health outcomes, depicted using a practical example. It outlines key methodological steps, including exposure assessment, cohort definition, outcome classification, and confounding control, along with selected SAS and R code snippets to support implementation.

The example is based on a REMEDIA study of air pollution and chronic obstructive pulmonary disease (COPD) in Denmark but can be adapted to other exposures, outcomes, and contexts. It is intended for researchers, analysts, and technical staff, and aims to support transparency and reproducibility within the REMEDIA framework.

Objective of this specific study:

To estimate the association between long-term, source-specific air pollution ($PM_{2.5}$, NO_2, UFP, and EC) and the risk of developing COPD using harmonized cohort data, high-resolution air pollution modeling, and national health registers.

The following sections present the main steps followed in this study, along with some of the results obtained.

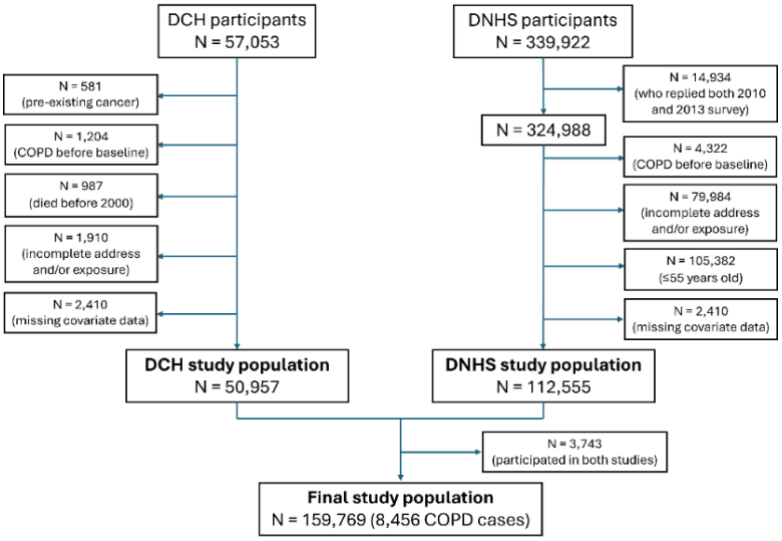
---

1. Study Population

- Identify cohorts suited for answering the research questions
- In the current example, harmonized data from two large Danish cohorts was used:
  - **DCH (Diet, Cancer and Health)**: Adults aged 50–64 recruited in 1993–1997 from Copenhagen and Aarhus.
  - **DNHS (Danish National Health Survey)**: Nationwide survey from 2010 and 2013, restricted to individuals aged ≥55 for comparability.
- Exclude participants with:
  - Prevalent COPD
  - Incomplete address or exposure history (>20% missing)
  - Missing covariate data

Example result:

Flowchart illustrating the number of participants included in the final study population:



---

2. Outcome Assessment

- Define the outcome in question based on a standard classification method, such as the International Classification of Diseases (ICD). For COPD this corresponds to ICD-8: 491, 492 and ICD-10: J42–J44.
- In the current example, incident COPD was identified using national registry linkage.
- Exclude all prevalent COPD cases at baseline.

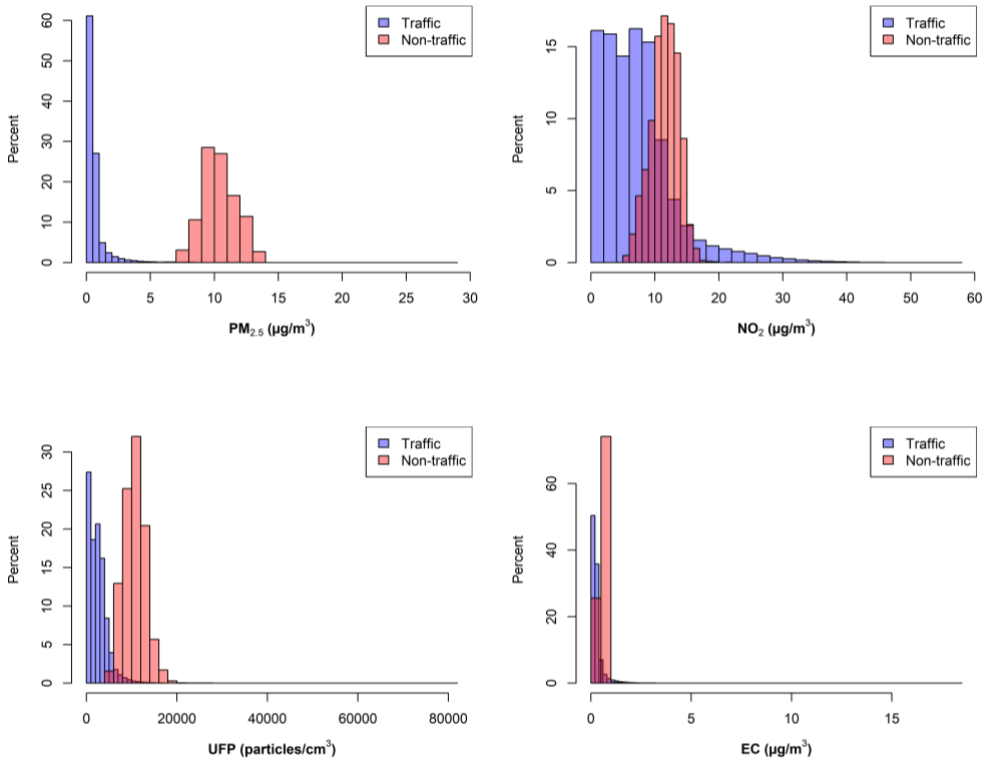
---

3. Exposure Assessment

- Identify addresses for all cohort participants, preferably for a longer time-period
- In the current example, full residential address history from 1990–2017 was found for all cohort participants.
- Model address-specific environmental pollution. In the current example outdoor air pollution concentrations from traffic (local traffic emissions) and non-traffic (industry, agriculture, long-range transport, etc.) using the **DEHM/UBM/AirGIS** system was modelled:
    - DEHM (regional scale)
    - UBM (urban scale)
    - OSPM (street-level)
- Calculate long-term time-weighted exposure, here **10-year time-weighted means**.

Example result:

Distribution of $PM_{2.5}$, NO_2, UFP, and EC concentrations as 10-year time-weighted averages for the entire study population (N=159,769), apportioned by traffic and non-traffic sources.



---

4. Covariates

- Identify potential confounders to be used in adjusted statistical models, based on a literature search, availability of information in the cohort(s) and construction of a DAG, e.g. in DAGitty (https://www.dagitty.net/)

- In the present example, the following covariates were selected:
    - Education, income, occupation, cohabiting status, area-level SES indicators (register-based)
    - Smoking status and intensity, alcohol intake, diet, physical activity (questionnaire-based)
- If more cohorts are used, then harmonize data for all cohorts, as shown for the present example in the table below:

Covariate	DCH cohort	DNHS cohort	Pooled cohort
Smoking status	1) Have you ever smoked cigarettes, cigars, cigarillos or pipes regularly, i.e. at least one per day for a year? Possible answers: yes, no 2) Do you smoke daily at the moment? Possible answers: yes, no	Do you smoke? Possible answers: yes, everyday; yes, at least once a week; yes, rarely one every week; no, I quit smoking; no, I never smoked.	Categorized into: never, former, current
Smoking intensity among current smokers	1) How many cigarettes, cigars, cigarillos or pipes do you smoke daily? 2) What types of cigarettes do you smoke? Possible answers: do not smoke cigarettes, cigarettes with filter, cigarettes without filter, both cigarettes with and without filters.	1) How many cigarettes, cigars, cigarillos or pipes do you smoke per day, in average?	Harmonized to g of tobacco/day
Alcohol intake and alcohol abstainers	How often do you drink common alcoholic beverages (beer, wine, fortified wine or spirits)? Possible answers: I do not drink, less than once a month, 1-3 a month, once a week, 2-4 a week, 5-6 a week, every day.	1) Have you drunk alcohol for the last 12 months? 2) How many drinks do you typically drink on each day of the week?	Harmonized to g/day for alcohol intake and yes/no for alcohol abstainers
Fruit and vegetable intake	For different types of food, participants were asked "How often do you eat this specific food?" Possible answers: never, less than once a month, once a month, 2-3 a month, once a week, 2-4 a week, 5-6 a week, once a day, 2-3 a day, 4-5 a day, 6-7 a day, 8 or more a day.	1) How many servings of fruit do you usually eat? Possible answers: more than 6 a day, 5-6 a day, 3-4 a day, 1-2 a day, 5-6 a week, 3-4 a week, 1-2 a week, none. 2) How often do you eat vegetables? Possible answers: more than once a day, 5-7 a week, 3-4 a week, 1-2 a week, rarely/never	Categorized into: no intake/very low, low, medium, high
Physical activity	How many hours a week you spend on the following activities: walk, cycle, housework activities, DIY activities in the house, gardening, sports? Out of those, how many hours the activity makes you out of breath? Answers are given for winter and summer months.	If you look at the past year, what would you say best describes your physical activity in your free time? Possible answers: train hard regularly; do physical exercise or gardening at least four hours a week; walk, cycle or do light exercise at least four hours a week; read, watch tv or do other sedentary activities.	Categorized into: none/low, medium, high

---

5. Statistical Analysis

- Identify follow-up time. In the current example from Jan 2000 (DCH) and from baseline (2010/2013) for DNHS until end of follow up defined as either COPD diagnosis, death, emigration, missing address, or Dec 31, 2017 (whatever came first)
- Use **Cox proportional hazards models** with age as the time scale.
- Select main exposure time-window, here 10-year time-weighted averages
- Report hazard ratios (HR). In our example, HRs are reported per:
    - **Interquartile change increase** (main analysis)
    - **Fixed increments** (e.g., per 10 $\mu g/m^3$ $PM_{2.5}$)

- Select adjustment models. In the present example:
  Model 1: Age, sex, calendar year, cohort (strata)
  Model 2: Further adjusted for socio-demographics
  Model 3: Further adjusted for smoking variables
  Model 4: Further adjusted for lifestyle (Main model)

Example result:
Associations between 10-year mean residential exposure to air pollution (per interquartile change) and risk for COPD (8,456 cases). Cohort study pooling data from two Danish cohorts.

Air pollutant exposure (per IQR) [a]	**IQR**	**Model 1** [a, b] **HR (95% CI)**	**Model 2** [a, c] **HR (95% CI)**	**Model 3** [a, d] **HR (95% CI)**	**Model 4** [a, e] **HR (95% CI)**
$PM_{2.5}$ ($\mu g/m^3$)					
Total	2.33	1.24 (1.18, 1.30)	1.19 (1.12, 1.25)	1.12 (1.06, 1.19)	1.11 (1.05, 1.17)
Traffic [f]	1.85	1.06 (1.05, 1.07)	1.03 (1.02, 1.05)	1.02 (1.00, 1.03)	1.01 (1.00, 1.03)
Non-traffic [g]	0.48	1.14 (1.06, 1.22)	1.20 (1.11, 1.29)	1.17 (1.08, 1.26)	1.17 (1.09, 1.26)
NO_2 ($\mu g/m^3$)					
Total	9.25	1.19 (1.15, 1.23)	1.15 (1.10, 1.19)	1.09 (1.05, 1.14)	1.08 (1.04, 1.13)
Traffic [f]	6.52	1.14 (1.11, 1.17)	1.10 (1.07, 1.13)	1.06 (1.03, 1.10)	1.05 (1.02, 1.09)
Non-traffic [g]	3.02	1.05 (1.02, 1.09)	1.12 (1.06, 1.17)	1.08 (1.03, 1.14)	1.08 (1.03, 1.14)
UFP (particles/cm^3)					
Total	5737	1.12 (1.07, 1.18)	1.12 (1.06, 1.18)	1.06 (1.01, 1.12)	1.05 (0.99, 1.11)
Traffic [f]	2570	1.18 (1.15, 1.23)	1.13 (1.09, 1.18)	1.08 (1.04, 1.13)	1.07 (1.03, 1.12)
Non-traffic [g]	3308	0.96 (0.92, 1.01)	1.01 (0.96, 1.06)	0.99 (0.94, 1.04)	0.98 (0.94, 1.03)
EC ($\mu g/m^3$)					
Total	0.34	1.06 (1.05, 1.08)	1.05 (1.03, 1.07)	1.03 (1.00, 1.05)	1.02 (1.00, 1.05)
Traffic [f]	0.22	1.07 (1.05, 1.09)	1.04 (1.02, 1.06)	1.02 (1.00, 1.04)	1.02 (1.00, 1.04)
Non-traffic [g]	0.12	0.97 (0.94, 1.00)	1.00 (0.98, 1.02)	1.00 (0.98, 1.03)	1.00 (0.98, 1.03)

[a] IQR, interquartile range; CI, confidence interval; HR, hazard ratio; $PM_{2.5}$, particulate matter with a diameter <2.5 µm; NO_2, nitrogen dioxide; UFP, ultrafine particles; EC, elemental carbon.
[b] Adjusted for age (by design), sex and calendar-year.
[c] Further adjusted for cohabiting status, education, income, occupational status, and area-level socioeconomic variables (i.e. percent population with low income, with only basic education, and with a criminal record).
[d] Further adjusted for smoking (smoking status and intensity (g tobacco/day) measured at baseline).
[e] Further adjusted for physical activity, dietary habits (i.e. intake of fruit and vegetable), and alcohol consumption (intake g/day and abstainers) measured at baseline.
[f] Local traffic sources.
[g] Non-traffic sources and non-local traffic sources.

---

6. Additional Analyses

- **Shape of associations:** Natural cubic splines (3 df)
- **Two-pollutant models:** Within same source type
- **Effect modification:** Test interactions, e.g. by:
  - Sex
  - Smoking status
  - Education

---

7. Software Used and Code Snippets

- Main analyses: **SAS 9.4**
- Spline models and correlations: **R 4.3.2**

SAS CODE FOR SUMMARY STATISTICS

```
libname mydata "INCLUDE FILE PATH HERE";

data descriptive01;
	set mydata.population; *(obs=100000);
	by id;
	if first.id;
run;

title "Descriptive categorical variables - all population";
proc freq data = descriptive01;
	tables sex cohab occup income3cat education3 smoking alko_abst physact cohort fruit_cat
	totveg_cat/ missing ;
run;

title "Descriptive continuous variables - all population";
proc means n median p5 p95 mean std data = descriptive01 ;
	var age_start age_end PI_lowincome PI_basiceducation  PI_crime EC_total_10 EC_nontraffic_10
	EC_traffic_10 NO2_total_10 NO2_nontraffic_10 NO2_traffic_10 PM25_total_10
	PM25_nontraffic_10 PM25_traffic_10 UFP_total_10 UFP_nontraffic_10 UFP_traffic_10 alko
	gram_nu;
run;
```

SAS CODE FOR COX REGRESSION ANALYSIS

```
libname mydata "INCLUDE HERE FILE PATH";
option compress=binary;

* SINGLE POLLUTANT MODELS
* Example: COPD and PM2.5 total;

data analyse01;
	set mydata.population; *(obs=100000);
	by id;
run;

title "PM2.5 total- Model 1 (adjusted for age, sex and calendar-year)";
	proc phreg fast data = analyse01 (keep=age_start age_end case_copd PM25_total_10_iqr sex
	cal_cat cohort);
	class sex cal_cat cohort;
	model (age_start,age_end)*case_copd(0) = PM25_total_10_iqr cal_cat sex
	/rl ties=breslow ;
	Strata cohort;
run;

title "PM2.5 total- Model 2";
proc phreg fast data = analyse01 (keep=age_start age_end case_copd PM25_total_10_iqr sex cal_cat
cohort cohab occup income3cat education3 PI_lowincome PI_basiceducation  PI_crime);
	class sex cal_cat cohort cohab occup income3cat education3;
	model (age_start,age_end)*case_copd(0) = PM25_total_10_iqr sex cal_cat cohab occup
	income3cat education3 PI_lowincome PI_basiceducation PI_crime
	/rl ties=breslow ;
	Strata cohort;
run;

title "PM2.5 total- Model 3";
proc phreg fast data = analyse01 (keep=age_start age_end case_copd PM25_total_10_iqr sex cal_cat
cohort cohab occup income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking
gram_nu);
	class sex cal_cat cohort cohab occup income3cat education3 smoking;
	model (age_start,age_end)*case_copd(0) = PM25_total_10_iqr sex cal_cat cohab occup
	income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking gram_nu
	/rl ties=breslow ;
	Strata cohort;
run;

title "PM2.5 total- Complete model";
proc phreg fast data = analyse01 (keep=age_start age_end case_copd PM25_total_10_iqr sex cal_cat
cohort cohab occup income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking
gram_nu alko alko_abst fruit_cat totveg_cat physact);
	class sex cal_cat cohort cohab occup income3cat education3 smoking alko_abst fruit_cat
	totveg_cat physact;
```

```
        model (age_start,age_end)*case_copd(0) = PM25_total_10_iqr sex cal_cat cohab occup
        income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking gram_nu alko
        alko_abst fruit_cat totveg_cat physact
        /rl ties=breslow ;
        Strata cohort;
run;

* TWO-POLLUTANT MODELS;
* Example: PM2.5 and NO2

title "PM2.5 + NO2 - Complete model";
proc phreg fast data = analyse01 (keep=age_start age_end case_copd PM25_total_10_iqr
NO2_total_10_iqr sex cal_cat cohort cohab occup income3cat education3 PI_lowincome
PI_basiceducation  PI_crime smoking gram_nu alko_abst fruit_cat totveg_cat physact);
        class sex cal_cat cohort cohab occup income3cat education3 smoking alko_abst fruit_cat
        totveg_cat physact;
        model (age_start,age_end)*case_copd(0) = PM25_total_10_iqr NO2_total_10_iqr sex cal_cat
        cohab occup income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking
        gram_nu alko alko_abst fruit_cat totveg_cat physact
        /rl ties=breslow ;
        Strata cohort;
run;
```

SAS CODE FOR EFFECT MODIFICATION ANALYSIS

```
libname mydata "INCLUDE HERE FILE PATH";

* Example: effect modification analysis by sex – models with PM2.5 total contribution;

data analyse01;
	set mydata.popopulation; *(obs=100000);
	by id;

	male = 0;
	if SEX = "M" then male = 1;

	female = 0;
	if SEX = "K" then female = 1;

	male_PM25tot = male*PM25_total_10_iqr;
	female_PM25tot = female*PM25_total_10_iqr;

run;

proc freq data = analyse01;
	where case_copd = 1 ;
	tables male female;
run;

title "PM25 total - Effect modification SEX";
proc phreg fast data = analyse01 (keep=age_start age_end case_copd male male_PM25tot
female_PM25tot cal_cat cohort cohab occup income3cat education3 PI_lowincome PI_basiceducation
PI_crime smoking gram_nu alko alko_abst fruit_cat totveg_cat physact);
	class cal_cat cohort cohab occup income3cat education3 smoking alko_abst fruit_cat totveg_cat
	physact male;
	model (age_start,age_end)*case_copd(0) = male male_PM25tot female_PM25tot cal_cat cohab
	occup income3cat education3 PI_lowincome PI_basiceducation  PI_crime smoking gram_nu alko
	alko_abst fruit_cat totveg_cat physact
	/rl ties=breslow ;
	Strata cohort;
	testsexPM25: test male_PM25tot = female_PM25tot ; * Testing for interaction;
run;
```

R CODE FOR COX REGRESSION ANALYSIS AND BUILDING SPLINES (DOSE-RESPONSE RELATIONSHIPS)

```
# Opening libraries (not all are needed)
library(foreign)
library(survival)
library(ggplot2)
library(magrittr)
library(Hmisc)
library(lattice)
library(Formula)
library(survminer) #til ggcoxzph
library(rms)
library(tibble)
library(splines)
library(svglite)
#Plotting smooth terms
require(survival)

#Creating work directory:
WD <- "INCLUDE HERE YOUR WORKING DIRECTORY PATH"
setwd(WD)

#Saving work directory:
save.image("Splines_dchdnsp_copd")
load("Splines_dchdnsp_copd")

#Importing dataset:
copd_pop <- read.csv('FILE PATH', header = T, sep=',')
str(copd_pop)
```

#EXAMPLES FOR COPD AND PM2.5 TOTAL

```
# Cox regression analysis in R
res.cox1a_01 <-coxph(Surv(age_start, age_end, case_COPD) ~ PM25_total_10 +
                     as.factor(sex) + as.factor(cal_cat) + strata(cohort) +
                     as.factor(cohab) + as.factor(occup)+ as.factor(income3cat)+ as.factor(education3) +
                     PI_lowincome + PI_basiceducation + PI_crime +
                     as.factor(smoking) + gram_nu +
                     alko + as.factor(alko_abst) + as.factor(fruit_cat) + as.factor(totveg_cat) +
                     as.factor(physact), data=copd_pop)
# Checking proportional hazards
res.cox1a_01
res.zph1a_01 <- cox.zph(res.cox1a_01)
res.zph1a_01$table

### SPLINES FOR PM2.5 TOTAL
PM25.ns <- coxph(Surv(age_start, age_end, case_COPD) ~ ns(PM25_total_10, df=3) +
                 as.factor(sex) + as.factor(cal_cat) + strata(cohort) +
                 as.factor(cohab) + as.factor(occup)+ as.factor(income3cat)+ as.factor(education3) +
```

```
            PI_lowincome + PI_basiceducation + PI_crime +
            as.factor(smoking) + gram_nu +
            alko + as.factor(alko_abst) + as.factor(fruit_cat) + as.factor(totveg_cat) +
            as.factor(physact), data=copd_pop)
PM25.ns
#Predicted terms
pred_solo_PM25 <-predict(PM25.ns, type="terms", se.fit=TRUE, terms=1) # for the single pollutant
model with PM 2.5 total

#Making percentiles (this will be used later for setting up limits when plotting the splines)
pm25_pctl1<-quantile(copd_pop$PM25_total_10,0.01)
pm25_pctl99<-quantile(copd_pop$PM25_total_10,0.99)
pm25_pctl50<-quantile(copd_pop$PM25_total_10,0.50)
pm25_pctl5<-quantile(copd_pop$PM25_total_10,0.05)
pm25_pctl95<-quantile(copd_pop$PM25_total_10,0.95)
pm25_pctl10<-quantile(copd_pop$PM25_total_10,0.10)
pm25_pctl90<-quantile(copd_pop$PM25_total_10,0.90)

#Spline for total PM2.5
png('PM25_ns_total.png', width = 4200, height = 3200, res = 600, bg = 'transparent')
plot(0.1, type = "n", xlab=expression(bold("Total PM"[2.5]*" (?g/m"^3*")")),
ylab=expression(bold("Hazard ratio")),
xlim=c(pm25_pctl5,pm25_pctl95),xaxs="i",ylim=c(0.8,1.3),log="y")
main = expression(bold("Total PM"[2.5]*" (?g/m"^3*")"))
lines(smooth.spline((copd_pop$PM25_total_10), exp(pred_solo_PM25$fit)), col="dodgerblue3",
lwd=1.6)
lines(smooth.spline((copd_pop$PM25_total_10), exp(pred_solo_PM25$fit +
1.96*pred_solo_PM25$se)), col="dodgerblue3", lty = 2, lwd=0.8)
lines(smooth.spline((copd_pop$PM25_total_10), exp(pred_solo_PM25$fit -
1.96*pred_solo_PM25$se)), col="dodgerblue3", lty = 2, lwd=0.8)
dev.off()
```