Study Protocol

ESCAPE Statistical Tasks Working Group (WG)

Version 7

Wednesday, February 20, 2012

Table of Contents

Introduction

During the ESCAPE plenary meeting of June 14-16, 2010, a working group was set up to harmonize statistical analysis within ESCAPE, where needed. During the meeting various suggestions were made for tasks. In July 2010, a first text was sent around to define the issues. Comments were received from the members of the WG, which were accommodated in a revised text, distributed December 22, 2010. That text was adapted based upon a telephone meeting on January 5, 2011 and written comments by Brian Miller and Barbara Hoffmann, and developed further incorporating comments and discussions upon telephone meetings of the group on April 12th, June 9th, June 30th, August 29th, October 27th, November 22nd, December 14th, 2011, and January 31st, 2012, where more specific issues about analyses strategy and specific code development were discussed. Section 1: 'Overall data analyses strategy', starts with a description of the overall strategy within ESCAPE of data analysis, based upon the study manual prepared in 2008 and the WP-specific data analysis manuals prepared later. Section 2 lists the members of the team with contact info.

This text remains general as the more specific points are addressed in the WP-specific (or even paper-specific) analysis plans. This text only serves to harmonize data analysis within the WPs where necessary. It is not the complete guidebook for data analysis within ESCAPE. Some sections have expanded discussions of the relevant issues, as a support to decisions made by Stat groups. These are sections *1.4.1 Modeling of Area-level confounders* and *1.5.1 Considerations of different methods for handling missing data in ESCAPE.*

Some of the code writing is still ongoing within the WPs, but because of the start of epidemiological we wanted to have a final document to serve the Health WP analyses.

1. Overall data analysis strategy

The main analysis selected for ESCAPE is cohort-specific analysis of the association between air pollution and health outcomes followed by a meta-analysis of the effect estimates for the individual cohorts. Cohort-specific analyses will be conducted locally. There will be no transfer of individual cohort data , except for selected outcomes in WP3. For a description of the cohorts we refer to the study manual available on the ESCAPE website and the detailed WP plans. Briefly, cohorts differ widely in their geographical coverage, ranging from entire countries (UK, France), to more typically an urban area with neighbouring rural areas. For selected outcomes, pooled analyses will be conducted, which we will not cover here. Study-specific analyses with subsequent pooling of the regression coefficients applying methods for meta-analyses will be the basis of analysis. This will allow using optimized city-specific confounder models and assessing regional heterogeneity across Europe.

In the design of the study we decided not to make use of exposure contrasts between cohorts in different countries in the primary analysis, based upon experiences in other European studies such as the ECRHS. It has been suggested to revisit this decision, e.g. by the External Advisory Committee. Christian Schindler has suggested to make use of exposure contrasts in the assessment of the shape of the concentration response function.

The analysis will include several phases:

1. Main phase (all a priori selected exposures and endpoints)
2. Additional analyses (selected exposures, endpoints):
    - Sensitivity analysis (e.g. missing confounder)
    - Analysis of the impact of included participants versus those lost to follow-up, moving during follow up and possibly other potential biases
    - Concentration response analysis (a priori cutpoints, spline models).
    - Effect modification (e.g. men vs women, use of medication)

1.1 Statistical Software

Code for statistical analyses for cohort-specific analyses were developed in SAS and Stata. Concentration-response analyses using splines for visualizing the functional form of the association between air pollution and health outcomes was developed in R statistical package and STATA (?). Code for meta-analyses will be provided for R and Stata.

1.2 Endpoints

ESCAPE has been designed to study the association between exposure to air pollution and four ‘effect’ categories:

- Adverse pregnancy outcomes and outcomes available from birth cohort studies (a.o. length of gestation, birth weight, birth length, various cognitive function test scores, lung function, asthma symptoms, sensitization to major allergens) included in WP3.
- Respiratory biomarker and morbidity endpoints in adults (prevalence of chronic respiratory symptoms, prevalence / incidence COPD, lung function, incidence of asthma) included in WP4.
- Cardiovascular biomarker and morbidity endpoints in adults (inflammation markers, markers of atherosclerosis such as IMT, blood pressure and hypertension, incident coronary events) included in WP5.
- Non-accidental all-cause and cause-specific (cardiovascular, respiratory) mortality, and cancer (lung, stomach, and brain) incidence included in WP6.

There are continuous and binary outcomes and time to event outcomes, requiring different link functions. All data are individual-level data from cohort studies. In several cohorts, measurements are available at multiple time points. Endpoints will be/have been defined in more detail within the specific WPs. As discussed in the plenary June 2010 meeting, prior hypothesis need to be provided to support the choice of performing specific analysis (e.g. in analyzing specific histological types of lung cancer or specific cardiovascular causes of mortality in addition to the overall set).

1.3 Exposures

For all endpoints we will use (where available) the annual average concentration at the residential address of the following components in the first phase of analyses (with between brackets the variable names as used in the dataset)

- $PM_{2.5}$ (PM25)

- PM_{10} (PM10)
- PM_{coarse} (PM_{10}-$PM_{2.5}$) (PMcoarse)
- Absorption coefficient $PM_{2.5}$ (PM25abs)
- NO_x (NOx)
- NO_2 (NO2)
- Traffic intensity of the nearest street (TRAFNEAR) combined with background NO_2 (NO2_BG)
- Total traffic on all major roads in a 100m buffer (TRAFMAJORLOAD_100) combined with background NO_2 (NO2_BG)

The main analysis will be conducted with the estimates directly taken from the ESCAPE LUR model. A sensitivity analysis will be conducted, taking into account changes in time using routine monitoring data (back-extrapolation). In addition to using the developed models to estimate exposure, the epidemiological analysis will also take the percentage of explained variation of the models (from cross-validation) into account, as one way of dealing with measurement error.

Within the WPs, consideration to latency aspects needs to be given. There will be limited possibilities to actually assign different exposure estimates for different time periods. Moving may however give rise to different estimates, and will be considered as a sensitivity analyses.

In a second phase, we will analyze the elemental composition data. This will be conducted after the summer of 2012. In ESCAPE we will develop LUR models for selected elements.

Each cohort should ensure that the air pollution exposure data contain the following additional info:

a) area-level SES confounding variables (e.g. % low-income in an geographical area);

b) definition variables of which neighbourhood/census tracts that the participant belong to.

1.4 Model specification

Both individual-level and area-level confounders will be used and will be selected and defined a priori. Therefore it is crucial that confounder data is standardised across all cohorts/studies. We will consider three levels of adjustment:

1. Model 1 - Crude Model: Crude estimates, which means adjusted only for sex and age, provided that age is not the time axis in the Cox model. In that case we should adjust for calendar time instead. Some groups may prefer crude model not to include adjustment for sex or age. Paper leads will make a final decision of what crude model is, and accordingly, number of total models may be 5, and not 4.

2. Model 2 - Minimum common adjusted model. A common set of potential confounders available in most studies in standardized format, *excluding* variables that could be on the pathway from air pollution to health outcome (e.g. blood pressure in the analysis of air pollution and atherosclerosis). Model 2 is defined a-priori based upon causal models, taking evidence from previous studies and the assessment and quality of available data within the ESCAPE cohorts into account. The use of DAGs could be useful here. If covariates included in the minimal adjustment are measured with acceptable precision, the minimal adjustment set can be used for the main analysis; otherwise a non-minimal adjustment set will be employed. Other DAG-based or "conventional" adjustment sets will be used in sensitivity analysis. If a potentially important confounder is lacking in e.g. one cohort (e.g. passive smoking), we would prefer to include this variable in the set of confounders for the other cohorts, in order to adjust as fully as possible. The exact definition of the set of confounders is a task of the separate WPs and specific aim leaders. Work is already ongoing to prepare a codebook to define the variables in the WPs. We decided that within ESCAPE we should not embark on building confounder models, by e.g. evaluating change in air pollution effect estimates, given the number of potential models we need to evaluate, the number of cohorts, and the observation that some of the cohort-analyses will be conducted by analysts whose primary interest is not in air pollution research and who thus need clear instructions (preferably as code).

3. Model 3 - Minimum common adjusted model with area-level confounders. Model 3 will expand the Model 2 to include area-level confounder and take account of potential clustering of individuals within areas. Studies have shown that both

individual (education, income, etc.) and area-level socioeconomic characteristics affect morbidity and mortality, and have therefore used area-level characteristics (such as socioeconomic status at the neighborhood, or even at larger spatial scales, such as income at municipality level) in addition to individual-level confounders. This is necessary in ESCAPE to evaluate effect of socioeconomic status at area-level in main cohort analyses. If information on socio-economic status at area level is not available for a cohort, indicator for area level (neighborhood, municipality, etc.), will be used alone for an evaluation of an effect of spatial clustering in that cohort. As standard regression models assume independence between observations, inclusion of area-level confounders and assigning the same value of income/socio-economic level to several individuals, will require extension of standard regression model, to accommodate the spatial clustering and correlation (non-independence) between persons within the same cluster. Clustering will be taken account of by introducing a random intercept – for this simply a variable indentifying the different areas is required (Model 3a). In addition, we will use area-level variables for socioeconomic status to account for confounding by these area-level factors (Model 3b). If no such socioeconomic variables are available Model 3a will be run. Random effects models and mixed models for linear/logistic regression models and frailty models for Cox proportional hazards models were proposed. Basically, in Model 3 tools will be provided for the evaluation of the spatial heterogeneity in each cohort. For example, in Cox regression, the spatial correlation of individuals at neighborhood levels may be evaluated as 'shared-frailty,' in which case Frailty Cox model for time-to-event data will be used. The equivalent linear and logistic models are random intercept models in mixed models.

The first step is to take account of clustering by introducing random intercepts (Model 3a). Then neighbourhood variables will be introduced (Model 3b).

Depending on the amount of clustering detected, as well as on the availability of appropriate data, the Health WP will decide whether models 3a (or b) or 2 will be used as the main model.

Generally it is recommended to present also a meta-analysis of model 3. See section 1.4.1 for more guidelines.

4. Model 4 - Maximum adjusted model or 'Best' model, which will be an extension of Model 2 (standard Cox and Linear/Logistic regression) or Model 3 (Frailty Cox or Mixed Linear/Logistic regression), depending on the spatial heterogeneity in a cohort.

Model 4 will adjust for an extended set of potential confounders, such as variables that could also be on the pathway from air pollution to health (e.g. blood presssure). Model 4 is specified for additional control of confounding and to increase power if such a variable, say X, is a strong predictor of the outcome (e.g. blood pressure and atherosclerosis). If air pollution associations in Models 2 or 3 and 4 differ widely, one might use the residual of X after regression on the other covariates of the model (including air pollution) instead of the variable itself. Then a potential indirect effect of air pollution on the outcome of interest mediated by X would be attributed to the air pollution variable and not to the X. The rationale for including these strong predictors despite the potential for overadjustment is that quite likely only a small fraction of its impact includes the air pollution pathway.

Models 2 and 4 will be defined by the specific aim leader after discussion in the respective WP. Study-specific deviations from these models (in case of missing covariates or inadequate quality of the data) will be reported by the study representative/analyst to the specific aim leader.

As default, results of the air pollution associations from Model 2 or Model 4 will be used in the meta-analysis. Some additional analyses are necessary to document the impact of missing specific information for a certain confounder in certain cohorts, by comparing the associations of model 2 and a model 4 excluding this variable in those cohorts that have more complete information.

1.4.1 Modeling of Area-level confounders

Health outcomes are often clustered and individuals in the same neighborhood and in spatially close neighborhoods are generally more similar to each other, therefore leading to a spatial pattern of the health outcome. The literature has shown many cases for neighborhood effects on health outcomes which can therefore also be important confounders (e.g. Diez-Roux *et al.* 1997, Kaufman *et al.* 2003).

Concerning modeling, ignoring the correlation between individuals by performing ordinary regression leads generally to higher probabilities of type I errors manifest by erroneously large confidence intervals for within-cluster covariates (Rabe-Hesketh, p.129) i.e. in our case exposure. This correlation can only be ignored if the included ecologic variable can account fully for the correlation (Guo and Zhao, 2000) and therefore leads to independent residuals. However, in most cases, this is an unlikely scenario. One solution is to include indicator variables for each neighborhood which however leads to numerous additional variables and may cause trouble with small cells and convergence of the models. It also does not allow appropriate modeling of neighborhood level variables.

US studies such as that by Burnett et al. 2001, incorporated the factors that lead to the spatial pattern or model the spatial patterns between neighborhoods in an ecological 2nd stage model. However, this approach is not appropriate for the ESCAPE-study: in the US study (the ACS), air pollution has been measured on the ecological level (in ESCAPE on the individual level) and spatial patterns are also modeled on this level i.e. the correlations between county death rates are modeled. Furthermore this method necessitates a number of decisions to be made during the spatial modeling process – this seems to be beyond the scope given that these analyses would need to be performed by the individual cohorts and cannot be standardized easily by providing a simple macro.

However, an appropriate approach to incorporate this neighborhood clustering and effects that is more feasible is modeling the individual outcome and taking account of the correlation between individuals by incorporating a random effect for the neighborhood level (i.e. classical multilevel or hierarchical models). This has also the advantage to model the hierarchical structure of the data taking into account explicitly which level the respective explanatory variable belongs to.

Random effects models generally estimate conditional models, i.e. they estimate the mean effect within one cluster and not the marginal effect over the whole population, i.e. averaging over all neighborhoods, that neglects the correlation within neighborhoods. Both levels of effect may be of interest to EU and local policy makers, respectively.

For linear regression the random intercept is equivalent to an exchangeable correlation structure (Vittinghoff). Logistic and Survival random effect models by default estimate conditional models. However, marginal models can also be derived (Rabe-Hesketh).

Standardized macros for doing multilevel analysis will be provided by the statistical working group.

The local exposure group responsible for exposure modeling is responsible for obtaining the data. See appendix 3 of the Exposure assessment manual, version July 2010. The definition of the variables and scale is part of the Health WP but guidance will be given below. During the plenary ESCAPE meeting of June 2010 we decided to focus on the neighborhood and metropolitan scale to represent potential area-level effects. We cannot define one variable across Europe, each center should collect the locally available variables. In case no external data are available, neighbourhood level variable can be approximated from individual data. However, care has to be taken that a sufficient number of individuals are available in the areas and that it is reasonable to assume that they constitute a random sample from the respective area with regard to socioeconomic status.

For the procedures to be adopted within the health WP, the following guidelines are given

- *To define the area identifier:*

- Wherever possible areas shall be defined as those areas for which the corresponding area level variables are available. From those you can generally also easily deduce the areas/neighbourhoods which you can then assign numbers from 1 to n
- Failing that, other variables such as zip codes or administrative units may be usedA third possibility is to define areas based on grids. The same applies if from the data available there are substantiate reasons to assume that random intercept and the area-level variable are best represented by different area levels. This is because the random intercept is not only needed to reflect the natural structure of the data, but is also especially useful to take into account residual heterogeneity in the spatial distribution of the outcomes, unexplained by individual and area level confounders.

-*To judge the importance of the random intercept based on the AIC criterion:*

if addition of the random intercept does improve the AIC (smaller AIC), then it is recommended to continue with this kind of model unless other considerations speak against it.

-*If more than one level of areas (e.g. neighbourhood, municipality, etc.) are available: which area level shall be incorporated as a random intercept?*

Use main model (generally Model 2 with all its individual covariates) WITHOUT exposure but WITH area level variables if available →test each area level and take the one that gives the smaller AIC

-*Which area level variables shall be used in case several options are available?*

- While there will be only one area-level for the random intercept, it is possible to include area-level variables from more than one level e.g. for neighbourhood and municipality.
- Supposedly many centres will not have unemployment rate – therefore take as first choice area-level income, as 2nd choice percentage of low income, as 3rd choice unemployment rate. In case neither is available take the available variable e.g. education.
- Use at most 2 area level variables for each area level, preferably one income related variable and one for education.

References:

- Burnett, R., Ma, R., Jerrett et al (2001). The Spatial Association between Community Air Pollution and Mortality: A New Method of Analyzing Correlated Geographic Cohort Data. *Environmental Health*, *109 (suppl 3*), 375-380.
- Diez Roux, A. V., Auchincloss, A. H., Franklin, T. G., Raghunathan, T., Barr, R. G., Kaufman, J., Astor, B., et al. (2008). Long-term exposure to ambient particulate matter and prevalence of subclinical atherosclerosis in the Multi-Ethnic Study of Atherosclerosis. *American journal of epidemiology*, *167*(6), 667-75. doi:10.1093/aje/kwm359
- *Guo* G, *Zhao* H. *Multilevel* modeling for binary data. Annual Review of Sociology. 2000;26:441–462.
- Kaufman, J. S., Dole, N., Savitz, D. a, & Herring, A. H. (2003). Modeling Community-level Effects on Preterm Birth. *Annals of Epidemiology*, *13*(5), 377-384. doi:10.1016/S1047-2797(02)00480-5
- Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata.* (Second Edition). College Station, TX: Stata Press.s
- Vittinghoff, E., S. C. Shiboski, D. V. Glidden, and C. E. McCulloch. 2005. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer.

1.4.2 Rural/urban gradient

In some study areas, rural as well as urban areas are included. This might lead to confounding, if the background exposure is lower in the rural areas, but outcomes are more frequent (or more severe in continuous outcomes) due to for example reduced access to medical care and less healthy lifestyle characteristics. Since this issue regards a larger spatial scale than clustering and confounding on the neighborhood level (see above), these methods will not alleviate the problem. The following steps should be taken to evaluate the problem in each cohort and to correct for possible confounding:

- Each cohort representative checks whether his or her cohort contains two areas which can be characterized as urban and rural, respectively. Since this requires intimate knowledge of the cohort and the country, this can only be done by the cohort representative.
- If a rural and an urban area can be identified within one cohort, a border must be identified to yield two distinct areas and to construct a binary indicator variable for urban versus rural living. This border can be an administrative border if available (within city limits and suburbia versus rural surroundings) or it can be defined based on distributional characteristics (for example median or first quartile of population density as appropriate). When using a distributional definition, the proportion of the resulting urban/rural samples should be in line with the geographical characteristics of the study area and the geographical distribution of the study population.
- Descriptive characteristics for the study population (exposure, outcome and covariates), dichotomized by the uban/rural indicator, will be supplied.
- The indicator variable is added in a separate step in the sensitivity analysis and results are reported to the specific aim lead.
- The specific aim lead may decide together with the cohort representative to include the urban/rural indicator variable in the main model, if they agree that relevant confounding takes place.

1.4.3 Differences in time of follow-up

Some studies have differences in the length/time of follow-up due to different recruitment periods or timing. To properly adjust for differences in follow-up time a term for follow-up time could be included in the linear or logistic regression analyses.

This could be done either by including a standard length of follow-up (e.g.10 year, 8 years etc depending on the follow-up period) or by creating a length of follow up variable using baseline minus follow up date divided by 365.25

In case the time of the interview and the time of the examination differs new variables defining the age of interview and the age of examination should be included in the analyses. This variable could be created by using the date of visit /or examination minus the date of birth divided by 365.25.

1.5 Confounders

Adjustment for covariates for each health outcome should be decided by specific aim lead a-priori, based on current literature. This will avoid too many different modeling approaches that may arise in individual cohorts, if individual cohort analyst should decide the best model according to best model fit, etc. For example, an issue of handling covariates that have been identified as potential confounders in the DAG, but which do not improve model fit differently in each cohort is best avoided by a-priori decided adjustment set (Model 2). Furthermore, the common set of covariates in individual-cohort is preferred method when pooling respective estimates into meta-analyses, to minimize the differences between cohorts. Modeling approach of non-linear relationships of the covariates, such as BMI (which may have U-shape with mortality) will be decided by specific aim leaders and may include splines, polynomials or categories, and will develop specific codes in collaboration with statistical group. Decisions about transformation of continuous variables, categorizations, etc will be taken by specific paper lead, and code developed together with stat group. It is important that same criteria are used for Model 2, to be used for meta-analyses, where some individual cohort analyses and decisions can be allowed in development of Model 4 (fully adjusted 'Best' model).

1.6 Missing values

A careful evaluation of the impact of missing values needs to be made within each cohort. Complete case analyses are recommend method for dealing with missing values in ESCAPE (see 1.5.1 bellow). Participants with missing values in covariates which are a priori chosen for Models 2 and 4 should be excluded before fitting of the

Model 1, and consequent models, to ensure comparability between the models. If the number of missing values is large, we may have to deviate from this principle. As a guideline variables should be included in the model only if there is less than 10% missing values. In case of a larger number of missing values, it may be better to exclude the variable. Paper leads will finalize the strategy for dealing with missing values in each specific aim.

1.6.1 Different methods for handling missing data in ESCAPE

In several of the cohorts, missing values may be present in a sizable number of observations. Multiple imputation is the preferred method for dealing with missing data,[1-3] but not feasible in ESCAPE, because we do not have access to all the cohort data and it is time consuming. Using a missing value indicator is attractive, as it saves observations, and is intuitive to use, by adding an extra variable to the statistical model to indicate that the value of a certain variable is missing. Complete case analyses is easy to apply as multivariate modeling in standard software packages usually exclude persons with a missing value on any of the variables in the model. However, complete case analyses imply loss of statistical power and can lead a selection bias when missing values are related to other observed subject characteristics. Literature suggests that bias may be introduced by using both the missing-value indicator and complete case analyses approaches, when data is assumed missing at random (MAR), which is a reasonable assumption for ESCAPE data, where many variables are assessed in the participants, and therefore the chance that he missing values depend on the observed data is quite high.[1] The direction and degree of bias depends on the how missing values are related to exposure and outcome and on the percentage of missing data.[2] Recent literature[2,3] discourages from using missing indicator method to handle missing data, due to bias of unpredictable direction when data is MAR, and even when data can be assumed missing completely at random (MCAR).[2] Complete case analyses method gives unbiased estimates when data are MCAR, and performs equally well as missing indicator method when data are MAR, but is easier to perform.[2] Thus, complete case analyses is recommended as a preferred method for handling missing data in ESCAPE.

1. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087-91.
2. Knol MJ, Janssen KJ, Donders AR, Egberts AC, Heerdink ER, Grobbee DE, Moons KG, Geerlings MI. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol* 2010;63:728-36.
3. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006;59:1102-9.

1.7 Data analysis

Data analysis should include several descriptive steps before the actual statistical models will be estimated. These need to be further specified by specific aim leaders, but will certainly include:

1. Description of distribution of exposures, endpoints, and confounders
2. Descriptive association between exposure and health e.g. as scatterplots

Statistical analyses are described in detail in other sections of this protocol. Statistical analyses entail application of three types of regression models for cohort cross-sectional or prospective cohort data outcomes:

1. Linear regression for continuous outcomes (WP3, WP4, WP5)
2. Logistic regression for binary outcomes (WP3, WP4, WP5)
3. Cox regression for time-to-event outcomes (WP5, WP6)

Corresponding alternative model that allow for incorporation of spatial clustering via inclusion of are-level variables (socioeconomic status, or indicator for are level) are mixed/random effects linear regression for continuous or binary outcomes (WP3, WP4, WP5) and frailty Cox Model for time-to-event outcomes (WP5, WP6).

1.8 Statistical Code

Stat group will provide a statistical code for each of three general types of data used (linear regression for continuous outcomes, logistic regression for binary outcomes, and Cox regression for time-to-event type outcomes), with general codes for 4 types of models proposed, including standard code for individual-level data (Model 1 and 2) and code incorporating adjustment for area-level variables (Model 3). Several code writing groups were developed, and code work is ongoing. Descriptive statistics code is written by Tamara in SAS and tested by Katheryna, and currently being translated into Stata. Linear/logistic regression codes are being developed by Kateryna in SAS and Tamara, Christian, Martin who tested the codes, and Gudrun. Linear regression codes in STATA are available by statistical group Swiss (M. Adam, L. Perez, Schikowski T, C. Schindler). Zorana has developed codes for Cox regression analyses in Stata, Massimo and Xavier have tested the code and will help with translating codes into SAS. Evi has developed codes for meta-analyses.

Stat group will provide codes that include model diagnostics, and set of criteria and suggestions for deciding model fit, and for dealing with model assumption violations. Stat group will provide codes which may be adapted by specific aim leaders. Paper/aim lead in can consultation with the stat working group decide which specific tests and criteria will be used for model diagnostics, model assumptions check, and model fit. Codes provided by Stat group will be providing a great detail of documented (supported by an example), will be general and flexible to allow changes by specific aim leaders.

1.9 Model diagnostic

Model diagnostics for two general types of models used in ESCAPE, generalized linear models, including linear and logistic regression, and Cox regression models are discussed separately here. Codes for model diagnostics presented here are provided in statistical codes for main cohort analyses, and intended to be applied to main/final model, Model 2 or Model 4 (depending on decision whether to include area-level confounders in the final model). Additional model diagnostics can be applied to extended models (Model 3), as decided by pacific paper leads for each aim/analyses.

1.9.1 Model Diagnostics for Linear/Logistic regression

Assumptions for linear regression will be tested in the main adjustment model. Collinearity of predictor variables will be checked with correlation matrix and variance inflation factor. Residuals' diagnostics will be performed, such as Q*Q plots or histograms of residuals' distribution, and plots of residuals against predicted values. Presence of influential values will be identified using Cook's distance or leverage value. A program to check for model assumptions will be prepared by the aim lead and sent to studies.

All regression models will be tested for the goodness-of-fit. Adjusted R^2 value will be reported for each adjustment model for linear regression. For the logistic regression we will apply the Hosmer-Lemeshow goodness-of-fit test and the Pearson's chi^2 test. A good fit as measured by the Hosmer-Lemeshow test will yield a large p-value, therefore a comparison of the p-values will be conducted. The goodness-of-fit test will be included in the logistic regression model codes.

1.9.2 Model Diagnostics for Cox regression

Cox proportional hazards model assumes that the hazard ratio is constant over time. Model check for Cox proportional hazards model is thus necessary to ensure valid interpretation of the results. Graphical evaluation of proportionality of the hazards can be executed in two ways. So called "log-log" plots plot -ln{-ln(survival)} curves for each category of a nominal or ordinal covariate versus ln(analysis time). Optionally, these estimates can be adjusted for covariates. The proportional-hazards assumption is not violated when the curves are parallel. Another way to visually check proportional hazards assumption is by plotting Kaplan–Meier observed survival curves and comparing them with the Cox predicted curves for the same variable. The closer the observed values are to the predicted, the less likely it is that the proportional-hazards assumption has been violated. Finally, the test of proportional hazards assumption is performed by test of nonzero slope in a generalized linear regression of the scaled Scheinfeld residuals on time. The null hypothesis in this test is of zero slope, which is equivalent of testing that the log hazard-ratio is constant over time. The rejection of the null hypothesis of a zero slope indicates deviation from the proportional-hazards assumption.

For variables that do not fulfill the proportional hazards assumption, stratification by that variable is recommended. Final strategy for dealing with variables that do not fulfill the proportional hazards assumption will be decided by each paper lead.

1.10 Effect-modification

For all major endpoints we will evaluate effect modification by, at least:

- gender
- education (low, medium, high)
- smoking status (current, ever, never)

Effect-modification will not necessarily be investigated for all available exposure variables. Paper leaders will make a decision of which variables will be tested for interaction with air pollution for each specific aim. Focus on the most interesting variables is efficient to reduce the number of analyses. In the case of examining interaction with continuous variables (BMI, fruit consumption, etc), categorization is recommended. It is important to a priori define within the WPs which grouping variables will be assessed, and which categories of continuous variables are relevant, to reduce number of analyses. The modification of an effect between air pollution and specific health outcomes by selected confounders will be evaluated by introducing interaction terms into the model and tested by the Wald test.

1.11 Linearity of air pollution effects

Exact definition of this work is still ongoing.

Air pollution and traffic variables will initially be evaluated as linear. Because of an interest in potential thresholds for air pollution, we will investigate the shape of the relationship using semi-parametric smoothing methods. Hence, we will apply a) a linear spline model, to investigate threshold effects and changes in slopes defined by cut off points set for example at limit values of the analyzed pollutants and /or b) a cubic spline with pre-specified knots (5 or more depending on the range of the pollutant) at common values for all participating cohorts. c) Penalized splines. d) Simple indicator variables. Common pre-specified knots are necessary in both models a) and b) for the final stage of pooling cohort-specific results. In the final stage, the cohort-specific spline-estimates will be pooled together using a multivariate meta-analysis approach in order to present a common European curve. This stage will depend on the investigation of the cohort-and outcome- specific shapes in order to assess whether such a combination into a common curve is justified.

We will also consider the use of penalized splines without pre-specified knots to allow more flexibility. One application of this approach would be to meta-analyze adjusted outcome estimates at certain fixed grid points of exposure across the different studies. This would also provide a way of visualizing effects as a function of exposure across the whole exposure range defined by the different studies. One inherent problem with this method is that the degree of heterogeneity may depend on the values to which the covariates / exposure are adjusted. This problem might be faced by applying each study specific model to predict outcomes for all subjects of the other studies as well, using individual covariate values, but varying exposure levels on a fixed grid. Then, the predicted outcomes would be averaged across all subjects for each of these exposure levels. Since models should not be applied outside their domain of definition, the subjects used in this process should belong to the common domain of definition of all models considered, i.e., have covariate values that occur in all participating studies.

For all methods we will provide a test to test statistical deviation from linearity.

Splines will be used for exposure variables only; for confounders we will a priori specify potential non-linear relationships e.g by polynomials or indicator variables.

1.12 Meta-analyses

Exact definition of this work is still ongoing.

These analyses will be conducted by the responsible authors for the defined papers, thus a smaller group of very involved people, compared to the cohort-specific analyses. The heterogeneity of the effect estimates between the studies will be assessed using X^2 test. In the absence of heterogeneity, fixed-effect models will be used to calculate the effect estimates; in the instance that heterogeneity is found however, random-effects models will be used instead. In addition, the I^2 statistic for quantifying heterogeneity will be calculated. We will carefully assess the contribution from each cohort to the overall effect estimate. In the cardiovascular and the mortality WPs, the cohort from Vorarlberg is very large and this cohort may therefore be very influential. It is also a cohort that differs in study area from many of the other more urban study areas. Hence combined estimates with and without influential cohorts will be performed.

We will report the meta-analysis graphically in Forest plots including the combined effect estimate and in tabular format making the exact quantitative estimate available for further use.

Potential explanatory variables for the observed heterogeneity between studies will be assessed and where feasible, meta-regression models will be fitted. Candidates for these cohort-level variables are:

- Region in Europe (South, west, north)
- Study area type (urban area, large region)
- Mean age of the cohort at baseline
- Prediction error of air pollution estimates
- Length of follow-up
- Pollution mixture
- Mean elemental composition of the study area

Evi Samoli has prepared the meta-analyses codes in R based on DerSimonian and Laird. One, that was developed within the framework of APHEA mainly for meta-analysis (and meta-regression) and produces fixed and random estimates (as lists) as well as the Q and I2 statistics, and an additional program which produces 2 forest plots with the individual cohort and the pooled fixed and random estimates. The first forest plot is per IQR for each cohort and per median of the IQR range for the pooled estimates as agreed, and the other per 10 ug/m^3.

It will be evaluated whether meta-analyses can be performed in Stata as well, and codes will be provided.

2.Members of ESCAPE WG statistics

Several new members joined the group in June, including Kathrin Wolf and Regina Hampel from Munich, Germany, Massimo Stafoggia and Giulia Cesaroni from Rome, Italy, and Martin Adam from Basel, Switzerland.

Name	e-mail
Gerard Hoek	g.hoek@uu.nl
Klea Katsouyanni	kkatsouy@med.uoa.gr
Kateryna Fuks	kateryna.Fuks@uk-essen.de
Kathrin Wolf	kathrin.wolf@helmholtz-muenchen.de
Regina Hampel	regina.hampel@helmholtz-muenchen.de
Martin Adam	Martin.Adam@unibas.ch
Massimo Stafoggia	stafoggia@asplazio.it
Giulia Cesaroni	cesaroni@asplazio.it
Ulrike Gehring	u.gehring@uu.nl
Christian Schindler	Christian.Schindler@unibas.ch
Barbara Hoffmann	barbara.hoffmann@uk-essen.de
Gudrun Weinmayr	gudrun.weinmayr@uni-ulm.de
Zorana Andersen	zorana@cancer.dk
Xavier Basagaña	xbasagana@creal.cat
Brian Miller	brian.miller@iom-world.org
Laura Perez	L.Perez@unibas.ch
Evi Samoli	esamoli@med.uoa.gr
Tamara Schikowski	Tamara.Schikowski@unibas.ch