Grant application for a nested project to the Swiss National Cohort:

SAPALDIA and the Swiss National Cohort Study : probabilistic record linkage with anonymized data

PI: Dr Christian Schindler – 16.6.2008

Short description of the SAPALDIA-study

The Swiss Cohort Study on Air Pollution and Lung Diseases in Adults (SAPALDIA) has been initiated in 1991 and could recruit 9651 healthy subjects from eight different areas (selected to represent the broadest possible range of exposure to air pollutants and of climatic conditions in Switzerland). The same study population was re-examined in 2002 with a response rate of more than 80 percent. At both time points respiratory and allergic status were assessed by questionnaire and clinical measurements (i.e., lung function, methacholine test, skin prick test and measurements of specific and total IgE's); data on life style and other potential determinants of respiratory health and allergies were collected by interview. In the second survey, 24-hr-Holter-ECG measurements were performed in 1813 subjects aged between 50 and 70 years. Moreover, data on cardiovascular blood markers were collected in about 6300 persons. Area specific environmental data were originally obtained from fixed monitoring stations, but individual exposure assignment for PM_{10} and NO_2 became possible in 2002 through the availability of dispersion model data for both PM_{10} and NO_2 and the development of a statistical model for NO_2 building on the respective dispersion model and additionally involving seasonal components and land use and traffic exposure characteristics of the subjects' homes.

SAPALDIA has carefully followed the address histories of its participants through regular mailings of news letters, inquiries with local registries of inhabitants and by administering a respective questionnaire at the second survey in 2002. This is both a prerequisite for a valid estimation of cumulative exposure to air pollution and for identifying cases of death in the cohort.

Today, more than 50 collaborators (experts in pneumology, cardiology, epidemiology, allergology, genetics, clinical chemistry, statistics, air pollution monitoring, and others) collaborate closely to investigate the interaction of various exposures and molecular markers in the etiology and progression of asthma, allergies, cardiovascular and respiratory symptoms and diseases.

Methods of data protection

Twofold anonymisation of SAPALDIA personal data

SAPALDIA personal and health data are anonymized in a twofold way.
The principle is that sensitive data (i.e., health data and data on personal exposure characteristics and behavior) are stored only in Basel and non-sensitive data (i.e., names, addresses and telephone numbers) only in Geneva. Moreover, sensitive data and non-sensitive data carry different ID-numbers (a public and a secret one) in

order to keep the two kinds of data strictly separated. The key linking the two types of ID-numbers is stored at the leading center in Geneva and is accessible only by the Geneva data manager. This key is stored in the address data base and there is an in-built function in the respective software enabling the exchange of the two ID's in a data file that has been docked to the address data base.

Exchange of non-sensitive data is always done using the public ID and exchange of sensitive data using the secret ID. Sensitive and non-sensitive data are never combined in the same document, with the sole exception when subjects receive results of health assessments from the Geneva center or provide new sensitive information to the Geneva center by mail. Immediately after receiving sensitive information from subjects, the Geneva data manager cuts off the part with the non-sensitive information (i.e., name, address and public ID) and labels the document with the secret ID.
In this way, it is virtually impossible for personnel processing and analyzing sensitive data to identify the underlying subject.

Securing data protection during the exchange of information with the data center of the Swiss National Cohort Study

For the planned record linkage with the Swiss National Cohort Study demographic data including address history will be provided using the public ID by the Geneva center. After having linked census data and data from death certificates to the SAPALDIA records, the team of the National Cohort study will delete all the variables containing non-sensitive information (i.e., sex, date of birth and addresses) from the records. The new information will then be sent back to Geneva carrying the public ID. Before sending the files to the health data center in Basel, the Geneva data manager will replace the public ID by the secret one.

First project proposal to the Swiss National Cohort Study

Update of information related to vital status and mortality of subjects of the SAPALDIA study

The validity of a cohort study can only be guaranteed if information on vital status can be updated at regular intervals for the vast majority of the subjects. To minimize loss to follow-up such updates should be done every few years. Moreover, the address history of participant needs to be followed carefully. Information on vital status has not been updated since the SAPALDIA2-study in 2002. It is thus of utmost importance that a new update can be done as soon as possible. Moreover, it is important to know the exact date and main cause of death of every deceased study participant.

The only efficient way of gaining such information is by using anonymous record linkage. This is the expertise of the Swiss National Cohort Study (SNC-Study). If the specialists of the SNC-study get an anonymized file with the exact birth dates, the

indication of gender and individual residential histories at the community level, they can search for matches of these records with records from the censuses in 1990 and 2000. This enables

a) to update vital status of study participants and to identify the dates and causes of death of deceased participants.

b) to gain additional socio-economic information on the study participants

The aim of the first project is just to achieve the objective described under b).

Second project proposal to the Swiss National Cohort Study

Are effects of air pollution on respiratory and cardiovascular outcomes modified by socio-economic factors?

Background

The SAPALDIA study has repeatedly documented effects of air pollution on respiratory and cardiovascular health outcomes [1] [2]. In the respective analyses, socio-economic influences were taken into account using the level of education and/or the professional position. However, it was never tested whether associations between health parameters and air pollution variables were modified by socio-economic factors. In recent years, several studies have documented socio-economic gradients in health effects from air pollution [3] [4] [5] [6] [7].

To conduct such analyses, SAPALDIA would need more detailed information on the socio-economic status of its subjects. Data from the 1990 and 2000 census would fill this gap. They could be provided by the Swiss National Cohort Study (SNC) using their established record linkage procedures. The availability of these data would allow SAPALDIA to address socio-economic gradients in health and potential modifications of air pollution effects by socio-economic status.

[1] Downs *et al.*. (2007) New England Journal of Medicine. 357:2338-2347
[2] Bayer-Oglesby *et al.* (2006) American Journal of Epidemiology. 164:1190-1198
[3] Charafeddine *et al.* (2008) Environmental Research. 106:81-8
[4] Oftedal *et al.* (2007) Clinical and Experimental Allergy. 37:1632-40
[5] Neidell *et al.* (2004) Journal of Health Economics. 23:1209-36
[6] Oftedal *et al.* (2008) Epidemiology 19: 129-137
[7] Beelen *et al.* (2008) Environ Health Perspect.116: 196-202

Hypotheses to be investigated

(1) Cross-sectional and longitudinal effects of air pollution on respiratory symptoms and lung function are modified by the participant's socio-economic status.

(2) Lower socio-demographic status aggravates the effect of air pollution on cardio-vascular health status.

Objectives

(1) To assess whether socio-economic factors modify cross-sectional associations of respiratory and cardiovascular parameters with average levels of air pollutants (PM_{10}, NO_2).

(2) To assess whether socio-economic factors modify associations of changes in respiratory parameters between the two surveys with variables derived from individual air pollution exposure histories between surveys.

Methods

For the probabilistic record-linkage between the census and the SAPALDIA database, data of roughly 9,500 SAPALDIA participants is provided. This data include socio-demographic characteristics (sex, date of birth, civil status) and the addresses at the two censuses in 1990 and 2000. In this way, different factors describing the socio-economic status of the participants can be established.

Census records provide additional information on many socio-demographic characteristics that do not strongly touch on privacy (i.e., information on salaries or capitals). Of particular importance for SAPALDIA is the additional information on

1. place or country of birth
2. level of education
3. profession
4. actual professional activity
5. degree and type of employment
6. actual working branch
7. community of the work place
8. commuting frequency and time spent commuting, means of transportation used for commuting
9. socio-professional category
10. civil status and date of its last change

11. number of children and years of birth of children
12. Existence of a second domicile

This information will help SAPALDIA to better characterize

a) its study subjects' socioeconomic status

and

b) their air pollution exposure situation (i.e., by being able to estimate exposure to ambient air pollution at the work place).

The complete data with socio-economic status, measured health outcomes and air pollution exposure variables is analysed in order to address the hypotheses stated above. These analyses will mainly involve mixed linear or logistic regression models.

Time frame

Preprocessing of data for the probabilistic record-linkage, the probabilistic record-linkage, quality control and integration of data into existing database:

3 Months, 1 Person, 50%

Data analysis and writing of research paper:

9 Months, 1 Person, 50%

Total: 12 Months, 1 Person, 50%

Principal Investigator

Dr. C. Schindler
senior statistician
Institute of Social and Preventive Medicine
Steinengraben 49
4051 Basel
e-mail: Christian.schindler@unibas.ch
Tel: 061 267 65 15

Tentative Budget (without potential overheads)

Salary/Social Security/family supplements: 54,700 CHF