

## International Lung Cancer Consortium: Instruction of Data Submission

This is the instruction of data submission for the Medical Condition Project

In support of the pooled analyses, we ask that you send the three files described to Dr. Eric Duell (duelle@iarc.fr)

- I. **The study questionnaire.** This is to ensure that the variables are comparable, and to facilitate data checking.
- II. **A data file** containing the following variables:
  1. ID number
  2. Case/Control status
  3. Histology code
  4. Sex
  5. Age (Date of birth, date of interview, date of diagnosis. If dates are not available, please provide the year of birth, interview and diagnosis)
  6. Ethnicity
  7. Education
  8. Height
  9. Weight
  10. Geographical location (as country, states or cities, depending on the study)

And *original variables* on the following:

11. **Detailed History of Tobacco Smoking:** including all relevant variables such as age of start, duration, intensity for different tobacco products (e.g. cigarettes, cigars, pipes) as in the questionnaires
12. **Detailed History of Passive Smoking:** including all relevant variables as in the questionnaires
13. **Detailed History of Medical History and Medication Usage:** including *all variables* related to the following items in the questionnaires
  - i. tuberculosis, lung fibrosis, chronic obstructive pulmonary disease (separately for chronic bronchitis, emphysema, asthma): e.g. ICD codes and age of diagnosis, etc
  - ii. Pneumonia: e.g. number of episodes, date of most recent diagnosis (or age at most recent diagnosis), etc
  - iii. a first primary cancer (separately for tobacco- and non tobacco-related cancers): e.g. ICD code and age of diagnosis, etc
  - iv. long-term use of aspirin and other NSAID: e.g. age or duration of usage, etc
  - v. Hay fevers, eczema and other allergy condition: e.g. age of diagnosis, etc

**14. Detailed Family History of Cancer:** including all relevant variables as in the questionnaires

We are asking for original variables in order for us to ensure that summary variables to be used in the analysis are constructed in a standard fashion. If standard calculated variables are also available (ex. Packyears), please include those variables also.

All personal identifiers should be removed, but please keep the ID number in order to facilitate queries on the data.

if some of the variables have been submitted previously for previous ILCCO pooled projects, then those variables do not need to be submitted again

**File format:** A SAS or STATA file is preferable, but Excel or ASCII file formats are also acceptable if they are more convenient for the sender. Please contact us (hung@iarc.fr) if you'd like to send in another format.

**III. A data dictionary listing:**

- a. the names of each variables being sent  
e.g., tob\_py, fam\_his, mthfr\_677, etc
- b. a brief description of each variable  
e.g. number of cigarette smoked per day;  
e.g. relatives who had cancer ; the cancer site of the affected relatives;  
e.g. XRCC1 R399Q (for genetic variables please specify the dbSNP no. whenever possible. eg. rs25487), etc
- c. definitions of data codes used for the variables  
e.g. 0=never, 1=ever  
e.g. 1= father; 2=mother; 3=brother; 4=sister; 5=son; 6=daughter  
e.g. 0=R/R, 1=R/Q, 2=Q/Q (for genotype data)

Also, if there are any specific issues associated with the use of particular variables, it would be helpful to include comments describing these issues. (This is optional.)

If the information requested in (b) and (c) is included as part of the data file and there are no issues with the data that warrant mention, then it is not necessary to send the data dictionary.