A SAS® PROGRAM FOR SELECTING SUBJECTS FOR A CASE-CONTROL STUDY

Terry L. Leet, Monsanto Company
Gerald A. Kientzel, Monsanto Company

Summary

The Case-Control Selection Program (CCSP) was developed to assist occupational epidemiologists in selecting subjects for case-control studies. This SAS program identifies the cases for the study and matches each case with R controls, based on certain specifications. The controls can be matched on sex, race, birth date, hire date, and/or length of employment, and can be selected with or without replacement from a pool of eligible controls. The controls can also be restricted to those employees who have not developed other exposure-related diseases. Demographic and disease-related information for the cases and controls selected for the study is stored in a permanent SAS data set for further analysis, and appropriate printouts are generated.

Purpose and Description

Case-control studies are often conducted among employee populations to measure the health risks associated with prior occupational exposures. A hybrid study design - a case-control study nested within a cohort study - can be used to estimate the relative risk of developing disease in a cohort study by obtaining exposure data on all cases and on a randomly selected subset of controls (Kupper et al., 1975; McMichael et al., 1976). This study design has a major advantage of reducing overall costs by eliminating the need to capture exposure data for all employees who may be included in the cohort study.

Once the cohort has been identified and followed to determine the incidence of disease within the study population, the cases and controls can be selected for subsequent analysis. The cases represent those employees who have developed the disease of interest, and usually the controls represent those employees who did not develop the same disease. In many case-control studies, the controls are matched to the cases based upon a set of criteria - similar age, sex and race, and sometimes age when hired by the company and length of employment. Until recently, controls were selected without replacement from the "disease-free" employees in the cohort. However, some have argued that this and other restrictions that force the selection of controls without replacement may bias the estimates of relative risk in the study (Lubin and Gail, 1984). Therefore, the CCSP was designed to allow the epidemiologist the flexibility of selecting controls using one of two models: 1) model I - cases are eligible to be controls at any time before disease onset and the controls are selected with replacement or 2) model II - cases are not available to be controls at any time and the controls are selected without replacement. Both models require that the control live as long as the case to allow for the same amount of latency.

The CCSP offers the flexibility of selecting a specific number of controls per case based upon any combination of matching variables, which may include race, sex, age and length of employment. The controls that most closely match the cases are selected using a formula that was originally developed by the National Institute for Occupational Safety and Health (1980). The following formula generates a weighted summary score of differences between the cases and controls with regard to the matching variables:

$$Score = w_1(DOB_{ca} - DOB_{co}) + w_2(DOH_{ca} - DOH_{co}) + w_3(Service_{ca} - Service_{co})$$

where:
- $DOB_{ca}$ = date of birth for the case
- $DOB_{co}$ = date of birth for the control
- $DOH_{ca}$ = date of hire for the case
- $DOH_{co}$ = date of hire for the control
- $Service_{ca}$ = length of employment for the case
- $Service_{co}$ = length of employment for the control
- $w_1$, $w_2$, $w_3$ = weighting factor for date differences
- $w_4$, $w_5$, $w_6$ = weighting factor for length of employment differences

If the epidemiologist prefers to emphasize or exclude any of the variables in the formula, this can be done using the weighting factors. Furthermore, the program will eliminate those controls with a summary score greater than a designated maximum value.

Program

Before the CCSP can be executed, the epidemiologist must create a data set containing demographic and disease-related information for each employee in the cohort. The required fields are:
- employee name
- social security number
- race
- birth date
- hire date
- employment status
- termination date
- diagnosis date if disease incidence is the outcome, or death date if mortality is the outcome
- diagnosis if disease incidence is the outcome, or underlying cause of death if mortality is the outcome

In addition to the above data set, the epidemiologist also needs to select the appropriate...
options for the control card, which includes the following:

- **Model type**: 1 = use cases as controls, or 2 = do not use cases as controls
- **Weighting factor** for birth date differences for summary score calculation
- **Weighting factor** for hire date differences for summary score calculation
- **Weighting factor** for length of employment differences for summary score calculation
- **Maximum summary score (DAYS)** for selecting controls
- **Maximum number of controls (R)** for each case
- **Study ending date**.

Once the program accepts the control card, it uses the input to generate two case files and two cohort files. The case files contain employees with the disease of interest for the case-control study. The cohort files contain potential controls. A case and a cohort file are used to generate summary scores and the other case and control files contain demographic data that are used later in the program. The cases are identified using the appropriate code(s) from the Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death. Some employees may be eliminated from the pool of eligible controls if they have developed diseases that may be related to the exposure of interest. Furthermore, if model type 2 is selected on the control card, the cohort files will not include the cases as potential controls.

During the initial phase, the program checks for several items. If a termination date is missing, the program assumes that the employee is still working as of the study ending date. Cases that occurred after the study ending date are modified so that the employee is considered disease-free.

Cases and controls are matched on race and sex, and a summary score is generated for each employee from the pool of eligible controls. The summary score is calculated using the formula described above, with weighting factors for birth date, hire date, and length of employment from the control card. If the summary score for the potential control is greater than the maximum summary score, then the control is dropped for this case. The POINTER feature of the SAS SET statement is used to accomplish this calculation.

Controls are then selected using the summary scores generated previously. This step uses a series of arrays to select R controls. A total of R controls with the best summary scores - the smaller the score the fewer the differences between the case and control - are selected for each case. A case/control index number is assigned to allow the program to label those controls that have been matched to each case in the study. The selected case and control score records are then merged with demographic records created in the first step. These records are stored in a permanent SAS data set for further analysis. Finally, three printouts are generated: 1) cases with matched controls, 2) cases and controls sorted by social security number, and 3) cases and controls sorted by employee name.

**Acknowledgements**

The authors would like to express appreciation to William Gaffey and Marcie Straus for assistance in developing the Case-Control Selection Program and to Julie Franklin for assistance in preparing this manuscript.

**References**


A copy of the complete program is available upon request to:

- Terry L. Leet
- Monsanto Company - 800 N. Lindbergh Boulevard
- Department of Medicine and Health Sciences
- St. Louis, Missouri 63167

©SAS is the registered trademark of SAS Institute Inc. Cary, NC, USA