Paper 061-2010

Using SAS® to Perform Individual Matching in Design of Case-Control Studies

Greg Grandits, Division of Biostatistics, University of Minnesota, Minneapolis, MN
Jacqueline Neuhaus, Division of Biostatistics, University of Minnesota, Minneapolis, MN

Abstract:

In the design of case-control studies it is often desired to match each case with one or more controls based on a set of variables and maximum differences between case and control in these variables. Case-control studies can arise within cohort studies where it is desired to compare subjects experiencing an event (cases) to subjects not experiencing the event (controls) for some risk factor not yet measured, but available upon further data extraction (e.g., blood determination from stored specimens, extraction of information from medical records, etc.). Typically, all cases are measured, but because of feasibility and cost only a subset of controls are measured, which are often matched to individual cases. Computerized methods to carry out the matching, although conceptually straightforward, are not readily available. In the Division of Biostatistics at the University of Minnesota we have written a SAS® utility program to carry out the matching. It has been used in several case-control studies where frozen sera aliquots from cases and matched controls have been selected from freezers and analytes determined. The utility program has recently been incorporated into a macro for easier and more generalized use. This paper describes the use of the macro and gives an example of its use for a study where Prostate-Specific Antigen (PSA) levels from stored sera were measured and compared between cases dying of prostate cancer up to 25 years after blood collection and matched controls.

Introduction:

A case-control study is a study in which subjects with a disease or condition (cases) are compared to a group of subjects from the same or similar population who don't have the disease or condition (controls). In this way risk factors for developing the disease or condition can be identified. Case-control studies can arise within cohort studies where it is desired to compare subjects experiencing an event (e.g., cancer) to subjects in the cohort which did not experience the event for some risk factor not yet measured, but available upon further data extraction (e.g., blood determination from stored specimens, extraction of information from medical records, etc.). Typically, all cases are measured, but because of feasibility, cost, or specimen savings only a subset of controls are measured, which are often matched to individual cases. This limitation usually has minimal impact on power.

Matched case-control studies employ matching of each case to one or more controls based on a set of factors that wish to be controlled for. Examples of matching variables are age, gender, smoking, prior disease, and clinical site. The matching factor is either matched exactly between the case and the control or within an acceptable range, depending on the factor and the level of comparability desired. In case-control studies designed within cohort studies it is also usually desired to require the control to be free of the disease for follow-up duration exceeding the case's follow-up prior to disease. Computerized methods to carry out the matching, although conceptually straightforward, are not readily available. It requires comparing each case with a pool of potential controls for the matching variables, and once a match is found, removing the found control from the list of potential controls for subsequent cases. It also needs to keep track of cases where no match is found.

In the Division of Biostatistics at the University of Minnesota we have written a SAS macro to carry out the matching. It has been used in several case-control studies where frozen sera aliquots from cases and matched controls have been selected from freezers and analytes determined.

Macro Use

To use the macro, the user creates two data sets, one data set containing each case and a second data set containing all potential controls. Each data set must contain the variables that will be used for matching. The user also supplies the list of matching variables and the maximum difference allowed between case and control for each variable. The program selects one case at a time and then searches through the control data set (randomly ordered)

for a control that satisfies the criteria for each of the matching variables. When a match is found the case and matched control data are written to a data set and the potential controls data set is updated removing the used control. If no match is found then the case is written to a separate – no-match – data set. The process is repeated for all cases (See flow diagram). A summary report of the matching process is then generated. Keys to the program are the flexibility of the DATA step and the ability to access randomly a SAS data set (point=option). The APPEND procedure is also used to accumulate the matching result data.

```
%MACRO MATCHCC (parameters)
```

Main parameters to the macro are:

casedata -	SAS data set of all cases
controldata -	SAS data set of potential controls (pool of all controls).
matchvar -	List of matching variables
matchval -	List of values giving the maximum difference allowed for the matching variables. A value of zero indicates an exact match is required.
fopvar -	an optional variable giving the follow-up time of each subject. If supplied then the control must have been followed at least as long as the case to be considered as a match.
controlspercase-	number of controls to be matched with each case
id -	patient identifier are both datasets

Example:

At screening into the Multiple Risk Factor Intervention Trail (MRFIT) blood was drawn from participants and sera stored in freezers for potential future case-control studies. Participants were men, aged 35-57 at time of screening. In 2002, approximately 25 years after specimen collection, a case-control study was undertaken to determine whether prostate-specific antigen (PSA) levels at time of screening was predictive of mortality due to prostate cancer [1]. As part of regular ongoing National Death Index (NDI) searches, 63 prostate cancer deaths were identified. Specimens for these cases and 63 matched controls were retrieved and PSA levels determined. Matching criteria were age ± 1 year and clinical site. Since age is a very strong risk factor for prostate cancer it was desired to match tightly on age at blood draw. Matching was also done on clinical center (1 of 22) to control for any differences in methodology used when taking and processing the specimens. Follow-up time was also computed for each participant. For cases this was the duration from initial screening to date of death from prostate cancer; for controls this was duration from screening to death (from causes other than prostate cancer) or date of last NDI search. Controls were required to have follow-up exceeding that of the case.

The call to the macro is as follows:

```
%matchcc (    casedata = psacases,
              controldata = psacontrols,
              matchvar = age clinic,
              matchval = 1 0,
              fopvar = fopdays,
              id = ptid,
              controlspercase = 1);
```

psacasedata	data set of 63 subjects that died of prostate cancer
controldata	data set of several thousand potential controls
matchvar	age clinic (matching on age and clinical center)
matchval	maximum age difference of 1; exact match on clinic
fopvar	fopdays (follow up time for each subject)
id	ptid (patient identifier)
controlspercase	1 (number of matched controls per cases)

The program was run with the following summaries. Matches for all 63 cases were found. A variable called setnumber is added to link the case with the matched control(s). A variable called ccstat is also defined as 1 for a case and 2 for a control.

Case Values for Cases Matched

Variable	N	Mean	Std Dev	Minimum	Maximum
setnumber	63	32.0	18.3	1.0	63.0
age	63	50.5	5.9	35.0	57.0
fopdays	63	6617.6	1603.0	2133.0	9189.0
clinic	63	11.2	6.0	2.0	22.0
ccstat	63	1.0	0.0	1.0	1.0

Control Values for Cases Matched

Variable	N	Mean	Std Dev	Minimum	Maximum
setnumber	63	32.0	18.3	1.0	63.0
Age	63	50.2	5.9	35.0	57.0
fopdays	63	8796.4	810.8	4756.0	9450.0
clinic	63	11.2	6.0	2.0	22.0
ccstat	63	2.0	0.0	2.0	2.0

Differences in Matching Variables Between Case and Control for Cases Matched

Variable	N	Mean	Std Dev	Minimum	Maximum
D_clinic	63	0	0	0	0
D_age	63	0.2857143	0.8314110	-1.0000000	1.0000000
D_fopdays	63	-2178.79	1513.83	-6744.00	-7.0000000

List of Matched Cases and Controls (First 10)

Obs	setnumber	PTID	AGE	fopdays	nclinic	ccstat
1	1	B049403	46	4329	2	1
2	1	B190124	45	8942	2	2
3	2	B111989	52	8565	2	1
4	2	B015594	51	9352	2	2
5	3	B150888	50	8894	2	1
6	3	B003897	49	9450	2	2
7	4	B197830	52	4107	2	1
8	4	B185363	53	4756	2	2
9	5	C050732	57	5801	3	1
10	5	C038091	56	8612	3	2
11	6	C121368	55	8025	3	1
12	6	C113191	55	9145	3	2
13	7	D022111	56	7684	4	1
14	7	D103895	55	8637	4	2
15	8	D076174	56	5668	4	1
16	8	D050856	57	8008	4	2
17	9	D112615	51	5429	4	1
18	9	D006197	51	9359	4	2
19	10	E000893	56	5707	5	1
20	10	E146274	57	8891	5	2

Discussion:

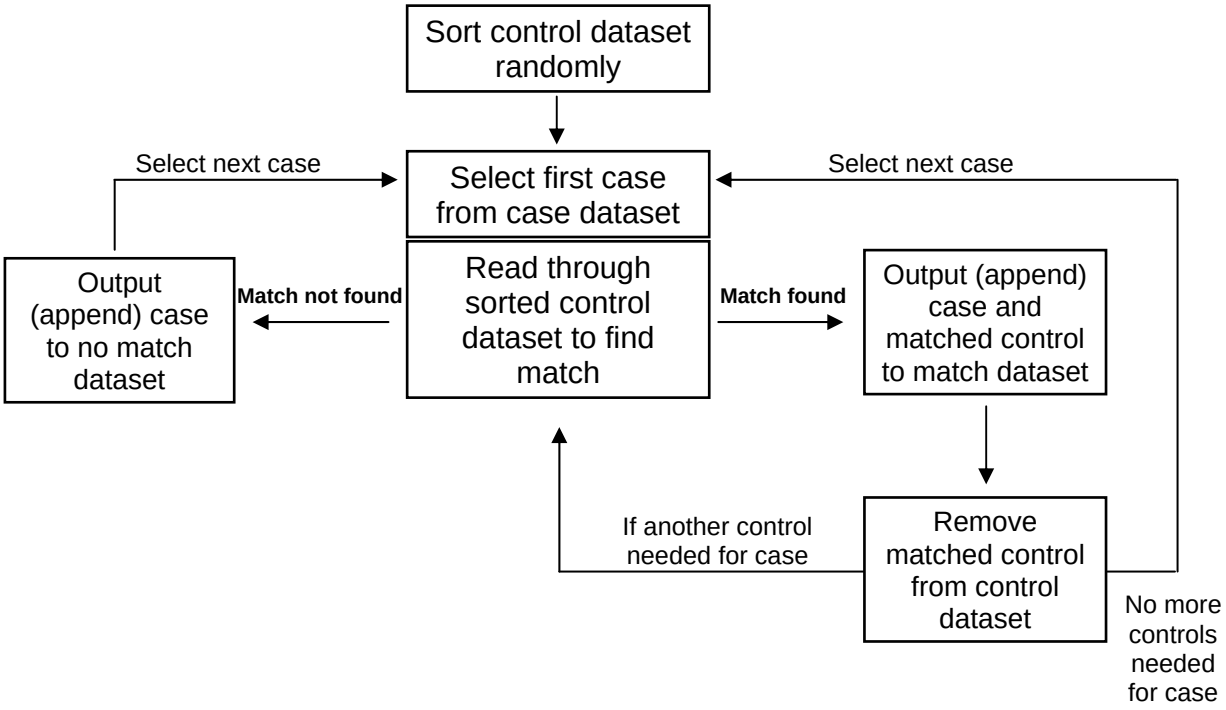
This program has been frequently used in the Division of Biostatistics at the University of Minnesota for case-control studies arising out of prospective studies involving analyses of stored blood specimens. It has also been used in case-control studies related to Type A behavior pattern assessments done at entry into MRFIT, where audio tapes of recorded structured interviews were audited and assessed for voice characteristics thought to be related to risk of CVD [2]. It was not feasible to listen to all the audio cassettes so a matched case-control design was employed.

This macro can be easily used by setting up companion data sets for the cases and potential controls. The report generated can yield information as to whether the matching was successful and to what level. If too few of the cases find a matched control then the program can be rerun changing the matching variables and/or the matching criteria. Keys to the program are the utility of the DATA step and the point= option on the SET statement that allows random access to SAS data sets. The control data set is ordered randomly so the first control that matches the case is a random match. While the program requires multiple reads of the control data set, it is only read through the occurrence of the first match. The program runs quickly for even a moderate number of cases and a large set of controls. In the MRFIT example there were 63 cases and a pool of over 10 thousand potential controls. The program took just 10 CPU seconds to run on a UNIX server. The program assumes that each variable is matched based on the absolute difference between case and control values. The macro code can be easily modified if a more complicated matching algorithm is required.

References:

1. Kuller LH. Thomas A, Grandits G, Neaton JD. MRFIT Research group. Elevated prostate-specific antigen levels up to 25 years prior to death from prostate cancer. Cancer Epidemiol Biomarkers Prev 2004; 13:373-7
2. Scherwitz L, Graham LE, Grandits G, Billings J. Speech characteristics and coronary heart disease incidence in the multiple risk factor intervention trail. J Behav Med 1990: 13:75-91/

Matched Case-Control Program Algorithm



Appendix (Code generated from example call of macro):

```
/******************  Code Generated From Call of Macro      *************
 Assume we have working data set named cases and controls
*********************************************************************/

* Sort control dataset by a random number;
proc sql;
create table random_controls as select *,
       ranuni(12345) as random from controls order by random;
quit;

* Rename control variable names - put c_ at beginning (code not shown);

* Select the first case- the step is repeated for all cases;
data active;
 setnumber = 1;
 set cases point=setnumber ;
 output;
 stop;
run;

* Main section of the program. Create dataset for matches, non-matches ;
data match   (keep = ptid nclinic age yrmorand fopdays setnumber ccstat)
     nomatch (keep = ptid nclinic age yrmorand
     used    (keep= c_ptid);
 set active ; * This has the current case;

* Read through control dataset with random access and pointer option;
 do i = 1 to totobs;
  set random_controls point=i nobs=totobs;

* Check if case and control data match;
  if abs (nclinic - c_nclinic) <= 0 and
     abs (age - c_age) <= 1 and
     abs (yrmorand - c_yrmorand) <= 2 then do;
* Make sure control lived as long as case if so we have a match!;
  if c_fopdays >= fopdays then do;
* We have a match!;
   ccstat= 1;
   output match; * This is the case data, aadding variable ccstat = 1;
* Store control values in variables with same name as case;
* Then output again to same dataset;
   nclinic = c_nclinic;
   age = c_age;
   yrmorand = c_yrmorand;
   fopdays = c_fopdays;
   ptid = c_ptid;
   ccstat= 2;
   output match;
* Output control data values to USED dataset;
   output used;
   stop; * Need to end DATA step since we have a match;
   end;
   end;
 end; * ends I loop;
 output nomatch; * If I loop is exhausted then no match;
 run;
```

```
 proc append data=match base=matchall;
 proc append data=nomatch base=nomatchall;

* Need to re-sort control dataset by subject id;
  proc sort data=random_controls; by c_ptid;
* Remove used control from control dataset;
  data random_controls;
   merge random_controls used (in=used); by c_ptid;
   if used ne 1;
  run;
* Need to resort control dataset by random number for next iteration;;
  proc sort data=random_controls;
   by random;
```