Paper 152-30

SAS® Programs to Select Controls for Matched Case-Control Studies

Robert Matthews and Ilene Brill
University of Alabama at Birmingham, Birmingham, AL

ABSTRACT

For epidemiologic matched case-control study designs, SAS programs were developed to match cases (persons with disease/event of interest) to controls (comparison group of persons without disease/event of interest). Matched case-control studies control for confounding by introducing stratification in the design phase of a study. Cases are individually matched to a set of controls (1:n ratio); n can vary from 1 to the desired number of controls for each case. Numbers of matched controls may vary dependent on the available controls possessing matching criteria. Selection of matched controls requires cases and controls to have similar values for important confounding variables defined as the matching criteria. Types of matching variables included within these programs are gender, race, year of birth/age at outcome. The programs illustrate nested matched case-control study design where cases and controls are chosen from larger retrospective cohort study populations. Study design elements featured are: persons eligible to be controls for >=1 case; cases eligible to be controls for another case; persons eligible to be cases multiple times. The second macro modification of the first program displays the inherent adaptability for other study populations, matching criteria, #controls:case ratio, and optional replacement/without replacement of cases and controls among matched case-control sets.

INTRODUCTION

Epidemiology research studies disease occurrence and its association with various risk factors in human populations. Population studies are the primary research tool employed to help describe and explain patterns and trends between risk factors and disease outcomes. A selected group of people is followed prospectively or retrospectively over time and information on disease outcomes and related factors are collected in these studies. Prospective studies collect information on the exposure of interest at the beginning of the follow-up period. The outcome events of prospective studies are relatively frequent occurrences (Kleinbaum, et al, 1982). Retrospective studies gather available previous exposure information for a clearly defined population and the outcome event is determined for all members of the cohort. The outcome events of retrospective studies are more likely to be rare occurrences or diseases with long latency periods. Alternatively, cross-sectional studies can be done which measure prevalence of disease at a point in time and associated exposures of interest. Case-control studies, another type of epidemiologic study design, either retrospectively identify cases and non-cases over time or use prevalent cases identified from a larger candidate population. Cases are defined as those with the disease or outcome event and the comparison group consists of the controls who are without disease or outcome event. The investigator selects the cases and controls from separate populations of available cases and non-cases (Kleinbaum, et al., 1982). An advantage of case-control studies is that they can be used to study infrequent disease and outcome events or diseases with long latency periods. The case-control study is designed to compare cases to controls with respect to a currently or previously assessed exposure variable of interest. Data from case-control studies are usually displayed in a standard 2x2 table where each of the four cells represents the frequency of either cases or controls classified according to the presence or absence of an exposure. The table is illustrated as follows:

	Exposed	Unexposed
Cases	a	b
Controls	c	d

The SAS programs described within this paper were both written for nested matched case-control studies where both cases and controls were sampled from larger retrospective study cohort populations.Matched study designs are often more financially feasible as the investigator decides upon the ratio of cases individually matched to a set of controls and thus the total number of subjects to be analyzed. In

numerical terms, the case:controls ratio is a 1:n ratio where n can vary from 1 to the desired # of controls for each case. Numbers of controls matched to each case may also vary within a study dependent on the available number of controls possessing the necessary matching criteria. Matched case-control studies control for confounding by introducing stratification in the design phase of a study (Breslow and Day, 1980). This design strategy prevents strata from having a large imbalance of cases and controls. The matching of cases to controls optimizes statistical efficiency when using the appropriate analytic procedures to control for confounding. In epidemiology confounding, defined as the mixing of the effect of another factor (the confounder) with that of the outcome and the exposure with both of which it is associated, is a major threat to validity. Selection of matched controls requires cases and controls to have similar values for important confounding variables defined as the matching criteria. Types of matching variables included within these programs are gender, race, and year of birth or age at outcome.

In both programs described below, incidence density sampling is used to sample controls from the larger cohort study population (Kleinbaum, et al, 1982). Each case is evaluated individually; the program selects one or more controls for each case from the set of all persons who were at risk for the outcome at the time of case identification (Checkoway, et al, 2004). Further specific details are provided in the following paragraphs.

The programs described in this paper are derived from different studies and consequently have different outcomes. The outcome event of Program #1 is death attributed to Prostate cancer. The outcome event of Program #2 is injury not work-related. In the latter program, an eligible working cohort has been subset from a larger study population. The worker did not have to be employed at the time of injury; however, the injury must have occurred during the worker's follow-up period. The beginning and end of follow-up have been previously defined for both programs. To be eligible to be a control for a case, the outcome event must always occur within the subject's follow-up period. In other words, the end of follow-up for the control will never precede the case's date of outcome event.

Additional study design elements featured in both of these example programs are: (1) persons eligible to be controls for more than one case and (2) cases eligible to be controls for another case. Only the second example program has the feature: (3) persons who are eligible to be cases multiple times. Death is the outcome in program #1 and thus only occurs once; however, cases are eligible to be a control for another case if their follow-up period extends after another case's death date. In program #2 non-work related injuries can occur multiple times within the same individual. Also, immediately following a non-work related injury the individual can be at risk for another injury; consequently, the person is easily eligible to be a control for another case. This occurs as long as the potential control possesses the relevant matching criteria and the case's date of injury falls within the potential control's follow-up period.

The second SAS example program modified several features of the first program. The study population differed, matching was now on age at outcome event rather than year of birth, and # of controls desired changed from 4 to 5. In addition, the second program was enveloped within a macro to enable looping through cases one at a time and comparing each case to the potential control set to match on age at outcome unique for each case. This necessitated recalculating age of the potential control at each case's outcome event.

The first example program can be modified very easily to conduct a search within a given study population to find potential duplicate records by matching on the study identification number. For example, we routinely use this method to search for duplicate subjects based on a 7 or 8 digit match on social security number. Each case is compared to all potential controls to ascertain possible duplicate SSNs masked by 1 or 2 incorrectly entered digits. We create a variable called MATCHSUM that can be used to determine the number of matching digits between two SSNs.

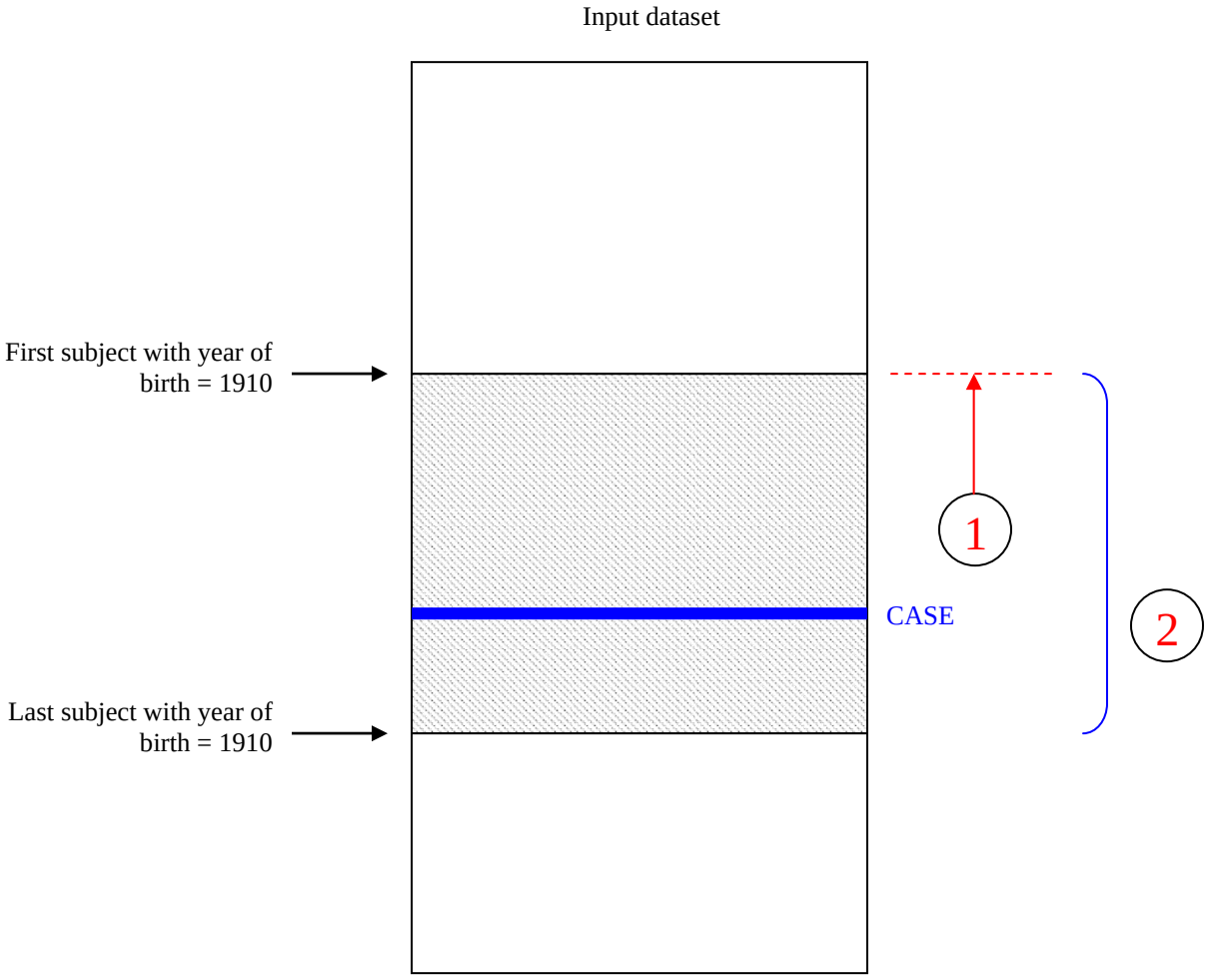
```
Matchsum = sum(substr(caseid2, 1, 1) = substr(ssn, 1, 1),  substr(caseid2, 2, 1) = substr(ssn, 2, 1),
               substr(caseid2, 3, 1) = substr(ssn, 3, 1),  substr(caseid2, 4, 1) = substr(ssn, 4, 1),
               substr(caseid2, 5, 1) = substr(ssn, 5, 1),  substr(caseid2, 6, 1) = substr(ssn, 6, 1),
               substr(caseid2, 7, 1) = substr(ssn, 7, 1),  substr(caseid2, 8, 1) = substr(ssn, 8, 1),
               substr(caseid2, 9, 1) = substr(ssn, 9, 1));
```

Below are more detailed descriptions and illustrations of Programs #1 and #2.

PROGRAM #1 – PROSTATE CANCER CASES AND MATCHED CONTROLS

Example showing how the program works for one case

- Case's date of birth = September 4, 1910
- All cases and potential risk set members are combined into a single dataset, which is then sorted by year of birth. Cases are identified by an appropriate variable, e.g. CASE=1



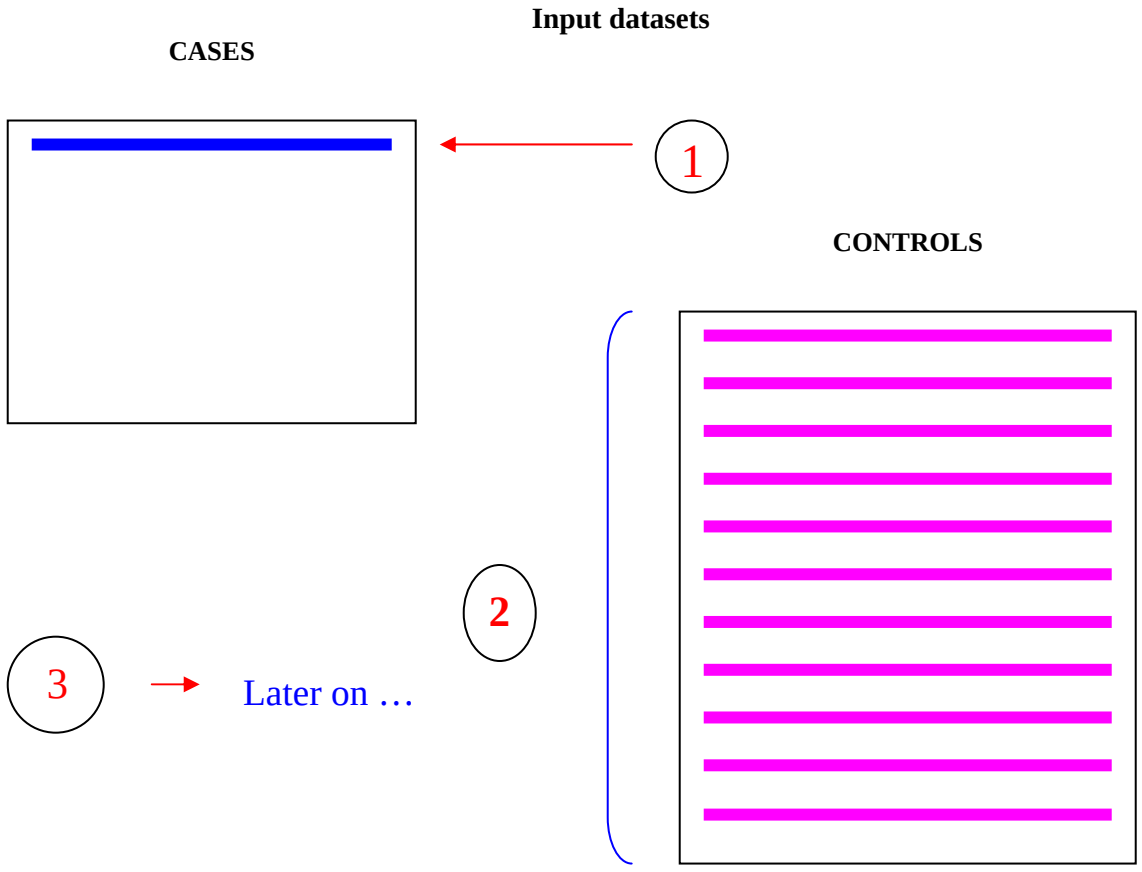
1. Back up through the input dataset to locate the first occurrence of the case's year of birth. This is accomplished by using the POINT= option on the SET statement to access dataset observations in a random fashion as opposed to the normal sequential access methods.

2. Iterate over all the observations that are equal to the case's year of birth. Depending on the size of the input dataset and the total number of cases, the absolute value function can be used to create a larger window of potential risk set members by allowing the year of birth to be different by a set amount, e.g. +/- 2 years. The final selection step is to apply any other matching criteria, such as race and gender, to each potential risk set member and then to select all matches for the resulting risk set for this particular case.

PROGRAM #2 – INJURY NOT WORK-RELATED CASES AND MATCHED CONTROLS

Example showing how the program works for one case

- Create case data set: injury is not work-related and ICD9 code 800-999 and injury date is non-missing
- Merge case data set to demographic data set by identification variables
- Retains only cases for whom the case's injury date occurs during case's previously defined follow-up period
- Calculate case's age at outcome: (injury date – birth date) / 365.25 (SAS date variables)
- Each match set's identification number - SETID - is set to the case's observation number. Cases in the match sets are identified as CACO=1
- Macro variable &num_obs is assigned the value of the total # of case observations



1. Beginning of macro MATCHED: The iterative loop cycles through the first observation to the last observation (&num_obs) in the case data set. Macro variable &obsno equals the value of the index macro variable &i. FIRSTOBS=&obsno and OBS=&obsno thus retrieving only one observation at a time from the case data set. The pictorial example above illustrates accessing only the 1st observation in the case data set.

2. Iterates over all the control data set observations. Calculates age of control at case's injury date. Selects and outputs all controls matching to case: age of outcome is +/-1 year, same race and gender, control ID# not equal to case ID# (not same person) and injury date occurs during control's follow-up period. Assigns a random number and CACO=0 to all controls output for each case. Controls keep variable SETID – its given value is unique for case and its match set.

3. Sorts controls chosen for each case by random number variable. Keeps only first 5 controls in data set after sorting. Appends controls to all controls data set. At end of iterative loop (end of macro MATCHED), appends all selected controls data set to all case data set.

CONCLUSION

Two example programs for choosing cases and controls for nested matched case-control studies have been described within this paper. They require effortless programming to be adapted for other similar studies. The second SAS program was a modification of the first program thus displaying the inherent flexibility and adaptability of these programs for other study populations, matching criteria, desired number of controls for each case, and optional with replacement or without replacement of cases and controls among matched case-control sets. When these programs were written it was not known to the authors that any examples of case-control matching programs were available for review. Thus we pioneered our own efforts in response to the researchers' requests. The ability of the SAS system to have multiple SET statements accessing multiple data sets or accessing the same data set in multiple locations simplified the writing of these programs. The macro facility and iterative looping in SAS also enabled the modifications necessary for the second program to retrieve data information for each case separately and thus a recalculation of the age at outcome matching criteria for controls dependent on each case's information. These programs have been extraordinarily useful to us as they have been repeatedly and easily adapted for numerous studies or different outcomes within the same study, and they have allowed us the flexibility to revise matching criteria within the same study populations.

REFERENCES

Breslow, N.E. and Day, N.E. (1980), *Statistical Methods in Cancer Research, Volume I - The Analysis of Case-Control Studies. Chapter V: Classical Methods of Analysis of Matched Data,* Lyon, France: IARC Scientific Publications, No. 32. International Agency for Research on Cancer.

Kleinbaum, D.G., Kupper L.L. and Morgenstern, H. (1982), *Epidemiologic Research, Principles and Quantitative Methods*, Belmont, CA: Lifetime Learning Publications.

Checkoway, H., Pearce, N., Kriebel, D. (2004), *Research Methods in Occupational Epidemiology, second edition*, New York, NY: Oxford University Press, Inc.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Colleen Beall, epidemiologist, for reviewing and commenting on the epidemiology-related sections of this paper. We would also like to thank Dr. Elizabeth Delzell and Dr. Fabio Barbone for initiating and directing the studies for which these programs were designed. Robert would like to thank Dr. Delzell and Ilene would like to thank Dr. Maurizio Macaluso and Dr. Susan Allen, for whom we work, for encouraging us always to write and present papers for SAS conferences, motivating us to constantly challenge ourselves with new tasks and allowing us the work-time to pursue these endeavors.

CONTACT INFORMATION

Robert Matthews
Department of Epidemiology
University of Alabama at Birmingham
Ryals School of Public Health, Room 517C
1665 University Boulevard
Birmingham, AL 35294-0022
Work Phone: (205) 934-1201
Fax: (205) 975-7058
E-mail: rsm@uab.edu

Ilene Brill
Department of Epidemiology
University of Alabama at Birmingham
Ryals School of Public Health, Room 533B
1665 University Boulevard
Birmingham, AL 35294-0022
Work Phone: (205) 934-7160
Fax: (205) 975-7058
E-mail: ibrill@ms.soph.uab.edu

APPENDIX 1

PROGRAM #1

This example is designed to select risk set members for prostate cancer cases .

```
proc sort data=in.fordpca out=studydata ; by ybirth; run;

data risksets;
   if restart < 1 then restart = 1;
   retain restart 0 ssn2 ybirth2 elig2 sexrace2 dod2 setid 0 seed 0;

   set studydata point=restart nobs=last; if _error_=1 then stop;

   if icd = 185 and ydeath < 88 then do;
        setid+1;

        * Output the observation for the case and save certain variables for determining valid
        * controls.  The SETid variable is used to link the control observations to the case.;
        caco     = 1;
        ssn2     = ssn;
        ybirth2  = ybirth;
        elig2    = eligdate;
        sexrace2 = sexrace;
        dod2     = dod;
        output;                 * Output the case;

        caco = 0;               * Reset CACO to indicate a control;
        index = _n_;

         *  Back up through the dataset until first ybirth.;
         do while (abs(ybirth2 - ybirth) <=1 and index > 1);  * Allows for control to be ;
            index = index-1;                                  +/- 1yr from case's birth year;
            set studydata point=index; if _error_=1 then stop;
         end;

         if abs(ybirth2 - ybirth) > 1 then do;
            index+1;
            set studydata point=index; if _error_=1 then stop;
         end;

         *  Read through all of the possible controls for a given case, determine which
         *  ones meet the necessary criteria, and output them.;
         do while (abs(ybirth2-ybirth) <=1);
            index+1;

            if  abs(ybirth2-ybirth) <= 1
                and sexrace = sexrace2
                and elig2 < eligdate
                and hiredate < dod2
                and ssn ^= ssn2
            then do;
                    *  The variable CHOOSE is a randomly generated number which is
                    *  used in the next DATA step to randomly pick out controls.;
                     call rannor (seed,choose);
                     output;
                     if index > last then goto out;
                 end;

            set studydata point=index; if _error_=1 then stop;
         end;
out:  end;
   restart = _n_+1;
   keep setid lname finit ssn ybirth sexrace hiredate eligdate vistat dod icd caco choose;
run;

data in.risksets;   * Create permanent dataset with n=4 controls per case;
   set risksets; by setid;
   if first.setid then count=0;
   if caco ne 1 then count+1;
   if count <= 4;
run;
```

PROGRAM #2

This example is designed to select risk set members for non-work related injury cases.

```
*** Selects cases and controls from a previously defined age-eligible worker population.

Begin process for creating injury not at work cases data set. Selects only conditions for an
injury occurring not at work and having ICD9 code 800-999. A subject can have more than 1
injury so a subject can be a case multiple times. Injury date is a SAS date value.

Condition due to injury: ICD9CODX=800-900 and injury is nonwork-related:
Selects only conditions with ACCDNWRK not equal to 1.;

libname dem v8 'G:\MEPS\data\Demographic file';
libname med v8 'G:\MEPS\data\Medical Conditions file';

data injury_nonwork;
set med.medical_conditions;
where ACCDNWRK ne 1 and
('800' <= icd9codx <= '999' or '800' <= icd9codx_2 <= '999');
run;


*** Deletes conditions with missing injury date due to missing values for injury month and/or
    year values;
data injury_nonwork;
 set injury_nonwork;
where accdentdt ne .;
run;

proc sort data=injury_nonwork; by dupersid condidx; run;

data injury_nonwork1;
  merge dem.demographic(in=ind) injury_nonwork(drop=duid pid in=ini);
    by dupersid;
if ini;
if ini and ind=0 then put dupersid=;
run;


*** In addition, the case's injury date must occur during his/her valid follow-up period.
    This is checked using the begin and end follow-up dates (first period and second period also
    when present) from the demographic file. Also, the potential case with the condition is not
    required to have a job on the date of the injury.;
data injury_nonwork2;
 set injury_nonwork1;
where startfu1 <= accdentdt <= endfu1 or
      (startfu2 ne . and endfu2 ne . and startfu2 <= accdentdt <= endfu2);

*** DOBYY are 4 digit years, calculates birth date;
if dobmm>0 and dobyy>0 then birthdate=mdy(dobmm,dobdd,dobyy);

*** Calculates age at injury date;
age_outcome = (accdentdt-birthdate)/365.25;
run;

*** Chooses 5 matched controls for each accident not at work case. Matches subjects on age at
    accident date +/- 1 year, same race and gender.  The potential controls must also have
    follow-up during the case's injury date. A case is eligible to be a control for another case
    if he/she meets the above criteria. Unique controls (sampling without replacement) are
    chosen for each case.  Cases cannot be their own controls.;

data contset;
 set dem.demographic;

*** DOBYY is a 4 digit year, calculates birth date;
if dobmm>0 and dobyy>0 then birthdate=mdy(dobmm,dobdd,dobyy);
run;

data caseset; set injury_nonwork2; run;

proc contents data=caseset out=contents noprint; run;
```

```
*** Creates macro variable for total # of cases in case data set;
%global num_obs;

data _null_;
 set contents(obs=1);
call symput('num_obs',trim(left(put(nobs,7.))));
run;

%put _global_;

*** Assigns case observation number to SETID that will be a unique number defining each
    case-control risk set.  CACO=1 for all cases.;
data caseset_temp;
 set caseset;
setid=_N_;
caco=1;
run;

*options mprint symbolgen mtrace;

*Begin macro MATCHED;
%macro matched;
%do i=1 %to &num_obs;
%let obsno=&i;


* Loops interatively through case data set selecting one case at a time, then evaluates all
  observations in control data set to select potential controls;

data contset1;
 if _N_=1 then set caseset_temp(firstobs=&obsno obs=&obsno
                                keep=dupersid condidx age_outcome race gender
                                     accdentdt setid
                                     rename=(dupersid=case_dupersid
                                             age_outcome=case_age_outcome
                                              race=case_race
                                               gender=case_gender));
   set contset;

    age_outcome = (accdentdt-birthdate)/365.25;
    if abs(case_age_outcome-age_outcome) <= 1
       and race=case_race
       and gender=case_gender
       and dupersid ne case_dupersid
       and (startfu1 <= accdentdt <= endfu1 or
         (startfu2 ne . and endfu2 ne . and startfu2 <= accdentdt <= endfu2))
    then do;
     choose=ranuni(0);
     caco=0;    *** CACO=0 for all controls;
     output contset1;   *** Output controls meeting matching criteria;
    end;
 label case_dupersid='Case: Person ID (DUID+PID)';
run;

*** Randomly chooses a maximum of 5 controls for each case and appends to a final selected
    control data set;
proc sort data=contset1; by choose; run;

data contset2(drop=choose);
 set contset1(obs=5);
run;

 %if &obsno=1 %then %do;
  data controls;
   set contset2;
  run;
 %end;
  %else %do;
   data controls;
    set controls contset2;
   run;
  %end;
%end;  *** ends loop do i=1 %to &num_obs;
%mend matched;   * End of macro MATCHED;
%matched    *** Invokes macro MATCHED;
```

```
*** Appends all cases and selected control data sets together;
data matchset;
  set caseset_temp controls;
   by setid;
run;

proc sort data=matchset; by setid descending caco; run;

*** Assigns unique IDNUM to each case and control in all risk sets;
data cacoset;
  set matchset;
by setid descending caco;
if first.setid then count=0;
if caco=1 then idnum=setid*10000;
 else if caco ne 1 then do;
   count+1;
   idnum=(setid*10000)+count;
 end;
run;

data less5(keep=setid count);
 set cacoset;
by setid;
if last.setid and count<5;
run;

proc freq data=less5;
tables count;
title2 'Cases with less than 5 controls after matching';
title3 'Frequency of # of controls';
run;

data less5cont;
  merge less5(in=inl drop=count) cacoset;
by setid;
if inl;
run;

proc print data=less5cont noobs n double;
var case_dupersid condidx dupersid caco setid idnum case_age_outcome
    age_outcome race gender accdentdt startfu1 endfu1 startfu2 endfu2;
title2 'Cases with less than 5 controls after matching';
title3 ' ';
run;

*** Creates final permanent data set with all case-control risk sets;
data anal.cacoset_nowork;
 set cacoset(drop=count);
run;
```