Frequency Matching case-control techniques: an epidemiological perspective.

Authors:

Hai Nguyen, MS
(1st and Corresponding Author)
Research Assistant
Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago
Phone: 1-312-355-4471
Email: hnguye72@uic.edu

Trang Pham, MS
Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago
tpham30@uic.edu

Garth Rauscher, PhD
Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago
garthr@uic.edu

Abstract

In many cohort and case-control studies, subjects are matched to intend to control confounding and to improve study efficiency by improving precision. An often-used approach is to check to see that the frequency distributions in each study group are alike. Being alike in the frequency distributions of key variables would provide evidence that the groups are comparable. However, there are instances where the overall distributions could be alike, but the individual cases could vary substantially. While there are no methods that can guarantee comparability, individual case matching has often been used to provide assurances that the groups are comparable.

We propose an algorithm build a macro in SAS to match control for the case given a set of matching criteria, including match exact on site, on the year, exact or 5-year interval match on age. Cases will be matched with a large number of controls. A large dataset from 2000-2017 from Metropolitan Chicago Breast Cancer Registry with more than 485000 women obtaining breast screening or diagnostic imaging (more than 15000 cases:30000 controls) was applied.

Introduction

Matching is a common technique in epidemiology to equalize the distribution of certain factor(s) among groups of individuals. It is applied in cohort and case-control studies. Nested case-control within existing cohort data can reduce cost, time and effort in data collection and analysis; furthermore, as exposure for cases and controls in a cohort are collected prior to diagnosis, causal inference can be applied without the concern of recall bias (Ernster, 1994).

Cancer registry is a fruitful source of cancer patient data for researchers to identify patterns and trend in various population groups (Izquierdo, 2000). However, as it was designed to collect specific data elements of the patients, it might not be sufficient to provide information to answer the research questions. Linking data from several sources is a way to gather information of interest for the investigators.

To answer the questions about the relationship between residential histories and risk of breast cancer subtypes, a dataset of 15000 women diagnosed with breast cancer matched with 30000 women without it from the MCBCR cohort was generated. From this set of 485000 women, validated residential histories from LexisNexis was obtained, and survival information was linked with the National Death Index.

This paper describes the algorithm to generate a nested case-control design using linked data from the Metropolitan Chicago Breast Cancer Registry (MCBCR).

Methods

Data from MCBCR included all women aged 18 – 70 who were screened and followed up to diagnosis and treatment from 2005 – 2016. Each individual had more than one screening cycles and pathologic results. Two sub-datasets were generated from the original data. From hereinafter, these two sub-datasets will be referred as case and control datasets. The MCBCR dataset has two IDs, one for study subject, which is unique for each person in the dataset; another one is exam_ID, unique for each exam date. One subject has only one study_ID and many exam_ID.

The case set had all cases diagnosed with breast cancer (confirmed by pathology) with the first diagnosis date from 2005-2016 and aged 18–70 at time of first diagnosis. Information of site of first cancer diagnosis was used to identify the matching control. The remaining individuals who were cancer-free for the whole duration of time, who was at the same site at time of diagnosis of the case served as controls. Age was matched within 5 years of the case (e.g. a case at 25 years old can be matched with any control aged 20-30).

Figure 1 illustrates the algorithm to generate the final dataset:

(1) Created the initial matched case-control dataset, in which, controls were matched to cases based on the desired characteristics: year, site, and age at first diagnosis. Noteworthily, a control subject in this dataset can serve as control for many cases. This step is completed using PROC SQL (Kawabata, 2004).
(2) Randomly selected one (or n depending on matching ratio) control for each case, given that one control can only be used once to match with one case. SAS's macro was used for this step. Due to the nature of cancer registry and screening cycles, a control individual might have multiple disease-free exam date in one year at one site to match with the case. The utmost goal of this step was to select one (or more) person(s) to serve as control(s) for a cancer patient in the final dataset. Thus, a DO LOOP was used inside this macro.

SAS 9.4 was used to perform the above processes.

Results

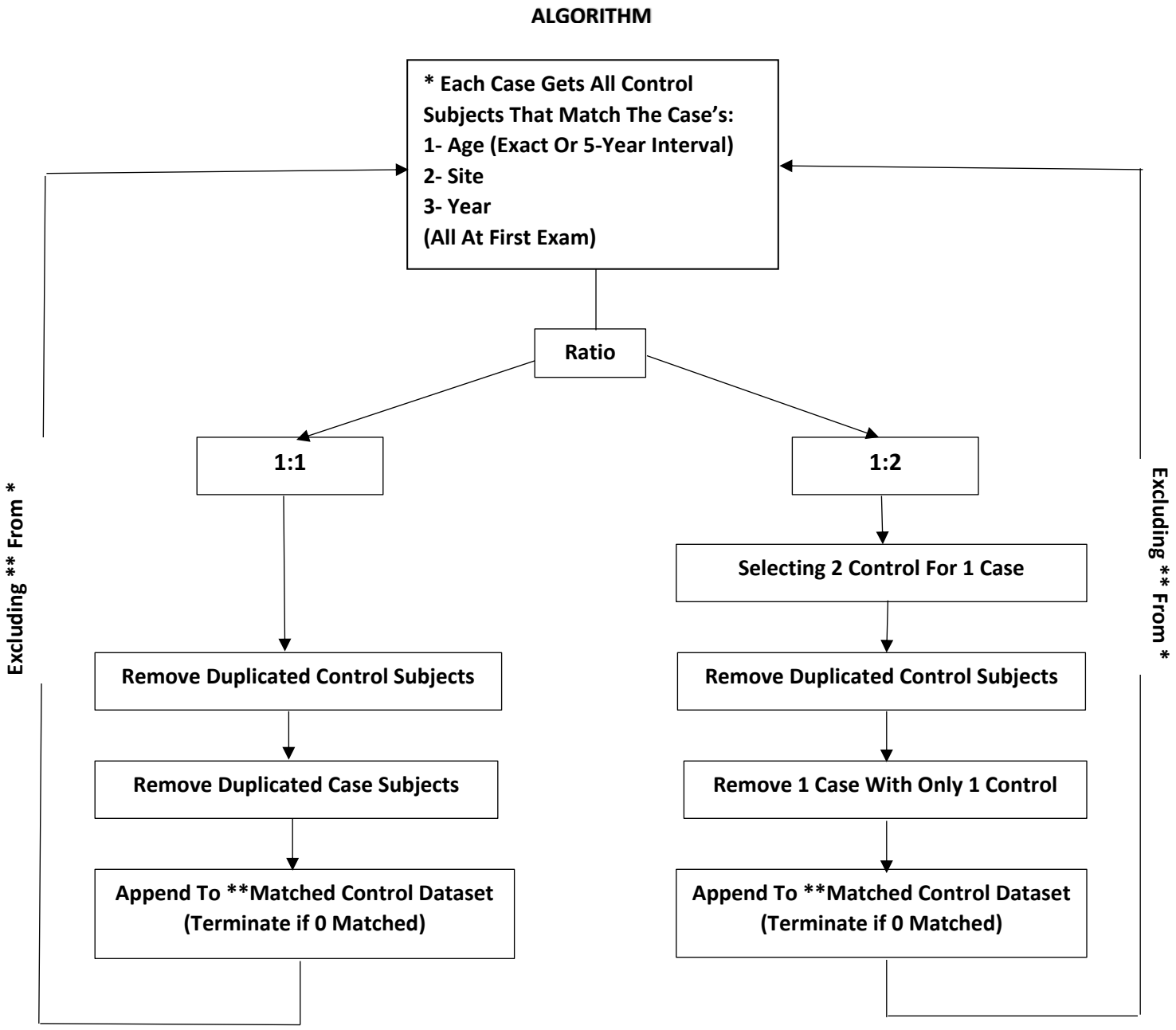

Figure 1. Working procedure to create the final dataset

From the two separated dataset of cases and controls (“case_dataset” and “control_dataset”), the initial matched dataset (controls_ID) was created using SQL procedure. The WHERE statement was used to set criteria for matching.

```
proc sql;
create table controls_ID
        as select
                one.studyid as case_id,
                two.studyid as control_id,
                one.age_first as case_age,
                two.age_first as control_age,
                one.site_first as case_site,
                two.site_first as control_site,
                one.year_first as case_year,
                two.year_first as control_year,
                1 as caco
        from case_dataset one, control_dataset two
        where (one.age_first=two.age_first and
                        one.site_first=two.site_first and
                        one.year_first=two.year_first);
quit;
```

The output dataset is “controls_ID”, including all potential controls for each case and all cases of the final dataset (1437895 observations). The “caco” variable was set to 1 for the purpose of creating a copy of case control pool dataset in the first loop running of macro.

From the “controls_ID” dataset, syntax for macro program was used to generate the final dataset. Case_ID was sort before generating macro program “create_final” based on the sorted dataset.

```
1   %macro create_final;
2   %do match_num = 1 %to 28;
3
4
5   data controls_ID_&match_num;
6       set  controls_ID;
7       if caco = &match_num;
8   run;
9
10
11  %let num = %SYSEVALF(&match_num-1);
12
13  data controls_ID_&match_num;
14      set cc_ID_&num;
15      if caco ^= &match_num;
```

```
16  run;
17
18  data controls_ID2 not_enough;
19      set controls_ID_&match_num;
20      by case_id;
21      retain ratio;
22      if first.case_id then ratio = 1;
23      if ratio le 2 then do;
24              output controls_ID2;
25              ratio = ratio + 1;
26              end;
27      if last.case_id then do;
28              if ratio le 2 then output not_enough;
29              end;
30  run;
31
32
33  data sample;
34      merge controls_ID2
35                      not_enough(in=e);
36      by case_id;
37      if e then delete;
38  run;
39
40
41
42  proc sort data=sample nodupkey;
43      by control_id;
44  run;
45
46
47
48  proc append base=work.final_dataset data=sample;
49  run;
50
51  proc sql;
52      create table cc_ID_&match_num
53              as select * from controls_ID_&match_num
54                      where case_ID not in (select case_ID from sample) and
55                              control_ID not in (select control_ID from sample);
56  quit;
57
58  %end;
59  %mend;
```

The first data step (line 5-8) in the MACRO function execute only if the macro variable match_num = 1. When this condition satisfied (as caco was set as 1 in the previous step), it created a copy of the case-control pooled dataset to serve for the first do-loop; however, from the second loop forward, SAS would not execute this data step but return an error message. After this step, the DO LOOP will continuously run for match_num > 1.

Lines 13-30 told SAS to create datasets for each match_num and num value:

- Another copy of the pooled dataset as in line 5-8 but to serve from the second loop forward (as the first data step would not be executed): For match_num=1 (the first case), there was no dataset name “cc_ID_0” to create “controls_ID_1” (match_num=1 and num=0), and SAS log returned a warning message. However, as the DO LOOP executed the whole macro program, “cc_ID_1” started to be used from match_num=2 as the result of PROC SQL at line 51-56.
- Two sets include the “controls_ID2” which were cases with exact number of controls (line 18-26) and the “not_enough” cases with less than the desired number of controls (line 27-30). The number “2”, which is the amount of control for a case, in line 23 and 28 is the number of controls to match with a case.

The DO LOOP continued to run from 1 to 28 to execute one sample dataset (line 42-44) and to stack up to the sample data to the final dataset (line 48-49).

The final dataset includes:

- Exact Match on Age, Site, and Year of First Exam (1 case: 1 control)
  - Matched 15,798 case subjects
  - Not matched 150
- Exact Match on Age, Site, and Year of First Exam (1 case: 2 control)
  - Matched 15,609
  - Not matched 339
- Matching on Five-Year Interval of Age, Exact Site and year of First Exam (1 case: 2 control)
  - Matched 15,899
  - Not matched 49

Conclusion

Matching case and control in longitudinal data is feasible using SAS Macro in addition to PROC SQL. Data in cancer registries include multiple screening and diagnosis records, which might over-represent the participation of one individual in the dataset. Given that one subject can have many repeated measurements, DO LOOP within macro should refine the random selection procedure to ascertain that one individual can serve as control for only one case. The proposed solution still needs to be improved due to the limit of selection of that lack of macro’s stop signal for loops to extend case selection

after reaching the maximum cases selected and to reduce manually selection activities after running the macro.

Reference

Ernster, V. L. (1994). Nested case-control studies. *Prev Med, 32*(5), 587-590.

Izquierdo, J. N., Schoenbach, V.J. (2000). The Potential and Limitations of Data From Population-Based State Cancer Registries. *American Journal of Public Health, 90*(5), 695-698.

Kawabata, H., Tran, M., Hines, P. (2004). *Using SAS to Match Cases for Case-Control Studies*. Paper presented at the Twenty-Ninth Annual SAS Users Group International Conference, Cary, NC.