

Programming Challenges of Sampling Controls to Cases from the Dynamic Risk Sets in Nested Case Control Studies

Victor A Kiri, PhD, Cstat, CSci^{1,2}

1 FV&JK Consulting Ltd, Guildford, United Kingdom

2 Centre of Biostatistics, University of Limerick, Limerick, Ireland

PhUSE, Budapest 14-17 October 2012



- Introduction
- Objective
- Design
- Method- SAS code
- Conclusion



- Trials are generally inadequate for determining rates of adverse effects because of low frequencies
- Pharmacoepidemiological studies on drug safety are increasingly popular because they reflect real-life utility of drugs
 - However, absence of randomization makes them liable to selection bias due to confounding
 - Confounding arises when the effect of a factor associated with both the outcome and treatment is mixed up with the actual treatment effect
- When properly conducted, these studies are capable of providing compelling data about the safety of drugs in certain settings
- Cohort studies are the most common in epidemiological research
 - But they are based on certain simplistic assumptions which may fail
- The Nested case-control design is a viable alternative in drug safety



- To highlight the main programming challenges of the nested case-control design, especially sampling from the dynamic risk sets
- Describe and demonstrate approaches for resolution and appropriate implementation based on easy-to-use SAS codes

METHOD- Background I



FV&JK CONSULTING

The Cohort vs Nested case-control designs:

- The cohort is a simple design, simple data analysis and results easy to interpret
 - We follow the entire cohort and measure the rate of occurrence of new cases to assess relationship between exposure and outcome
 - However, it may require large and lengthy study to command adequate statistical power
 - It involves exposure measurement that is based on assumptions which may not reflect real-life
 - When sophisticated models are used to handle time-varying exposure, the results may be difficult to interpret
- The nested case-control design avoid most of these problems
 - We identify the subjects who develop the event (the cases) and match each to controls sampled from all those at risk of having the event
 - Matching is done to control for potential effects of confounding by key factors in preference over covariate adjustment or modelling
 - The programmer is the most important team member!

METHOD- Background II



FV&JK CONSULTING

The Nested case-control design:

- The strength of the nested case-control design rests largely on the appropriateness of the controls matched to the cases
 - Can be quite daunting from programming perspective
 - We need to sample the controls from the dynamic risk sets and ensure the sample is representative of the at-risk population
 - But there are various methods available!
 - Incidence density sampling has been identified as the method of choice for obtaining unbiased results
 - Controls are selected without replacement from all persons at risk at the time of case occurrence, excluding the index case itself
- Incidence density sampling- Programmer requires a study cohort with all the relevant variables that define on each member, the:
 - Case status
 - Treatment exposure status
 - Time

METHOD- Background III



FV&JK CONSULTING

Incidence density Sampling: involves series of 2x2 contingency tables- each identified in SAS as a stratum

	Exposed	Unexposed
Case	a	b
Control(s)	c	d

- Each case is matched to n controls ($n \geq 1$)
- Controls can be matched to multiple cases
- A control can be a case afterwards
- Matching is usually on categorical variables such as age, gender, etc
- In general, the more variables involved, the smaller the matched study sample size
- The SAS PROC SQL offers a robust, simple approach

METHOD- SAS Codes



FV&JK CONSULTING

Based on “fullset” as cohort with pateid (unique patient identifier), tcens (follow-up time), ageindex & gender – last two for matching:

- First step- create a separate file “cases” comprising of only the cases (their “pateid” renamed as “pateid_case”).
- The INNER JOIN option ensures that
 - Any person in “fullset” is linked to all possible cases in “cases”
 - Those not matched to any case are removed
- “a.tcens <= b.tcens” ensures that the time at risk is at least as long for the control (b.pateid) as it is for the case (a.pateid_case)
- “a.pateid_case ne b.pateid” ensures that a patient does not get matched to itself
- “rand_num” to choose random controls from the possible pool
- “casecon”=1 for a case

METHOD- SAS Codes I



FV&JK CONSULTING

* The Matching procedure- matching on ageindex and gender *;

```
proc sql;  
create table join1 as select distinct  
  a.pateid_case, a.tcens, a.ageindex, a.gender,  
  b.pateid, ranuni(12345) as rand_num  
from cases a inner join fullset b  
on (a.tcens) <=b.tcens  
and a.pateid_case ne b.pateid  
and a.ageindex-1 <= b.ageindex <= a.ageindex+1  
and a.gender = b.gender  
order by pateid_case, rand_num;
```

quit;

* Macro that selects n matches for each case *;

```
%Macro match(controlsnum);  
data match(keep=pateid_case pateid);  
  set join1;  
  by pateid_case rand_num;  
  retain mcount;  
  mcount+1;  
  if first.pateid_case then mcount=1;  
  if mcount le &controlsnum;  
run;  
%mend match;  
quit;
```

METHOD- SAS Codes II



FV&JK CONSULTING

```
%match(n);          /* Input the number of controls (n) to match to each case */
* Output full dataset of matched cases and controls *;
data matched_cases_controls_all;
  set match;
run;
proc printto;
run;
*Select cases only from matched file and extract tcens, other covariates from full cohort *;
data matched_cases_controls;
  merge matched_cases_controls_all(in=a)
        pnsurv.study_all(in=b keep=pateid tcens rename=(pateid=pateid_case));
  by pateid_case;
  if a=1 and b=1;
run;
proc sort data=matched_cases_controls;
  by pateid_case;
run;
data case_id (keep=pateid pateid_case tcens); /* to make pateid_case the identifier (pateid) */
  set matched_cases_controls;
  by pateid_case;
  if first.pateid_case; /* select only the 1st obs to identify a case */
  pateid=pateid_case; /* make the case_id as the identifier (pateid) */
run;
```

METHOD- SAS Codes III



FV&JK CONSULTING

```
data id (keep= pateid pateid_case matchset tcens); /*generate id to link cases with their  
matched sets */
```

```
set case_id;
```

```
by pateid_case;
```

```
matchset=_N_;
```

```
run;
```

```
proc sort data=matched_cases_controls;
```

```
by pateid;
```

```
run;
```

```
proc sort data=case_id;
```

```
by pateid;
```

```
run;
```

```
*To obtain follow-up time and if needed, generate updates of covariates for each matched  
control *;
```

```
data matched_cases_controls1; /* first merge your cases to matched controls using control id  
and cases id */
```

```
merge matched_cases_controls case_id(keep= pateid pateid_case tcens);
```

```
by pateid pateid_case;
```

```
run;
```

METHOD- SAS Codes IV



FV&JK CONSULTING

```
proc sort data=matched_cases_controls1;
by pateid_case;
run;

data final_matched_temp;          /* Assign the matchset identifier to all records */
merge matched_cases_controls1 id(keep=pateid_case tcens matchset
rename=(matchset=statu));
by pateid_case;
run;

proc sort data= final_matched_temp;  /* reordered by original identifier */
by pateid;
run;

data pnsurv.final_matched_all(drop=pateid_case);      /* ready for analysis */
set final_matched_temp;
if pateid=pateid_case then casecon=1;
else casecon=0;
run;
```

CONCLUSIONS



FV&JK CONSULTING

- Observational studies involving the nested case-control design are now commonplace, especially in drug safety
 - most published pharmacoepidemiological safety studies are based on the design
- The increasing trend in uptake is likely to accelerate over time for a number of reasons
- The problem of time-varying treatment arguably provides the main rationale
- To minimize the effect of bias, we match cases to controls on at least, the most important confounding covariates in preference over modelling their effects
- The process involves sampling of potential controls and incidence density sampling is the most preferred- because it gives the least bias
- The programmer's role is the most vital- the main challenges depend on:
 - The size of the cohort, the extend of confounding variables and
 - Appropriate implementation of the recommended sampling approach

QUESTIONS



FV&JK CONSULTING

