## ELECTRONIC PAPER

# An incidence density sampling program for nested case-control analyses

## D B Richardson

Correspondence to:
Prof. D Richardson,
Department of
Epidemiology, School of
Public Health, University of
North Carolina, Chapel
Hill, NC 27599-8050,
USA; david_richardson@
unc.edu

Accepted 24 June 2004

**Background:** The nested case-control design can be a very efficient approach to an epidemiological investigation. In order to obtain unbiased estimates of relative risk, controls should be selected by incidence density sampling, which involves matching each case to a sample of those who are at risk at the time of case occurrence.

**Methods:** This paper presents a simple computer program for incidence density sampling. This program was evaluated using data derived from a cohort study of mortality among workers employed in the nuclear weapons industry. Controls were selected for cases via incidence density sampling; an estimate of the exposure-mortality association was obtained via conditional logistic regression. After 100 iterations of this procedure, the average effect estimate was compared to the risk estimate obtained via proportional hazards regression. The same methods were used to evaluate a program for incidence density sampling that was proposed previously by Pearce in 1989.[5]

**Results:** Relative risk estimates obtained from nested case-control analyses conducted using the incidence density sampling program reported in this paper are unbiased. In contrast, the program for incidence density sampling proposed by Pearce[5] tended to produce biased relative risk estimates; the magnitude of bias increased with increasing numbers of controls selected per case.

**Conclusions:** The computer program described in this paper offers a simple approach to incidence density sampling for nested case-control analyses with exact matching on attained age and appropriate enumeration of the pool of eligible controls for each case. This method overcomes problems of bias inherent in a previously proposed program for incidence density sampling.

The nested case-control design is an efficient approach for investigating exposure-disease associations in a study population.[1] This approach is particularly useful in studies of large cohorts, since the time and cost involved in collecting exposure and covariate information for all members of a large study cohort may be substantial. By drawing a sample of controls for each case, the number of study subjects for whom exposure information needs to be obtained is reduced. The nested case-control design is also useful as a method for computational reduction, which is achieved by drawing a sample of the eligible controls and organising the study data in a form for analysis by conditional logistic regression.

While various methods have been used for sampling controls in nested case-control analyses, incidence density sampling is the method of choice for obtaining unbiased results.[2] An efficient incidence density sampling scheme is one in which controls are selected without replacement from all persons at risk at the time of case occurrence, excluding the index case itself.[3]

Two programs for incidence density sampling have been described previously in the epidemiological literature. Beaumont et al described a system for incidence density sampling developed by the National Institute for Occupational Safety and Health (NIOSH).[4] The NIOSH system encompasses a series of programs written to operate in the IBM mainframe environment. Controls are matched to cases on attained age, with the option of additional matching on sex, race, year of birth, and duration of employment. Pearce proposed a simpler program that could be run a personal computer and easily accommodate an investigator's choice of matching variables.[5] While the program proposed by Pearce has the advantages of being easy to implement on a personal computer and flexible in accommodating an

investigator's needs, it is essentially a modified program for person-time tabulation and consequently suffers some limitations. Since an observation is created for each unit of follow up time (for example, person-year), the program is best suited to approximate matching of controls to cases on a timescale such as attained age, rather than exact matching on that timescale. Furthermore, in a tabulation of person-time, an individual contributes a single observation for each unit of follow up regardless of the number of cases that arise during that time interval; in contrast, in incidence density sampling, a person is eligible to serve as a control for multiple cases at a given moment in time. Pearce's program does not allow for this possibility.

This paper presents a simple program for incidence density sampling. This program is evaluated using empirical data for a large occupational cohort.

## METHODS

In incidence density sampling, controls are selected from among those persons under study who survived at least as long as the index case. Survival may be defined in terms of attained age, time since treatment, or some other timescale. Attained age is often the timescale of interest since disease rates typically are strongly associated with age; creating risk sets that are matched on attained age achieves perfect control for potential confounding by this factor.[6]

Table 1 describes the basic cohort data needed to select controls by incidence density sampling on attained age using the Statistical Analysis System (SAS) program shown in the Appendix.[7] The requirements are simple enough to apply to a variety of epidemiological studies in which a cohort is enumerated and followed over time to identify incident disease or mortality.

**Main messages**

- Incidence density sampling for nested case-control analyses is a useful tool for epidemiological research.
- This paper shows a problem of bias inherent in a previously published incidence density sampling program and presents a simple algorithm that leads to unbiased results, yet is flexible enough to be applied in a wide variety of study settings.

**Table 1** Required variables in the source file used by the incidence density sampling program shown in the Appendix

| Variable | Comment |
|---|---|
| study_id | Unique identification number for cohort member |
| age_entry | Age at starting date of follow up (in days)* |
| age_dlo | Age at ending date of follow up (in days)† |
| censor | Indicator of case status (1 = case of study disease, 0 = non-case) |

*The difference, in days, between the person's date of first observation and date of birth.
†The difference, in days, between the person's date of last observation and date of birth.

Four variables are required on the source file used by the program. The variable *age_entry* indicates a person's age, in days, at entry in the study. The variable *age_dlo* indicates a person's age, in days, at last observation (that is, case occurrence, censoring, or end of study). The variable *censor* is a binary indicator of case status; and the variable *study_id* is a unique numeric study identifier.

In the first two lines of the program the user specifies whether sampling of controls will occur, and, if so, the ratio of controls to cases. The variable *sampling* may be assigned a value of "yes" or "no". If "no" is specified then each index case is matched to all eligible controls. If "yes" is specified then a specified number of controls are randomly selected for each case from the pool of all eligible controls; the variable *ratio* specifies the number of controls to select per case.

Next, the number of cases in the study cohort is enumerated. The number of cases in the study cohort is equal to the number of risk sets that need to be created by incidence density sampling. For each case, the age at time of failure for the index case is identified and denoted, *age_rs*. A temporary data set is constructed of all eligible controls for the index case. The pool of eligible controls includes all cohort members (with the exception of the index case itself), whose age at entry into the study was less than or equal to the attained age of the index case and whose age at end of study was greater than or equal to the attained age of the index case. Controls may be drawn by random sampling from this pool of eligible controls to form the incidence density sampled risk set for the index case. These risk sets, indexed by the variable *rs*, are appended together to generate the final analytical data set, *final*.

Covariate information that is fixed (for example, sex, race) may be retained as additional variables associated with each observation output to a risk set. Covariate information that is time dependent (for example, cumulative exposure level) can be calculated after the risk sets are created by reference to the exact date that each person in a risk set reached the attained age of the index case (that is, *age_rs* days after the person's date of birth).

An investigator may wish to control for potential confounding by matching on one or more covariates (rather than modelling the covariate effects). Incidence density sampling with matching on a covariate (for example, sex) is easily accommodated by this program (see http://www.unc.edu/~davidr/id). Matching on a covariate when selecting controls via incidence density sampling will lead to conditional logistic regression estimates of association that approximate those obtained in a Cox proportional hazards regression analysis that is stratified on that covariate.

**Pearce's incidence density program**

The algorithm for incidence density sampling that Pearce proposed is an adaptation of a program for generating person-time data.[3][8] Pearce's incidence density sampling program generates a data set that has a unique observation for each unit of person-time; person-time may be partitioned into units of person-years or into smaller units (for example, person-months or person-days). An individual contributes observations to the pool of eligible controls through the penultimate unit of person-time. The observation for the final unit of follow up time, however, does not contribute to the pool of eligible controls (for any case); it is solely a case record.

There are several limitations to the program proposed by Pearce.[5] Firstly, regardless of whether person-time is partitioned into units of person-years or person-days, the attained age of each case is rounded down to the nearest year (the case's age on their last birthday). Consequently, risk sets are not matched exactly on age, and therefore time dependent covariates may not be correctly specified. For example, calendar year at risk (or cumulative exposure level) assigned to a control based on the birthday when they reach the age of the case is not necessarily the same value that would be assigned to that control if matching were done according to the exact age of the case.

Secondly, a case is excluded from all risk sets enumerated during the time interval spanned by their final period (for example, person-year) of follow up. Therefore, the pool of controls for a given case does not include any other case that attained that age in their final period of follow up.

Thirdly, a control is excluded from serving in more than one risk set that is enumerated at a given year of age. If multiple cases occur at a given year of age, a person who is eligible to serve as a control can be selected for only one of the cases. For example, if five cases occur at 76 years of age, then a person who is alive and under study at age 76 should be eligible to serve as a control for each of these cases. However, under Pearce's program, the person who is alive and under study at age 76 may serve as a control for only one of these cases. Since the attained age of each case is rounded downward to the nearest year (that is, the case's age on their last birthday), multiple cases may occur at the same age.

Finally, under Pearce's program, a case may be selected to serve as its own control. Again, this occurs because the attained age of each case is rounded down to the nearest year, so the case may have contributed a unit of person-time at risk to the pool of controls for the period spanning their last birthday. Consequently, if follow up time is partitioned into units smaller than a person-year then the pool of controls for a given case may include the case itself.

In order to illustrate the differences between the incidence density sampling program proposed in this paper and the program proposed previously by Pearce,[5] a hypothetical cohort that included eight people (and three cases) was constructed. Using these data, incidence density sampling was used to construct a nested case-control study with three controls per case, and a nested case-control study with four controls per case. Incidence density sampling was performed using the program shown in the Appendix and using the program proposed by Pearce (1989). The risk sets enumerated

by each approach were compared (see http://www.unc.edu/~davidr/id).

## Statistical methods

In order to illustrate the consequences of the differences between these incidence sampling algorithms for epidemiological risk estimates, nested case-control analyses were conducted using data for a cohort of 8307 white males hired from 1943 and 1972 who worked at least 30 days at Oak Ridge National Laboratory (ORNL).[9] Vital status through 1990 was ascertained through Social Security Administration, National Death Index, and employer records. At the close of follow up, 5879 (70.8%) members of the cohort were still alive, 2110 (25.4%) members had died, and 318 (3.8%) members had been lost to follow up. The exposure variable of interest in these analyses is pay code, which has been used previously as an indicator of socioeconomic status and shown to be an important predictor of mortality risk.[10] [11] For these analyses, a fixed dichotomous variable was defined by the worker's pay code at date of hire; a comparison was drawn between weekly paid workers and other workers.

Controls were selected for cases by incidence density sampling using the program presented in the Appendix of this paper. Conditional logistic regression, via SAS PROC PHREG, was used to derive an estimate of the pay code-mortality association, $\hat{\phi}_{cl}$.[7] [12] The regression model included a single parameter for pay code, with attained age exactly controlled for via stratification. The process of selecting controls and deriving an estimate of the exposure-mortality association was repeated over 100 iterations after which the average effect estimate, $E(\hat{\phi}_{cl})$, was calculated.

Cox proportional hazards regression, via SAS PROC PHREG, was used to derive an estimate of the pay code-mortality association, $\hat{\phi}_{ph}$, using data for the full cohort.[7] [12] The proportional hazards regression model also included a single parameter for pay code, with attained age as the time scale. The ratio of $E(\hat{\phi}_{cl})$ to $\hat{\phi}_{ph}$ is referred to as bias; a value of unity indicates that results obtained via conditional logistic regression analyses of nested case-control data, on average, equal the result obtained via Cox proportional hazards regression.

These analyses were repeated under sampling ratios (number of controls per case) of 10:1, 50:1, and 100:1. The same methods were used to evaluate the incidence density sampling program proposed by Pearce[5] with follow up time tabulated in units of person-years and in units of person-months.

## RESULTS

Analyses were first conducted using the incidence density sampling program shown in the Appendix (method I). The estimate of the relative risk of all cause mortality when comparing weekly paid workers to other workers was derived via conditional logistic regression analyses of nested case-control data. Ten controls were selected for each case. After 100 iterations of this procedure, the average effect estimate $((E(\hat{\phi}_{cl}) = 0.41))$ approximated the result obtained via proportional hazards regression using data for the full cohort $((\hat{\phi}_{ph} = 0.41))$.

Table 2 reports the approximate bias in the average parameter estimate derived via conditional logistic regression after 100 repetitions of nested case-control analyses with 10, 50, and 100 controls per case. A value of unity indicates an absence of bias in conditional logistic regression results (relative to the estimate obtained via proportional hazards regression). For analyses conducted using the incidence density sampling method I, values for this coefficient of bias were equal to unity under each of the sampling ratios evaluated. Figure 1A presents a histogram of the conditional

logistic regression parameter estimates obtained from 100 nested case-control analyses conducted with 10, 50, and 100 controls per case. Estimates of association are distributed relatively symmetrically around the value of the parameter estimate obtained via proportional hazards regression.

Table 2 also presents coefficients of bias for conditional logistic regression estimates obtained from nested case-control analyses with incidence density sampling performed using the program proposed by Pearce (method II). For analyses in which 10 controls were selected per case, there was little of evidence of bias in the average effect estimate obtained via conditional logistic regression (relative to the estimate obtained via proportional hazards regression). In contrast, for analyses in which 50 or 100 controls were selected for each case, there was clear evidence that the resultant parameter estimates were positively biased. Figure 1B presents a histogram of the conditional logistic regression parameter estimates obtained from 100 nested case-control analyses conducted with 10, 50, and 100 controls per case. With incidence density sampling conducted using method II, the resultant estimates of association are progressively skewed away from the value for the parameter estimate obtained via proportional hazards regression as the sampling ratio increases.

We repeated the analyses in table 2 with incidence density sampling performed using Pearce's program with person-time partitioned into units of person-months (rather than person-years). Results were similar to those reported in table 2; average coefficients of bias for conditional logistic regression parameter estimates obtained from 100 nested case-control analyses conducted with 10, 50, and 100 controls per case were 1.01, 1.09, and 1.25, respectively.

## DISCUSSION

A case-control study nested within a cohort can be a highly efficient method for epidemiological investigation. If exposure data are costly or difficult to collect, then use of a nested case-control approach may facilitate an investigation by reducing the number of subjects for whom exposure data are needed. However, even when exposure data are available for all cohort members, analysis of cohort data via the nested case-control approach offers a useful method for computational reduction when compared to Cox regression.[1] In Cox proportional hazards regression, the risk set for an incident case includes all persons at risk at the moment instantaneously prior to case occurrence (therefore including the case itself). Incidence density sampling is a method of forming a sub-sample of that risk set. The index case is matched to a specified number of controls drawn from the full risk set. The index case—and only the index case—should be excluded from the pool of eligible controls if sampling of controls is done without replacement.[3]

These considerations are directly relevant to understanding the bias observed in analyses (table 2) when incidence density sampling was performed using the program proposed by Pearce.[5] Bias arose as a result of the inappropriate exclusion of subjects from the pool of eligible controls. Under the incidence density sampling program proposed by Pearce, a case is excluded from all risk sets enumerated during the time interval spanned by their final period (for example, person-year) of follow up. Also, under the incidence density sampling program proposed by Pearce, if multiple cases arise at a given year of age, a person is excluded from serving as a control for more than one of the cases. The inappropriate enumeration of risk sets when using the incidence density sampling program proposed by Pearce can be illustrated by means of a simple example using hypothetical cohort data (see http://www.unc.edu/~davidr/id).
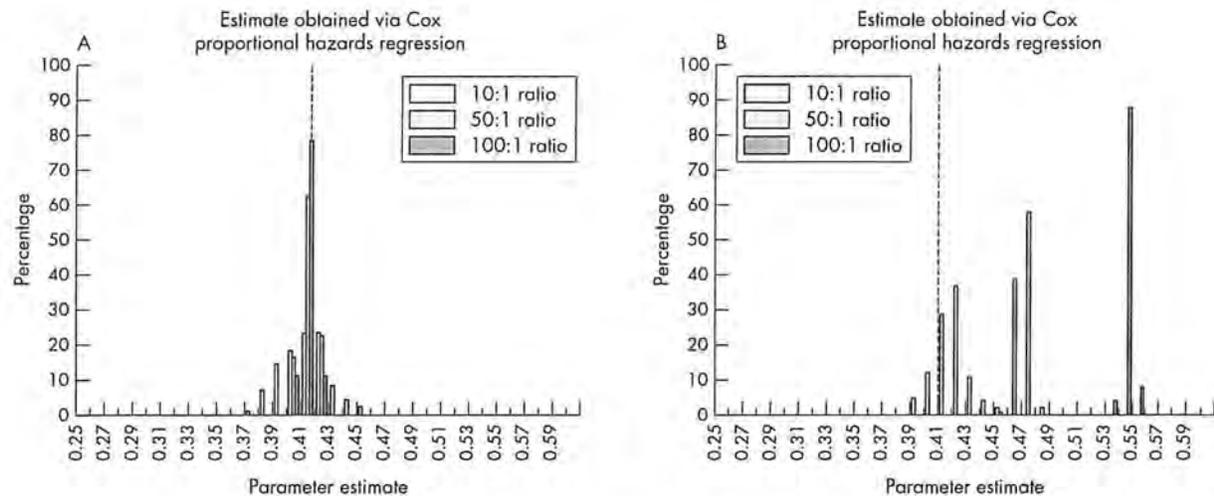
**Figure 1** Histograms of the distribution of conditional logistic regression parameters estimates for the association between weekly pay code and all cause mortality in a cohort of 8307 white male nuclear industry workers. Values obtained after 100 replications of nested case-control analyses conducted using two methods for incidence density sampling with 10, 50, and 100 controls per case. Dashed line indicates the estimate of association obtained via Cox proportional hazards regression using data for the full cohort. (A) Method I for incidence density sampling (the incidence density sampling program shown in the Appendix of this paper). (B) Method II for incidence density sampling (the incidence density sampling program proposed by Pearce[5]).

The analyses of empirical data (table 2) suggest that the degree of bias arising due to these problems with the incidence density sampling program proposed by Pearce may be minimal if the number of controls selected per case is small relative to the total size of the risk set. In situations where a nested case-control study design is used to reduce the cost and time involved in data collection, it is unusual to select more than 10 controls per case. However, incidence density sampling is also used as a means of computational reduction. Cox proportional hazards regression is a standard approach to analyses of continuous cohort data, but conditional logistic regression with incidence density sampling offers an efficient alternative to Cox regression analyses. When using the latter approach, an investigator may select a relatively large number of controls for each case (for example, 50 or more) in order to preserve the precision of risk estimates while still benefiting from the computational efficiency afforded by incidence density sampling. The analyses in this paper suggest that under these conditions,

the program proposed by Pearce may lead to highly biased risk estimates (while the program reported in the Appendix will produce unbiased risk estimates). In the example presented in table 2, the direction of bias was away from the null; however, inflation, attenuation, or reversal of exposure-disease associations would be possible as a result of the problems in the incidence density sampling program proposed by Pearce.

Given the advances in the processing speed of personal computers, exact incidence density sampling can be achieved via this relatively simple program. This approach allows an investigator to take advantage of the efficiencies associated with the nested case-control design without producing biased risk estimates.

## APPENDIX

### SAS program for incidence density sampling

```
%macro caseset;

%let sampling = 'yes';
%let ratio = 10;

** Enumerate Cases **;
data cases;
set source ;
if censor = 1;
run;

data cases;
set cases end = eof;
if eof then call symput ('ncases', put(_n_,6.));
run;

** Create Risk Set **;
%do iter = 1 %to &ncases;

data temp_case;
set cases;
```

**Table 2** Approximate bias in conditional logistic regression estimates of the association between weekly pay code and all cause mortality in a cohort of 8307 white male nuclear industry workers

| Sampling ratio (Controls:case) | Method I[†] Bias | Method II[‡] Bias |
|---|---|---|
| 10:1 | 1.00 | 1.01 |
| 50:1 | 1.00 | 1.14 |
| 100:1 | 1.00 | 1.32 |

Values obtained after 100 replications of nested case-control analyses conducted using two methods for incidence density sampling with 10, 50, and 100 controls per case.
Note the value reported as bias is the ratio of the average parameter estimate derived via conditional logistic regression (obtained after 100 iterations of the nested case-control analysis) to the parameter estimate derived via Cox proportional hazards regression using data for the full cohort.
[†]The incidence density sampling program shown in the Appendix of this paper.
[‡]The incidence density sampling program proposed by Pearce[5] with follow up time partitioned in units of person-years.

```
if _n_ = &iter ;
call symput ('rs', put(_n_,6.));
call symput ('age_rs', put(age_dlo,8.)); call symput ('case_id',
put(study_id,8.));
run;

data temp_control;
set source;
if age_entry < = &age_rs < = age_dlo;
** Exclude Index Case **;
if study_id = &case_id then delete;
number = ranuni(0);
age_rs = &age_rs;
censor = 0;
run;


**Sample Controls **;
%if &sampling = 'yes' %then %do;
proc sort data = temp_control;
by number;

data temp_control;
set temp_control;
by age_rs;
retain m;
if first.age_rs then m = 0;
m = m+1;
if m< = &ratio then output temp_control;
run;
%end; * End If Sampling = yes;

** Combine Case with Controls **;
data rs_&iter;
set temp_case
temp_control;
rs = &rs;
age_rs = &age_rs;
run;
DM "Output; Clear; Log; Clear";
```

```
%end; * End Loop Creating Risk Set;

** Append Risk Sets **;
%do j = 2 %to &ncases;
proc append base = rs_1 data = rs_&j;
run;
%end;
data final; set rs_1; run;
%mend ; ** End Macro **;

** Invoke Macro **;
%caseset;
```

## REFERENCES

1  Prentice RL. On the design of synthetic case-control studies. *Biometrics* 1986;42:301–10.
2  Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. *Biometrics* 1984;40:63–75.
3  Robins JM, Gail MH, Lubin JH. More on "Biased selection of controls for case-control analyses of cohort studies". *Biometrics* 1986;42:293–9.
4  Beaumont JJ, Steenland K, Minton A, *et al.* A computer program for incidence density sampling of controls in case-control studies nested within occupational cohort studies. *Am J Epidemiol* 1989;129:212–19.
5  Pearce N. Incidence density matching with a simple SAS computer program. *Int J Epidemiol* 1989;18:981–4.
6  Breslow NE, Day NE. *Statistical methods in cancer research: the design and analysis of cohort studies*, Vol. II. Lyon: International Agency for Research on Cancer, 1987.
7  SAS. *Version 8.01.* Cary, NC: SAS Institute, 1999.
8  Pearce N, Checkoway H. A simple computer program for generating person-time data in cohort studies involving time-related factors. *Am J Epidemiol* 1987;125:1085–91.
9  Richardson DB, Wing S. Greater sensitivity to ionizing radiation at older age: follow-up of workers at Oak Ridge National Laboratory through 1990. *Int J Epidemiol* 1999;28:428–36.
10  Wing S, Shy C, Wood JL, *et al.* Supplementary document: Mortality among workers at Oak Ridge National Laboratory: evidence of radiation effects in follow-up through 1984. National Auxillary Publication Service Document 04849. New York: Microfiche Publications, 1991:1–19.
11  Hoffmann W. *Social inequalities in mortality and findings in a large cohort of workers at four Department of Energy (DOE) facilities in Oak Ridge, TN.* University of North Carolina at Chapel Hill, 1995.
12  SAS Institute Inc. *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07.* Cary, NC: SAS Institute, Inc., 1992.