Case-cohort studies

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Design of Epidemiological Studies

The nested case-control study

- The **nested case-control study** aims to reduce the size of the cohort, by sampling a small number of controls at random, from each risk set
- This sampling scheme is referred to as **incidence density sampling**
- Incidence density sampling increases statistical uncertainty, but usually not very much

The case-cohort study

- The **case-cohort study** is an alternative design, which also aims to reduce the size of the cohort by subsampling
- Not as common as the nested case-control study, but has some important advantages

Outline

The case-cohort study

Choice of study design

Outline

The case-cohort study

Choice of study design

Intuition

- The easiest way to subsample from a cohort is to pick a sample from the cohort at baseline
- If this is done randomly, then the obtained sample is 'representative' for the whole cohort
  - e.g. unbiased estimates

Intuition, cont'd

- If the sample is small, relative to the full cohort, then the statistical uncertainty will be much bigger in the sample
  - e.g. much larger standard errors and wider confidence intervals
- When the rate of the outcome is small, the 'weak link' is usually the few number of subjects with an outcome

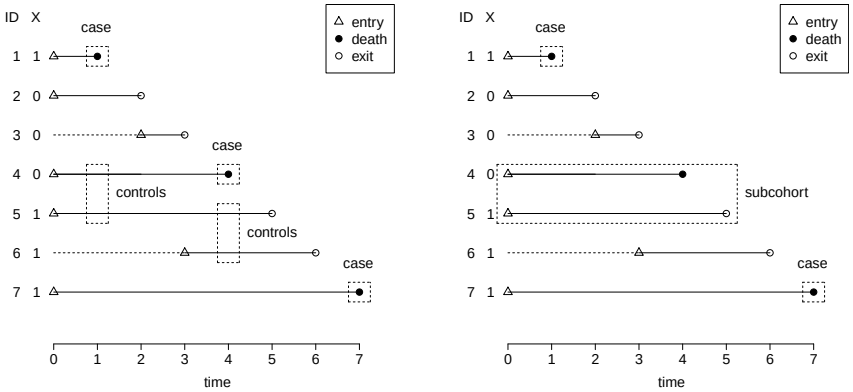
- Intuitively, we may reduce statistical uncertainty by adding all subjects who have an outcome to the sample

The case-cohort study

- In the **case-cohort study**, the sample consists of two parts
- First, a small (e.g. 10%) random sample of the cohort is selected at baseline
  - the **subcohort**
- Then, the full cohort is followed over time and all subjects who have an outcome are selected
  - the **cases**
- The subcohort also includes subjects with the outcome; they are just not called 'cases' in the case-cohort jargon
- Subjects who are not part of the subcohort, and are not cases, are discarded

Nested case-control study vs case-cohort study

- Nested case-control study
- Case-cohort study



Analysis?

- How should we analyze case-cohort data?
- Cannot use conditional logistic regression, since no matched strata
- Perhaps we can use ordinary Poisson/Cox regression?

Cohort study vs case-cohort study

- Cohort study:

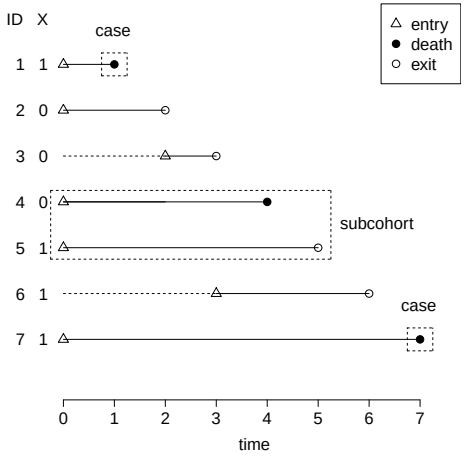
```
> fit <- coxph(formula=Surv(time, death)~
  alc.binary, data=cohort)
> summary(fit)
              coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.8790    2.4086   0.1578 5.572 2.52e-08 ***
```

- Case-cohort study:

```
> fit <- coxph(formula=Surv(time, death)~
  alc.binary, data=case.cohort)
> summary(fit)
              coef exp(coef) se(coef)     z Pr(>|z|)
alc.binary 0.5761    1.7791   0.1586 3.633 0.000281 ***
```

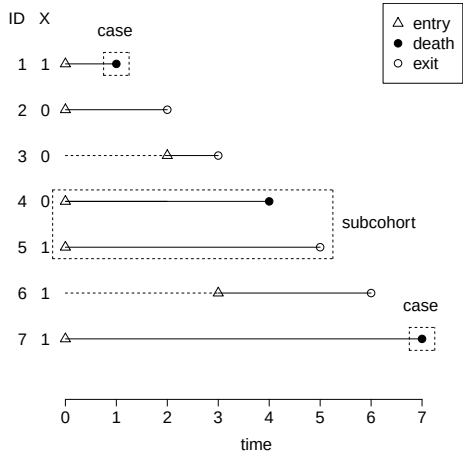
- The case-cohort estimate is much smaller than the full cohort estimate
- *Why does this happen?*

Example



- *Compare the exposure distribution among the controls in the full cohort and in the case-cohort sample, in each risk set*

Solution



Risk set	controls in cohort	controls in case-cohort
1	50% exposed	66.7% exposed
2	100% exposed	100% exposed

Bias in the case-cohort sample

- In the case-cohort sample, both subcohort and cases contribute to the risk sets
- The subcohort is taken at random, and is thus representative for the full cohort
- By definition, all cases have the outcome before end of follow-up, and are thus not representative for the full cohort
- The risk sets in the case-cohort sample are not representative for the risk sets in the full cohort
- This leads to bias

Appropriate analysis of case-cohort data

- There are two ways to avoid bias when analyzing case-cohort data
- 1) Use only subcohort in the risk sets, and use standard Poisson/Cox regression
  - Cox regression by `cch` function in `survival` package, with `method="SelfPrentice"`
- 2) Use both subcohort and cases in the risk sets, and use weighted Poisson/Cox regression to correct for bias
  - Cox regression by `cch` function in `survival` package, with `method="LinYing"`
- Often similar results

Subcohort only in risk sets

```
> fit.cch <- cch(Surv(time, death) ~ alc.binary,
  data=case.cohort,  subcoh=~subcoh, id=~id,
  cohort.size=1000, method="SelfPrentice")
> summary(fit.cch)
Coefficients:
       Coef   HR  (95%   CI) p
Value 1.022 2.78 1.688 4.576 0
> sqrt(fit.cch$var)
          [,1]
[1,] 0.2543443
```

- Similar estimate as with Cox regression in cohort study ($= 0.88$)
- Larger standard error than with Cox regression of full cohort ($= 0.1578$)

Subcohort + cases in risk sets; weighted analysis

```
> fit.cch <- cch(Surv(time, death) ~ alc.binary,
  data=case.cohort,  subcoh=~subcoh, id=~id,
  cohort.size=1000, method="LinYing")
> summary(fit.cch)
Coefficients:
       Coef    HR  (95%   CI) p
Value 1.035 2.816 1.722 4.603 0
> sqrt(fit.cch$var)
          [,1]
[1,] 0.2508102
```

- Similar estimate as with Cox regression in cohort study ($= 0.88$)
- Larger standard error than with Cox regression in cohort study ($= 0.1578$)

Outline

The case-cohort study

Choice of study design

Two strategies for subsampling

- The nested case-control study and the case-cohort study are strategies for subsampling from the full cohort
- Subsampling is useful when
  - it is difficult or expensive to collect data on the exposure and confounders
  - the full cohort is very large, and takes a long time to analyze
- How to chose between a nested case-control study and a case-cohort study?

Statistical uncertainty

- Both the nested case-control study and the case-cohort study give higher statistical uncertainty than the full cohort
  - e.g. larger standard errors, wider confidence intervals
- The increase in uncertainty depends on several factors
- No general rules for determining which design gives smallest increase in uncertainty

Matching

- The nested case-control study is implicitly matched on the underlying time scale
- It is easy to extend this design to allow for matching on other factors
- The case-cohort study requires, by design, an unmatched subcohort and a set of unrelated cases
- Not trivial to allow for matching within this design

Analysis

- The nested case-control study can be analyzed with conditional logistic regression
- This technique is standard, and implemented in all major statistical software
- The case-cohort study requires non-standard analysis techniques
- Implemented in R, but may require some programming in other software

Choice of parameter

- In nested case-control studies, the special ‘case-control’ sampling scheme only allows for estimation of odds ratios
  - ... which, in this design, may be interpreted as rate ratios
- In case-cohort studies, the subcohort is drawn at random and can be used to estimate any parameter of interest
  - e.g. rate ratio, risk ratio, risk difference, survival functions

Multiple outcomes

- By design, the nested case-control study is ‘tied’ to one particular outcome
- In case-cohort studies, the subcohort is drawn at random and can be used to analyze any outcome of interest

Summary

- The case-cohort study is a design for subsampling from the full cohort
- It includes
  - a subcohort, which is drawn randomly at baseline
  - all cases, which occur during follow-up
- Analyzing case-cohort data requires special techniques
  - use only subcohort in the risk sets, and use standard Poisson/Cox regression, or
  - use both subcohort and cases in the risk sets, and use weighted Poisson/Cox regression to correct for bias