Epidemiology for PhD students

Bendix Carstensen Steno Diabetes Center Copenhagen
Gentofte, Denmark
http://BendixCarstensen.com/EpiPhD/F2017

Department of Biostatistics, University of Copenhagen, Spring 2019

From /home/bendix/teach/Epi/KU-epi/slides/slides.tex

Tuesday 19th February, 2019, 10:49

Case-control studies

Tuesday 19 February 2019

Epidemiology for PhD students
Department of Biostatistics, University of Copenhagen,Spring 2019

http://BendixCarstensen.com/EpiPhD/F2018

cc-lik

Relationship between follow–up studies and case–control studies

In a **cohort study**, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups.

The follow–up allows the investigator to register those subjects who develop the disease during the study period and to identify those who remain free of the disease.

Case-control study

In a **case-control study** the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up, and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease.

Rationale behind case-control studies

- In a follow-up study, rates among exposed and non-exposed are estimated by:

$$\frac{D_1}{Y_1} \qquad \frac{D_0}{Y_0}$$

- and hence the rate ratio by:

$$\frac{D_1}{Y_1} \Big/ \frac{D_0}{Y_0} = \frac{D_1}{D_0} \Big/ \frac{Y_1}{Y_0}$$

- In a case-control study we use the same cases, but select controls to represent the distribution of risk time between exposed and unexposed:

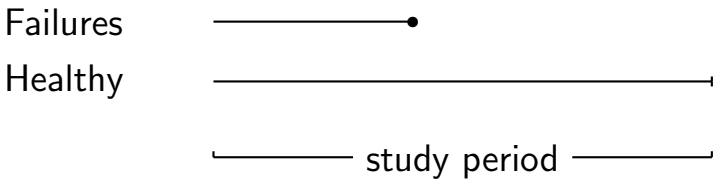
$$\frac{H_1}{H_0} \approx \frac{Y_1}{Y_0}$$

- Therefore the rate ratio is estimated by:

$$\frac{D_1}{D_0} \Big/ \frac{H_1}{H_0}$$

- Controls represent risk time, **not** disease-free persons.

Choice of controls (I)

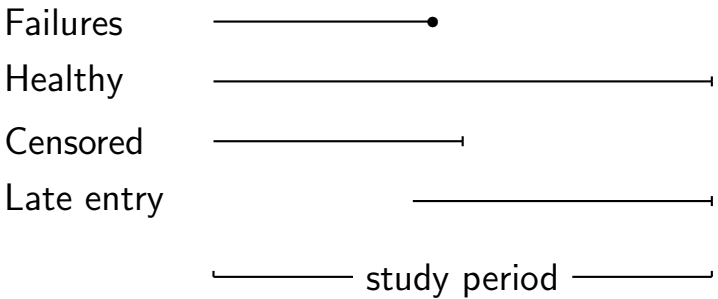


The period over which failures are registered as cases is called the study period.

A group of subjects who remain healthy over the study period is chosen to represent the healthy part of the source population.

— but this is an oversimplification...

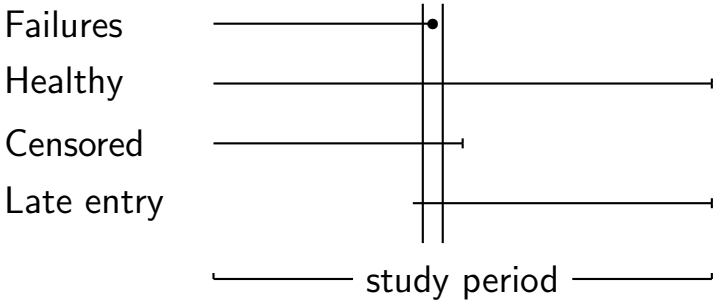
What about censoring and late entry?



Choosing controls which remains healthy throughout takes no account of censoring or late entry.

Instead, choose controls who are in the study and healthy, at the times the cases are registered.

Choice of controls (II)

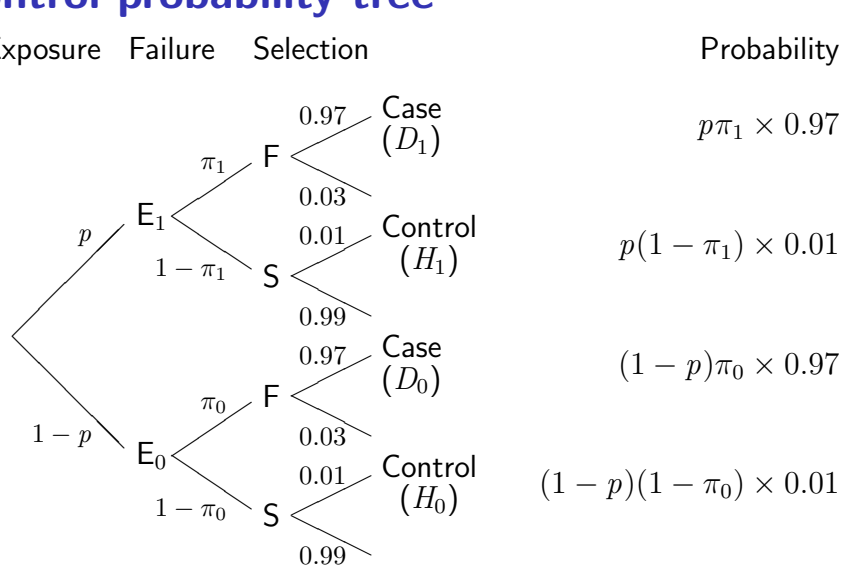


This is called **incidence density sampling**.

Subjects can be chosen as controls more than once, and a subject who is chosen as a control can later become a case.

Equivalent to sampling observation time from vertical bands drawn to enclose each case.

Case-control probability tree

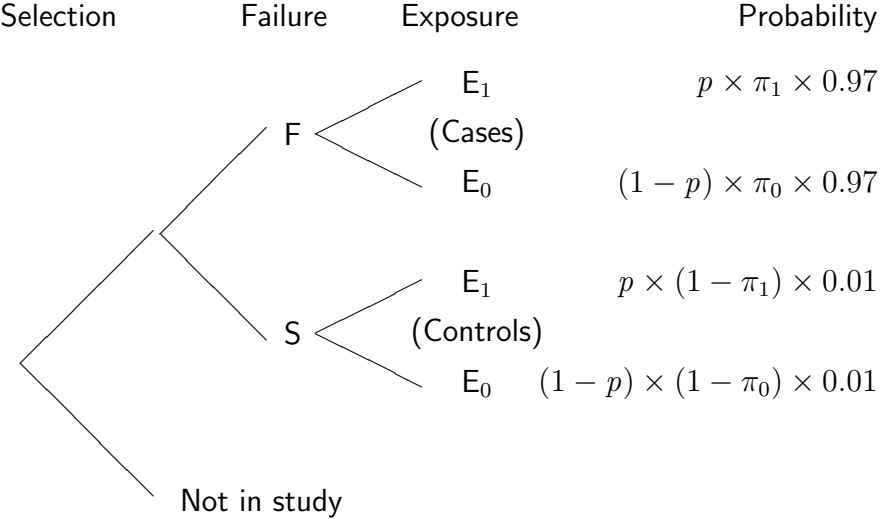


Retrospective analysis of case-control studies

Compare the distribution of exposure between cases and controls.

- How does exposure vary between cases and controls?
- The proportion of cases who smoke compared to controls
- The mean age of cases compared to controls
- Looks at the study backwards:
  — using case/control as explanatory variable
- Only works properly for binary explanatory variables

The retrospective argument



Note: Parameters in the previous tree not on these branches.

Odds of exposure for cases resp. controls:

$$\Omega_{\text{cas}} = \frac{p \times \pi_1 \times 0.97}{(1-p) \times \pi_0 \times 0.97} = \frac{p}{1-p} \times \frac{\pi_1}{\pi_0}$$

$$\Omega_{\text{ctr}} = \frac{p \times (1-\pi_1) \times 0.01}{(1-p) \times (1-\pi_0) \times 0.01} = \frac{p}{1-p} \times \frac{1-\pi_1}{1-\pi_0}$$

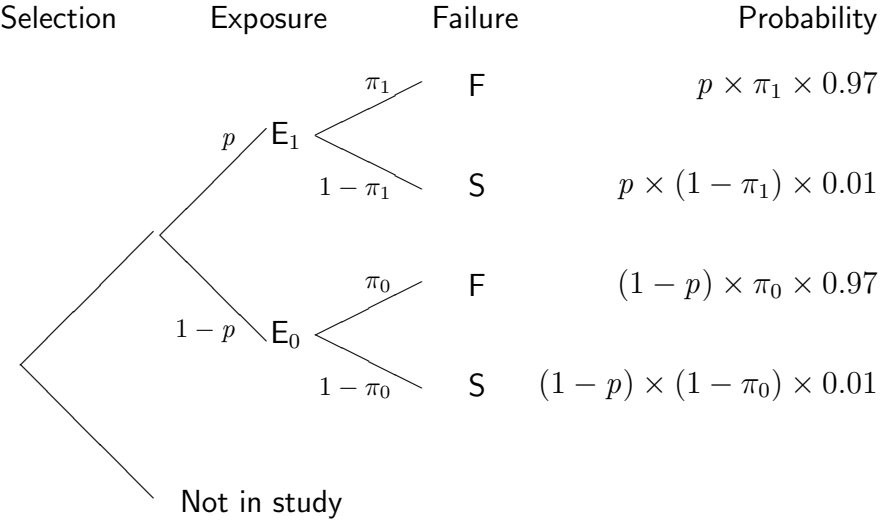
Odds-ratio of exposure between cases and controls:

$$\Omega_{\text{cas}} / \Omega_{\text{ctr}} = \frac{\pi_1}{\pi_0} \Big/ \frac{1-\pi_1}{1-\pi_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \text{OR}(\text{disease})_{\text{population}}$$

Prospective analysis of case-control studies

- Compare the case/control ratio between exposed and non-exposed subjects — or more general:
- How does case-control ratio vary with exposure ?
- The point is that **in the study** it varies in the same way as in the population
- Argument similar to retrospective, but more intuitive

The prospective argument



$$\text{Odds of disease} = \frac{\text{P}\left\{\text{Case } \textit{given} \text{ inclusion}\right\}}{\text{P}\left\{\text{Control } \textit{given} \text{ inclusion}\right\}}$$

$$\omega_1 = \frac{p \times \pi_1 \times 0.97}{p \times (1-\pi_1) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_1}{1-\pi_1}$$

$$\omega_0 = \frac{(1-p) \times \pi_0 \times 0.97}{(1-p) \times (1-\pi_0) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_0}{1-\pi_0}$$

$$\text{OR} = \frac{\omega_1}{\omega_0} = \frac{\pi_1}{1-\pi_1} \Big/ \frac{\pi_0}{1-\pi_0} = \text{OR}(\text{disease})_{\text{population}}$$

What is the case-control ratio?

$$\frac{D_1}{H_1} = \frac{0.97}{0.01} \times \frac{\pi_1}{1-\pi_1} = \left(\frac{s_{1,\text{cas}}}{s_{1,\text{ctr}}} \times \frac{\pi_1}{1-\pi_1} \right)$$

$$\frac{D_0}{H_0} = \frac{0.97}{0.01} \times \frac{\pi_0}{1-\pi_0} = \left(\frac{s_{0,\text{cas}}}{s_{0,\text{ctr}}} \times \frac{\pi_0}{1-\pi_0} \right)$$

$$\frac{D_1/H_1}{D_0/H_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \text{OR}_{\text{population}}$$

— but only if the sampling fractions are identical:
$s_{1,\text{cas}} = s_{0,\text{cas}}$ and $s_{1,\text{ctr}} = s_{0,\text{ctr}}$.

Log-likelihood for case-control studies

- Log-Likelihood (conditional on being included)
- . . . is the log-likelihood for two binomials with odds-parameters ω_0 and ω_1:

$$D_0 \log(\omega_0) - N_0 \log(1 + \omega_0) + D_1 \log(\omega_1) - N_1 \log(1 + \omega_1)$$

  where $N_0 = D_0 + H_0$ and $N_1 = D_1 + H_1$
- Exposed: D_1 cases, H_1 controls
- Unexposed: D_0 cases, H_0 controls

Log-likelihood to derive s.e.

Odds-ratio (θ) is the ratio of the odds ω_1 to ω_0, so:

$$\log(\theta) = \log\left(\frac{\omega_1}{\omega_0}\right) = \log(\omega_1) - \log(\omega_0)$$

Estimates of $\log(\omega_1)$ and $\log(\omega_0)$ are just the empirical odds:

$$\log\left(\frac{D_1}{H_1}\right) \qquad \text{and} \qquad \log\left(\frac{D_0}{H_0}\right)$$

The standard errors of the odds are estimated by:

$$\sqrt{\frac{1}{D_1}+\frac{1}{H_1}} \qquad \text{and} \qquad \sqrt{\frac{1}{D_0}+\frac{1}{H_0}}$$

Exposed and unexposed form two independent bodies of data (they are sampled independently), so the estimate of $\log(\theta)$ $[= \log(\mathrm{OR})]$ is:

$$\log\left(\frac{D_1}{H_1}\right) - \log\left(\frac{D_0}{H_0}\right),$$

$$\text{with} \quad \mathrm{s.e.}\big(\log(\mathrm{OR})\big) = \sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}}$$

Confidence interval for OR

First a confidence interval for $\log(\mathrm{OR})$:

$$\log(\mathrm{OR}) \pm 1.96 \times \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}$$

Take the exponential:

$$\mathrm{OR} \overset{\times}{\div} \underbrace{\exp\left(1.96 \times \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}\right)}_{\text{error factor}}$$

BCG vaccination and leprosy

Does BCG vaccination in early childhood protect against leprosy?

New cases of leprosy were examined for presence or absence of the BCG scar. During the same period, a 100% survey of the population of this area, which included examination for BCG scar, had been carried out.

The tabulated data refer only to subjects under 35, because vaccination was not widely available when older persons were children.

Exercise I

BCG scar	Leprosy cases	Population survey
Present	101	46 028
Absent	159	34 594

Estimate the odds of BCG vaccination for leprosy cases and for the controls. Estimate the odds ratio and hence the extent of protection against leprosy afforded by vaccination.

Give a 95% c.i. for the OR.

Use SAS for this: Exercise from the notes.

Exercise II

BCG scar	Leprosy cases	Population controls
Present	101	554
Absent	159	446

The table shows the results of a computer-simulated study which picked 1000 controls at random.

What is the odds ratio estimate in this study?

Give a 95% c.i. for the OR.

Use SAS for this: Exercise from the notes.

More levels of exposure (William Guy)

Physical exertion at work of 1659 outpatients:
341 pulmonary consumption, 1318 other diseases.

Level of exertion in occupation	Pulmonary consumption (Cases)	Other diseases (Controls)	Case/ control ratio	OR relative to (3)
Little (0)	125	385	0.325	1.643
Varied (1)	41	136	0.301	1.526
More (2)	142	630	0.225	1.141
Great (3)	33	167	0.198	1.000

The **relationship** of case-control ratios is what matters.

The retro/prospective argument

- **Retrospective:** Four possible outcomes (little/varied/more/great),
- **Prospective:** Two possible outcomes (case/control), but a large number of comparisons (between any two exposure levels).
- But the probability model is still a binary model, and the argument for the analysis is still the same as before.
- Prospective argument applicable in deriving a logistic regression model for case-control studies.

Odds-ratio and rate ratio

- If the disease probability, π, in the study period is small:

$$\pi = \text{cumulative risik} \approx \text{cumulative rate} = \lambda T$$

- For small π, $1 - \pi \approx 1$, so:

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0} \approx \frac{\lambda_1}{\lambda_0} = \text{RR}$$

π small $\Rightarrow$ OR estimate of RR.

Important assumption behind rate ratio interpretation

The entire "study base" must have been available throughout:

- no censorings.
- no delayed entries.

This will clearly not always be the case, but it may be achieved in carefully designed studies.

Avoiding censoring and delayed entry

- Can be achieved simultaneously with small π by *incidence density sampling*:
  - Subdivide calendar time in small time bands.
  - New case-control study in each time band.
  - Only one case in each time band.
  - No delayed entry or censoring.
- If the fraction of exposed does not vary much over time, all the small studies can be analysed together as one.
- This is effectively matching on calendar time.

The rare disease assumption

Necessary to make the approximation:

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0}$$

This is more appropriately termed:

"**The short study duration assumption**"

— each of the small studies we imagine as components of the entire study should be sufficiently short in relation to disease occurrence, so that the π (disease probability) if small.

Nested case-control studies

- Study base = "large" cohort
- Expensive to get covariate information for all persons. (expensive analyses, tracing of histories,. . .)
- Covariate information only for cases and *time matched* controls:
- To each case, choose one or more (usually ≤ 5) controls from the risk set.

How many controls per case?

The standard deviation of $\log(\mathrm{OR})$:

Equal number of cases and controls:

$$\begin{aligned}\sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} &= \sqrt{\frac{1}{D_1}+\frac{1}{D_1}+\frac{1}{D_0}+\frac{1}{D_0}} \\ &= \sqrt{\left(\frac{1}{D_1}+\frac{1}{D_0}\right)\times(1+1)}\end{aligned}$$

Twice as many controls as cases:

$$\begin{aligned}\sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} &= \sqrt{\frac{1}{D_1}+\frac{1}{2D_1}+\frac{1}{D_0}+\frac{1}{2D_0}} \\ &= \sqrt{\left(\frac{1}{D_1}+\frac{1}{D_0}\right)\times(1+1/2)}\end{aligned}$$

m times as many cases as controls:

$$\sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} = \sqrt{\left(\frac{1}{D_1}+\frac{1}{D_0}\right)\times(1+1/m)}$$

How many controls per case?

- The standard deviation of the $\log[\mathrm{OR}]$ is

$$\sqrt{1+\frac{1}{m}}$$

  times larger in a case-control study, compared to the corresponding cohort-study.
- Therefore, 5 controls per case is normally sufficient. (Only relevant if controls are "cheap" compared to cases).
- **But** if cases and controls cost the same — and are available — the most efficient is to have the same number of cases and controls.

Remember for next time:

Read:

Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichlorethe. J Cancer Res Clin Oncol, 1998, pp 374–382.

— available at homepage