Epidemiology for PhD students

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

Department of Biostatistics, University of Copenhagen, Spring 2022

From `/home/bendix/teach/Epi/KU-epi/slides/slides.tex` Monday 7th February, 2022, 10:40

Case-control studies

Epidemiology for PhD students
Department of Biostatistics, University of Copenhagen, Spring 2022

`http://BendixCarstensen.com` `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:

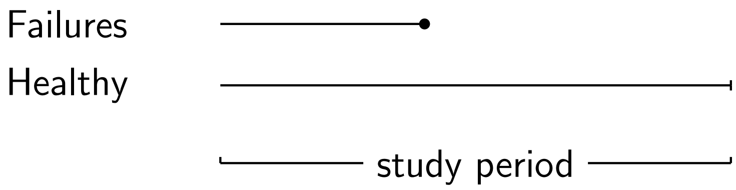
$$\frac{H_1}{H_0} \approx \frac{Y_1}{Y_0}$$

- Therefore the rate ratio can be 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)



Failures

Healthy

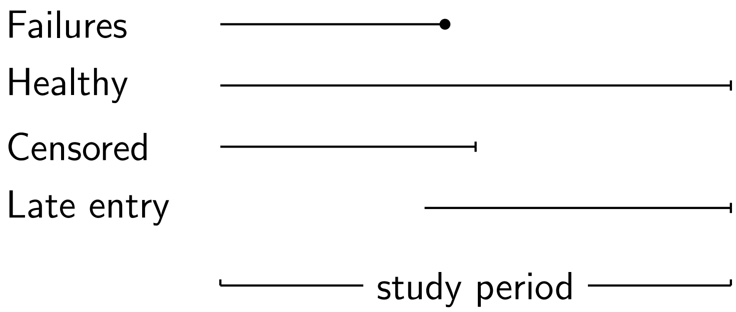
study period

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?



Failures

Healthy

Censored

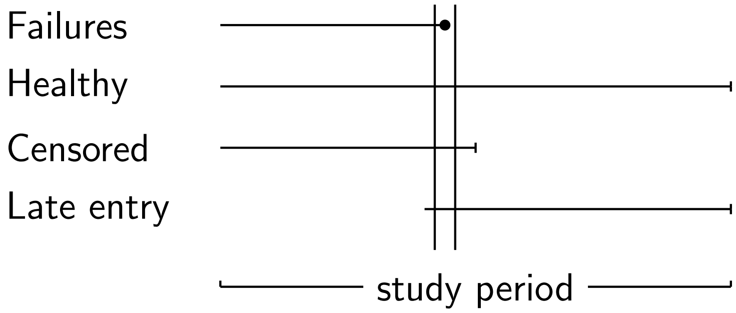
Late entry

study period

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)



Failures

Healthy

Censored

Late entry

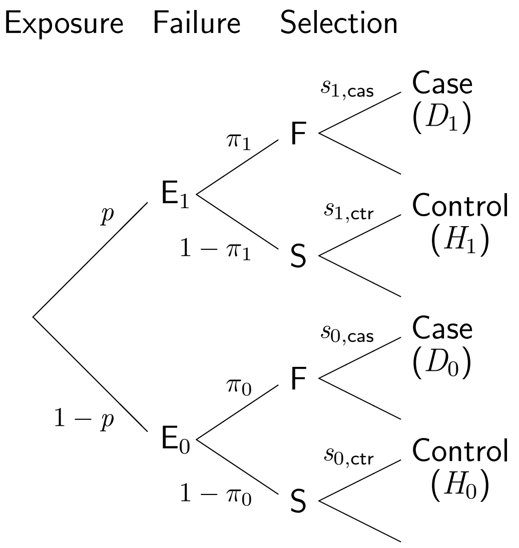
study period

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



Probability

$p\pi_1 \times s_{1,\text{cas}}$

$p(1-\pi_1) \times s_{1,\text{ctr}}$

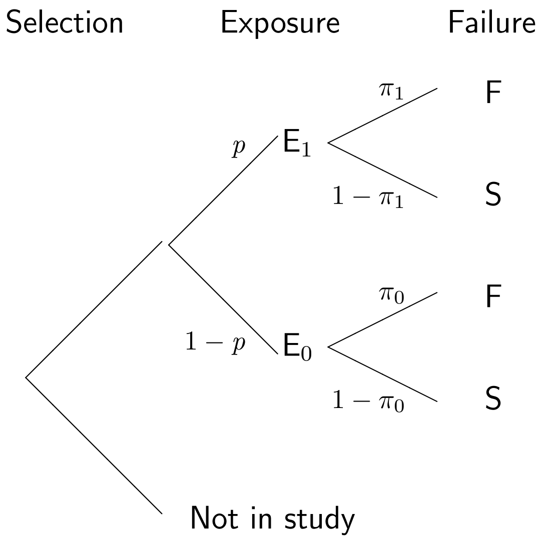
$(1-p)\pi_0 \times s_{1,\text{cas}}$

$(1-p)(1-\pi_0) \times s_{1,\text{ctr}}$

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



Probability

$p \times \pi_1 \times s_{1,\text{cas}}$

$p \times (1-\pi_1) \times s_{1,\text{ctr}}$

$(1-p) \times \pi_0 \times s_{0,\text{cas}}$

$(1-p) \times (1-\pi_0) \times s_{0,\text{ctr}}$

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

$$\omega_1 \quad = \quad \frac{p \times \pi_1 \times s_{1,\text{cas}}}{p \times (1-\pi_1) \times s_{1,\text{ctr}}} = \frac{s_{1,\text{cas}}}{s_{1,\text{ctr}}} \times \frac{\pi_1}{1-\pi_1}$$

$$\omega_0 \quad = \quad \frac{(1-p) \times \pi_0 \times s_{0,\text{cas}}}{(1-p) \times (1-\pi_0) \times s_{0,\text{ctr}}} = \frac{s_{0,\text{cas}}}{s_{0,\text{ctr}}} \times \frac{\pi_0}{1-\pi_0}$$

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

What is the case-control ratio?

$$\frac{D_1}{H_1} = \frac{s_{1,\text{cas}}}{s_{1,\text{ctr}}} \times \frac{\pi_1}{1-\pi_1}$$

$$\frac{D_0}{H_0} = \frac{s_{0,\text{cas}}}{s_{0,\text{ctr}}} \times \frac{\pi_0}{1-\pi_0}$$

]

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

— 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.

Solution to I

$$\mathrm{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{101/46028}{159/34594} = \frac{0.002194}{0.004596} = 0.48$$

$$\begin{aligned} \mathrm{s.e.}(\log[\mathrm{OR}]) = & \quad \sqrt{\tfrac{1}{D_1} + \tfrac{1}{H_1} + \tfrac{1}{D_0} + \tfrac{1}{H_0}} \\ = & \sqrt{\tfrac{1}{101} + \tfrac{1}{46028} + \tfrac{1}{159} + \tfrac{1}{34594}} = 0.127 \end{aligned}$$

The 95% limits for the odds-ratio are:

$$\mathrm{OR} \overset{\times}{\div} \exp(1.96 \times 0.127) = 0.48 \overset{\times}{\div} 1.28 = (0.37, 0.61)$$

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.

Solution to II

$$\mathrm{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{101/554}{159/446} = \frac{0.1823}{0.3565} = 0.51$$

$$\begin{aligned}\text{s.e.}(\log[\mathrm{OR}]) &= \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}} \\ &= \sqrt{\frac{1}{101} + \frac{1}{554} + \frac{1}{159} + \frac{1}{446}} = 0.142\end{aligned}$$

The 95% limits for the odds-ratio are:

$$\mathrm{OR} \overset{\times}{\div} \exp(1.96 \times 0.142) = 0.51 \overset{\times}{\div} 1.32 = (0.39, 0.68)$$

More levels of exposure (William Guy)

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

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:

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

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 πs (disease probabilities over the study period) is 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 the course homepage

Case-control studies: Stratification

Epidemiology for PhD students
Department of Biostatistics, University of Copenhagen, Spring 2022

http://BendixCarstensen.com cc-str

Age-stratified odds-ratio

Exposure: BCG

Potential confounder: age

- Age and BCG-scar correlated.
- Age is associated with leprosy.
- Bias in the estimation of the relationship between BCG-scar and leprosy.

How do we control the confounding?

Stratify the analysis by age.

Analysis stratified by age

	Leprosy cases		Population		OR
BCG	–	+	–	+	estimate
Age					
0–4	1	1	7593	11719	0.65
5–9	11	14	7143	10184	0.89
10–14	28	22	5611	7561	0.58
15–19	16	28	2208	8117	0.48
20–24	20	19	2438	5588	0.41
25–29	36	11	4356	1625	0.82
30–34	47	6	5245	1234	0.54
				Overall	0.58

Analysis stratified by age

- Assume odds-ratios are equal across strata.
- Allow disease-odds (odds of being a case) to vary across strata.
- Model:

$$\omega_{a1} = \theta \omega_{a0}$$

- This model assumes:
  - incidence rate / disease probability **varies** by age.
  - effect of exposure is the **same** regardless of age.

Matching and efficiency

- If some strata have many controls per case and other only few, there is a tendency to "waste"
  - controls in strata with many controls
  - cases in strata with few controls
- The solution is to *match* or *stratify* the study; i.e make sure that the ratio of cases to controls is approximately the same in all strata (e.g. age-groups).

BCG-example

Without age-stratification:

		Cases		Controls	
	BCG	−	+	−	+
Age	0–4	1	1	101	137
	5–9	11	14	91	115
	10–14	28	22	82	101
	15–19	16	28	28	87
	20–24	20	19	25	69
	25–29	36	11	63	21
	30–34	47	6	56	24

BCG-example

With age stratification (1:4 case/control ratio):

		Cases		Controls	
	BCG	−	+	−	+
Age	0–4	1	1	3	5
	5–9	11	14	48	52
	10–14	28	22	67	133
	15–19	16	28	46	130
	20–24	20	19	50	106
	25–29	36	11	126	62
	30–34	47	6	174	38

Analysis, controlled for age:

Analyzing the two datasets gives:

	Non-stratified	Stratified
Estimate (θ)	0.578	0.564
$\mathrm{s.d.}[\log(\theta)]$	0.160	0.155
Error factor	1.369	1.354
Lower 95% limit	0.422	0.417
Upper 95% limit	0.792	0.764

No dramatic difference: the number of controls is in both cases sufficient to produce a reasonably precise estimate.

Matching: BIAS!

- If the study is stratified on a variable, this variable **must** enter in the analysis too:

	Cases		Controls		Odds
Stratum	+	−	+	−	ratio
1	89	11	80	20	2.0
2	67	33	50	50	2.0
3	33	67	20	80	2.0
Total	189	111	150	150	1.7

- The bias from ignoring matching will always be toward 1.

Incidence density sampling

- Incidence density matching. Not because calendar time is associated to exposure, but mostly of practical reasons.
- The calendar time (of matching/inclusion) need not enter in the analysis.

Incidence density sampling

- Theoretically controls may later appear as cases. They should appear twice in the study — first as control with the set of covariates relevant to the control sampling date.
- Definition of exposure in relation to case-diagnosis — when a person is included as control, exposure status is at time of diagnosis of the corresponding case.
- If he later is included as a case, exposure status is at date of diagnosis. So the person appears twice but with different exposure.

Exercises

- BCG-exercises:
  1. Simple 2×2 tables (already done)
  2. Stratified analysis by `proc freq`
- Renal cancer exercise:
  1. Discussion
  2. Replicate the analysis.
  3. Use logistic regression.

Case-control exercise

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

1. What is the primary aim of the study?
2. How was cases sampled?
3. How was controls sampled?
4. Are they comparable; i.e. what assumptions are needed?
5. What is the (actual) study base?
6. What study base is the intended? (for generalization).
7. Is this incidence density sampling?
8. Can the age-effect on the occurrence renal cancer be estimated?
9. Is age a confounder?
10. What is the main result?
11. Key in the numbers in table 6 (p.380), and verify the analysis using SAS.

Stratified by age (table 6 in the paper):

	Cases		Controls			
Exp.	+	−	+	−	OR	95% c.i.
Age						
<40	2	0	1	21	∞	(1.64; ∞)
40–50	2	1	4	11	4.92	(0.21; 352.2)
50–60	10	12	2	25	9.89	(1.73; 106.8)
60–70	1	17	0	14	∞	(0.02; ∞)
≥70	4	9	0	6	∞	(0.31; ∞)
Total	19	39	7	77	5.29	(1.93; 16.2)
				MH-estimate	13.73	(3.08; 61.2)

(Estimates and c.i.s based on a hypergeometric likelihood.)

The logit-estimate (Adding 0.5 to tables with 0s)

Age	Exp.	Ca	Co	$\log(\mathrm{OR}_a)$	$\mathrm{var}[\log(\mathrm{OR}_a)]$
<40	+	2.5	1.5	$\log\left(\frac{2.5\times21.5}{0.5\times1.5}\right)$	$\frac{1}{2.5}+\frac{1}{1.5}+\frac{1}{0.5}+\frac{1}{21.5}$
	−	0.5	21.5	$= 4.27$	$= 3.11$
40–50	+	2.0	4.0		
	−	1.0	11.0	1.70	1.84
50–60	+	10.0	2.0		
	−	12.0	25.0	2.34	0.72
60–70	+	1.5	0.5		
	−	17.5	14.5	0.91	2.79
≥70	+	4.5	0.5		
	−	9.5	6.5	1.82	2.48

The common odds-ratio is calculated, using the inverse variances as weights ($w_a = \mathrm{var}[\log(\mathrm{OR}_a)]$):

$$\begin{aligned}\mathrm{OR}_{\text{logit}} &= \exp\left(\sum_a (\log(\mathrm{OR}_a)/w_a) \Big/ \sum_a (1/w_a)\right) \\ &= \exp\left(\frac{4.27/3.11 + 1.70/1.84 + \cdots}{1/3.11 + 1/1.84 + \cdots}\right) \\ &= 8.96\end{aligned}$$

Are the odds-ratios really equal?

The assumption behind both the MH-estimate and the logit-estimate is that the odds-ratio **is** the same in all strata.

This can be tested by the **Breslow-Day test:**

- Compares the observed numbers in the table with the expected assuming the the odds-ratio is equal to OR_{MH} in all strata.

NE Breslow & NE Day: Statistical Methods in Cancer Research, Volume 1: The analysis of case-control studies. IARC, Lyon 1980, pp. 142 ff.

Using SAS proc freq

Enter data one line per cell entry: `renal.sas`
Use `weight` to tell SAS the numbers in each cell:

```
data a ;
 input age tri ck n ;
cards ;
30   1    1    2
40   1    1    2
50   1    1   10
60   1    1    1
70   1    1    4
30   0    1    0
40   0    1    1
50   0    1   12
60   0    1   17
70   0    1    9
30   1    0    1
40   1    0    4
50   1    0    2
```

```
proc freq  data = a ;
 table age * tri * ck
    / norow nocol
     nopct cmh ;
 weight n ;
run ;
```

Output from `proc freq`:

```
Table 1 of tri by ck
Controlling for age=30

tri       ck
Frequency        0         1    Total
----------------------------
       0       21         0       21
----------------------------
       1        1         2        3
----------------------------
Total          22         2       24

osv...
```

```
         Estimates of the Common Relative Risk (Row1/Row2)

Type of Study   Method              Value    95% Conf. Limits
--------------------------------------------------------------
Case-Control    Mantel-Haenszel  13.7285    3.5989  52.3684
  (Odds Ratio)  Logit **          8.9623    2.8949  27.7466
...
** These logit estimators use a correction of 0.5 in every
   cell of those tables that contain a zero.

Breslow-Day Test for Homogeneity of the Odds Ratios
------------------------------
Chi-Square              2.8440
DF                           4
Pr > ChiSq              0.5843

Total Sample Size = 142
```

Analysis by logistic regression

- Assuming the odds ratio, θ, to be constant over strata, each stratum adds a separate contribution to the log likelihood function for θ.
- The log likelihood can be analyzed in a model where odds is a product of age-effect and exposure effect.
- This is a **logistic regression** model:

  $$\text{case-control odds}(a) = \mu_a \times \theta$$

  — a multiplicative model for **odds**.
- additive model for log-odds:

  $$\log(\text{odds}) = m_a + b$$

Recall the sampling fractions:

What is estimated by the case-control ratio?

$$\frac{D_1}{H_1} = \frac{0.97}{0.01} \times \frac{\pi_1}{1-\pi_1} = \left(\frac{s_1}{k_1} \times \frac{\pi_1}{1-\pi_1}\right)$$

$$\frac{D_0}{H_0} = \frac{0.97}{0.01} \times \frac{\pi_0}{1-\pi_0} = \left(\frac{s_0}{k_0} \times \frac{\pi_0}{1-\pi_0}\right)$$

Study valid only for equal sampling fractions: $s_1/k_1 = s_0/k_0 = s/k$.

Population odds **multiplied** ratio of sampling fractions for cases to controls.

Logistic regression for C-C studies

- Model for the population:

$$\ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Model for the observed data:

$$\ln\big(\text{odds}(\text{case}|\text{incl.})\big) = \ln\left[\frac{\pi}{1-\pi}\right] + \ln\left[\frac{s}{k}\right]$$
$$= \left(\ln\left[\frac{s}{k}\right] + \beta_0\right) + \beta_1 x_1 + \beta_2 x_2$$

Logistic regression for C-C studies

- Analysis of $\mathrm{P}\{\text{case} \mid \text{inclusion}\}$
  — i.e. binary observations:

$$Y = \begin{cases} 1 & \sim \ \text{case} \\ 0 & \sim \ \text{control} \end{cases}$$

- Effects of covariates are estimated correctly.
- Intercept is (almost always) meaningless.
  Depends on the sampling fractions for cases, s, and controls, k, which are usually not known.

Parameter interpretation in logistic regression

Model for persons with covariates x_A, resp. x_B:

$$\ln\big(\text{odds(case} \mid x_A)\big) = \left(\ln\left[\frac{s}{k}\right] + \beta_0\right) + \beta_1 x_{1A} + \beta_2 x_{2A}$$
$$\ln\big(\text{odds(case} \mid x_B)\big) = \left(\ln\left[\frac{s}{k}\right] + \beta_0\right) + \beta_1 x_{1B} + \beta_2 x_{2B}$$

$$\ln\big(\mathrm{OR}_{x_A \text{ vs. } x_B}\big) = \beta_1(x_{1A} - x_{1B}) + \beta_2(x_{2A} - x_{2B})$$

$\exp(\beta_1)$ is OR for a difference of 1 in x_1
$\exp(\beta_2)$ is OR for a difference of 1 in x_2
— assuming that other variables are fixed.

Stratified sampling

- We have different sampling fraction for each stratum (age-class, sex, . . .)
- Model for the observed data:

$$\ln\big(\text{odds(case}|\text{incl.)}\big) = \ln\left[\frac{\pi}{1-\pi}\right] + \ln\left[\frac{s_a}{k_a}\right]$$
$$= \left(\ln\left[\frac{s_a}{k_a}\right] + \beta_0\right) + \beta_1 x_1 + \beta_2 x_2$$

- Thus, an intercept for each stratum
- — but with no interpretation
- this is why the stratification variable must be in the model

SAS commands — data

```
data a1 ;
  input bcg alder cases cont rcont mcont ;
  total = cases + cont ;
 rtotal = cases + rcont ;
 mtotal = cases + mcont ;
cards;
1 7  1  7593 101   3
0 7  1 11719 137   5
1 6 11  7143  91  48
0 6 14 10184 115  52
1 5 28  5611  82  67
0 5 22  7561 101 133
1 4 16  2208  28  46
0 4 28  8117  87 130
1 3 20  2438  25  50
0 3 19  5588  69 106
1 2 36  4356  63 126
0 2 11  1625  21  62
1 1 47  5245  56 174
0 1  6  1234  24  38
```

SAS commands
— random sample of controls

```
proc genmod  data = a1 ;
  class alder bcg ;
  model cases / rtotal = alder bcg
        / dist = bin
          link = logit
         type3 ;
  estimate "+bcg" bcg  1 -1 / exp ;
  estimate "-bcg" bcg -1  1 / exp ;
run;
```

Random sample of controls

```
Deviance                 6        6.6268        1.1045

Analysis Of Parameter Estimates
Parameter        DF    Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT         1     -4.5008      0.7138     39.7577  0.0001
ALDER       1     1      4.2062      0.7333     32.9008  0.0001
ALDER       2     1      4.0452      0.7345     30.3339  0.0001
ALDER       3     1      3.9700      0.7363     29.0739  0.0001
ALDER       4     1      3.9233      0.7333     28.6209  0.0001
ALDER       5     1      3.4711      0.7282     22.7200  0.0001
ALDER       6     1      2.6685      0.7414     12.9538  0.0003
ALDER       7     0      0.0000      0.0000           .       .
BCG         0     1     -0.5475      0.1604     11.6557  0.0006
BCG         1     0      0.0000      0.0000           .       .
```

```
LR Statistics For Type 3 Analysis:

Source           DF   Chi-Square    Pr > ChiSq
alder             6     149.73        <.0001
bcg               1      11.78        0.0006

Contrast Estimate Results
                     Standard                          Chi-
Label      Estimate    Error    Conf. Limits     Square    Pr>ChiSq

+bcg        -0.5475   0.1604   -0.8619 -0.2332    11.66      0.0006
Exp(+bcg)    0.5784   0.0928    0.4224  0.7920
-bcg         0.5475   0.1604    0.2332  0.8619    11.66      0.0006
Exp(-bcg)    1.7290   0.2773    1.2626  2.3676
```

Matched sample of controls I

```
Deviance                 6        4.4399        0.7400

Analysis Of Parameter Estimates
Parameter        DF     Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT         1      -1.0667      0.7998      1.7786  0.1823
ALDER       1     1      -0.2380      0.8129      0.0857  0.7697
ALDER       2     1      -0.1628      0.8136      0.0400  0.8414
ALDER       3     1       0.0244      0.8160      0.0009  0.9761
ALDER       4     1       0.0713      0.8139      0.0077  0.9302
ALDER       5     1       0.0119      0.8116      0.0002  0.9883
ALDER       6     1      -0.0421      0.8271      0.0026  0.9594
ALDER       7     0       0.0000      0.0000           .       .
BCG         0     1      -0.5721      0.1547     13.6790  0.0002
BCG         1     0       0.0000      0.0000           .       .
```

Matched sample of controls II

```
 LR Statistics For Type 3 Analysis
                              Chi-
Source           DF         Square     Pr > ChiSq
alder             6           2.33         0.8867
bcg               1          13.89         0.0002

Contrast Estimate Results
                     Standard                          Chi-
Label     Estimate     Error    Conf. Limits          Square  Pr>ChiSq

+bcg       -0.5721    0.1547   -0.8752  -0.2689   13.68     0.0002
Exp(+bcg)   0.5644    0.0873    0.4168   0.7642
-bcg        0.5721    0.1547    0.2689   0.8752   13.68     0.0002
Exp(-bcg)   1.7719    0.2741    1.3085   2.3994
```

Matched sample of controls III

Standard deviation of $\ln(\mathrm{OR})$ shrinks from 0.160 to 0.155 by age-matching.

The age-BCG and the age-leprosy associations are not very strong.

Caveat: remember the matching variable

With age in the model:

```
Label       Estimate   StdErr     Conf. Limits     ChiSq  Pr>ChiSq
+bcg         -0.5721   0.1547   -0.8752  -0.2689   13.68    0.0002
Exp(+bcg)     0.5644   0.0873    0.4168   0.7642
```

Without age in the model:

(**wrong!**—OR biased toward 1):

```
+bcg         -0.4769   0.1416   -0.7543  -0.1994   11.35      0.0008
Exp(+bcg)     0.6207   0.0879    0.4703   0.8192
```

Change in $\ln(\mathrm{OR})$ is $0.0952 \approx 61\%\ \mathrm{s.e.}$!

Individually matched study

If strata are defined so finely that ony one case is in each, we have an individually matched study:

- Comparability between cases and controls.
- Control for ill-defined factors.
- Convenience in sampling.
- Controlling for age, calendar time, . . .

(incidence density sampling).

Individually matched study

- Conventional method for analysis (logistic regression) breaks down, because we get one parameter per case!
- If matching is on a well-defined variable as e.g. age, then broader stata may be formed *post hoc*, and age included in the model.
- If matching is on “soft” variables (neighbourhood, occupation, . . .) the original matching cannot be ignored: Matched analysis.

Matched studies

- $1:1$ matching:
  For each case select one matched control,
  - similar w.r.t. age / sex / place of residence / ...
  - in order to control for:
    - the matching variables
    - "undefined" variables associated with the matching.
- $1:m$ matching:
  For each case select m matched controls.
  m need not be the same for all matched sets.

Salmonella Manhattan study

Telephone interview concerning the food items eaten during the last three days:

- Case: Verified infection with *S.* Manhattan
- Control: Person from same geographical area.
- 16 matched pairs — $1:1$ matched study.
- Exposure: Eaten sliced saxony ham (hamburgerryg)

```
OBS  PAR  PK  KONTR   HAMB        OBS  PAR  PK  KONTR   HAMB
  1    1   P      0      0         17   12   P      0      0
  2    1   K      1      0         18   12   K      1      0
  3    3   P      0      1         19   14   P      0      1
  4    3   K      1      0         20   14   K      1      0
  5    4   P      0      1         21   16   P      0      0
  6    4   K      1      0         22   16   K      1      0
  7    5   P      0      1         23   17   P      0      1
  8    5   K      1      1         24   17   K      1      0
  9    7   P      0      1         25   18   P      0      0
 10    7   K      1      0         26   18   K      1      1
 11    8   P      0      0         27   19   P      0      1
 12    8   K      1      1         28   19   K      1      1
 13    9   P      0      0         29   20   P      0      1
 14    9   K      1      0         30   20   K      1      1
 15   11   P      0      1         31   23   P      0      1
 16   11   K      1      1         32   23   K      1      0
```

1:1 matched studies — Tabulation

1:1 matched case-control study can be tabulated as:

No. of pairs		Control exposure $+$	 $-$	
Case exposure	$+$	a	b	$a+b$
	$-$	c	d	$c+d$
		$a+c$	$b+d$	N

1:1 matched studies — Estimation

Remember: Exposure OR = Disease OR:

$$\text{OR} = \omega = \frac{\mathrm{P}\left\{\text{E}+ \mid \text{case}\right\} \mathrm{P}\left\{\text{E}- \mid \text{control}\right\}}{\mathrm{P}\left\{\text{E}- \mid \text{case}\right\} \mathrm{P}\left\{\text{E}+ \mid \text{control}\right\}}$$

estimated by:

$$\hat{\omega} = \frac{b}{c}$$

Standard error on the log-scale:

$$\text{s.e.}[\ln(\hat{\omega})] = \sqrt{\frac{1}{b} + \frac{1}{c}}$$

Salmonella Manhattan study

Exercise: Tabulate data:

No. of pairs		Control exposure $+$	 $-$	
Case	$+$			
exposure	$-$			

— and compute the OR with a 95% c.i.

		Control exposure	
		+	−
Case	+	4	6
exposure	−	2	4

$$\hat{\mathrm{OR}} = \frac{b}{c} = \frac{6}{2} = 3$$

$$\text{s.e.}[\ln(\hat{\mathrm{OR}})] = \sqrt{\frac{1}{b} + \frac{1}{c}} = \sqrt{\frac{1}{2} + \frac{1}{6}} = 0.816$$

Approximate 95% c.i. for OR:

$$3 \overset{\times}{\div} \exp(1.96 \times 0.816) = (0.61, 14.9)$$

1:1 matched studies: — Test

No. of pairs		Control exposure		
		+	−	
Case exposure	+	a	b	$a+b$
	−	c	d	$c+d$
		$a+c$	$b+d$	N

- McNemar's test of OR$=1$ compares b og c:

$$\frac{(b-c)^2}{b+c} \sim \chi^2(1)$$

- McNemar's test with continuity correction:

$$(|b-c|-1)^2 \sim \chi^2(1)$$

Test for $OR = 1$

- Compute McNemar's test for the *Salmonella* Manhattan data.

Test for $OR = 1$

- Compute McNemar's test for the *Salmonella* Manhattan data.
- Without continuity-korrektion:

$$\frac{(6-2)^2}{6+2} = \frac{16}{8} = 2, \qquad \mathrm{p} = 0.158$$

- With the continuity-correction:

$$\frac{(|6-2|-1)^2}{6+2} = \frac{9}{8} = 0.289, \qquad \mathrm{p} = 0.158$$

1:1 matched studies — Likelihood

Possible to derive a **contional** likelihood.

Analysis of regression models is then possible for matcehd studies — both $1:1$ and $1:m$ studies:

Conditional logistic regression.

Available in SAS, either as a variant of `proc phreg` or as an option `proc logistic`.

This is a topic of the Advanced Epidemiology course.