Matched and nested case-control studies

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
b@bxc.dk
http://BendixCarstensen.com

Department of Biostatistics, University of Copenhagen,
18 November 2016

http://BendixCarstensen.com/AdvEpi

Case-control studies

Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
http://BendixCarstensen.com/AdvEpi

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.

Relationship between follow–up studies and case–control studies

- In a **case-control study** the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up
- A group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease.
- Persons are selected on the basis of **disease outcome**.
- Occasionally referred to as "retrospective study".

Rationale behind case-control studies

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

$$\frac{D_1}{Y_1} \quad \text{and} \quad \frac{D_0}{Y_0}$$

- and the rate ratio by:

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

Rationale behind case-control studies

- Case-control study: same cases but controls represent the distribution of risk time

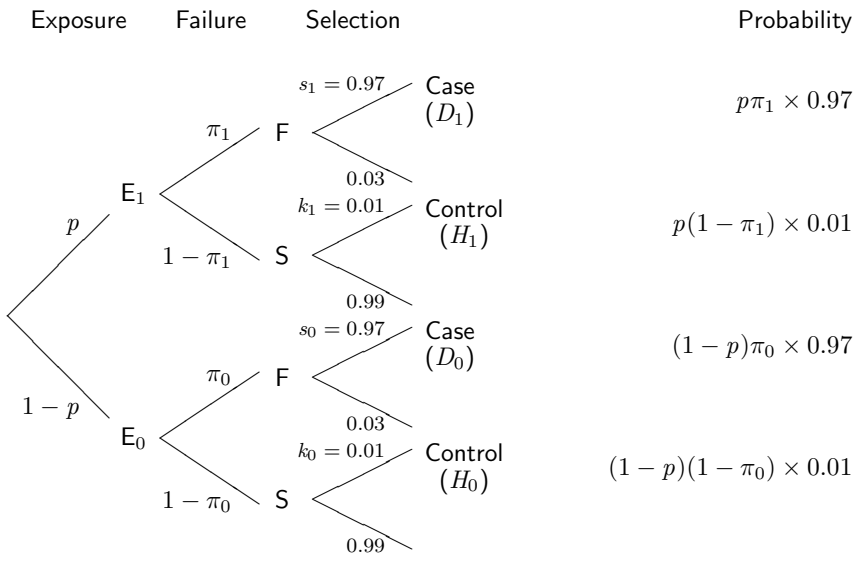
$$\frac{H_1}{H_0} \approx \frac{Y_1}{Y_0}$$

- . . . 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.

Case–control probability tree



What is estimated by the case-control ratio?

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

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

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

— but only for equal sampling fractions:

$$s_1/k_1 = s_0/k_0 \quad \Leftarrow \quad s_1 = s_0 \wedge k_1 = k_0$$

.

Estimation from case-control study

Odds-ratio of disease between exposed and unexposed **given inclusion**:

$$\mathrm{OR} = \frac{\omega_1}{\omega_0} = \frac{\pi_1}{1-\pi_1} \Big/ \frac{\pi_0}{1-\pi_0}$$

odds-ratio of disease (for a small interval)

between exposed and unexposed *in the study* is the same as odds-ratio for disease between exposed and unexposed in the "study base",

Estimation from case-control study

. . . under the assumption that:

- inclusion probability is the same for exposed and unexposed cases.
- inclusion probability is the same for exposed and unexposed controls.

The selection mechanism can **only** depend on case/control status.

Disease OR and exposure OR

- The **disease**-OR comparing exposed and non-exposed given inclusion in the study is the same as the population-OR:

$$\frac{D_1}{H_1} \Big/ \frac{D_0}{H_o} = \frac{\pi_1}{1-\pi_1} \Big/ \frac{\pi_0}{1-\pi_0} = \mathrm{OR}_{\mathrm{pop}}$$

- The **disease**-OR is equal to the **exposure**-OR comparing cases and controls:

$$\frac{D_1}{H_1} \Big/ \frac{D_0}{H_o} = \frac{D_1}{D_o} \Big/ \frac{H_1}{H_o} = \frac{D_1 H_0}{D_0 H_1}$$

Log-likelihood for case-control studies

The **observations** in a case-control study are

- Response: case/control status
- Covariates: exposure status, etc.

Parameters possible to estimate are
odds of disease
conditional on inclusion into the study.

and therefore also

odds **ratio** of disease **between groups**
conditional on inclusion into the study.

Log-likelihood for case-control studies

The log-likelihood is a binomial likelihood with odds of being a case (conditional on being included):

- odds ω_0 for unexposed and
- odds ω_1 for exposed

  or
- odds ω_0 for unexposed and
- the odds-ratio $\theta = \omega_1/\omega_0$ between exposed and unexposed.

Only the odds-ratio parameter, θ, is of interest

Log-likelihood for case-control studies

Case/control outcome and exposure $(0/1)$:

- unexposed group:
  N_0 persons, D_0 cases, $N_0 - D_0$ controls,
  case-odds ω_0
- exposed group:
  N_1 persons, D_1 cases, $N_1 - D_1$ controls,
  case-odds $\omega_1 = \theta\omega_0$

Binomial log-likelihood:

$$D_0\ln(\omega_0) - N_0\ln(1+\omega_0) + D_1\ln(\theta\omega_0) - N_1\ln(1+\theta\omega_0)$$

— logistic regression with case/control status as outcome and exposure as explanatory variabale

Log-likelihood for case-control studies

Binomial outcome (case/control) and binary exposure $(0/1)$

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

$$\ln(\theta) = \ln(\omega_1/\omega_0) = \ln(\omega_1) - \ln(\omega_0)$$

Estimates of $\ln(\omega_1)$ and $\ln(\omega_0)$ are:

$$\widehat{\ln(\omega_1)} = \ln\left(\frac{D_1}{H_1}\right) \quad \text{and} \quad \widehat{\ln(\omega_0)} = \ln\left(\frac{D_0}{H_0}\right)$$

Log-likelihood for case-control studies

Estimated log-odds have standard errors:

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

Exposed and unexposed form two independent bodies of data, so the estimate of $\ln(\theta) \quad [= \ln(\mathrm{OR})]$ is

$$\ln\left(\frac{D_1}{H_1}\right) - \ln\left(\frac{D_0}{H_0}\right), \quad \text{s.e.} = \sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}}$$

BCG vaccination and 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.

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

The tabulated data refer only to subjects under 35. What are the sampling fractions in this study?

Odds ratio with confidence interval

$$\mathrm{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{101/46,028}{159/34,594} = 0.48$$

$$\begin{aligned}\text{s.e.}(\ln[\mathrm{OR}]) &= \sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} \\ &= \sqrt{\frac{1}{101}+\frac{1}{46,028}+\frac{1}{159}+\frac{1}{34,594}} \\ &= 0.127\end{aligned}$$

$$\mathrm{erf} = \exp(1.96 \times 0.127) = 1.28$$

$$\mathrm{OR} \overset{\times}{\div} \mathrm{erf} = 0.48 \overset{\times}{\div} 1.28 = (0.37, 0.61) \quad \text{(95\% c.i.)}$$

Unmatched study with 1000 controls

BCG scar	Leprosy cases	Controls
Present	101	554
Absent	159	446

What are the sampling fractions here?

$$\text{OR} = \frac{101/554}{159/446} = \frac{0.1823}{0.3565} = 0.51$$

$$\text{s.e.}(\ln[\text{OR}]) = \sqrt{\frac{1}{101} + \frac{1}{554} + \frac{1}{159} + \frac{1}{446}} = 0.142$$

$$\text{erf} = \exp(1.96\text{s.e.}(\ln[\text{OR}])) = 1.32$$

95% c.i.: $0.51 \overset{\times}{\div} \text{erf} = (0.39, 0.68)$

Frequency matched studies

Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
`http://BendixCarstensen.com/AdvEpi`

Age-stratified odds-ratio: BCG data

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.

Estimate an OR for leprosy associated with BCG in each age-stratum.

Combine to an overall estimate (if not too variable between strata).

This is called stratified analysis (by age):

	Cases		Population		OR
BCG	−	+	−	+	estimate
Age					
0–4	1	1	7,593	11,719	0.65
5–9	11	14	7,143	10,184	0.89
10–14	28	22	5,611	7,561	0.58
15–19	16	28	2,208	8,117	0.48
20–24	20	19	2,438	5,588	0.41
25–29	36	11	4,356	1,625	0.82
30–34	47	6	5,245	1,234	0.54
				Overall	0.58

The simulated cc-study, stratified by age

	Cases		Population	
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
Total	159	101	446	554

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 **design**:
- Make sure that the ratio of cases to controls is approximately the same in all strata (e.g. age-groups).

Simulated cc-study (group-matched)

	Cases		Population	
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

4 times as many controls as cases.

What are the sampling fractions here?

Simulated cc-study (group-matched)

- **Not** possible to estimate effect of age.
- Age **must** be included in model.
  But estimates of age-effects do not have any meaning.
- Testing of the age-effect is irrelevant.
- If a variable is used for matching (stratified sampling) it **must** be included in the model.

Matching: BIAS!

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

Stratum	Cases		Controls		Odds
Exp	+	−	+	−	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.

Interaction with the matching variable

- How age influences the risk of leprosy cannot be estimated from an age-matched study.
- **Age-effect** cannot be estimated from an age-stratified study.
- But the exposure×age **interaction** can be estimated:
- How does the BCG-effect vary with age:
  - The OR of leprosy between BCG yes/no is not same in all age-classes.
  - The OR of leprosy between BCG yes/no decreases from age-class to age-class.

Confounding and matching

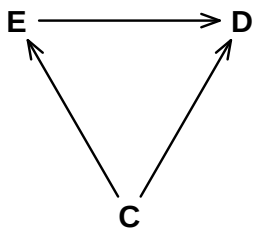
Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
`http://BendixCarstensen.com/AdvEpi`

Confounding definition

- Exposure effect estimated wrongly because a factor is associated both with exposure and disease.
- Age and sex are the most common confounders.
- Confounder characteristics:
  - Associated to exposure
  - Risk factor by itself (associated to disease).
- Associated to exposure only: Irrelevant
- Associated to disease only: Independent risk factor

Confounding and causal chain:



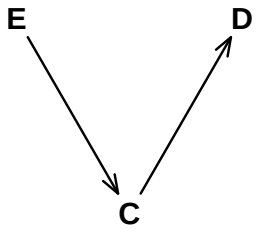
Confounding:

Ignoring **C** gives biased estimate of the effect of **E**.

Control of the confounding effect of **C** is necessary.

BMI — Age — DM
Should we match on **C** (age)?
If we do, should it be included in analysis?

Confounding and causal chain:



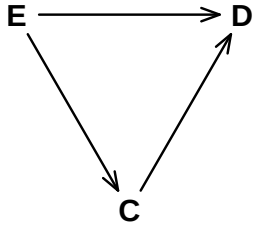
Intermediate variable:

Control of the effect of **C** is not wanted:

C is a stage in the development of **D**.

Genotype — BMI — Insulin resistance
Should we match on **C** (BMI)?
If we do should it be included in analysis?

Confounding and causal chain:



Intermediate variable **and** direct effect of **E**:

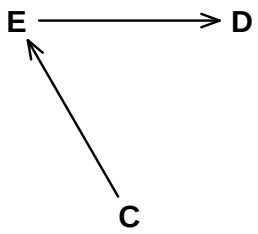
Control of the effect of **C** is not wanted:

Cannot be distinguished from confounding.

Genotype — BMI — Insulin resistance
Should we match on **C** (BMI)?
If we do should it be included in analysis?

Mediation analysis — outside this lecture.

Confounding and causal chain:



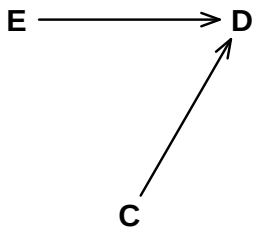
Preceding exposure:

Control of the effect of **C** is not necessary.

It will just decrease the precision of the effect estimate.

BMI — Genotype — Insulin resistance
Should we match on **C** (genotype)?
If we do should it be included in analysis?

Confounding and causal chain:



Separate risk factor (independent of **E**):

Control of the effect of **C** is not necessary.

But it will probably be useful to estimate the effect of both **E** and **C**.

Should we match on **C**?
If we do should it be included in analysis?

Confounding and causal chain:

- Do not include variables **preceding** exposures of interest
- Do not include **intermediate** variables, on the causal chain from exposure to outcome
- — neither in stratification or analysis
- Otherwise sensible it is to include (potential) confounders / exposures in a statistical model.
- The causal structure is **assumed** and cannot be inferred from data.
- There is no way to test for confounding
- . . . or for intermediate effects

Logistic regression in CC-studies

Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
`http://BendixCarstensen.com/AdvEpi`

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:

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

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:

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

$$= \left(\ln\left[\frac{s_a}{k_a}\right] + \beta_0\right) + \beta_1 x_1 + \beta_2 x_2$$

- 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
;
run ;
```


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

Label	Estimate	Standard Error	Conf. Limits		Chi-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

Source	DF	Chi-Square	Pr > ChiSq
alder	6	2.33	0.8867
bcg	1	13.89	0.0002

Contrast Estimate Results

Label	Estimate	Standard Error	Conf. Limits		Chi-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
+bcg	-0.5721	0.1547	-0.8752	-0.2689	13.68
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
Exp(+bcg)	0.6207	0.0879	0.4703	0.8192	

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

Interpretation and study design

Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
`http://BendixCarstensen.com/AdvEpi`

Odds-ratio and rate ratio

- If the disease probability, π, in the study period (length of period: T) is small:

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

- 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.

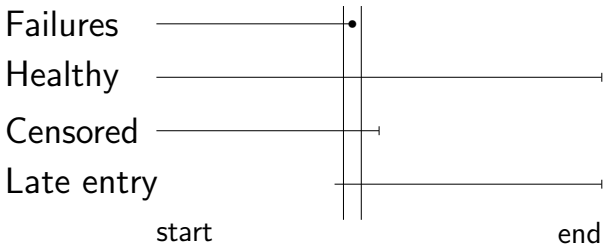
Important assumption behind rate ratio interpretation

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.

Choice of controls (I)



Instead, choose controls from members of the source population who are in the study and healthy, at the (calendar) times cases are registered.

This is called **incidence density sampling**

Incidence density sampling

- The method is equivalent to sampling observation time from vertical bands drawn to enclose each case.
  — this is how controls are chosen to represent risk time. ($H \propto Y$).
- New case-control study in each time band.
- No delayed entry or censoring
- Can be analysed together if no confounding by calendar time:
  - If disease risk does not vary over time
  - or
  - If the fraction of exposed does not vary over time

Incidence density sampling

Implications for sampling:

- a person can be a control more than once
- a person chosen as a control can be a case later
- each person is sampled at a specific **time**
- **covariates** refer to this time
- if the same person included multiple times, it will typically with different covariate values
- — representing the non-diseased **risk time**
- — and not the non-diseased **persons**

Nested case-control study

- Case-control study nested in cohort:
- Controls are chosen from a cohort from which the cases arise.
- Controls are chosen among those at risk of becoming cases at the time of diagnosis of each case.
- In Scandinavia, most case-control studies are nested in the entire population, because this is available as a cohort in the population registers.

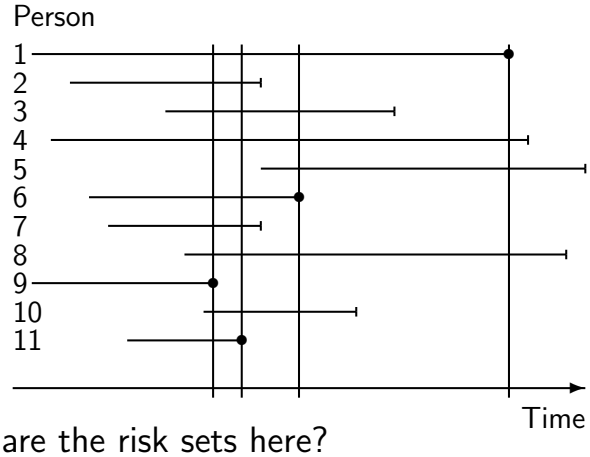
Reasons for nested case-control study

- Collection of data on covariates:
  - not measured in the cohort study
  - but available for measuring
  - *e.g.* stored blood samples
- Data collection only for cases and matched controls.
- Alternative would be collecting data on the entire cohort at risk at each failure time (=diagnosis of case).
- Any cohort study can be used as basis for generating a nested case-control study.

Nested case-control study

The technical term is to **sample the risk set**, i.e. instead of collecting exposure information on all individuals in the risk set, we only do it for a subsample of them.

Sampling the risk set



What are the risk sets here?

Draw two controls at random from the risk sets, and list the resulting matched sets.

The risk sets

Defined at each event time (•):

Event	Risk set	Sample
1		
2		
3		
4		

The risk sets

Event	Risk set	Controls
1	1,2,3,4,6,7,8,**9**,10,11	4,1
2	1,2,3,4,6,7,8,10,**11**	2,1
3	1,3,4,5,**6**,8,10	8,3
4	**1**,4,5,8	4,5

- Individuals 4 and 1 are used twice as controls.
- Individual 1 eventually becomes a case.
- Perfectly OK, because they are at risk at the time where they are selected to represent the risk set.

How many controls per case?

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

Equal number of cases and controls:

$$\sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} \approx \sqrt{\frac{1}{D_1}+\frac{1}{D_1}+\frac{1}{D_0}+\frac{1}{D_0}}$$

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

How many controls per case?

Twice as many:

$$\sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}} \approx \sqrt{\frac{1}{D_1}+\frac{1}{2D_1}+\frac{1}{D_0}+\frac{1}{2D_0}}$$

$$= \sqrt{\left(\frac{1}{D_1}+\frac{1}{D_0}\right)\times(1+1/2)}$$

m times as many:

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

- The standard deviation of the $\ln[\mathrm{OR}]$ is (approximately) $\sqrt{1+1/m}$ times larger in a case-control study, compared to the corresponding cohort-study.
- Therefore, 5 controls per case is normally sufficient: $\sqrt{1+1/5}=1.09$.
- Only relevant if controls are "cheap" compared to cases.
- If cases and controls **cost the same**, and cases are available the most efficient is to have the **same number** of cases and controls.

Individually matched studies

Bendix Carstensen

Matched and nested case-control studies
18 November 2016
Department of Biostatistics, University of Copenhagen
http://BendixCarstensen.com/AdvEpi

Individually matched study

- If strata are defined so finely that there is only one case in each, we have an **individually matched** study.
- The reason for this may be:
  - Comparability between cases and controls
  - Convenience in sampling
  - Controlling for age, calendar time (incidence density sampling)
  - Control for ill-defined factors

Individually matched study

- Pitfall in design:
- Overmatching (cases and controls are identical on some risk factors).
- Problem in analysis:
- Conventional method for analysis (logistic regression) breaks down, because we get one parameter per set (one parameter per case)!

Individually matched study

- If matching is on a well-defined **quantitative** variable as e.g. age, then broader stata may be formed *post hoc*, and age included in the model.
- $\Rightarrow$ assuming effect of age (matching variable) is continuous.
- If matching is on "soft" variables (neighborhood, occupation, ...) the original matching cannot be ignored:
- ... no way to have a continuous effect of a non-quantitative variable.
- $\Rightarrow$ matched analysis.

Salmonella Manhattan study

Telephone interview concerning the food items ingested 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	PARNR	KONTROL	HAMBURG
1	1	0	0
2	1	1	0
3	3	0	1
4	3	1	0
5	4	0	1
6	4	1	0
7	5	0	1
8	5	1	1
9	7	0	1
10	7	1	0
11	8	0	0
12	8	1	1
13	9	0	0
14	9	1	0
15	11	0	1
16	11	1	1

OBS	PARNR	KONTROL	HAMBURG
17	12	0	0
18	12	1	0
19	14	0	1
20	14	1	0
21	16	0	0
22	16	1	0
23	17	0	1
24	17	1	0
25	18	0	0
26	18	1	1
27	19	0	1
28	19	1	1
29	20	0	1
30	20	1	1
31	23	0	1
32	23	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

This is a table of **pairs**.

Remember: Exposure OR = Disease OR:

$$\text{OR} = \omega = \frac{\text{P}\{\text{E}+|\text{case}\}\,\text{P}\{\text{E}-|\text{control}\}}{\text{P}\{\text{E}-|\text{case}\}\,\text{P}\{\text{E}+|\text{control}\}}$$

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 the *Salmonella* data:

No. of matched pairs		Control exposure +	Control exposure −	
Case exposure	+			
	−			

OR estimated by:

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

Standard error on the log-scale:

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

Find approximate 95% c.i. for the OR:

Solution to exercise:

OR estimated by:

$$\hat{\omega} = \frac{b}{c} = \frac{6}{2} = 3.0$$

Standard error on the log-scale:

$$\text{s.e.}[\ln(\hat{\omega})] = \sqrt{\frac{1}{b} + \frac{1}{c}} = \sqrt{\frac{1}{6} + \frac{1}{2}} = 0.8165$$

Approximate 95% c.i. for OR:

$$3.0 \overset{\times}{\div} \exp(1.96 \times 0.8165) = (0.6055, 14.8636)$$

1:1 matched studies: — Test I

		Control exposure		
Pairs		$+$	$-$	
Case	$+$	a	b	$a+b$
exposure	$-$	c	d	$c+d$
		$a+c$	$b+d$	N

- McNemars test of OR$= 1$ compares b and c:

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

Problems of 1:1 matched studies

- If a single control is missing, the corresponding case is also lost.
- Large loss of information from trivial reasons.
- Normally more than one control per case is selected.
- But the $1:1$-matched study is useful for understanding the mechanics of the $1:m$-matched study.

1:1 matched studies: Parameters

What we really try to model is:

$$\text{odds(disease)} = \omega_P \theta_i \quad \Leftrightarrow \quad \mathrm{P}\{\text{disease}\} = \frac{\omega_P \theta_i}{1 + \omega_P \theta_i}$$

- ω_P — baseline odds for pair P
- — this is the irrelevant (nuisance) parameter
- θ_i — covariate effects for person i in the pair.
- Two persons in a pair — based on pair (P) and covariates:
- person $i = 1$: $\omega_1 = \omega_P \theta_1$
- person $i = 2$: $\omega_2 = \omega_P \theta_2$

1:1 matched studies: Likelihood

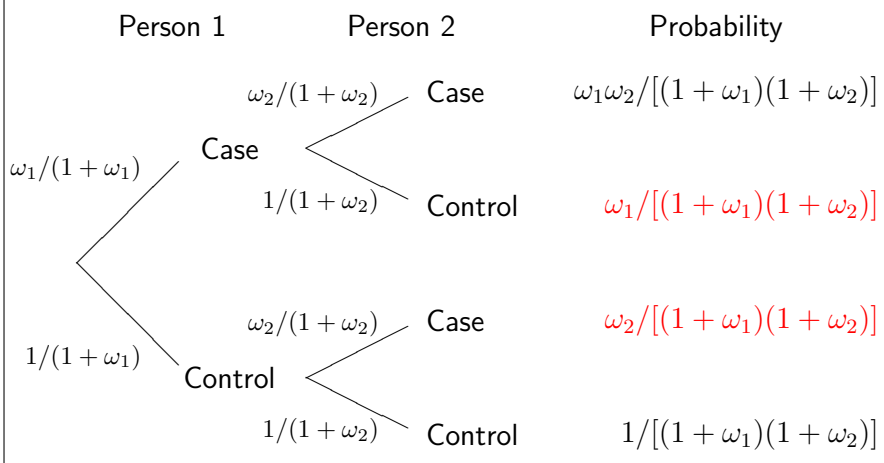
$$\text{odds(disease)} = \omega_P \theta_i$$
$$\ln[\text{odds(disease)}] = \ln[\omega_P] + \ln[\theta_i] = \boxed{\text{Cnr}_P} + \ln(\text{OR})$$

One parameter per pair: no. of parameters $\approx N/2$.

Profile likelihood approach breaks down, instead:

- Probability of data, **conditional** on design, i.e. on 1 case and 1 control per set.
- Distribution of covariates for case and control contains the information.

A set with 2 persons



Only the middle two outcomes need be considered.

Likelihood from one matched pair

$$L = \mathrm{P}\{\text{subj. 1 case} \mid \text{1 case, 1 control}\}$$

$$= \frac{\omega_1}{\omega_1+\omega_2} = \frac{\omega_P\theta_1}{\omega_P\theta_1+\omega_P\theta_2} = \frac{\theta_1}{\theta_1+\theta_2}$$

Log-likelihood contribution from one matched pair:

$$\log\left(\frac{\theta_{\text{case}}}{\theta_{\text{case}}+\theta_{\text{control}}}\right)$$

Independent of the parameters ω_P.

$1:m$ matching

Odds for disease in one matched set:

$$\begin{array}{lll} \text{person } 1: & \omega_P\theta_1 = & \omega_1 \\ \text{person } 2: & \omega_P\theta_2 = & \omega_2 \\ & \cdots & \\ \text{person } m+1: & \omega_P\theta_{m+1} = & \omega_{m+1} \end{array}$$

Probability that person 1 is the case, and the others are the controls:

$$\frac{\omega_1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \cdots \times \frac{1}{1+\omega_{m+1}}$$

$1 : m$ matching

Probability that person 2 is the case, and the others are the controls:

$$\frac{1}{1+\omega_1} \times \frac{\omega_2}{1+\omega_2} \times \cdots \times \frac{1}{1+\omega_{m+1}}$$

...

Probability that person $m+1$ is the case, and the others are the controls:

$$\frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \cdots \times \frac{\omega_{m+1}}{1+\omega_{m+1}}$$

Probability of 1 case and m controls:

$$\sum_i \frac{\omega_i}{(1+\omega_1) \times (1+\omega_2) \times \cdots (1+\omega_{m+1})}$$

$$= \frac{\sum_i \omega_i}{(1+\omega_1) \times (1+\omega_2) \times \cdots (1+\omega_{m+1})}$$

Conditional probability that person 1 is the case and persons $2, 3, \ldots, m+1$ are the controls, *given* one case and m controls:

$$\frac{\omega_1}{\omega_1 + \omega_2 + \cdots + \omega_{m+1}} = \frac{\theta_1}{\theta_1 + \theta_2 + \cdots + \theta_{m+1}}$$

— the ω_P is the same so it cancels

$1 : m$ matching

Log-likelihood contribution from one matched set:

$$\ell = \log\left(\frac{\theta_{\text{case}}}{\sum_{i \in \text{ cases \& controls}} \theta_i}\right)$$

Log-likelihood for the total study:

$$\ell = \sum_{\text{matched sets}} \log\left(\frac{\theta_{\text{case}}}{\sum_{i \in \text{ cases \& controls}} \theta_i}\right)$$

$1 : m$ matching

- Number of controls can vary between sets.
- Variable constant **within** matched sets: **impossible** to estimate a multiplicative effect:

$$\frac{\exp(\beta x_{\text{case}})\theta_{\text{case}}}{\sum_i \exp(\beta x_i)\theta_i} = \frac{\exp(\beta x)\theta_{\text{case}}}{\sum_i \exp(\beta x)\theta_i} = \frac{\theta_{\text{case}}}{\sum_i \theta_i}$$

- Over matching: $x_i = x$ within strata.
- **Interactions** between such variables and other variable **can** be estimated.
- In particular, interaction with matching variables can be estimated.

$1 : m$ matching

The conditional log-likelihood for a $1 : m$-matched CC-study looks like a Cox-log-likelihood:

$$\ell = \sum_{\text{failure times}} \ln\left(\frac{\theta_{\text{case}}}{\sum_{i \in \text{ Risk set}} \theta_i}\right)$$

The matched case-control likelihood is of this form if at each death time:

- The case dies.
- Only controls from the same set are at risk.

Use of `proc phreg`

- Input is a dataset with one observation per person.
- "Survival time" for controls $>$ for cases.
- Cases events, controls censorings.
- Matched set variable required for `strata`-command.
- Ties handling = `discrete`.
  (not really necessary if only one case per matched set).

This is what traditionally is recommended for programs that can handle a stratified Cox-model.

Use of proc phreg I

```
proc phreg  data = manh11 ;
  model kontrol * kontrol (1) = hamb / ties = discrete ;
  strata parnr ;
run ;

The PHREG Procedure

 Model Information
Data Set                  WORK.MANH11
Dependent Variable        kontrol
Censoring Variable        kontrol
Censoring Value(s)        1
Ties Handling             DISCRETE

          Summary of the Number of Event and Censored Values
                                                        Percent
Stratum    parnr          Total       Event    Censored Censored

      1    1                  2           1           1   50.00
      2    3                  2           1           1   50.00
      3    4                  2           1           1   50.00
      4    5                  2           1           1   50.00
      5    7                  2           1           1   50.00
      6    8                  2           1           1   50.00
      7    9                  2           1           1   50.00
```


Use of proc phreg II

```
      8    11                 2           1           1   50.00
      9    12                 2           1           1   50.00
     10    14                 2           1           1   50.00
     11    16                 2           1           1   50.00
     12    17                 2           1           1   50.00
     13    18                 2           1           1   50.00
     14    19                 2           1           1   50.00
     15    20                 2           1           1   50.00
     16    23                 2           1           1   50.00
-------------------------------------------------------------------
  Total                      32          16          16   50.00

 Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio         2.0930        1         0.1480
Score                    2.0000        1         0.1573
Wald                     1.8104        1         0.1785


 Analysis of Maximum Likelihood Estimates
              Parameter   Standard                              Hazard
Variable       Estimate      Error  Chi-Square    Pr>ChiSq      Ratio

hamb            1.09861    0.81650      1.8104      0.1785      3.000
```


How the *S.* Manhattan study REALLY was

```
             KONTROL
            0      1
 PARNR
 1          1      2
 3          1      2
 4          1      1
 5          1      3
 7          1      3
 8          1      2
 9          1      3
 10         .      2
 11         1      3
 12         1      3
 14         1      3
 16         1      3
 17         1      3
 18         1      3
 19         1      3
 20         1      3
 22         .      2
 23         1      3
```

```
proc phreg  data = manh ;
  model kontrol * kontrol (1) = hamb
        / ties = discrete ;
  strata parnr ;
run ;
```

```
The PHREG Procedure

        Model Information

Data Set                  WORK.MANH
Dependent Variable        kontrol
Censoring Variable        kontrol
Censoring Value(s)        1
Ties Handling             DISCRETE


Number of Observations Read          63
Number of Observations Used          63


 Summary of the Number of Event and Censored Values

                                                          Percent
Stratum    parnr    Total       Event    Censored       Censored

      1        1        3           1           2          66.67
      2        3        3           1           2          66.67
      3        4        2           1           1          50.00
      4        5        4           1           3          75.00
      5        7        4           1           3          75.00
      6        8        3           1           2          66.67
      7        9        4           1           3          75.00
      8       10        2           0           2         100.00
      9       11        4           1           3          75.00
```



```
     10       12        4           1           3          75.00
     11       14        4           1           3          75.00
     12       16        4           1           3          75.00
     13       17        4           1           3          75.00
     14       18        4           1           3          75.00
     15       19        4           1           3          75.00
     16       20        4           1           3          75.00
     17       22        2           0           2         100.00
     18       23        4           1           3          75.00
-------------------------------------------------------------------
  Total               63          16          47          74.60

        Testing Global Null Hypothesis: BETA=0

Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio         5.8323        1         0.0157
Score                    5.6749        1         0.0172
Wald                     4.9411        1         0.0262

Analysis of Maximum Likelihood Estimates
                  Parameter  Standard                                 Hazard
Parameter    DF   Estimate     Error  Chi-Square   Pr > ChiSq       Ratio

hamb          1    1.52985   0.68824      4.9411       0.0262       4.617

               Hazard      95% Hazard Ratio
Parameter      Ratio      Confidence Limits

hamb          4.617       1.198      17.792
```


Using proc logistic I

```
proc logistic  data = manh ;
  class parnr hamb(ref="0") ;
  model kontrol = hamb ;
  strata parnr ;
run ;

...

            Strata Summary
            kontrol
Response    -------    Number of
 Pattern    0     1       Strata    Frequency
       1    0     2            2            4
       2    1     1            1            2
       3    1     2            3            9
       4    1     3           12           48

...

              Analysis of Maximum Likelihood Estimates
                                 Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
hamb      1     1      0.7649      0.3441        4.9411        0.0262
```

Using `proc logistic` II

```
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
hamb      1     1      0.7649      0.3441        4.9411        0.0262

The LOGISTIC Procedure
 Conditional Analysis

             Odds Ratio Estimates
                  Point          95% Wald
Effect         Estimate      Confidence Limits

hamb 1 vs 0       4.617       1.198      17.792
```

Obs: $0.7648 = 1.5296/2, \quad \exp(1.5296) = 4.617$ — estimates from `proc logistic` are using the so-called Helmert-contrasts; a leftover from pre-computing times, difficult to understand and largely irrelevant in epidemiology.

Using `clogit` in Stata I

```
. use manh
. gen case = (pk==2)

. clogit case hamburg, group(parnr)

note: 2 groups (4 obs) dropped because of all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -17.713566
Iteration 1:   log likelihood =  -17.70835
Iteration 2:   log likelihood = -17.708349

Conditional (fixed-effects) logistic regression   Number of obs   =         59
                                                  LR chi2(1)      =       5.83
                                                  Prob > chi2     =     0.0157
Log likelihood = -17.708349                       Pseudo R2       =     0.1414

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hamburg |   1.529847   .6882356     2.22   0.026     .1809297    2.878763
------------------------------------------------------------------------------
```


Using `clogit` in Stata II

```
. clogit case hamburg, group(parnr) or

note: 2 groups (4 obs) dropped because of all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -17.713566
Iteration 1:   log likelihood =  -17.70835
Iteration 2:   log likelihood = -17.708349

Conditional (fixed-effects) logistic regression   Number of obs   =         59
                                                  LR chi2(1)      =       5.83
                                                  Prob > chi2     =     0.0157
Log likelihood = -17.708349                       Pseudo R2       =     0.1414

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hamburg |   4.617468   3.177906     2.22   0.026     1.198331    17.79226
------------------------------------------------------------------------------
```

Using `clogistic` in R I

```
> library(foreign)
> manh <- read.dta("../data/manh.dta")
> library(Epi)
> mh <- clogistic( (pk=="P")*1 ~ hamb, strata=parnr, data=manh )
> mh
Call:
clogistic(formula = (pk == "P") * 1 ~ hamb, strata = parnr, data = manh)

     coef exp(coef) se(coef)    z     p
hamb 1.53      4.62    0.688 2.22 0.026

Likelihood ratio test=5.83  on 1 df, p=0.0157, n=48
> ci.exp(mh)
     exp(Est.)    2.5%    97.5%
hamb  4.617463 1.19833 17.79223
```


Matched studies in practice

- Think of the scenario where extensive follow-up and all measurements were available for all persons in the cohort.
- Use "history" of a person as predictor of mortality / morbidity.
- Definition of "history":
  - Original treatment allocation.
  - Profile of measurements over time.
  - Genotype.
  - ...

Definition of history

- Is the entire profile of measurements relevant:
  - Only the most recent.
  - Only measurements older than 1 year, say (latency).
  - Cumulative measures?
- What are the relevant summary measures of a persons history.
- Age (current age, age at entry)
- Calendar time (current or at entry)
- Exposure history

Selecting controls: Incidence density sampling

- Timescale:
  Controls should be alive when the corresponding case dies.
- More than one time-scale:
- e.g. age and calendar time:
- Match on:
  - date of event (calendar time)
  - date of birth (and hence age at event).
- Ensure comparability of covariates within matched sets.

Summary

- Case-control study:
  Select persons based on **outcome** status.
- Nested case-control studies saves money when **extra** information on persons must be collected.
  Logistic regression.
- If **all** information is in the cohort it is always better to analyze the full cohort.
- **Individually** matched case-control studies for control of ill-defined variables.
  Conditional logistic regression.