Διάλεξη 7.

MATCHED CASE-CONTROL STUDIES

1. Introduction:

- In case-control studies, controls may be randomly selected from the population of individuals free from the condition that defines the cases.
- Controls can be “matched” to cases with respect to factors that are related to the risk of disease.
- Variables that are usually used for matching are age, sex, place of recruitment and time of recruitment.
- Due to the special design, matched case control studies require special analyses.

2. Why match?

- Matching is a technique of selecting control subjects for the control of confounding at the design stage
- *Idea:* Place constraints on selection of controls to make two groups similar at least with respect to confounding variables.
    - In matched case-control studies, for each case or a fixed size group of cases, a fixed (or even variable) number of controls are identified who match the cases on a set of characteristics.
- The distribution of these characteristics will be the same (or at least similar) between cases and controls, so no associations are possible *by design*.
- During the analysis of the results: Post-stratification analysis

Why Match?

- **Deal with bias due to confounding**
  - Matching on *Confounder (C)* forces no association between *C* and Disease (*D*), so *C* cannot confound.

- This gain in precision occurs when the matching variable (C) is associated with both exposure status (E) and the disease occurrence (D) in the source population, so that we would need to control for C as the confounder even if matching were not done.
- Occasionally to test a particular pathway

Major Statistical Advantage of Matching

- In both case-control and cohort studies we aim to **reduce the variance** of adjusted estimators, at a given sample size.
  - This goal is especially important when there is a limited number of diseased individuals (cases).

- Balance cases/controls within strata to improve efficiency, i.e achieve a given performance using fewer observations
- Thus, the major statistical reason for matching is not to control for confounders, which can be done in the analysis, but to produce a **more efficient study** (one that yields an estimator with a smaller variance for a given sample size) than if we had not matched.

More on matching

- Controls can be individually matched or frequency matched.
- **Individual matching:** Search for one (or more) controls who have the required matching criteria. Paired or triplet matching is when there is one or two controls individually matched to each case.
- **Frequency matching:** select a population of controls such that the overall characteristics of the group match the overall characteristics of the cases. e.g. if 15% of cases are under age 20, 15% of the controls are also.
- Gain power by matching more than one control per case.
  - Number of controls should be < 4, because there is no further gain of power above four controls per case.

2.1. Example

Consider the following example: The BCG vaccination and leprosy:

New cases of leprosy examined for presence or absence of the BCG scar. Say we identified 260 cases of leprosy. Assume we use 1000 controls for the 260 cases. After stratification by age:

	BCG scar			
	Cases		Controls	
Age	Absent	Present	Absent	Present
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

Not very efficient! There are 238 controls for the 2 cases in the 0 - 4 age group!

2.2 Group matching

The optimal strategy is to maintain the same ratio of controls to cases in different age strata

For example in the previous study we could maintain the 1:4 case/control ratio as shown below

	BCG scar			
	Cases		Controls	
Age	Absent	Present	Absent	Present
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

This is a group-matched case-control study.

Caution!

- Controls are no longer representative of source population

- **Matching introduces bias if not taken into account in the analysis**!

2.3 Can we ignore matching in the analysis?

Indeed it was thought that matching is an alternative way of controlling for confounding - **this is not true**; see the example below:

	Cases		**Controls**		**Odds ratio**
Stratum	**exposed**	**unexposed**	**exposed**	**unexposed**	
1	89	11	80	20	2
2	67	33	50	50	2
3	33	67	20	80	2
Total	189	111	150	150	1.7

Odds ratio is biased towards 1, i.e., towards the null. This turns out to be a general result!

A case-control study introduces a new confounding structure in place of the original structure and this is why the estimate from an analysis that ignores matching is biased towards the null. Remember:

Matched design =======> «Matched» analysis

3. Advantages of a matched design

Precision / efficiency in a matched case-control
When the analysis of a study involves stratification on the basis of some confounding variable, the precision of the study will usually be maximal if the ratio of cases to controls is approximately the same across strata. We can succeed on this by a matched design.

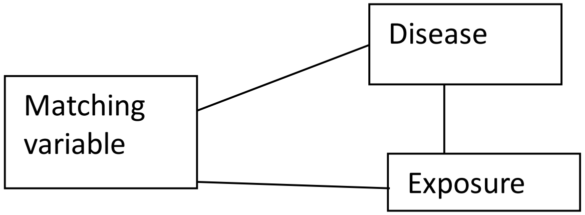
Study 1 case: control ratio = 1:4

	Exposed	Unexposed	Total	
Cases	30	10	40	= 3.0
Controls	80	80	160	(1.30,7.07)
	100	90	200	

Study 2 case: control ratio = 1:1

	Exposed	Unexposed	Total	
Cases	75	25	100	= 3.0
Controls	50	50	100	(1.58,5.72)
	125	75	200	

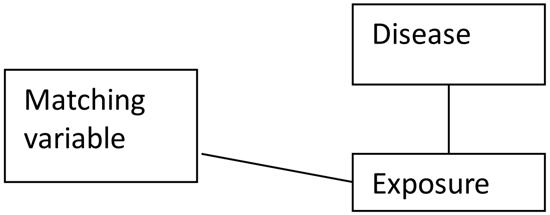
The power of a case-control study of total sample N to detect a difference in exposure rates between cases and controls is greatest if number of cases equals number of controls.


Disease

Matching variable

Exposure

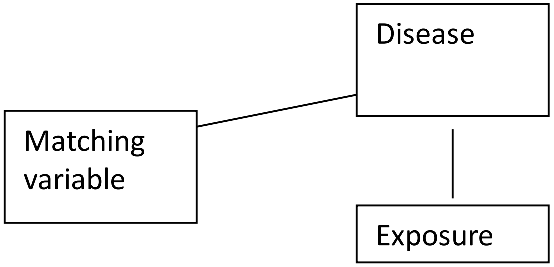
Matching variable is a confounder – matching will gain us precision in the exposure/disease relationship


Disease

Matching variable

Exposure

Overmatching – precision is lost

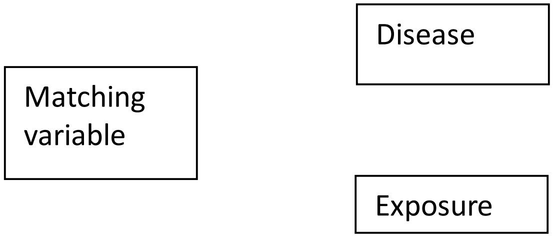

Disease

Matching variable

Exposure

UNNECASSARY MATCHING: Matching can be ignored in the analysis since its effect is neutral

If analysis with stratification→ reduction of power


Disease

Matching variable

Exposure

UNNECASSARY MATCHING: Matching can be ignored in the analysis since its effect is neutral

If analysis with stratification→ reduction of power

Overmatching

- Controls are supposed to provide an estimate of the distribution of the exposure in the source population.
- Matching by a factor associated with exposure makes the **controls more similar to the cases with respect to exposure**
  - **This biases the crude estimate towards the null no matter what the direction of the association between matching factor and exposure**!

Overmatching

- Matching on a variable which is associated with exposure but not with disease should be avoided because this in practice **will reduce power** – the more the association with exposure the more the reduction will be.

- In general it is only worthwhile matching on variables which are strong confounders.

- And do not forget that:
  - Matching must be taken into account in the *analysis*.
  - Attempting to match for more than a few variables usually inefficient.

4. Disadvantages of matched studies

- The association of the matching variable with the outcome cannot be studied: By definition the distribution of a matching variable is the same (or similar) in the case and control groups

- Logistically more difficult

- Data may be more difficult to present and analyze.

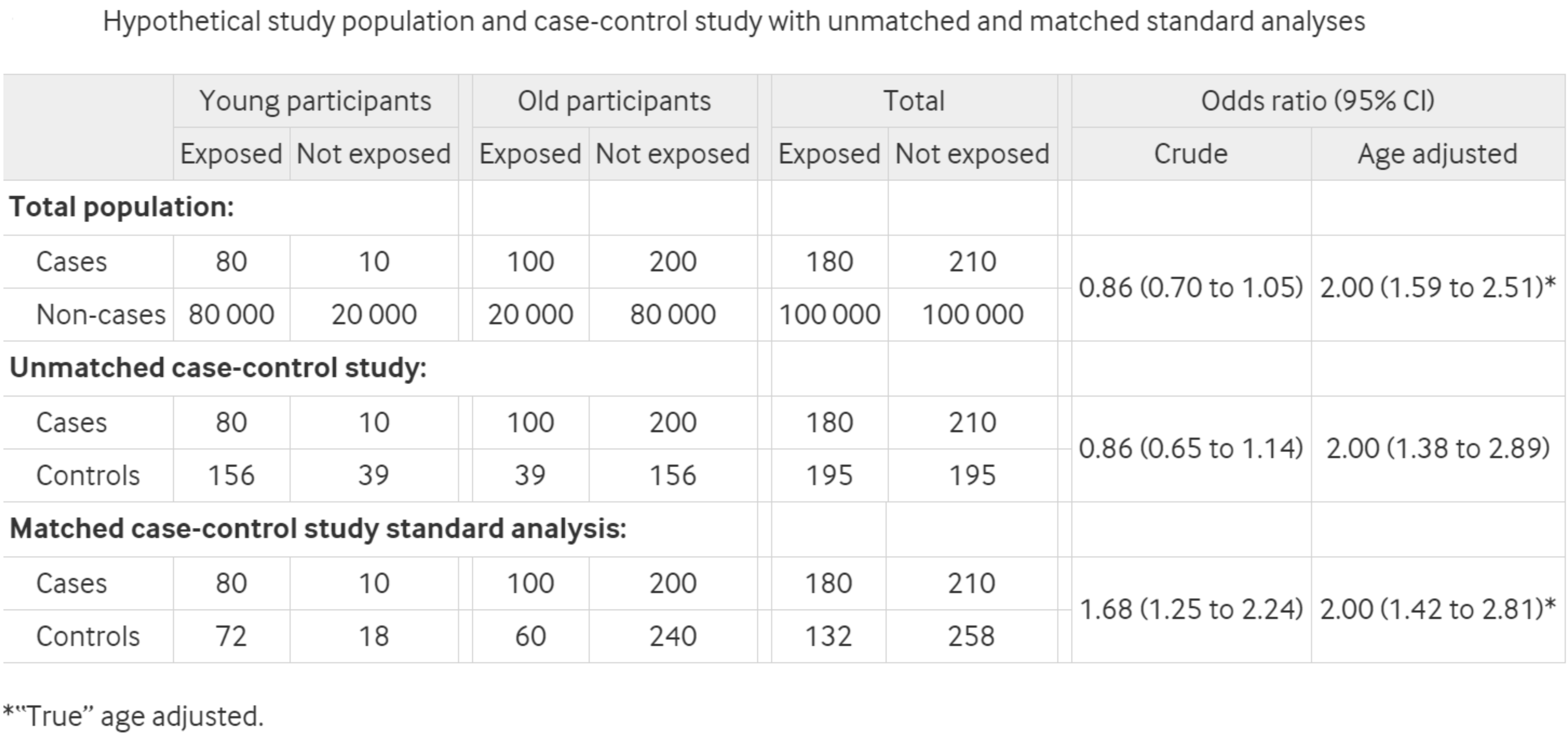
- May be difficult to find suitable matches. May reduce available sample size – many potential cases may be excluded because no match can be found

- Possibility of «overmatching».
  - on variable associated with exposure but not disease: power loss.
  - on variable in causal pathway: bias.

5. Analysis of grouped matched case-control studies

- A 1:1 matched design does not always requires a matched analysis.
- Control for the matching factors can be obtained, with no loss of validity and a possible increase in precision, using a “standard” (unconditional) analysis, and a “matched” (conditional) analysis may not be required or appropriate
  - Assumption: There are no problems of sparse data

Example



Hypothetical study population and case-control study with unmatched and matched standard analyses

	Young participants		Old participants		Total		Odds ratio (95% CI)	
	Exposed	Not exposed	Exposed	Not exposed	Exposed	Not exposed	Crude	Age adjusted
Total population:								
Cases	80	10	100	200	180	210	0.86 (0.70 to 1.05)	2.00 (1.59 to 2.51)*
Non-cases	80 000	20 000	20 000	80 000	100 000	100 000		
Unmatched case-control study:								
Cases	80	10	100	200	180	210	0.86 (0.65 to 1.14)	2.00 (1.38 to 2.89)
Controls	156	39	39	156	195	195		
Matched case-control study standard analysis:								
Cases	80	10	100	200	180	210	1.68 (1.25 to 2.24)	2.00 (1.42 to 2.81)*
Controls	72	18	60	240	132	258		

*"True" age adjusted.

Pearce N. Analysis of matched case-control studies. BMJ 2016;352

5. Analysis of grouped matched case-control studies

In case we use standard analysis, (i.,e. unconditional logistic regression):

Matching variables should be in the logistic regression model in order to get unbiased estimates of the effects of interest.

Example: Consider the previous example on leprosis and say we matched for age with age being a categorical variable with k levels.

The model

$$\log(\text{odds}_i) = \alpha + \Sigma\, \beta_{1k}\, \text{age}_{i\kappa} + \beta_2\, \text{BCG}_i$$

Parameter	**Estimate**	**SD**
Cons	-1.07	0.8
Age(1)	-0.04	0.83
Age(2)	0.012	0.81
Age(3)	0.07	0.8
Age(4)	0.024	0.82
Age(5)	-0.16	0.81
Age(6)	-0.24	0.81
BCG	-0.53	0.16

Note that because of matching the age effects are small and not interpretable. But can we remove age from the model?

Example (continued):

Removing age :

BCG scar	**Leprosy cases**	**Controls**
Present	101	526
Absent	159	514

The odds ratio is (101 x 514) / (159 x 526) = 0.621, so that the log of odds is -0.477 i.e, biased towards the null.

Note that the age parameters are really nuisance parameters but they are still estimated.

In case of many of nuisance parameters -- this approach does not work

e.g. when we match *individually*, which is effectively the perfect matching!

6. Matched pairs (1 : 1)

Suppose we have n matched pairs. **Each pair can be thought of as a stratum.** For each stratum (pair) there are four possible outcomes as follows:

	Exposure								Total
	+	-	+	-	+	-	+	-	
Case	1	0	1	0	0	1	0	1	
Control	1	0	0	1	1	0	0	1	
	2	0	1	1	1	1	0	2	
Total no. of pairs of each kind	n_{11}		n_{10}		n_{01}		n_{00}		n

Where n_{ij} corresponds to the number of pairs with exposure status i (0=unexposed, 1=exposed) for the case and j (0=unexposed, 1=exposed) for the control.

The results of an 1:1 matched, case-control study can therefore be presented in a table of the form:

		Control	
		Exposed	Unexposed
Case	Exposed	n_{11}	n_{10}
	Unexposed	n_{01}	n_{00}

From this table we can easily obtain the following table that contains *individuals.*

	Exposure		Total
	+	-	
Case	$n_{11} + n_{10}$	$n_{00} + n_{01}$	n
Control	$n_{11} + n_{01}$	$n_{00} + n_{10}$	n

6.1 Estimating the odds ratio from a 1:1 matched case control

Let’s see this in detail:

Exposure

	+	-	+	-	+	-	+	-	Total
Case	1	0	1	0	0	1	0	1	1
Control	1	0	0	1	1	0	0	1	1
Total	2	0	1	1	1	1	0	2	
# of such tables	n_{11}		n_{10}		n_{01}		n_{00}		2

We will consider a stratified analysis, where each matched pair is a stratum. So the Mantel Haenszel estimate will be:

$$MHOR = \frac{\Sigma D_{1j} H_{0j} / N_j}{\Sigma D_{0j} H_{1j} / N_j}$$

For the first case, where both case and control are exposed and where we have n_{11} pairs, the contribution of each of them to the MH estimate is: $n_{11}*D_{11}H_{01}/N_1$ for the numerator and $n_{11}*D_{01}H_{11}/N_1$ for the denominator

6.1 Estimating the odds ratio from a 1:1 matched case control

Thus, the Mantel-Haenszel estimate considering each stratum=matched pair is:

$$\boldsymbol{MHOR} = \frac{\Sigma \boldsymbol{D}_{1j}\boldsymbol{H}_{0j} / \boldsymbol{N}_j}{\Sigma \boldsymbol{D}_{0j}\boldsymbol{H}_{1j} / \boldsymbol{N}_j} = \frac{\boldsymbol{Q}}{\boldsymbol{R}} = \frac{[(\mathrm{n}_{11}\mathrm{x}0) + (\mathrm{n}_{10}\mathrm{x}1) + (\mathrm{n}_{01}\mathrm{x}0) + \mathrm{n}_{00}\mathrm{x}0)]/2}{[(\mathrm{n}_{11}\mathrm{x}0) + (\mathrm{n}_{10}\mathrm{x}0) + (\mathrm{n}_{01}\mathrm{x}1) + (\mathrm{n}_{00}\mathrm{x}0)]/2} = \frac{\boldsymbol{n}_{10}}{\boldsymbol{n}_{01}}$$

D_{1j}: Number of exposed cases in pair j

D_{0j}: Number of unexposed cases in pair j

H_{1j}: Number of exposed controls in pair j

H_{0j}: Number of unexposed controls in pair j

N_{0j}: Number of individuals in pair j

Pairs with case and control being both exposed or both unexposed do not contribute to the odds ratio estimate.

Example

- Suppose a matched case control study has been conducted to investigate risk factors for infant death from diarrhoea (Clayton and Hills).
- **Cases were defined as infants dying from diarrhoea** at less than 1 year of age.
- These cases were matched with 1 *neighborhood* control who had to be the same age group (0-2, 3-5, ≥6 months) as the case also (two matching variables).
- The study included 86 cases and 86 controls.
- Among other variables, information on social and environmental factors, birth weight and feeding mode were also collected.
- See in the following table this case control study with exposure being the breastfeeding mode.

		Control	
	Feeding mode	Breast fed	Not breast fed
Case	Breast fed	24	6
	Not breast fed	29	27

MH odds ratio from the matched table: 6/29 = 0.21

Example:

Ignoring matching

		Control	
	Feeding mode	Breast Fed	No Breast Fed
Case	Breast fed	30	56
	Not breast fed	53	33

Odds ratio ignoring matching (30*33)/(56*53) = 0.33 bias towards the null

6.2. Confidence interval for the MHOR for the 1:1 matched case control study

An approximate 95% confidence interval for the odds ratio may be calculated using the method given in previous lectures. Recall that the error factor was:

$$\boldsymbol{EF} = \exp(1.96\boldsymbol{xS}) \qquad \boldsymbol{where\ S}^2 = \frac{\boldsymbol{V}}{\boldsymbol{QR}}$$

Note that:

$$\boldsymbol{Q} = \frac{\boldsymbol{n}_{10}}{2}, \quad \boldsymbol{R} = \frac{\boldsymbol{n}_{01}}{2}, \quad \boldsymbol{V} = \frac{\boldsymbol{n}_{10}}{4} + \frac{\boldsymbol{n}_{01}}{4}, \qquad S^2 = 1/n_{10} + 1/n_{01}$$

$$\boldsymbol{EF} = \exp[1.96\sqrt{(1/\boldsymbol{n}_{10} + 1/\boldsymbol{n}_{01})}]$$

concordant pairs contribute nothing to the confidence interval.

This approximation brakes down when the number of discordant pairs is small (e.g. less than 20)→ «exact» 95% confidence intervals

6.3. Test of the null hypothesis that the true MHOR = 1

- Test of the null hypothesis that the true odds ratio is 1, based only on the discordant pairs.
- **When the true odds ratio is 1 the probability of a discordant pair to be of either type, should be 0.5→ E(n_{10}) = (n_{10}+n_{01})/2.**

➢ Instead of OR=1, test whether n_{10} differs from its expected value under the null hypothesis.

- For large numbers of discordant pairs (> 20) → Normal approximation to the Binomial distribution

Under the null hypothesis p=0.5, var(n_{10}) = np(1-p) = (n_{10}+n_{01})/4
Using the Normal approximation on the Binomial distribution gives:

$$\boldsymbol{x}^2 = \frac{(\boldsymbol{n}_{10} - \boldsymbol{E}(\boldsymbol{n}_{10}))^2}{\boldsymbol{Var}(\boldsymbol{n}_{10})} = \frac{(\boldsymbol{n}_{10} - (\boldsymbol{n}_{10} + \boldsymbol{n}_{01})/2)^2}{(\boldsymbol{n}_{10} + \boldsymbol{n}_{01})/4} = \frac{(n_{10} - n_{01})^2}{(n_{10} + n_{01})} \qquad \text{On 1 DF}$$

McNemar's test for matched pairs = MH χ^2 test, using pairs as strata.

No of discordant pairs ≤ 20 (say) → exact test based upon the Binomial distribution
Table of cumulative probabilities of the Binomial distribution with a value of p = 0.5 (null hypothesis value).

6 .4. Testing for heterogeneity of the odds ratio

Matching variable -- confounding variable.

Test whether matching factor is an effect modifier of the association of exposure with the outcome of interest.

Straight forward for group-matching variables.

1:1 matched study: levels of the matching factor (e.g. age groups) and estimate the odds ratio by the pairs in each subgroup.

Wide groups for the matching factor → enough number of discordant pairs in each

For example, the pairs may be closely matched for age (e.g. ± one year), but the subgroups may be defined by 10-year age groups.

	Matching factor						
	1	2	3	i		k	Total
No of Pairs with Case exposed and Control unexposed							
No of Pairs with Case unexposed and Control exposed							

χ^2 test for a 2 x k table → tests whether the odds ratio estimates vary according to the level of the matching factor.

If matching factor is on an ordinal scale then a test for trend can also be used.

Example (cont)

Say we want to assess the effect of birth weight (low vs. normal) on risk of death from diarrhoea. Look below the crude estimate of the odds ratio.

		Control	
	Birth weight	Low	Normal
Case	Low	12	25
	Normal	18	31

OR = 25/18 = 1.39 (0.76, 2.55)

We have matched for age because age is a confounding variable but we want to check whether low birth weight has a greater effect on the risk of death from diarrhea among younger infants than among older infants. Since the data are matched for age, we may stratify the pairs into three age groups as follows:

OR_1 = 7/6 = 1.17
OR_2 = 12/7 = 1.71
OR_3 = 6/5 = 1.20

		Age					
		0-2 months		3-5 months		≥ 6 months	
		Control		Control		Control	
	Birth weight	Low	Normal	Low	Normal	Low	Normal
Case	Low	4	7	4	12	4	6
	Normal	6	8	7	15	5	8

χ^2 = 0.35 on 2 df, p>0.5 → no evidence for a modifying effect of age on the odds ratios. Odds ratio of 1.39 is the association of low birth weight on the risk of death from diarrhoea adjusted for both neighborhood and age.

7. The analysis of 1:k matched case control studies

>1 controls per case recruited the number of possible outcomes increases.
E.g. 2 controls per case there are six possible outcomes for each triplet
Previous methods can be extended to the general case of 1:k matched case control studies.

$$\text{OR} = \frac{\text{Total no. of unexposed controls who hove an exposed case}}{\text{Total no. of exposed controls who have an unexposed case}}$$

Formulas for these situations and approximate confidence intervals have also been established.

8. Adjustment for other factors

NOT POSSIBLE through stratification (i.e. MH) since the data are already stratified into pairs of cases and controls so that no further stratification is possible.

Use statistical modeling techniques.

8. Analysis of matched case-control studies using statistical models

Use logistic regression with a separate parameter for each case-control set:

$$\log(\boldsymbol{odds}_i) = \boldsymbol{a} + \sum_{j=1}^{m} \boldsymbol{\gamma}_{\mathbf{j}}\, \mathbf{z}_{\mathbf{ij}} + \sum_{k=1}^{p} \boldsymbol{\beta}_{\mathbf{k}}\, \mathbf{x}_{\mathbf{ik}}$$

x_{ik} are the exposure and possible confounders, and z_{ij} are dummies with 1 if subject *i* is in matched set *j*, and 0 otherwise.

β_k are still interpreted as estimates of the population odds ratios associated with certain levels of the x_{ik} variables.

For large number of sets usual properties of MLEs do not apply; parameter estimates will not be consistent:

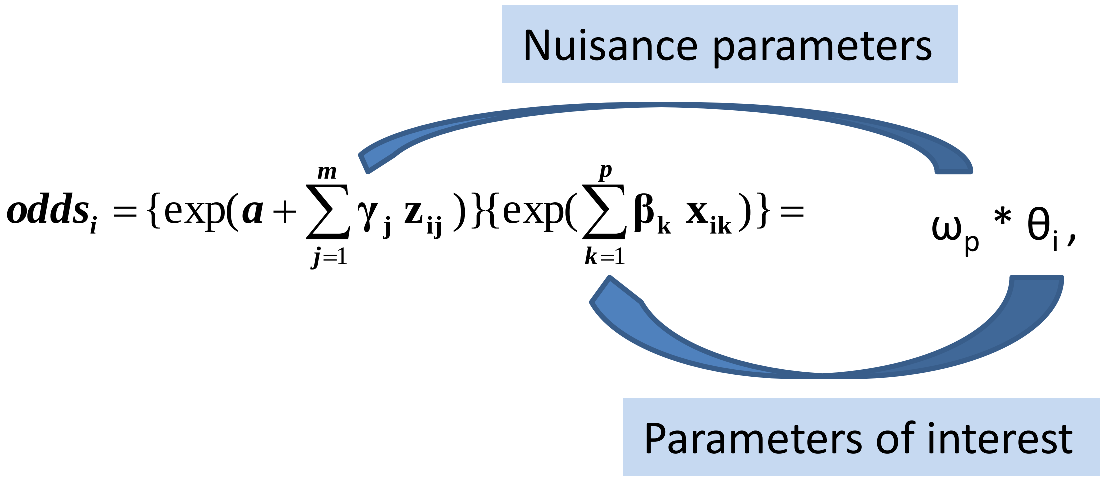
1) Assume that the matched set parameters γ_j are themselves a sample from some distribution - i.e, set up a *mixed (random effects) model*, or,

2) Perform conditional logistic regression

The model for conditional logistic regression

$$1st\ part = a + \sum_{j=1}^{m} \gamma_j \, \mathbf{z}_{ij} \qquad\qquad 2nd\ part = \sum_{k=1}^{p} \boldsymbol{\beta}_k \, \mathbf{x}_{ik}$$

Or,



Nuisance parameters

$$odds_i = \{\exp(a + \sum_{j=1}^{m} \gamma_j \, \mathbf{z}_{ij})\}\{\exp(\sum_{k=1}^{p} \boldsymbol{\beta}_k \, \mathbf{x}_{ik})\} = \quad \omega_p * \theta_i ,$$

Parameters of interest

Eliminate nuisance parameters using the *conditional likelihood:* the pair of the control/case matched set is used as the unit for the analysis.

Only β_k are estimated and reported
α is not estimated and not reported

1:1 matched studies - Parameters

odds(disease)=$\omega_P\vartheta_i$

$\boldsymbol{\omega_P}$: baseline odds of pair *P, s*pecific of each pair because of matching.

$\boldsymbol{\vartheta_i}$: covariate effects for subject *i* (a function of covariate values for subject *i*).

For each pair p we have the same baseline odds, different exposure level:

Disease odds for subject 1: $\omega_P\vartheta_1 = \omega_1$

Disease odds for subject 2: $\omega_P\vartheta_2 = \omega_2$

ln[odds(disease)]=ln[ω_P] + ln[ϑ_i]= C_P + ln(OR)

One parameter per pair, i.e. number of parameters =~ N/2.

Profile likelihood breaks down.

Conditional likelihood

Solution:

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.

1:1 matched studies – Conditional likelihood

Conditional on the design one case and one control in each set, a set would contribute:

L = P(subj. 1 case | 1 case, 1 control)

To the likelihood

Taking into account

1. P(1 **case**, 1 control |subj. 1 case)=P(subj 2 control)
2. P(disease)=$\omega_P\vartheta_i / (1+\omega_P\vartheta_i)$, P(no disease)=$1 / (1+\omega_P\vartheta_i)$
3. P(A|B)=P(B|A)*P(A)/(P(B|A)*P(A)+P(B|A-)*P(A-))

1:1 matched studies – Conditional likelihood

L = P(subj. 1 case | 1 case, 1 control)=

P(1 case, 1 control|subj. 1 case)*P(subj. 1 case)/(P(1 case, 1 control|subj. 1 case)*P(subj. 1 case)+P(1 case, 1 control|subj. 1 control)*P(subj. 1 control))

=P(subs 2 control)*P(subj. 1 case)/(P(subj. 2 control)*P(subj. 1 case)+P(subj 2 case)*P(subj. 1 control))=

$$= K\omega_1/(K\omega_1+K\omega_2) = K\omega_P\vartheta_1 \,/\, (K\omega_P\vartheta_1+K\omega_P\vartheta_2)$$
$$= \vartheta_1 \,/\, (\vartheta_1+\vartheta_2)$$

where

$K = [1/(1+\omega_1)]*[1/(1+\omega_2)] = 1/[(1+\omega_1)(1+\omega_2)]$

Log-likelihood contribution from one matched pair is:

$$\ln[\vartheta_{\text{case}}/(\vartheta_{\text{case}}+ \vartheta_{\text{control}})]$$

Independent of the corner parameters!

1:M matching

Odds for disease on one matched set:

subject 1: $\omega_P\vartheta_1 = \omega_1$

subject 2: $\omega_P\vartheta_2 = \omega_2$

subject *m*+1: $\omega_P\vartheta_{m+1} = \omega_{m+1}$

Probability that subject 1 is the case and the others are the controls: $[\omega_1/(1+\omega_1)]*[1/(1+\omega_2)]*\ldots*[1\,/(1+\omega_{m+1})]$

Probability to have 1 case and *m* controls:

$\Sigma_i\{\omega_i/[(1+\omega_1)*(1+\omega_2)*\ldots*(1+\omega_{m+1})]\}$

$= \ \Sigma_i\omega_i\,/[(1+\omega_1)*(1+\omega_2)*\ldots*(1+\omega_{m+1})]$

Conditional probability that subject 1 is the case and subjects 2, 3, …, *m*+1 are the controls, *given* one case and *m* controls:

$$\omega_1/(\omega_1+ \omega_2+\ldots+ \omega_{m+1}) = \vartheta_1/(\vartheta_1+ \vartheta_2+\ldots+ \vartheta_{m+1})$$

1:M matching

Log-likelihood contribution from one matched set:

$$l = \ln\left(\frac{\theta_{case}}{\sum_{i \in cases \& controls} \theta_i}\right)$$

Log-likelihood for the total study:

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

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

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

The matched CC likelihood is of this form if at each death time, the case dies and only controls of the same set are at risk.

Analysis of conditional likelihood by ordinary logistic regression

Likelihood contribution from one matched pair is:

$$\vartheta_{\text{case}}/(\vartheta_{\text{case}}+\vartheta_{\text{control}})=(\vartheta_{\text{case}}/\vartheta_{\text{control}})/(1+\vartheta_{\text{case}}/\vartheta_{\text{control}})=\omega/(1+\omega)$$

This is the likelihood contribution from one binary observation with odds of success $\omega = \vartheta_{\text{case}}/\vartheta_{\text{control}}$

Linear model for $\ln(\vartheta)$

$$\ln(\vartheta_{\text{case}}) = \text{Corner}+\text{Set}+A_{\text{case}}$$

$$\log(\boldsymbol{odds_i}) = \boldsymbol{a} + \sum_{j=1}^{m} \boldsymbol{\gamma}_{\mathbf{j}}\ \mathbf{z}_{\mathbf{ij}} + \sum_{k=1}^{p} \boldsymbol{\beta}_{\mathbf{k}}\ \mathbf{x}_{\mathbf{ik}}$$

leads to (for one matched pair)

$$\ln(\omega) = \ln(\vartheta_{\text{case}}) - \ln(\vartheta_{\text{control}})$$

$$= (\text{Corner}+\text{Set}+A_{\text{case}}) - (\text{Corner}+\text{Set}+A_{\text{control}})$$

$$= A_{\text{case}} - A_{\text{control}}$$

Corresponds to logistic regression without intercept.

One observation per matched set.

Covariates are: covariate-value for case – covariate-value for control

Logistic regression without intercept. “Through the origin”.

1:1 matched studies by ordinary logistic regression

- The information is in the covariates:
- Continuous covariate: $\text{Age}_{\text{case}} – \text{Age}_{\text{control}}$.
- Differences between dummies, value for case minus value for control.
- Categorical covariate, dummies replaced by variables with values -1, 0 or 1:
  - if case and control belong to the same category all are = 0
  - if case and control belong to different categories:
  - 1 for the category where the case is.
  - -1 for the category where the control is.
  - 0 for the other categories.
- ONLY possible for 1:1 matched studies.

Choice of controls

Source	Potential advantages	Potential disadvantages
Hospital/health facility	Controls likely to have been recruited as cases if ill Cheap?	Need to exclude controls with conditions that could be related to exposure of interest
Neighbourhood	Control a range of factors Simple rule	If wide range of care providers, need to ensure controls would have been recruited as cases Expensive if cases widely dispersed
Friends/siblings	Likely to be co-operative	Overmatching? As for neighbours
Telephone	Cheap	Excludes individuals without phones Bias towards people who stay in? Quality of data?