Cohort and Case-control studies

Lesley Rushton
Epidemiology and Public Health
l.rushton@imperial.ac.uk

Cohort and Case-Control Studies

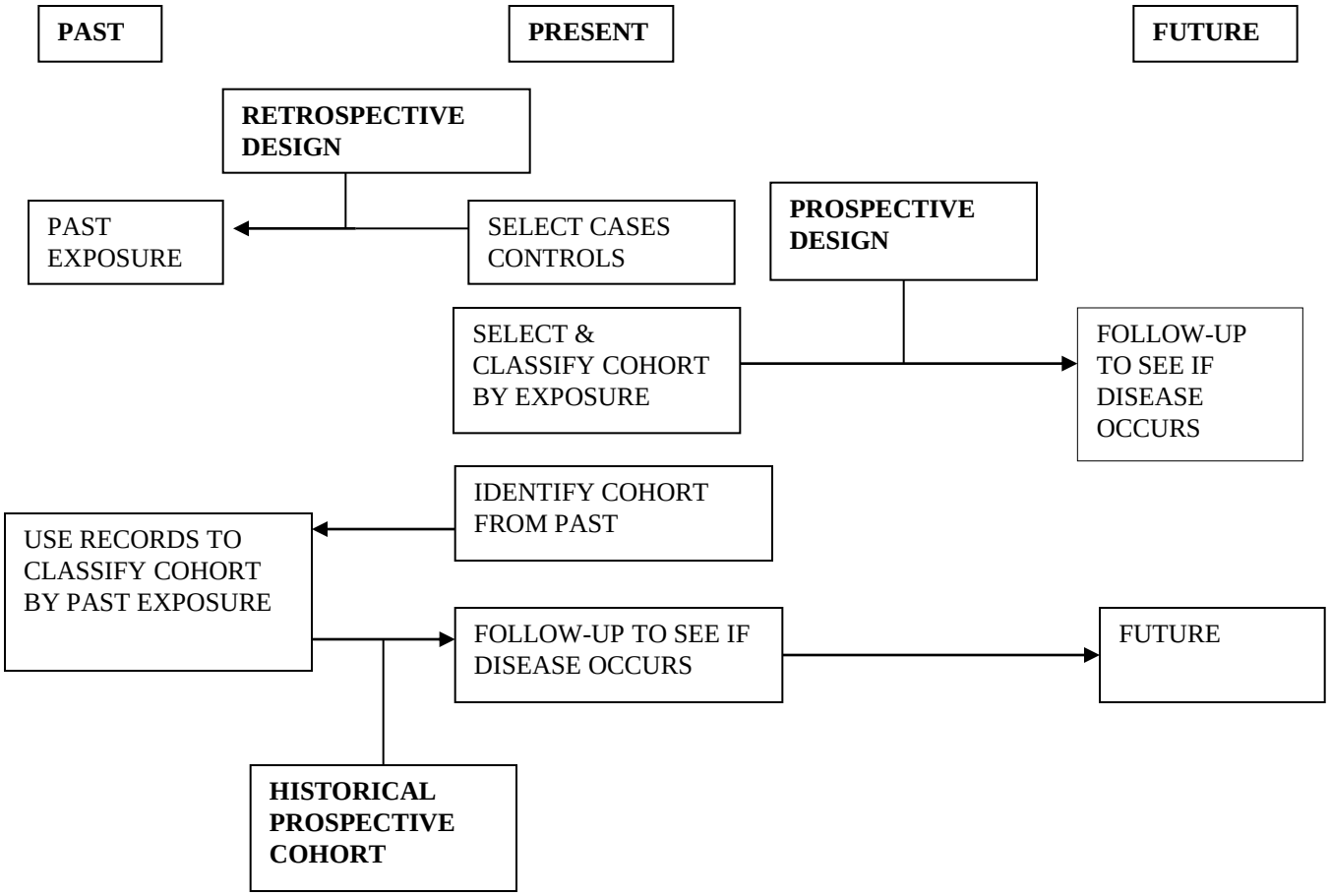
A COHORT is a group of people with particular characteristics in common, who are then observed over a period of time to see what happens to them.

COHORT STUDIES search for associations between previously defined characteristics of the cohort and the development of disease (NB sometimes called follow-up or prospective studies). The rate at which the disease develops in a group in which a certain characteristics is present is compared with a group where the characteristic is absent.

In a CASE-CONTROL study individuals with a particular condition or disease (the CASES) are compared with a group of individuals without the disease (the CONTROLS). Information on past exposure to possible risk factors is then obtained for both cases and controls. The amount of exposure in the cases is compared with that in the controls.

(NB sometimes called a retrospective study)

Retrospective and Prospective Study Design

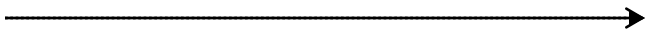


COHORT works from cause to effect
CASE-CONTROL works from effect to cause

EXAMPLE

To investigate the association between exposure to benzene in oil refinery workers and death from leukaemia

COHORT

Refinery workers  Leukaemia

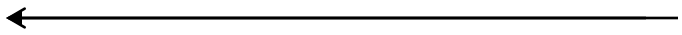
(exposed to benzene)

comparison group

CASE-CONTROL

Exposure to benzene

Leukaemia (cases)



or

Non exposure To benzene

Not leukaemia (controls)

Cohort Studies

Procedure For Carrying Out A Cohort Study

PROCEDURE	**EXAMPLE** (Doll and Hill)
1. Obtain representative sample of target population (excluding those with disease)	100% sample of UK medical profession
2. Obtain details of potential aetiological characteristics or exposures	Questionnaire about smoking habits
3. Collect information on new cases of the disease	Collected death certificates
4. Compare proportions developing the disease (incidence rates) in subgroups with or without the characteristics	Compared death rate from lung cancer in non smokers with rates in smoking groups

If each person in the study falls in

(a) either the exposed or non exposed group

or

(b) either the diseased or non diseased group

then the results can be put into a 2 x 2 table

EXAMPLE Doll and Hill

	Died of lung cancer	**Did not die of lung cancer**	**Total**
Smoker	36	21353	21389
Not smoker	1	3093	3094
Total	37	24446	24483

MORE GENERALLY

	Diseased	Not diseased	Total
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total	a+c	b+d	N

Where a is the number who develop the disease amongst those exposed, c is the number who develop the disease among those not exposed etc.

And N= a + b + c + d

= total number in the study

We can calculate the proportions of the exposed or non exposed who develop the disease.

Example Doll and Hill

Proportion of smokers who develop lung cancer $= \frac{36}{21389} = 0.0017$

Proportion of non smokers who develop lung cancer $= \frac{1}{3094} = 0.0003$

IN GENERAL

	Diseased	**Total**	**Proportion who are diseased**
Exposed	a	a+b	$P_E = a/(a+b)$
Not exposed	c	c+d	$P_0 = c/(c+d)$

NB Conventional to express these proportions as *an incidence rate of disease per 1000 people at risk of developing it per year* when the period of follow is lengthy.

We need to compare these proportions.

When the effect of exposure is to multiply the risk in the unexposed then:

risk of disease in exposed =

risk of disease in unexposed X

relative risk associated with exposure.

i.e. relative risk = $\frac{\text{risk of disease in exposed}}{\text{risk of disease in unexposed}}$

$= P_E/P_0$

Doll and Hill

Relative risk of Lung cancer $= \frac{0.0017}{0.0003}$

$= 5.67$

NB A relative risk of 1 corresponds to no increase in risk in the exposed group compared with the risk in the unexposed group.

Can calculate a confidence interval for the relative risk

Comparison Group Choice

Internal

Doll's study used an internal group i.e. he had a large number of exposed (smokers) and non exposed (non smokers).

External

Ideally, we want a group identical in all characteristics to the study group except for the exposure characteristics. Not always easy to obtain.

In occupational cohort studies comparison is often made with the national population (can have drawbacks) e.g. Deaths observed in a cohort of workers are compared with the numbers expected if the death rates in the national population had been experienced by the cohort. The ratio of these gives the **Standardised Mortality Ratio (SMR)**, which is equivalent to the relative risk.

Attributable Risk

Example

Exposure A multiplies risk of lung cancer by 10 (RR = 10)
Exposure B multiplies risk of lung cancer by 20 (RR = 20)
Does B have a greater effect on public health than A?

Suppose A is smoking, 40% of adults smoke
B is uranium mining, 0.04% miners
RR mining high but effect on community small

Attributable Risk

Combines relative risk and risk factor prevalence to reflect fraction of all cases associated with risk factor.

Can be defined for exposed group and total population.

When effect of exposure is to add to the risk in unexposed then

Risk of disease in exposed = risk of disease in unexposed + excess risk attributable to exposure.

Attributable risk = risk of disease in exposed – risk of disease in unexposed = $P_E - P_O$

Doll and Hill Example: AR=0.0017 – 0.0003=0.0014

EXAMPLE FROM DOLL AND HILL

Relative and attributable risks of death from selected causes associated with heavy cigarette smoking by British make physicians, 1951 to 1961*

Cause of death	**Death rate+ among non-smokers**	**Death rate+ among heavy smokers++**	**Relative risk**	**Attributable death rate+**
Lung cancer	0.07	2.27	32.4	2.20
Other cancers	1.91	2.59	1.4	0.68
Chronic bronchitis	0.05	1.06	21.2	1.01
Cardiovascular disease	7.32	9.93	1.4	2.61
All causes	12.06	19.67	1.6	7.61

* From Doll and Hill

+ Annual death rates per 1000

++ Heavy smokers are defined as smokers of 25 or more cigarettes per day

RR greater for lung cancer and chronic bronchitis

AR greater for cardiovascular disease

Generally

Size of RR better index of likelihood of causal relationship between exposure and disease. But if it is accepted that observed association is causal then attributable risk gives better idea of impact of a preventive program.

Advantages And Disadvantages Of Cohort Studies

Complete description of experience after exposure	Large numbers needed for rare diseases
Can study all effects of exposure (benefits and risks) and different outcomes	Lengthy time and costly
Can calculate rates of disease in exposed and unexposed	Changes in practice, exposure over time
	Maintenance of follow-up

CASE-CONTROL STUDIES
Procedure For Carrying Out A Case - Control Study

Procedure	Example
1 Select cases with disease controls without disease	Cases: leukaemia deaths Controls: from refinery population without leukaemia
2 Obtain information on past exposures and other relevant factors	Examine job history exposure records to classify each subject into high or low benzene actors exposure
3 Compare proportions with exposure in cases and controls	Compare proportion with high benzene exposure in cases and controls

Example

36 leukaemia deaths , 108 controls

18 leukaemia deaths , 36 controls had high benzene exposure

Can put this data into a 2 X 2 table

		Cases	Controls	Total
Exposure to Benzene	High	18	36	54
	Low	18	72	90
	Total	36	108	144

More generally

	Cases	Controls
Exposed	a	b
Not Exposed	c	d

Where a is no. of cases exposed
b is no. of controls exposed
c is no. of cases not exposed
d is no. of controls not exposed

We can calculate the odds of being exposed in the cases, a/c, and the odds of being exposed in the controls, b/d. The ratio of these two odds gives an estimate of the risk of being exposed and is equivalent to calculating the cross product ad/bc

This is known as the ODDS RATIO

Benzene and Leukaemia Example

$$\text{Odds Ratio} = 18 \text{ X } 72 \text{ / } 18\text{X } 36 = 2.0$$

$\Rightarrow$Twice the risk of leukaemia if high exposure to benzene compared with low exposure

A Confidence Interval can be calculated for the Odds Ratio

Relationship Between Relative Risk and Odds Ratio

In a case control study we cannot calculate the relative risk because we do not know the total numbers of exposed and non exposed in the study population. The groups are selected because they either had or did not have the disease of interest and are NOT a random sample from populations of all those with high or low exposure rates to the factor under investigation

However, we can use the odds ratio to estimate the relative risk.

	Diseased	Not Diseased	Total
Exposed	a	b	a + b
Not exposed	c	d	c + d

Relative risk $= \frac{a/(a+b)}{c/(c+d)}$

$= \frac{a(c+d)}{c(a+b)}$

This is approximately $= \frac{ad}{bc}$

If a is small compared with b and c is small compared with d i.e. the numbers developing the disease are small compared with those who do not develop the disease

Example Doll and Hill

	Lung cancer	not lung cancer	total
Smoker	36	21353	21389
Non smoker	1	3093	3094

$$\text{Odds ratio} = \frac{36 \times 3093}{1 \times 21353}$$

$$= 5.21$$

Relative risk = (36/21389) / (1/3094) = 5.67

The odds ratio is approximately the same as the relative risk it the outcome of interest is rare

Example of Odds Ratios for Several Categories of Exposure

Alcohol Consumption (g)	Cases	Controls	OR
0-39	29	386	(1)
40-79	75	280	2.63 ((75X386)/(29X280))
80-119	51	87	7.80 ((51X386)/(29X87))
120+	45	22	27.23 ((45X386)/(29X22))

Defining and selecting cases

- Ensure cases are as homogenous as possible. Establish strict diagnostic criteria (e.g. certain histologic characteristics).
- Sub-definitions of cases such as definite, probable or possible may be needed.
- Analysis can be conducted for each sub-group.

Prevalent vs. Incident (Newly Diagnosed) Cases

- Where possible avoid prevalent cases even though they may give more cases
- Prevalent cases may exclude those with short disease duration or rapid cure or death
- Determinants of disease duration may be related to the exposure such that the magnitude of the exposure (e.g. low vs. high) may be inaccurate.
- Prevalent cases with long disease duration may not accurately recall antecedent events.

Sources of cases

- Hospital based – readily accessible
- Population based – may give a better cross section of all cases and avoids biased referral patterns
- Screen detected

Ascertainment of Disease Status

- Case registries (i.e. cancer)
- Records of physicians e.g. GPs
- Hospital admission or discharge records
- Pathology department log books
- Self-report

Selecting Controls

- Selection of an appropriate comparison group is the most difficult and critical issue in the design of case-control studies.
- Controls are subjects free of the disease (or outcome of interest).
  - Controls are seldom subjected to a medical examination to rule out the disease of interest.
  - Usually, they are assumed disease free if they have not been diagnosed.

Selecting Controls ctd

- The prevalence of exposure among controls should reflect the prevalence of exposure in the source population.
- Controls should come from the same source population as cases (e.g. would have been cases if diagnosed with the disease).
- The time during which a subject is eligible to be a control should be the time in which the individual is also eligible to be a case.

Sources of controls

- General population
- Random digit dialing
- Neighborhood
- Friends/relatives
- Hospital or clinic-based

Matching cases and controls

- Cases and controls are often matched on age, sex, location etc
- The matching variables are chosen to be those not under study but that might affect the risk of exposure and/or the risk of disease
- May get overmatching i.e. cases and controls too similar with regard to the exposure of interest
- The effects of the variables that are used for matching cannot be explored

How many controls?

- the commonest case-control ratio is 1:1
- when the number of cases is small, the sample size for the study can be increased by using more than one control

  e.g. 1:2 1:3 1:4
- Gives more power

Ascertaining Exposure

- Sources of exposure data (cases and controls):
  - Study subjects (self-report). Particularly vulnerable to recall bias as cases may recall their exposure history more thoroughly than controls.
  - Records (preferably completed before the occurrence of outcome events).
  - Interviews with surrogates (spouses, siblings, etc.)

Ascertaining Exposure

- How far back should exposure be assessed?
  - Define a part of the person's exposure history considered relevant to the aetiology of disease (e.g. the "empirical induction" period).
  - Define the measurement variable for exposure (the exposure metric) in an aetiologically-relevant way (e.g. magnitude of exposure, years of exposure, ever exposed, etc.)

Advantages and Disadvantages of Case-Control Studies

Advantages	Disadvantages
Good for rare diseases	Problems of selection of controls
Quick and cost efficient	Recall bias
Can investigate many risk factors simultaneously	Poor for rare exposures
	Cannot estimate separate risk of disease among exposed and non-exposed

Confidence Interval for Relative Risk

The sampling distribution of the $\log_e$ RR is the Normal Distribution

A 95% confidence interval for $\log_e$ RR is

Log RR$\pm$1.96 X SE($\log_e$ RR)

Where SE($\log_e$ RR) = standard error ($\log_e$ RR)

$$= \textit{square root} \left(\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d} \right)$$

We obtain the 95% confidence interval for the relative risk by taking antilogs

Smoking Example

Log_e RR = 1.735

$$\text{se}(\log_e 5.67) = \textit{square root} \left(\frac{1}{36} - \frac{1}{21389} + \frac{1}{1} + \frac{1}{3094} \right)$$

= 1.014

95% CI = 1.735 $\pm$ 1.96 X 1.014

= -0252 to 3.722

Taking antilogs 95% CI for RR = 5.67 is 0.78 to 41.35

Confidence Interval for OR (Woolf's method)

Variance for logarithm of odds ratio

$$var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

95% confidence interval for OR

$$= \exp\left(\ln(OR) \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

Confidence interval for benzene leukaemia example: OR=2.0

$$var(\ln OR) = \frac{1}{18} + \frac{1}{36} + \frac{1}{18} + \frac{1}{72}$$

Var (In 2.0)=0.1527

95% CI for In OR is

$$\text{In } (2.0) \pm 1.96\sqrt{0.1527}$$

$$= 0.69 \pm 0.766$$

$$= -0.076 \text{ to } 1.456$$

95% CI for OR $= e^{-0.076}$ to $e^{1.456}$

$= 0.93$ to 4.29

Analysis of Matched Case-Control Studies

	Controls	
Cases	Exposed	Not Exposed
Exposed	n_1	n_2
Not Exposed	n_3	n_4

Where n_1 is the no. of pairs where both case and control where exposed. n_2 is the no. of pairs where the case was exposed and control was not exposed etc.

Odds Ratio $OR = \frac{n_2}{n_3}$

Relationship of Systolic Blood Pressure (SBP) to Myocardial Infarction (MI): Paired Case-Control Data Matched by Age

	Controls		
Cases	SBP ≥ 140	SBP < 140	Total
SBP ≥ 140	15	13	28
SBP < 140	11	17	28
Total	26	30	56 OR = 13 = 1.18 11

This table shows the results of 56 pairs of cases and controls matched On age. Each cell relates to pairs of individuals. We cannot learn Whether SBP ≥ 140 is associated with M1 if everyone has SBP ≥ 140. So the 32 pairs with cases and controls in the same BP category provide No basis for learning how SBP is related to MI.

$$OR = \frac{13}{11} = 1.18$$