Name of Faculty: Dr. Amit Kumar Nayak
Designation: Professor
Department: Pharmacy
Subject: Biostatistics and Research Methodology (BP 801T)
Unit: IV
Topic: Cohort Study

Concept of cohort

The word cohort is derived from the Latin “cohorts” meaning an enclosure, company, or crowd.

In Roman times a cohort was a body of 300–600 infantry.

In epidemiological terms the cohort is a group of people with something in common, usually an exposure or involvement in a defined population group.

Definition

- Cohort study is a type of analytical study which is undertaken to obtain additional evidence to refute or support existence of association between suspected cause and diseases.

- Other names of cohort study are Longitudinal study, Incidence study and forward looking study

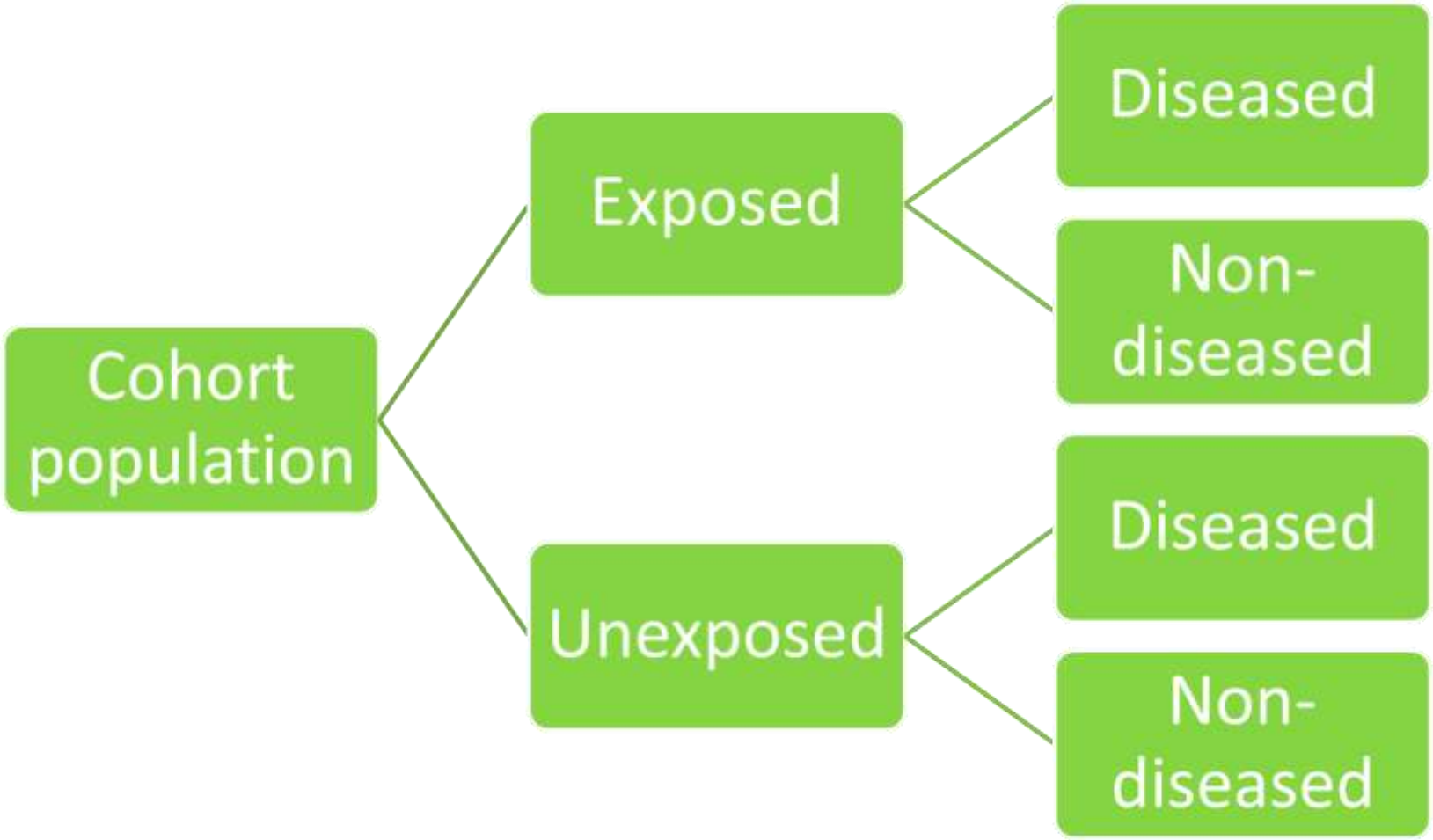
Features of cohort studies

- Cohorts are identified prior to appearance of disease under investigation
- The study groups are observed over a period of time to determine the frequency of disease among them
- The study proceeds from cause to effects

Indications for cohort study

- There is good evidence of an association between exposure and disease, from other studies.
- Exposure is rare.
- Attrition of study population can be minimized.
- Sufficient fund is available.

Framework of cohort study



Design of Cohort Study

First, identify

	Totals
Exposed	a+b
Not exposed	c+d

Then

Then, follow to see whether

Calculate and compare

	Disease develops	Disease does not develop	Totals	Incidence of disease
Exposed	a	b	a+b	$\frac{a}{a+b}$
Not exposed	c	d	c+d	$\frac{c}{c+d}$

(a+b) is called study cohort and (c+d) is called control cohort

Consideration during selection of Cohort

- The cohort must be free from disease under study.

- Insofar as the knowledge permits, both the groups should be equally susceptible to disease under study.

- Both the groups must be comparable in respect of all variable which influence the occurrence of disease

be defined beforehand.

Types of cohort study

- Prospective study
- Retrospective cohort study
- Ambi-directional cohort study

Prospective cohort study

- The common strategy of cohort studies is to start with a reference population (or a representative sample thereof), some of whom have certain characteristics or attributes relevant to the study (exposed group), with others who do not have those characteristics (unexposed group).

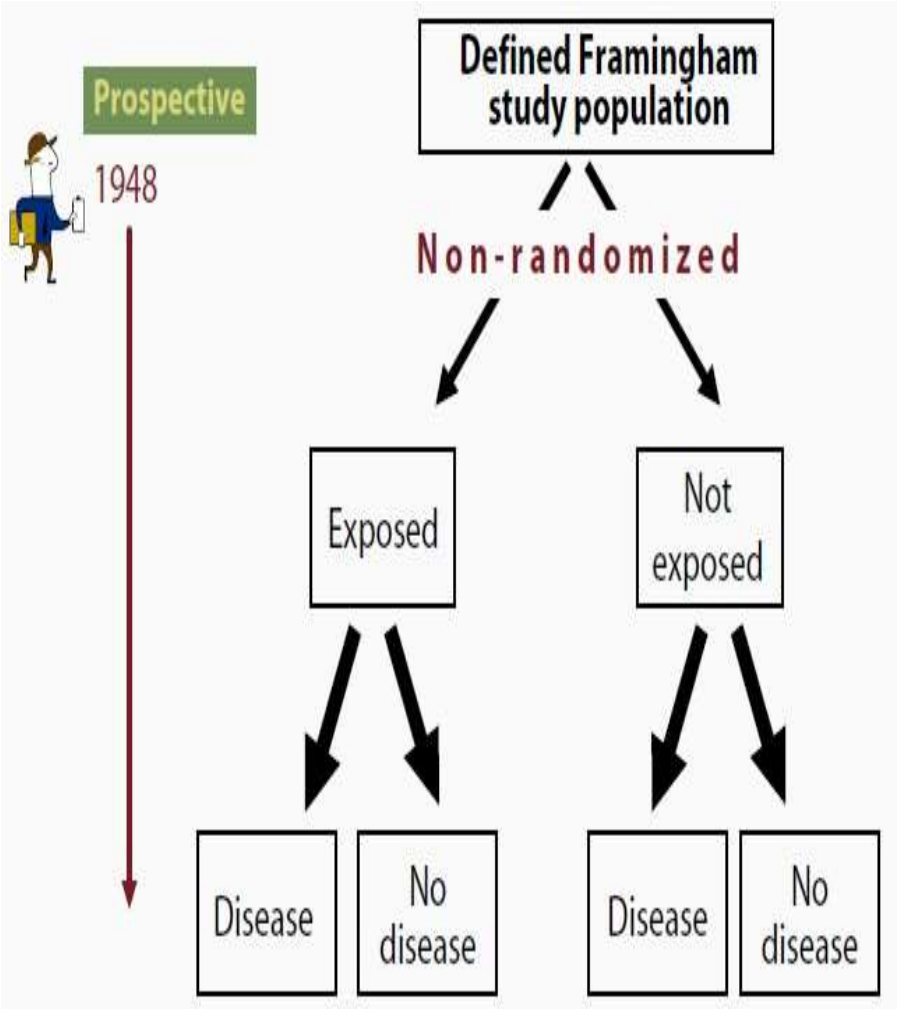
- Both groups should, at the outset of the study, be free from the condition under consideration. Both groups are then observed over a specified period to find out the risk each group has of developing the condition(s) of interest.

Example of Prospective Cohort Study

Framingham Heart Study

- Initiated in 1948 to study the relationship of a variety of factors to the subsequent development of heart disease with 5127 samples(30 to 59 yrs) at Framingham.
- Study subjects were examined every 2 yrs for 20 years.
- Daily Surveillance of hospitalization at Framingham hospital.
- Study found that Hypertensive, tobacco smoking, elevated blood cholesterol are associated to CHD
- Increased physical activity associated with decreased risk of CHD

Framework



Problem of prospective study

- Study might take long duration.
- Sufficient amount of funding for long period.
- Missing of study subjects.

Retrospective Cohort Study

- A retrospective cohort study is one in which the outcome have all occurred before the start of investigation.
- Investigator goes back to the past to select study group from existing records of the past employment, medical and other records and traces them forward through time from the past date fixed on the records usually to the present.
- Known with the name of Historical Cohort and noncurrent cohort

Example of Retrospective Study

- Suppose that we began our study on association between smoking habit and lung cancer in 2008
- Now we find that an old roster of elementary schoolchildren from 1988 is available in our community, and that they had been surveyed regarding their smoking habits in 1998.
- Using these data resources in 2008, we can begin to determine who in this population has developed lung cancer and who has not.

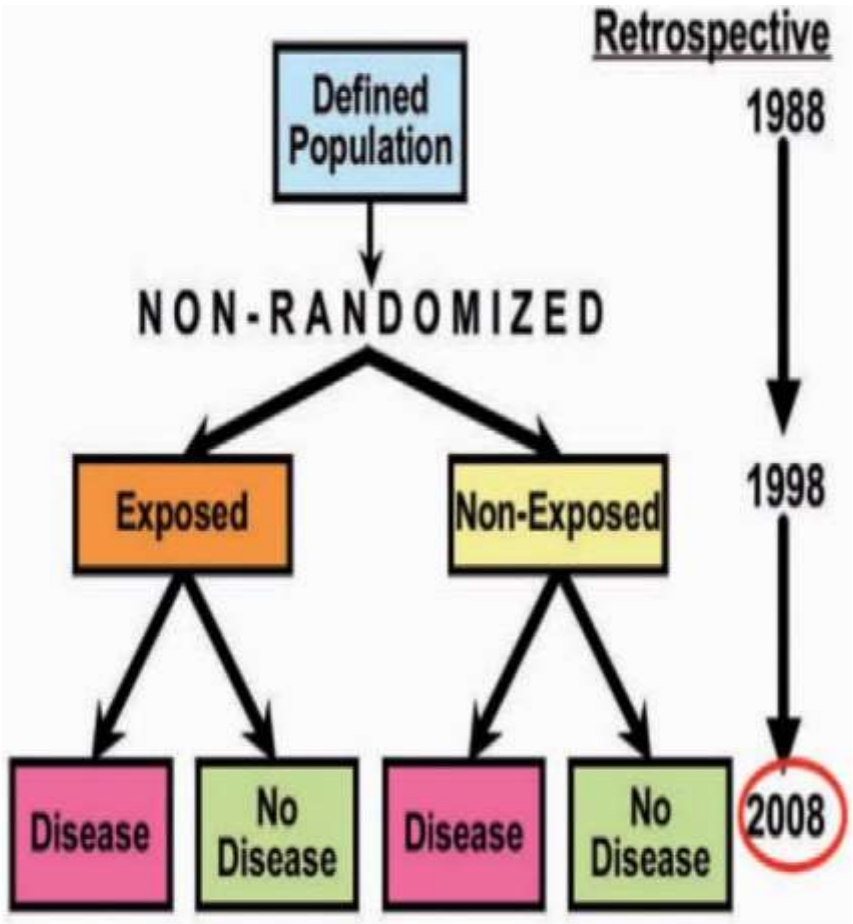


Figure 9-7 Time frame for a hypothetical retrospective cohort study begun in 2008.

Ambi-directional cohort Study

- Elements of prospective and retrospective cohort are combined.

- The Cohort is identified from past records and assesses of date for the outcome. The same cohort is the followed up prospectively into future for the further assessment of outcome

Example of Ambi-directional cohort study

- Curt- Brown and Dolls study on effects of radiation Began in 1955 with 13,352 patients who received large dose of radiation therapy for ankylosing spondylitis between 1934 to1954.
- Outcome evaluated was death from Leukemia or aplastic anemia between 1934 to 1954.
- A prospective component was added up in 1955 and surviving subjects were followed up to identify deaths in subsequent years

Comparison of retrospective and prospective cohort study

Attribute	Retrospective approach	Prospective approach
Information	< complete < accurate	> complete > acurate
Emerging new exposures	Not useful	Useful
Expense	Less costly	More costly
Completion time	Shorter	Longer

Prognostic cohort studies

Prognostic cohort studies are a special type of cohort study used to identify factors that might influence the prognosis after a diagnosis or treatment.

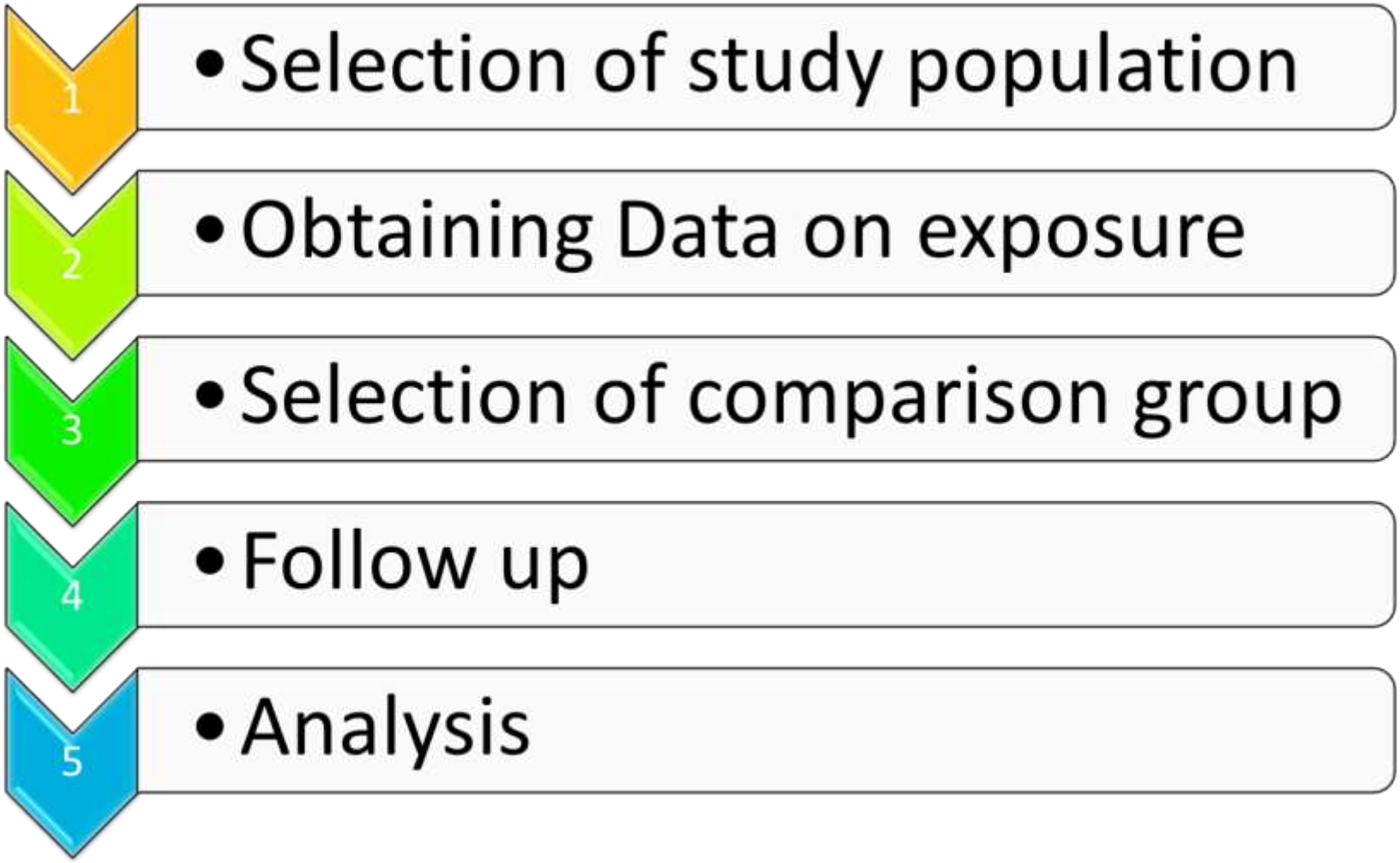
These follow-up studies have the following features:

The cohort consists of cases diagnosed at a fixed time, or cases treated at a fixed time by a medical or surgical treatment, rehabilitation procedure, psychological adjustment.

By definition, such cases are not free of a specified disease, as in the case of a conventional cohort

The outcome of interest is usually survival, cure, improvement, disability, or repeat episode of the illness, etc.

Steps of Cohort Study



1. Selection of study subjects

The usual procedure is to locate or identify the cohort, which may be a total population in an area or sample thereof. Cohort can be:

- community cohort of specific age and sex;
- exposure cohort e.g. radiologists, smokers, users of oral contraceptives;
- birth cohort e.g. school entrants;
- occupational cohort e.g. miners, military personnel;
- marriage cohort;
- diagnosed or treated cohort, e.g. cases treated with radiotherapy, surgery, hormonal treatment.

Open or dynamic cohort

- **Open population or dynamic population** describe a population in which the person-time experience can accrue from a changing roster of individuals.

- For example, in a study, the incidence rates of cancer reported by the Connecticut Cancer Registry come from the experience of an open population. Because the population of residents of Connecticut is always changing, the individuals who contribute to these rates are not a specific set of people who are followed through time.

Fixed and Closed Cohort

- Fixed Cohort :When the exposure groups in a cohort study are defined at the start of follow-up, with no movement of individuals between exposure groups during the follow-up, the groups are called fixed cohorts.

- If no losses occur from a fixed cohort, the cohort satisfies the definition of a closed population and is often called a closed cohort

2. Obtaining data on Exposure

- From Cohort Members : Personal interview, mailed questionnaire
- Review of Records : Certain kinds of information like dose of radiation, kinds of surgery received can only be obtained from medical records.
- Medical examination/ Special tests: In some cases information needs to be obtained from medical examination like in case of blood pressure, serum cholesterol,
- Environmental Survey of location where cohort lives

Information should be collected in a manner that allows classification of cohort according to

- whether or not they have been exposed to suspected factor
- According to level or degree of exposure
- Demographic variables which might influence frequency of disease under investigation

3. Comparison Group

Internal Comparison Group :

Single Cohort enters the study and its members on the basis of information obtained , can be classified into several comparison according to degree of exposure

Classification of exposure	No. of Deaths	Death rate
½ pack	24	95.2
½ to 1 pack	84	107.82
1-2 pack	90	229.2
+ 2 pack	97	264.2

Age Standardized death rate among 100000 men per year according to amount of cigarette smoking

External Comparison Group: when information on degree of exposure is not available.

if all workers at the factory had some degree of exposure, we would need to select a comparison group from another population, possibly another type of factory

Comparison with general population can also be used as comparison group

4. Follow UP

- The length of follow-up that is needed for some studies to reach a satisfactory end-point, when a large enough proportion of the participants have reached an outcome, may be many years or even decades.

- At the start of study, method should be determined depending on the outcome of study to obtain data for assessing outcome.

Procedure may be:

- Periodic medical examination of each member of cohort
- Reviewing physician and hospital records
- Routine surveillance of death records
- Mailed questionnaire, telephone calls and periodic home visits

5. Analysis

Data analyzed in terms of

- Incidence rate of outcome among exposed and non exposed

- Estimation of risk

Incidence rate

Choice between cumulative incidence and Incidence Density is a crucial issue

- Cumulative incidence: In cohort studies on acute diseases with short induction periods and a short time of follow-up, like outbreaks, the risk of disease can be estimated directly using the cumulative incidence, given a fixed cohort with fixed period of follow-up and a low fraction of drop-outs.

- Incidence Density: In cohort studies on chronic diseases with their long follow-up periods, however, the use of the cumulative incidence is not appropriate because usually disease-free follow-up periods differ strongly among cohort members. In such case incidence density is apposite measure

ANALYSIS OF COHORT STUDIES

	Outcome*			
	Death	**No death**	**Incidence rate**	**Total**
Exposed	**A**	**B**	**A/(A+B)**	**A + B**
Unexposed	**C**	**D**	**C/(C+D)**	**C + D**
Total	**A + C**	**B + D**		**A+B+C+D**

* Outcome : death/disease

A = Exposed persons who later develop disease or die
B = Exposed persons who do not develop diseases or die
C = Unexposed persons who later develop disease or die
D = Unexposed persons who do not develop diseases or die

The total number of exposed persons = A + B
The total number of unexposed persons = C + D

Incidence of disease(or death) among exposed= A/A+B
Incidence of disease(or death) among non-exposed= C/C+D

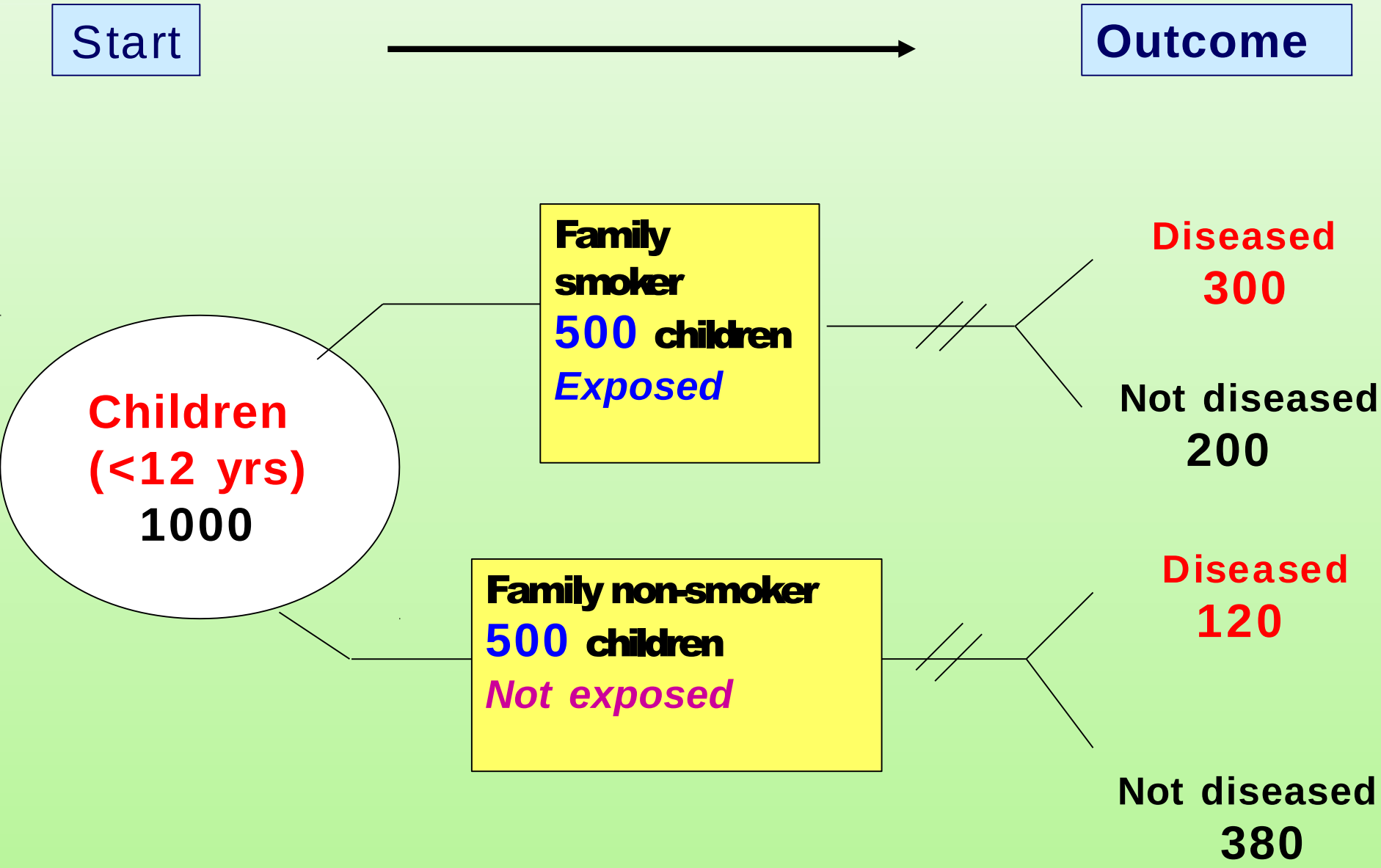
Relative Risk (RR)

- Estimates the magnitude of an association between exposure and disease
- Indicates the likelihood of developing the disease in the exposed group relative to those who are not exposed
- Ratio of risk of disease in exposed to the risk of disease in nonexposed

Relative Risk

$$RR = \frac{\text{Risk in exposed(Incidence in exposed group)}}{\text{Risk in non exposed(Incidence in non exposed group)}}$$

EXAMPLE


Start
Outcome
Children
(<12 yrs)
1000
Family
smoker
500 children
Exposed
Diseased
300
Not diseased
200
Family non-smoker
500 children
Not exposed
Diseased
120
Not diseased
380

Rate: Incidence rate

- **Incidence of Resp. Infection among exposed children:** $\frac{300}{500} = 60\%$

- **Incidence of Resp. Infect. Among non exposed children:** $\frac{120}{500} = 24\%$

Relative Risk: $\frac{\text{Incidence rate among exposed}}{\text{Incidence rate in non exposed.}}$

Risk Ratio

$$\frac{60}{24} = 2.5$$

Exposed individuals are 2.5 times more likely to develop disease than non exposed individuals.

Difference Measures

- **Attributable risk**
  - **No. of cases among the exposed that could be eliminated if the exposure were removed**

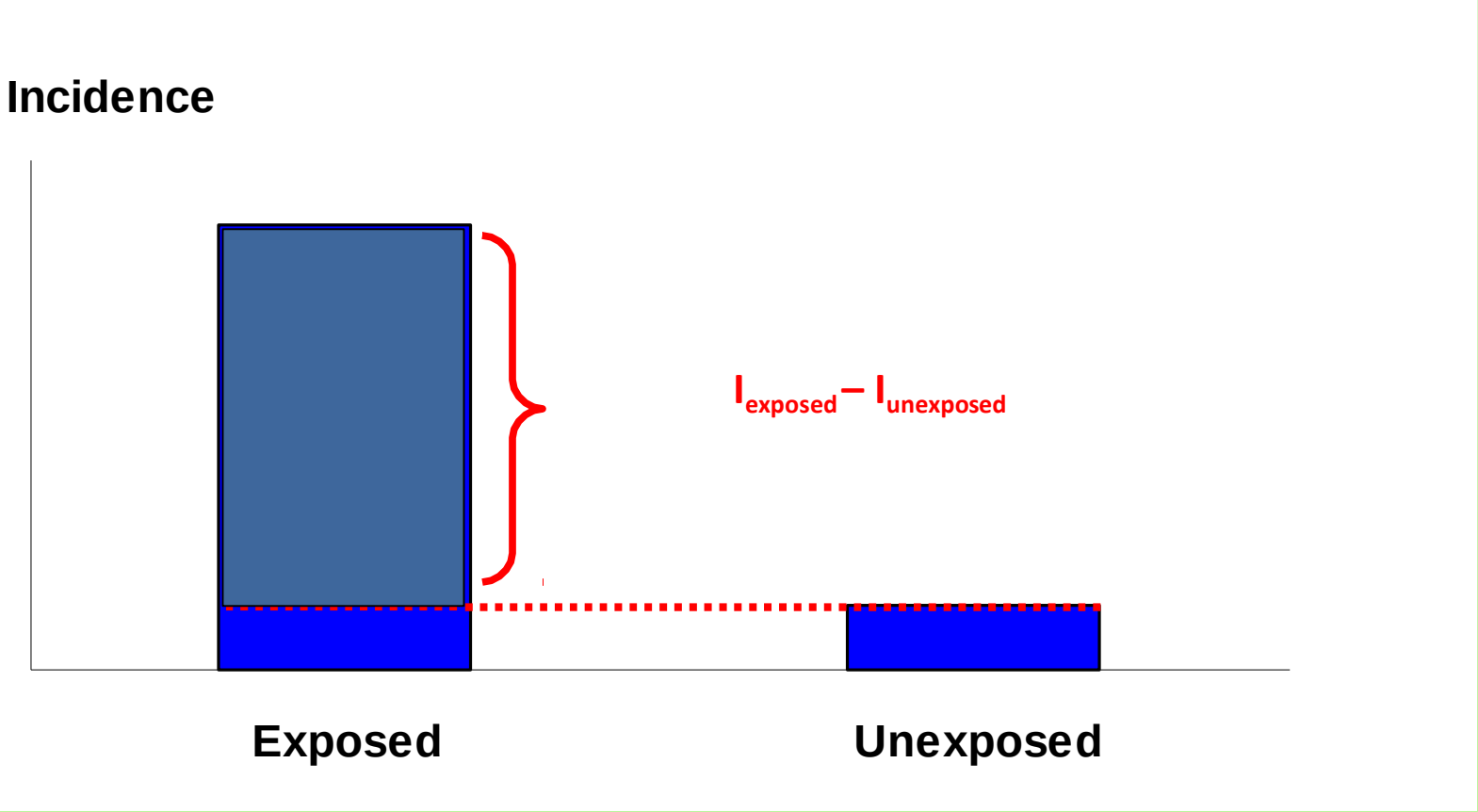
  = Incidence in exposed - Incidence in unexposed

- **Population Attributable Risk percentage:**

PAR expressed as a percentage of total risk in population

$$PAR\% = \frac{I_{population} - I_{unexposed}}{I_{population}} \times 100$$

Attributable Risk



I = Incidence

AR: Smoking and Lung cancer

	Lung Cancer				
Smoking	**Yes**	**No**		**Incidence**	**RD**
Yes	**100**	**1900**	**2000**	**0.05**	**0.04**
No	**80**	**7920**	**8000**	**0.01**	
	180	**9820**	**10000**		

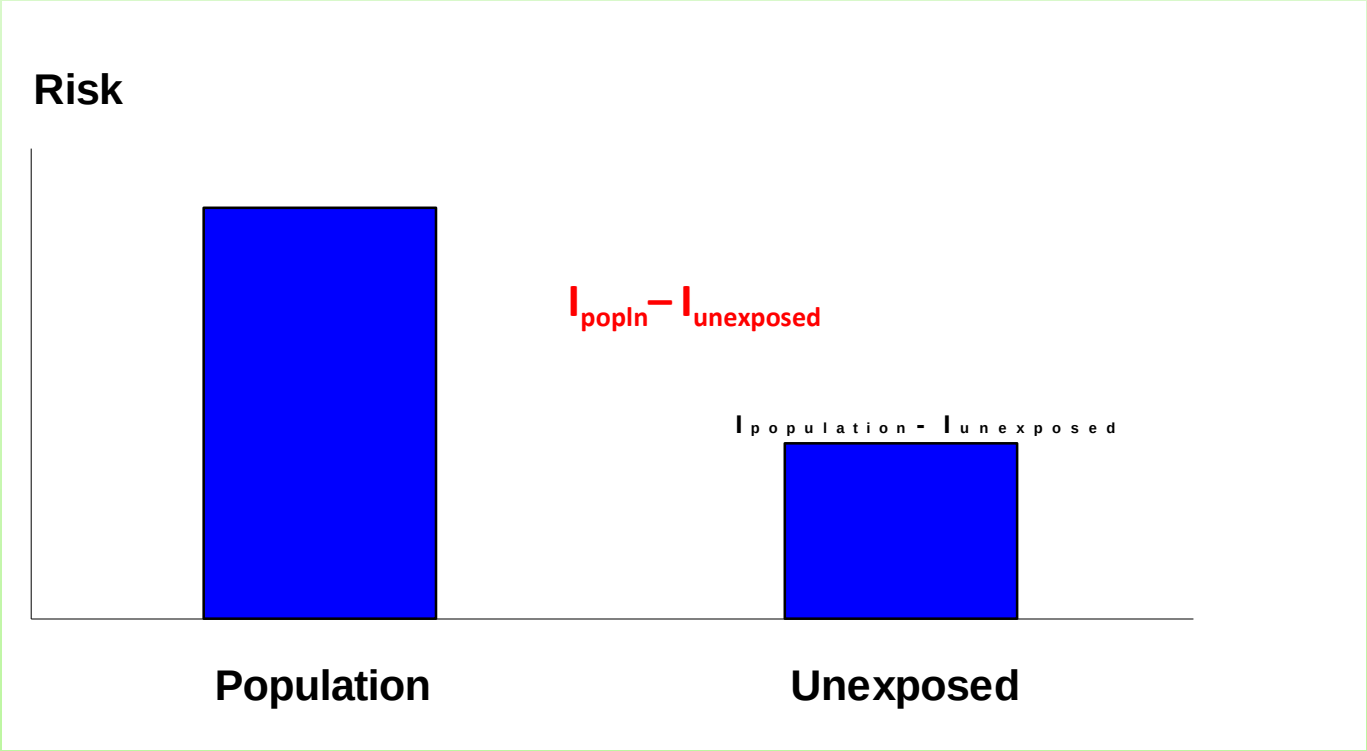
Attributable risk = **Incidence in exposed - Incidence in unexposed**
=0.5-0.1
=0.4

Population Attributable Risk (PAR)

- **Excess risk of disease in total population attributable to exposure**
- **Reduction in risk which would be achieved if population entirely unexposed**
- **Helps determining which exposures relevant to public health in community**

$$PAR = I_{population} - I_{unexposed}$$

Population Attributable Risk



PAR: Smoking

	Lung Cancer			
Smoking	**Yes**	**No**		**Risk**
Yes	**100**	**1900**	**2000**	**Incidence in exposed= 0.050**
No	**80**	**7920**	**8000**	**Incidence in unexposed=0.010**
	180	**9820**	**10000**	**Incidence in population=0.018**

$$PAR = 0.018 - 0.010 = 0.008$$

$$PAR\% = \frac{0.018 - 0.010}{0.018} \times 100 = 44\%$$

Conclusion:

44% of lung cancer in the population could be prevented if use of smoking were eliminated

But calculations are

not that simple in real Cohort studies

British Doctors Study

- In 1951, a prospective cohort study was set up among British doctors to investigate the relationship between smoking and mortality, particularly the association between smoking and lung cancer

- In 1951, a questionnaire on smoking habits was sent to 49,913 male and 10,323 female doctors , 34,440 male doctors and 6194 female doctors gave sufficient information to classify their smoking status.

- The causes of death of 10,072 male and 1094 female doctors who had died during this period were ascertained from death certificates.

- The rate of death from lung cancer among smokers was compared to that among non-smokers.

Table 6.1 Numbers of smokers and non-smokers by age group among the male doctors

Age group (years)	*Smokers*	*Non-smokers*	*Total*
20–29	3 568	1 693	5 261
30–39	8 057	1 884	9 941
40–49	6 310	1 030	7 340
50–59	5 144	610	5 754
60–69	3 082	288	3 370
70+	2 476	298	2 774
Total	28 637	5 803	34 440

Source: data based on Doll and Peto (1976)

Table 6.4 Distribution of doctors by age group among smokers and non-smokers

Age group (years)	*Smokers*	*Non-smokers*	*Total population*
20–29	3 568 (12%)	1 693 (29%)	5 261 (15%)
30–39	8 057 (28%)	1 884 (32%)	9 941 (29%)
40–49	6 310 (22%)	1 030 (18%)	7 340 (21%)
50–59	5 144 (18%)	610 (11%)	5 754 (17%)
60–69	3 082 (11%)	288 (5%)	3 370 (10%)
70+	2 476 (9%)	298 (5%)	2 774 (8%)
Total	28 637	5 803	34 440

Source: based on Doll and Peto (1976) and Doll *et al.* (1980)

Since ***mortality depends on age and the distribution of subjects by age group is different between the smokers and non-smokers, the effect of age on mortality has to be adjusted*** for when making comparison on lung cancer mortality between these two groups. A commonly used method to adjust for the age is direct standardization

It would not be rational to categorize individual smoking one cigarette per day and more than 25 cigarette in same category with equal emphasis

So

Its better we opt for stratification

Table 6.2 Death rates from lung cancer by status of smoking and sex

Sex of doctors	*Lung cancer death rates per 100 000 persons per year**			
	Non-smokers	*Smoking 1–14 cigarettes/day*	*Smoking 15–24 cigarettes/day*	*Smoking 25+ cigarettes/day*
Male	10	78	127	251
Female	7	9	45	208

* Adjusted for age by direct standardization.

Source: based on Doll and Peto (1976) and Doll *et al.* (1980)

Again its not only the dose of exposure that determines the frequency of disease, there are some other factors like duration of exposure and age at initiation of exposure that can influence occurrence of disease. We need to make adjustment for that too

Table 6.5 Relative risk of lung cancer death (adjusted for age) by status of smoking in male and female doctors

Sex	Non-smokers	Smoking 1–14 cigarettes/day	Smoking 15–24 cigarettes/day	Smoking 25+ cigarettes/day
Male	1	7.8	12.7	25.1
Female	1	1.3	6.4	29.7

Source: based on Doll and Peto (1976) and Doll *et al.* (1980)

The relative risk of lung cancer death increased with the level of smoking in both males and females. ***The relative risk in the men smoking 1–14 and 15–24 cigarettes per day is much higher than in the women; in the group smoking 25 or more cigarettes per day, the relative risk in men is marginally less than that in women. Does this mean that the effect of low levels of smoking is higher among men than among women?***

Table 6.3 Distribution of inhalation and mean age when started to smoke

	Smoking 1–14 cigarettes/day		*Smoking 15–24 cigarettes/day*		*Smoking 25+ cigarettes/day*	
Features of smoking	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Proportion inhaling smoke (%)	66	42	80	54	83	58
Mean age (years) when started to smoke	20	24	19	23	19	22

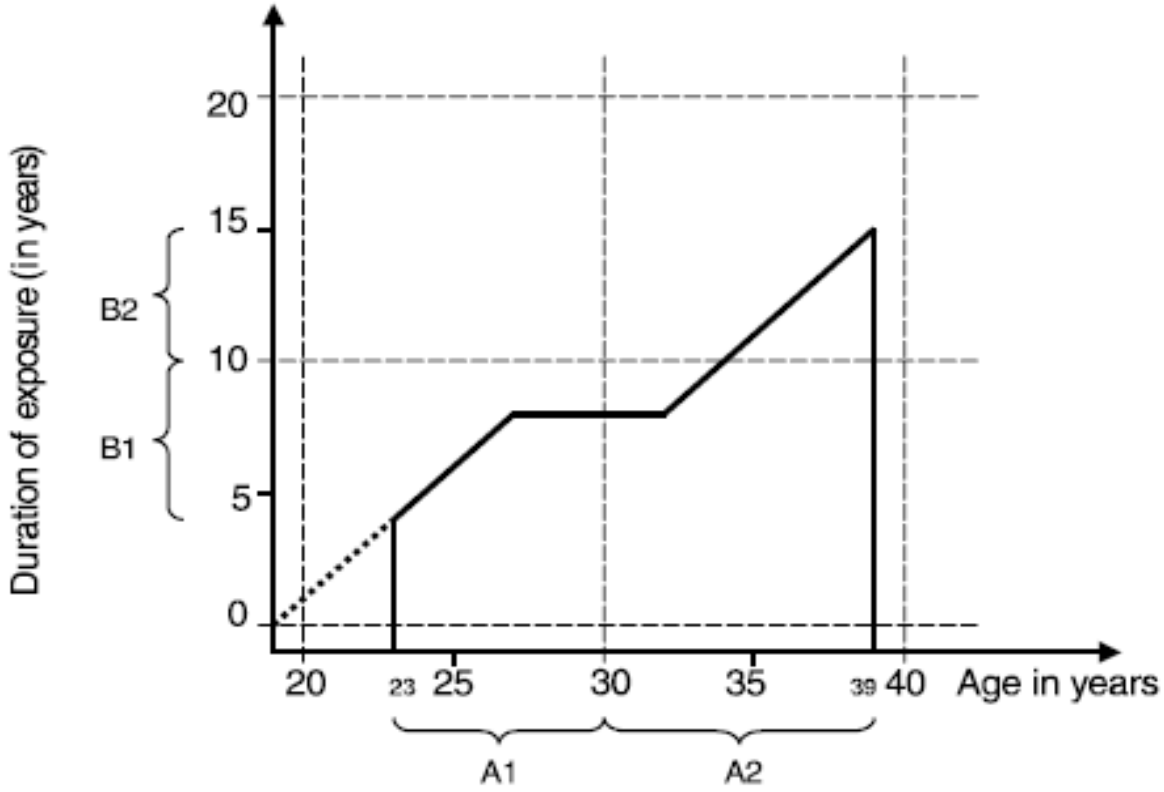
Source: based on Doll and Peto (1976) and Doll *et al.* (1980)

The **proportion of men inhaling smoke is higher than women in all three levels** of smoking. **Men seemed to have started to smoke at an earlier age** than women. Since these features of smoking may modify the effect of smoking on lung cancer**, their effects have to be adjusted for when comparing the association between smoking and lung cancer in men and women**.

…….. too complicated ????

But

Problem does not end here….



What if, a subject is followed up from age 23 but has been exposed from age 19 on, he/*she is exposed until age 27 followed by an unexposed 5 year period. He*/*she* is again exposed until age 39 at which time his/*her person-time at risk ceases either* because of disease diagnosis or because of end of follow-up.

For analyzing such data we use Poisson models and Cox Proportional Hazards

Specialized software packages exist to perform these computations such as Stata (Version 7 or later and Epicure

Advantage of Cohort Studies

- Temporality can be established
- Incidence ca be calculated.
- Several possible outcome related to exposure can be studied simultaneously.
- Provide direct estimate of risk.
- Since comparison groups are formed before disease develops certain forms of bias can be minimized like misclassification bias.
- Allows the conclusion of cause effect relationship

Disadvantage of Cohort Studies

- Large population is needed
- Not suitable for rare diseases.
- It is time consuming and expensive
- Certain administrative problems like loss of staff, loss of funding and extensive record keeping are common.
- Problem of attrition of initial cohort is common
- Study itself may alter people's behavior

Ethics in Cohort Study

- Classic example issues on research ethics is Tuskegee study on natural history of syphilis in which US Public health service recruited 399 poor black sharecroppers in Macon County as cohort.
- Study was lasted from 1932 to 1972.
- They were denied of treatment of syphilis although effective treatment was available. Government deceived by saying that they were being treated.

Ethics in Cohort Study

- On July 26, 1972, The *New York Times described the study as "the longest non therapeutic experiment on human beings in medical* history." The disclosure of this study by the press was a major scandal in the United States.

- Led to *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects in Research*

Ethics in Cohort Study

- These problems can be encountered in cohort study designed to study natural history of disease.
- What if treatment becomes available in the middle of research, should we continue research with treatment denial of abort research?
- Should we communicate the research finding to individuals are controversial issues.

Biases in cohort study

Differential loss of follow up

Differential follow-up between compared groups may be a major problem. Losses to follow-up, whether due to study withdrawals, unmeasured outcomes, or unknown reasons, are always a concern.

This is particularly true when more outcome data is missing in one group than another, as there is no way to be certain that the factor being studied is not somehow related to this observation.

Contamination

Subjects initially unexposed to the risk factor of interest may become exposed at a later date. Such “ contamination ” tends to reduce the observed effect of the risk factor.

Selection Bias

Perhaps the largest threat to the internal validity of a cohort studies is selection bias, also called *case-mix bias* .

Select participants into exposed and not exposed groups based on some characteristics that may affect the outcome

Information bias–

Collect different quality and extent of information from exposed and not exposed groups

Misclassification Bias

Differential misclassification

Non differential misclassification

- ***Differential misclassification*** **– Errors in measurement are one way only**
  - **Example: *Measurement bias* – instrumentation may be inaccurate, same cut off level of weight for male and female to determine malnourishment**

Misclassification Bias (cont.)

True Classification

	Disease +	Disease-	Total
Exposed	**100**	**50**	**150**
Nonexposed	**50**	**50**	**100**
	150	**100**	**250**

RR = a/(a+b)/c/(c+d) = 1.3

Differential misclassification - Overestimate exposure for 10 cases, inflate rates

	Disease +	Disease -	Total
Exposed	**110**	**50**	**160**
Nonexposed	**40**	**50**	**90**
	150	**100**	**250**

RR = a/(a+b)/c/(c+d) = 1.6

- ***Nondifferential (random) misclassification*** **– errors in assignment of group happens in more than one direction**
  - **This will dilute the study findings -** ***BIAS TOWARD THE NULL***

Misclassification Bias (cont.)

True Classification

	Disease +	Disease -	Total
Exposed	**100**	**50**	**150**
Nonexposed	**50**	**50**	**100**
	150	**100**	**250**

RR = a/(a+b)/c/(c+d) = 1.3

Nondifferential misclassification - Overestimate exposure in 10 cases, 10 controls – bias towards null

	Disease +	Disease -	Total
Exposed	**110**	**60**	**170**
Nonexposed	**40**	**40**	**80**
	150	**100**	**250**

RR = a/(a+b)/c/(c+d) = 1.3

Control of Bias

- Restriction
- Stratification
- Mathematical Modeling
  - -Poisson regression model
  - -Cox proportional hazard

When Is a Cohort Study Warranted?

- When the (alleged) exposure is known
- When exposure is rare and incidence of disease among exposed is high (even if the exposure is rare, determined investigators will identify exposed individuals)
- When the time between exposure and disease is relatively short
- When adequate funding is available
- When the investigator has a long life expectancy

Reference

- Park.k.,(2007), *Park's textbook of preventive and social medicine(19th edition)*

- *Gordis.L.,(2008), Epidemiology.,(4th edition)*

- *Framimgham Heart study retrived from www.framingham.com/heart/timeline.htm,on* 29th Jan,2011.

THANK YOU

Classic example of Cohort study : Study on London Cholera Outbreak

- The classical study on the London cholera epidemic of 1849 conducted by John Snow is an example of a cohort study on infectious diseases .

- Two different water companies (the Lambeth and the Southwark & Vauxhall) supplied households within various regions of London

Classic example of Cohort study : Study on London Cholera Outbreak

- The companies differed in one important feature, the location of the water intake. The Lambeth had moved their water intake upstream from the sewage discharge point in 1849; whereas, the Southwark & Vauxhall continued to obtain water downstream of the sewage discharge point.
- Dr. Snow classified households according to their exposure to the two distinct companies and showed a substantial difference in cholera mortality, 315 versus 37 cholera deaths per 10,000 households served by the Lambeth and Southwark & Vauxhall companies, respectively.

Summary of analysis

Measure	Strengths	Uses
Relative risk (RR)	Evaluates the *strength* of an association between exposure and disease	To help identify causes of disease
Attributable risk (AR)	Measures the burden of disease attributable to exposure in the *exposed* group	To assess the magnitude of a public health problem associated with an exposure *among those exposed*
Population attributable risk (PAR)	Measures the burden of disease attributable to exposure in the *population*	To assess the magnitude of a public health problem associated with an exposure *in the whole population*
Attributable fraction (AF)	Identifies the specific exposures that cause most disease *in those who are exposed*	To identify potential targets for prevention
Population attributable fraction (PAF)	Identifies the specific exposures that cause most disease *in a population*	To identify potential targets for prevention