Epidemiology is the study of the frequency, distribution and determinants of health and disease in a particular and defined population. The application of epidemiology is to the control of health issues. The aims are to understand the pattern of disease in order to intervene to prevent disease and/or death.

**DEFINITIONS**

| | |
|---|---|
| Determinant | Also called Exposure, risk factor, independent, predictor or explanatory variable, x-line<br>This is any factor that is suspected of causing the outcome of interest |
| Outcome | Also called the response or dependent variable<br>This could be health in the case of clinical trials/analytic studies, disease in the case of observatory studies, immune status in the case of vaccination campaigns |
| Health | The condition of being sound of body, mind or spirit and especially freedom from physical disease or pain |

**EPIDEMIOLOGY**

1. ***Descriptive epidemiology***

These studies gather information about the occurrence of disease, but with no attempt to establish an association between cause & effect. They characterise the outcome with respect to frequency and distribution. These studies are important for: surveillance, public health, outbreak investigation, hypotheses generation

Tools: counting, sampling, surveys (if no comparisons are being made)

2. ***Analytic epidemiology***

Establishes relationships between causative factors (determinants) and the outcome of interest (e.g. occurrence of disease) to characterise the determinants and to investigate the reasons for/behind the outcome.

Key tools: statistical analysis, measures of association, statistical models, type I & II errors

3. ***Clinical epidemiology***

Application of epidemiological principles and methods to the practice of clinical medicine. It is useful for: diagnostic test evaluation, evidence-based medicine, clinical decision-making e.g. evaluating diagnostic tools, animal health economics

Key tools: sensitivity and specificity, strength of evidence

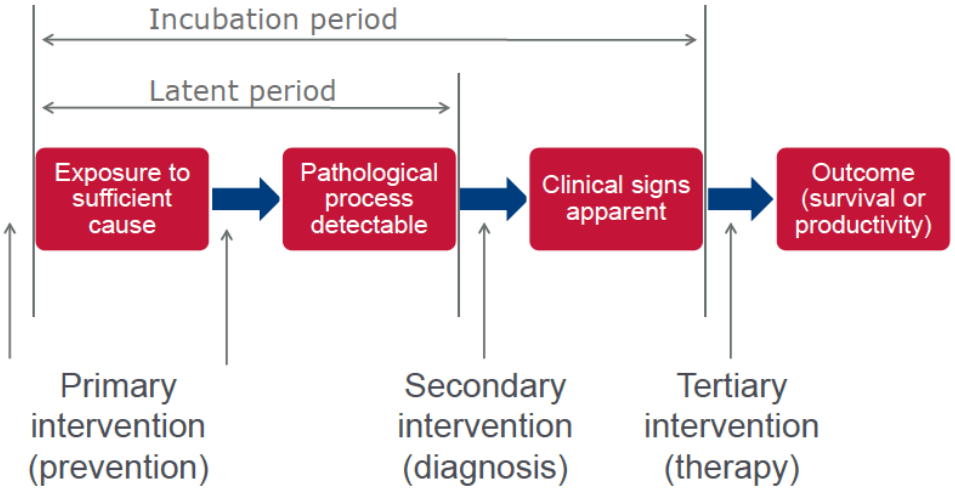
***Disease terminology***

The population being studied has to be defined.

| | |
|---|---|
| Sporadic | Disease occurs very infrequently in a population e.g. rabies in Britain that is sporadic vs. rabies in Bombay that is endemic |
| Endemic | Disease is habitually present at a defined level in a population e.g. flu, rabies in Bombay |
| Epidemic | Disease cases in excess of what would normally be expected in a region during a given time period e.g. ebola, CHIKV in Italy |
| Pandemic | An epidemic that crosses international borders, a worldwide epidemic e.g. HIV |
| Infected | Host is invaded by microorganisms, the organisms multiply and the host's immune system responds |
| Colonised | Infectious agent establishes itself in the host but the immune system is not stimulated to respond, thus the host is not considered "infected". These individuals are carriers of the pathogen and can spread it to uninfected people. |
| Diseased | Infection causes clinical signs/symptoms |
| Zoonosis | Any disease transmitted between vertebrate animals and humans e.g. swine flu, rabies; humans |

| | |
|---|---|
| | can infect animals too |
| Vector | An insect (or any living carrier) that transports an infectious agent from an infected individual to a susceptible individual e.g. *Anopheles*, *Aedes* |
| Reservoir | Any person, animal, arthropod, plant, or fomite in which an infectious agent normally lives and multiplies, on which it depends for its survival, and where it reproduces, so that it can be transmitted to a susceptible host i.e. Animate or inanimate sources that harbour disease-causing organisms and thus serve as potential outbreak sources.<br>E.g. snails for schistosomes |

***Disease development***




Latent period: the time from when an individual is exposed to an agent until infectious
Incubation period: the time from when an individual is exposed to an agent until clinical disease is detectable

*Disease prevention*
Primary prevention – Intervene before the onset/initial development of disease to prevent morbidity e.g. flu shot, vaccines, prophylactics.
Need to know the biological, clinical and epidemiological background of the condition to enact primary prevention programmes.

- Secondary prevention – Early detection/screening of disease to reduce severity and complications as a way to reduce mortality e.g. pap smear, prostate exam
- Tertiary prevention – Reducing the impact of the disease by rehabilitation e.g. antibacterial, antiviral medication, physiotherapy, stroke rehabilitation

**Measures of morbidity**

Prevalence

This is the number of existing cases at a point in time. It measures disease status at a given time (does not account for the duration of the disease) and is measured at a single point in time (point prevalence) or during a defined period of time (period prevalence).
Prevalence is a proportion and not a rate, and is a function of the duration and incidence of disease.

Incidence

This is the number of new cases of disease within a given time period. It requires two or more measurements to ensure individuals are disease free at start of observation and to incorporate time component. It also measures the number of disease "events" and is therefore a measure of risk (accounts for duration of disease). The cumulative incidence is the number of new cases in a vulnerable population.
Diseases with a higher incidence have a higher risk.

Incidence rate

Or incidence density. This is the average speed at which new cases occur per unit of person time at risk. Person time at risk is the total accumulated time that all individuals are at risk of disease e.g. 1 person observed for 10 years; 2 people observed for 5 years. This is a useful way to compare populations e.g. 2.5 cases per 10 person years at risk > 1.3 cases per 10 person years at risk.

Prevalence ~ Incidence X Duration
Incidence: cases per unit time
Duration: unit of time same as incidence

**Measures of Mortality**

All cause mortality rate
The incidence of death – the total number of deaths from all causes during a specified period of time in a population

$$\text{Annual Mortality Rate} = \frac{\text{Number of Deaths in one year}}{\text{Number of persons in the population at mid-year}}$$

*Group-specific mortality*
The total number of deaths within a specifically defined age based on a common characteristic e.g. gender, age, ethnic group, etc.
This is an incidence measure so those in the denominator have the potential for becoming part of the numerator.

$$\text{Annual mortality for children < 10 years old} = \frac{\text{Number of Deaths in one year among children < 10 years old}}{\text{Number of children < 10 years old in the population at mid-year}}$$

$$\text{Annual mortality from lung cancer} = \frac{\text{Number of deaths from lung cancer in one year}}{\text{Number of persons in the population at mid-year}}$$

Cause specific mortality
The total number of deaths as a result of a cause within the total population.

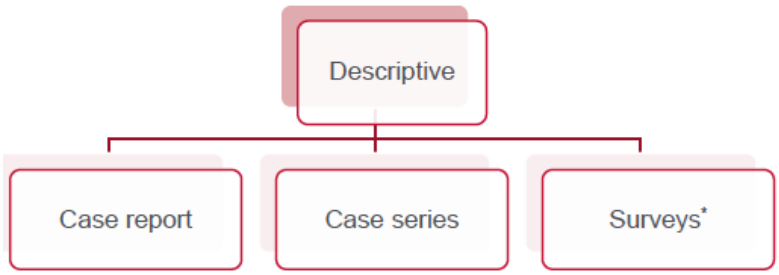
*Case-fatality rate*
The percentage of people diagnosed with a certain disease that die within a certain time after diagnosis.
The denominator is the population with the specific disease.

$$\text{Case-Fatality Rate} = \frac{\text{Number dying within a certain time period after disease onset}}{\text{Number with a specified disease}}$$

*The difference between Cause-specific mortality and Case-fatality rate* is:
Cause-Specific Mortality – the at risk population is both those who have developed the disease AND those who have not yet developed the disease
Case-Fatality Rate – the at risk population is those who have already developed the disease and is a measure of the severity of the disease

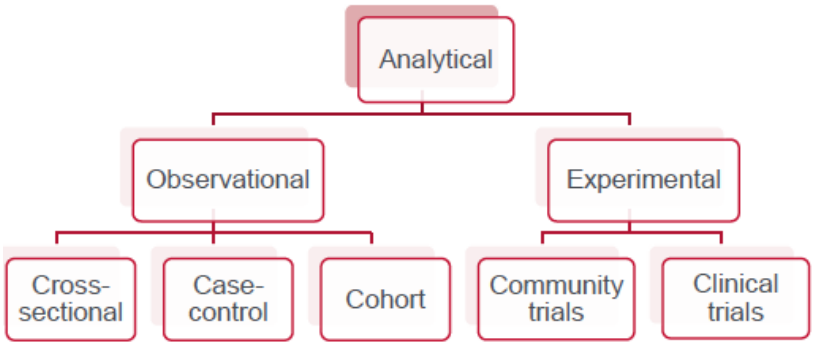
**EPIDEMIOLOGICAL STUDIES**

**I. Descriptive studies – these studies are used to generate hypotheses**




- Case reports – these studies describe a rare condition or an unusual case of a more common disease. They can be written by doctors who observe such cases.
- Case series – these studies describe an unusual clinical course.
- Surveys – when these studies are not making comparisons (this would make them observational studies), they are used to estimate the frequency and distribution of an outcome (disease) in a defined population.

## II. Analytical studies – these studies are used to try and explain




**Observational studies – 'we watch'**
Observational because there is no individual intervention, treatment, exposures occur in a "non-study" environment (i.e. not randomly)**.** Individuals can be observed prospectively, retrospectively or currently.
Data collected to see what is happening

*A. Cross-sectional studies*
Random sample of individuals taken and included in study and each person is classified according to their exposure and disease status after enrolment i.e. Participants are not selected on exposure or outcome status because no knowledge of the participants' exposure OR disease status is known at the start of the study. This means that you can investigate many exposures and many outcomes. The time frame may range from a point in time to a restricted time period. These studies provide prevalent data.

- Measures the prevalence of risk factor and disease
- Measurements of exposure and effect are made at the same time
- Deals with individuals
- Easy and economical to conduct
- Difficult to assess the reasons for associations
- Can investigate multiple exposures and outcomes
- Leads to investigation of the cause
- Helps to assess health care needs of population

| Advantages | Disadvantages |
|---|---|
| • Can study multiple exposures & outcomes<br>• Relatively fast & inexpensive<br>• Good for permanent risk factors<br>• Can determine prevalence of exposure and disease in population<br>• Can estimate almost all measures of association<br>• Less potential for bias than case-control studies | • Not good for rare exposures or outcomes<br>• Temporal issues/reverse-association (Did exposure or outcome come first?)<br>• Prevalent, not incident data<br>• Not good for diseases of short duration<br>• Statistical association does not necessarily mean causal association |

*B. Cohort studies*
Participants are enrolled in the study based on their exposure status (E+ or E-) and all participants must be disease free at the start of the study. They are then followed over a pre-determined time period to estimate the rate of disease in each cohort. These studies provide incidence data. Cohort studies are the only observational study design where it is possible to determine the temporal sequence with 100% certainty.

- Deals with individuals
- Longitudinal studies
- Prospective studies (exposure and effect)
- Prospective and retrospective (investigation time)
- Investigates late or chronic effects
- Alerts the risks of development of disease
- Expensive because large groups of people
- Losses to follow-up
- Follows those exposed and unexposed – what is the difference in disease? Allows sequence of events to be examined

| Advantages | Disadvantages |
|---|---|
| • Good for rare exposures<br>• Establish temporal sequence<br>• Incidence data<br>• Can study several outcomes of a single factor<br>• Can calculate all measures of association<br>• No recall bias<br>• Better control of exposure and disease definitions (if prospective) | • Not good for rare outcomes<br>• Often expensive<br>• Often long follow-up times<br>• Loss to follow-up<br>• Changes in diagnostic methods over time may bias results<br>• Can study only one or a few exposures (unless general population cohort) |

*C. Case-control studies*

Participants enrolled in the study based on their disease status (D+ or D-). These studies retroactively determine the frequency of exposure (E) in the cases (D+) compared to the controls (D-). This can be done with currently available information i.e. Data already exists. This is a good study option in rare diseases and outbreak investigation.

| Advantages | Disadvantages |
|---|---|
| • Good for rare outcomes<br>• Often fast & inexpensive<br>• Good for outbreak investigations<br>• Good for identifying multiple risk factors for a single disease | • Not good for rare exposures<br>• Temporal issues/reverse-association<br>• Most subject to bias<br>• Cannot estimate exposure or disease rates<br>• Cannot calculate all measures of association |

- Deals with individuals
- Longitudinal studies
- Investigates the causes of disease
- Simple and economical to conduct
- Diseased vs. non-diseased – what is different?
- Particularly efficient if disease is rare

**Cohort vs. Case-Control Studies**

| Cohort | Case-Control |
|---|---|
| Rates of disease in exposed and non-exposed | Proportion exposed in people with and without the disease |
| Prospective – minimizes risk of bias | Problem in selection of controls<br>Recall bias |
| Rare exposures | Rare diseases |
| Several outcomes (several exposures if population-based cohort) | Several exposures |
| Large populations, time consuming, losses to follow-up, expensive | Economical |

**Experimental studies – 'we play'**

Researchers (randomly) allocate test subjects to the groups being compared. Things are made to happen (interventions etc.)

The experiment condition is applied to existing patients to decide upon appropriate therapy OR those currently free of symptoms to decide on an appropriate preventative strategy.

*A. Community trials*

Treatment is allocated to individual people and is used most often by pharmaceutical companies to test efficacy of a novel drug, also for hospital procedures, vaccines and evaluate contraceptive practices

*B. Clinical trials*

Treatments are allocated to groups and are rare.

Advantages:

- Most efficient for investigating causality – can ensure "cause" precedes "effect"
- Ensure possible confounding factors do not confuse the results – patients are allocated to treatment in any way investigators choose (usually randomly)
- Ensure treatments are compared efficiently (i.e. Same number in each treatment group for maximum statistical power or for more specific combinations/interactions)

Disadvantages:

- Expensive & time-consuming due to monitoring of large set of subjects over a long time period
- Loss to follow-up may occur

- Ethical problems giving experimental treatments – equipoise and ethics
- Often intervention studies screen out "problem" subjects (e.g. elderly, pregnant women etc), who may have a special reaction to the treatment – may restrict generalisation of results

*To reduce bias:*

1. Use of a control group: need a comparison group for new treatment/intervention given a placebo or current standard treatment
2. Blinding: person unaware of which treatment group they are in = single-blind. If doctor/nurse/treatment assessor also blind = double-blind. Person interpreting the results also blind = triple-blind study.
3. Randomisation: subjects allocated to treatment group according to some chance mechanism – necessary to avoid systematic bias
4. Consent before randomisation: prevent bias in eventual composition of group, check individuals for eligibility and consent for each treatment

**MEASURES OF ASSOCIATION**

Ratio measures:
- Relative risk, Risk ratio, Rate ratio (RR)
- Odds ratio (OR)
- Incidence rate ratio (IRR)

Difference measures
- Risk difference (RD)
- Attributable risk (AR)
- Population attributable risk (PAR)

**Relative risk (RR)**

The ratio of the risk of disease in exposed individuals to the risk of disease in unexposed individuals (also called risk ratio).

| | Diseased | Not Diseased |
|---|---|---|
| **Exposed** | a | b |
| **Unexposed** | c | d |

$$RR = \frac{\text{Risk in Exposed}}{\text{Risk in Unexposed}} = \frac{a / (a + b)}{c / (c + d)}$$

Interpretation:
- RR = 1: No Association
- RR > 1: Positive association - Individuals in the exposed group have an increased risk of developing the disease than those in the unexposed group
- RR < 1: Negative association - Individuals in the exposed group have a decreased risk of developing the disease than those in the unexposed group

**Coun**

| | Diseased | Not Diseased |
|---|---|---|
| **Exposed** | a | b |
| **Unexposed** | c | d |

**Odds ratio (OR)**

Ratio of the odds of the disease in exposed individuals to the odds of disease in unexposed individuals

$$OR = \frac{\text{Odds of disease in Exposed}}{\text{Odds of disease in Unexposed}} = \frac{a*d}{b*c}$$

**CONFOUNDING & EFFECT MODIFICATION**

There is often a third factor, which may have an important influence on the apparent relationship between the two variables studies – independent and dependent variable. There could be more than one confounding factor.

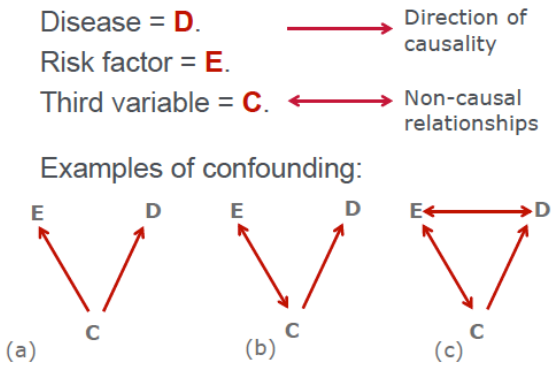
Confounding vs. Effect Modification

In the presence of confounding, the association between exposure and disease is the same, or similar, at each level of the third variable, but the crude and adjusted OR/RR differ.

In contrast, when effect modification (interaction) is present, the association between each exposure and outcome is different for different strata. The difference can be in direction or magnitude.

***Confounding***

A confounding variable is an extraneous factor that wholly or partially accounts for the observed effect of the risk factor on disease status. It is a form of bias.




Criteria for confounding:

1. A confounder has to be associated with the exposure
2. A confounder has to be associated with the outcome
3. The confounder should not be on the causal pathway between the exposure and the outcome (i.e. not be an intermediate variable)

If the third factor can (at least partially) explain the relationship between the risk factor of interest & disease status then confounding is present.

There are different ways to deal with confounding:

- Design
  - » Randomisation
  - » Restriction
  - » Matching
- Analyses
  - » Standardisation
  - » Stratification
  - » Multivariable analyses/statistical modelling

Perfect confounding occurs when the overall relative risk and relative risk of both strata (confounder present and absent) is 1. Often RR/OR differ slightly but are similar to each other. The summary overall relative risk (Mantel-Haenszel OR) adjusts for confounding. If RR/OR in strata are very different (~50%) to each other then confounding may not be present

***Effect modification***

Unlike confounding, if the third factor modifies the relationship between risk factor and disease, then interaction is present.

Example: the relationship between salt consumption and stroke is different for men and women – women need a lot of salt to elevate their risk of stroke while men only need a moderate intake. Sex interacts with salt consumption in determining propensity for a stroke.

**BIAS**

**Bias** – A systematic error introduced into data during data collection. Observational studies are prone to bias and this can reduce the internal validity of the results. A valid measure of association in the study population will have the same value as the true measure in the target population. Once introduced into a study, bias cannot be removed but it can be controlled for.

***Categories of bias***

I. *Selection/Participation bias* – Due to problems with how subjects were selected/their willingness to participate. Systematic differences between those selected to participate and those who are not.

Participation bias describes error arising from systematic differences in the characteristics (for example, sexual behaviour) of those who agree to participate in a study compared with those who do not.

*Types:*

Non-response bias
Loss to follow-up/Follow-up bias
Admission risk bias (Berkson's Bias)
Detection bias
Volunteer bias

Non-response bias produces bias if the association between exposure and outcome is different in responders than non-responders. Such bias is often understated. To assess this bias: it is important to ascertain if extent of non-response in each group (E+/E- or D+/D-) is roughly equal, compare responders and non-responders however possible and item non-response bias is similar
>> *Information bias* – Due to inaccuracies in the measures taken on exposure, outcome or other factors of interest
- Incorrect classification or measuring of exposure, outcome and/or extraneous factors
- Misclassification bias: categorical variables
- Measurement error: continuous variables e.g. Weight, height, etc.

Types:
Reporting and Recall bias
Even among respondents who attempt to "accurately" report their past behaviours have problems with recall.
Incidence reports are generally more reliably reported than frequency reports
In general, longer recall intervals result in either underreporting or inaccurate recall of sexual practices and partners.
In case-control studies, cases may better recall behaviours or events associated with the disease being studied than controls e.g. Parents of children who have died from sudden infant death syndrome are more likely to remember details about their child's behaviours than parents of healthy children.

Social desirability bias
People believe there is a social norm and they then alter their answers to better align with the conceived social norm e.g. Men report more sexual partners and/or women report fewer sexual partners than either have actually had.

>> Confounding bias

***Combating bias***
*Questionnaire design and delivery method* may affect response rate and reliability of data generated
- Delivery (face-to-face, self-administered, postal survey)
- Characteristics of interviewers/questionnaire administrator
- Setting of interview/presence of third parties
- Language (technical terms vs. normal language)
- Literacy & complicated skip & filter questions
- Order of questions
- Open- or closed- questions
- Repetition of questions through re-wording

*Piloting* – due to questionnaire design factors, it is advisable to (extensively) pilot you questionnaire to determine the right length of the questionnaire, and to avoid unnecessary complicated skip and filter questions and the right language – refining definitions/language used and to remove any questions deemed too threatening (e.g. Immigration status) that could be removed.

*Community links* – researchers should have strong community links for hard-to-reach communities
e.g. particular ethnic groups, sex workers, "hidden" populations. This could be done through the community advisory group to help recruit a group for piloting.

**VALIDITY**
**Validity** – the absence of systematic bias in results.
Internal Validity – do the study results relate well to the target population?
External Validity – can you extrapolate the study results to the external population?

*Generalisability of data*

- Comparison of standard demographic characteristics between study participants and those eligible to participate but refused to participate
- Compare data from studies to nationally reported data on standard demographics to see if non-participation may have affected the broader generalisability of the study findings on the target population
- Can then weight your data accordingly BUT this assumes that behaviours in each demographic class are the same, which may not be true

**SAMPLING**

Census: Whole population is surveyed. This can be time consuming and expensive because everyone needs to be contacted to participate.
Sample: proportion of population is surveyed
Descriptive study: take sample to describe characteristics of population
Analytical study: take sample to assess associations between factors (exposures) and outcomes (disease)

Main stages to sampling:

1. Determine when/what to sample
   - Based on research question
   - How subjects are selected will impact on validity
   - If subjects aren't representative of the population, then conclusions may be incorrect
   - Must establish inclusion and exclusion criteria before sampling
     »Inclusion criteria: e.g. healthy, age range, gender, etc.
     »Exclusion criteria: e.g. rule out people already on medication or who might have a bias
   - Define your populations
     »Study population: population of individuals chosen for study i.e. "who gets in"
     »Target population: who you want to know something about; larger population that you can extrapolate results to (internal validity)
     »External population: might be able to extrapolate results to this very large population (external validity)
   - Sampling units: each individual measurement; subjects in study population
   - Sampling frame: list of all subjects in study population

2. Determine how to sample the chosen sample population
   - Sampling strategy will determine the nature of the extrapolations you can make from the sample population

   - Non-probability sampling:
     »NO random selection process
     »Convenience sampling: sampling units chose because they're easy to get
     »Judgement sampling: investigator chooses what s/he determines to be units representative of population
     »Purposive sampling: units chosen on purpose because of their exposure or disease status (in an analytical study)

   - Non-probability sampling (cont.):
     »Pros: relatively easy to do, cheaper if choose subjects based on convenience, appropriate for a homogenous population
     »Cons: can produce biased results if subjects not representative of target population; can limit how far you can extrapolate results

   - Probability sampling:
     »Uses some form of random selection process

»All individuals in the population have the same probability of being included in the sample
»Simple random sampling: a fixed percentage of the sample population is chosen using a formal random process
-- Sample should be representative of target population and this requires that the sampling frame be known
»Systematic random sampling: sampling interval ("j" or "k") is computed as the study population size divided by the required sample size
-- Starting point in first interval is selected on a formal random basic
-- Don't need to know the complete sampling frame
-- If interval is related to sampling frame in any way then can get bias

»Stratified random sampling: before choosing participants, the sampling frame is broken down into strata based on some factor likely to influence the characteristic being measured, then simple or systematic random sampling is conducted within each strata
-- The percentage sampled in each stratum does not have to be equal, but must be weighted according to actual distribution

3. Determine sample size

With too few subjects, valid results may not be found and with too many subjects can waste time and money, along with possible ethical considerations (depending on study).

- Basic steps:

i) Establish expected variation in variable of interest

From expected proportion (p) if estimating a proportion or from population variance ($\sigma 2$) if estimating a mean

ii) Select level of confidence that your estimate will include the true value in the population (usually 95% & $\alpha=0.05$)

iii) Specify desired precision (total width) of confidence interval

- "Allowable error" i.e. within X% of true value
- 5% typical unless $p<0.1$, then use $p/2$

iv) Use appropriate formula to calculate sample size, sample mean, comparing two proportions and two means.

- Error:

»Type I error ($\alpha$): rejecting null hypothesis (H0) when it's true
»Type II error ($\beta$): accepting H0 when it's false
»Power ($1-\beta$): probability of finding a significant effect if there's really one to be found (i.e. a real difference exists)

- Larger associations are easier to detect than smaller ones and the less the study groups differ in your character of interest, the harder it will be to separate them