EPIDEMIOLOGY

Training module

1. Scope of Epidemiology

Definitions
Clinical epidemiology
Epidemiology research methods
Difficulties in studying epidemiology of Pain

2. Measures used in Epidemiology

Disease frequency
Disease risk
Disease rate

3. Study of Causation

Pitfalls
Types of epidemiological studies
Measures: Risk, Odds ratios, Numbers- needed-to-treat

4. Accuracy of Clinical tests and Investigations

Validity
Reliability: measurement of agreement
Sensitivity and Specificity
Likelihood ratios
Predictive value

5. Presentation and analysis of Epidemiological Data

Types of variables
Data presentation
Summary statistics
Standard error
Z scores
95% Confidence intervals
Comparing means and proportions
Probabilities and significance
Tests of significance: parametric, non-parametric
Tests of correlation
Survival curves
Power and sample size
Multivariable statistics
Choosing a statistical test

6. Evidenced-based Medicine

Levels of evidence
Meta-analyses
Systematic reviews
Web-based sources of evidence

1. Scope of Epidemiology

1. Definition: Epidemiology is the study of *distribution* and *determination* of health related states or events in a specified population and the application of this study to the control of health problems (Last, 1988)

Note that this definition is broad and covers more than "disease" and thus encompasses the entity of Pain. This definition covers the incidence, prevalence of painful condition and the characteristics of who develops the condition. Knowledge of frequency and distribution of a health state is used to test epidemiological hypotheses which results in *determinants* of the condition. The process of *determining* true associations or causes requires complex research methodology.
Classical Epidemiology is population- oriented and addresses the community origin of health problems e.g. nutrition, behaviour, psychosocial state of the population.
Clinical Epidemiology is patient-orientated to health settings and aimed at improving diagnosis, treatment and prognosis of patients who already suffer a medical condition.

2. Clinical Epidemiology: This is the application of Epidemiology to clinical medicine: If applied to the study of 'pain', clinical epidemiology includes health service planning, evaluating the effectiveness and availability of pain management/ treatment services, evaluating long-term management and natural history of painful conditions, identifying syndromes, diagnostic features and most importantly developing preventive measures against development of pain. Prevention includes primary prevention (stopping pain from ever occurring), secondary prevention (early intervention to limit further development of the condition) and tertiary prevention (minimizing the disability and impairment that could arise from the condition).

3. Epidemiological Research Methods: Three principle methods used: Surveys, Cohort studies, Case-control studies. Note preventative and clinical trials are essentially prospective cohort studies in which exposure to disease prevention or control measure is randomly allocated. The systematic collection and analysis of data involves determination as to whether a statistical association exists. This association is based on determining the probability of this association occurring by chance. Other explanations are sort for the association including bias and confounding factors. It is crucial to consider whether the findings are valid and generalizable. Finally a judgement is made as to whether the association represents a cause-effect relationship. Judgment criteria include the magnitude of the association, consistency of findings from other studies and biological credibility, time sequence of the association and dose-response relationships.

4. Difficulties in studying Epidemiology of Pain:
The multi-dimensional features of "pain" has made it difficult to classify and measure. The risk factors are numerous and may act in varying and interactive ways.
Pain is not constant but changes in many directions over time.
Patients with pain may be hard to identify and follow over time.

- Be familiar with the IASP classification of Pain and its deficiencies.
- Be aware with the mechanistic approach to classification of pain.
- Be aware of selected classification systems and their strengths and deficiencies eg International Headache Society, American Rheumatological Association diagnostic criteria of fibromyalgia.

2. Measures used in Epidemiological studies

- **Measures of Disease frequency:**

Incidence: the number of new cases of a condition that develops during a given period in a population at risk.

Prevalence: (sometimes called *point prevalence*) is the number of persons of a defined population who have a disease or condition at a given point in time.
Note the term *period prevalence* refers to the number of persons who have the disease over a time period, i.e., at a point in time plus the number who develop the condition during that period (incidence). It is thus mixed measure and not useful for scientific work.
Prevalence increases with the incidence and the duration of the disease. Assuming low prevalence and steady state conditions: the prevalence approximates the incidence multiplied by average duration of condition.

- **Measures of risk**

Risk is defined in epidemiology as the proportion of persons who are unaffected at the commencement of a study and who are exposed to the risk event. The risk condition may be a disease, injury or death. The persons at risk are a *cohort*. Risk rate can be confusing, because only those truly at risk in the population should be included in the denominator. This is sometimes very difficult to determine. It is also difficult to define what they are at risk of, as there are a number of subsets of the outcome.

- **Measures of rate**

A rate is a frequency occurring over a defined time divided by the average population at risk.
Incidence density incidence is frequency of new events per person time (person months or years).
Cumulative Incidence

$$= \frac{\text{number of new cases of a disease during a given period of time}}{\text{total population at risk}}$$

Incidence rate $= \frac{\text{number of new cases of a disease during given time period}}{\text{total population at risk at mid-point of study}}$

Prevalence rate is actually a proportion of people with condition at the time studied.

Standardization of rates To compare rates across populations, crude rates may need to be standardized.

3. Study of causation in epidemiological investgation and research

In discovery of cause and effect relationships there are several steps.
Risk factors or protective factors are sort.
Common pitfalls encountered in identifying these factors are:

- **Bias**
  (1) inadequate sampling techniques and selection bias.
  (2) Recall bias.
  (3) Measurement error.

- **Random error**: Random errors are usually in both directions. This type of error decreases the statistical power of the study making it less likely to find a real association or effect.
- **Confounding**: Confusion of 2 or more causal variables.
- **Synergism** Interaction of 2 or more causal variables such that the total effect is greater than the sum of individual effects.
- **Effect modification**: A third variable alters the direction or strength of the association between 2 other variables.
-

Types of epidemiological studies

(A) Descriptive studies:

Concerned with distribution of disease including determining which sub-groups of the population do and don't have the condition.

(1) Population (**correlational)**

(2) Individuals

case reports

case series

cross-sectional surveys

(B)Analytical studies

(1) Observational: used to detect associations

Case-Control studies

Case-control: all subjects have condition and matched with group who don't have condition. Risk factors compared.
Advantage: cheaper and of shorter duration.
Disadvantage: recall bias, matching samples.

Cohort studies;

Types of cohort: (prospective and retrospective).
Design and conduct: one group is exposed to a factor, the control group is not
Disadvantage: expensive, long duration, matching groups limited (only factors collected at the start can be measured) Retrospective subject to recall bias.

(2) Intervention studies (clinical trial studies):used to test hypotheses.
Types: Randomized controlled trials +/- blinding +/- active control +/- placebo
Unique problems: external validity, cost, often prolonged time to completion
Sampling methodology varies for different types of studies.

Note: RCTs use randomization techniques. Despite the use of these techniques which should eliminate bias between experimental and control group (internal validity), there can still be bias such that the study population is not representative of the true patient population. This can result in the study having poor generalisability or poor external validity.
e.g. - patients that are prepared to take part in clinical trials may be different to the normal population.

- The study patients may have more extreme parameters of an abnormality which makes them candidates for the trial. (This often results in the phenomenon of regression to mean).
- The reporting of the study may not include those patients who elected to abandon the study. It is thus important to know whether a study reports outcome in terms of "intention to treat", and states reasons for patient withdrawal from the study.
- The conditions of the study environment may be far removed from the usual clinical setting.

Measures of Risk

Often data is presented as two-by-two tables called contingency tables.
Absolute risk
Relative risk
Attributable risk
Population attributable risk
Odds ratio
Numbers Needed to Treat

The **Relative Risk (RR)** is an expression of the risk of having a specified condition after exposure to the specified variable compared to the risk in the control condition eg RR of an ulcer in NSAID group is 16/100 cp to 12/200 i.e., 0.16/0.06 =2.7
Remember that this will change over time: the cumulative chance of dying whether or not one is exposed to treatment X has a risk ratio of 1 over time! It is important that the time period is specified.

In a cohort the RR is the cumulative incidence of a condition amongst those exposed compared to the non-exposed.
In a case control study it is not possible to calculate the rate of development of disease since subjects are chosen who have the presence or absence of the disease. The relative risk can be estimated by calculating the ratio of the odds (odds ratio or OR) of exposure among the cases as compared to the controls. It is incorrect to use the term "relative risk" in this situation. Estimating RR from OR can be inaccurate if the risk of the disease in the general population is greater than 5% however this is very unusual.

The Odds Ratio (OR) is the ratio of the odds of exposure to a factor in those that have a condition compared to those without the condition. It is a mathematically stable and unbiased estimate of RR in case-controlled studies.

The Attributable Risk or AR is the risk difference and reflects the absolute risk difference of developing a condition between those exposed and not exposed.
AR = Incidence of exposed -incidence of those not exposed. This may be expressed as a % as AR/Incidence of exposed.

In case control studies it can be estimated by

$$\frac{(RR-1)}{RR} \quad X100\%$$

Population Attributable Risk (PAR) = AR x prevalence of exposure

Note: The RR provides a measure of strength of association whereas the AR provides a measure of impact of condition on the population in general. A patient factor can result in a higher risk of having a condition but due to the relatively low incidence of the condition attempts to treat the factor have little impact and visa versa.

Absolute risk is the actual probability that an individual will experience a specified outcome during a specified period. The change in absolute risk (e.g. **absolute risk reduction**) the same as the attributable risk when applied to treatment effects. e.g. a drug caused a 0.44% reduction in absolute risk when subtracting incidence of an event in a control group from a treatment group, e.g., 1.91-1.47%= -0.44%
The problem with this measure is that it is difficult to understand the clinical meaning of 0.44% reduction. For that reason the change in risk is often expressed as a **relative risk reduction** and expressed as a ratio of risk in intervention group to risk in control group

e.g. $\frac{1.47}{1.91}$ = 0.77 (relative risk when on treatment) or1-0.77=0.23 (23%risk reduction)

Note, that this is misleading as it does not convey how many are effected by this risk reduction, which is better expressed by absolute risk reduction. Furthermore, to make this figure relevant to public health, the prevalence is used to compute the decrease in risk to the population.

Numbers Needed to Treat: NNT
This is the number of patients that you would need to treat with a specific intervention over a given time for one person to benefit. To calculate: express absolute risk reduction as a reciprocal

e.g. $\frac{1}{0.0044} = 227$

A NNT of 227 means that 227 people would need to be treated to have one success

Note that the NNT applies to each study and will be different for different patient populations.
The NNT will be lower for those with a higher absolute risk.
If the odds ratio or relative risk is not statistically significant, the NNT is infinite, i.e., reporting the 95%CI of the NNT is important.
The NNT can also be calculated from the odds ratio. When the control event rate is high >50%, odds ratio is unreliable for estimating NNT.
Another similar measure is the **Numbers Needed to Harm** (NNH).

4. Clinical Epidemiology: Accuracy of clinical tests and investigations

- Normal values for a test i.e., (values in patients without the condition in question) generally follow a normal distribution. This may be skewed to right or left, be more or less distributed, and sometimes have two peaks rather than one. The operator or the instrument may have an *error*, which may be random, or biased in one direction, or may be a result of normal biological variation.
- Generally, there is an overlap of test results with those that have a condition and those who are well. The more extreme the value of the test result, the greater the likelihood that it is diagnostic. The higher the incidence of the disease or condition, the more likely there will be overlap with some of those who are healthy with the same test value as those who have the condition.

- Test validity, and reliability, determine the sensitivity and specificity of the test. Sensitivity and specificity are factors that predict the presence or absence of a disease or condition.
- The predictive value of a test is also influenced by the prevalence of a condition in the population being tested.

Validity is the extent to which a measurement corresponds to the true state of the phenomenon being measured.
Reliability is the extent to which repeated measures by a particular observer in a stable environment produce the same result. Reliability is a necessary pre-requisite to validity.

Types of Validity:
Criterion validity is established when the result of the test conforms with a known gold standard for diagnosis e.g., tissue obtained surgically.
Construct validity is when the result is consistent with other manifestations of the disease e.g., pain scale measures higher when patient is moaning and sweating.
Content validity is if on close scrutiny the test includes all characteristics of the phenomenon that are generally agreed upon, but not others for which there is no agreement.
Face validity similar to content validity, but assessed on more superficial grounds, i.e., on face value test seems to be in agreement with what it is measuring.
Concurrent (convergent) validity: shows how well a test correlates with other well-validated tests of the same parameter administered at about the same time.
Predictive validity: How well the test can forecast some future criterion such as job performance or recovery from illness.
Therapeutic utility: Clinical usefulness.
See this web site for an excellent discussion of validity

http://www2.chass.ncsu.edu/garson/pa765/validity.htm

Types of Reliability
Scalability
Reproducability
Repeatability

- Reliability may be expressed by correlation co-efficients. Note that in the behavioural sciences it is rare to find reliabilities >0.9. The square of the correlation co-efficient expresses the percentage of true shared variance e.g., if the correlation co-efficient is 0.9, the 2 measures have 0.81 percent in common. If the reliability or correlation coefficient drops to 0.8 the 2 measures have only 64% in common, with the error component being about 1/3.

Reliability may be measured by test-retest method,
internal consistency methodology (Cronbach's alpha)
split half method
parallel form method
Measuring agreement between raters is another form of reliability:
e.g., **Percent agreement**. This is easily measured. However,
- does not tell the prevalence of the finding and

- does not tell how the disagreements occurred e.g., if the positive and negative results equally distributed between 2 rater, or if one consistently found more positives than the other.
- does not tell how much agreement has occurred by chance.

M**easures of Agreement**

The principles of measuring agreement are quite complex. There is an excellent web site related to this topic in the reading list.

There is overall no agreement as to which is the best way to analyze agreement between statisticians. The basic issues are however quite simple. The most appropriate test will depend on,

(1) **Goals** of analyzing agreement e.g., validating a new or old rating scale will require different approaches.
(2) **Type of data, assumptions, distribution**
(3) **Reliability (consistency) vs Validity (accuracy)**
(4) **The method of rating descriptive or model-based** on an explicit or implicit model of how and why raters choose.
(5) **Components of disagreement:** e.g., definition or rating scale

Some approaches to this are:

(1) **Percentage of agreement** (there is an argument of a high level of chance agreement when one or two categories predominate).
(2) **Pairwise Correlations** e.g., Pearson's, but problem is that this will not measure consistent difference between judges that correlate. (e.g., one judge always scores 5 points above the other judge).
(3) **Intra class correlation coefficient**. Intraclass correlation (ICC) is used to measure inter-rater reliability. Although Pearson's r may be used to assess test-retest reliability, ICC is preferred when sample size is small (<15), or when there are more than two tests (one test, one retest) to be correlated. ICC may be conceptualized as the ratio of between-groups variance to total variance. The equation is a ratio of the variance of the "effect in question" divided by the sum of variances of individual judge, the variance of subject, the variance of interaction of subject and judge and the variance of the error. The measure calculated is similar to the r^2 of Pearson's. Because the judge only interacts once with each subject, it is not possible to estimate the error or the interaction separately.

To use intra-class correlation for inter-rater reliability:

(1)Construct a table in which the column variable is the raters (A, B, C, ...).
The row variable is some grouping variable, which is the target of the ratings, such as persons (Subject1, Subject2, etc.)
The cell entries are the raters' ratings of the target on some interval variable or interval-like variable, such as some Likert scale.
(2)Assess the inter-rater (column) effect in relation to the grouping (row) effect, using two-way ANOVA.

For interest the formula for intra class correlation can be derived. See figure 1 excerpt from http://www2.chass.ncsu.edu/garson/pa765/reliab.htm#intraclass below:

Derivation of the ICC formula, following Ebel (1951: 409-411): Let A be the true variance in subjects' ratings due to the normal expectation that different subjects will have true different scores on the rating variable. Let B be the error variance in subjects' ratings attributable to inter-rater unreliability. The intent of ICC is to form the ratio,

ICC = A/(A + B). That is, intra-class correlation is to be true inter-subject variance as a percent of total variance, where total variance is true variance plus variance attributable to inter-rater error in classification. B is simply the mean-square estimate of within-subjects variance (variance in the ratings for a given subject by a group of raters), computed in ANOVA. The mean-square estimate of between-subjects variance equals k times A (the true component) plus B (the inter-rater error component), since each mean contains a true component and an error component. Given B = mswithin, and given msbetween = kA + B, substituting these equalities into the intended equation (ICC =A/[A+B]), the equation for ICC reduces to the formula for the most-used version of intraclass correlation

(Haggard, 1958: 60): ICC = rI = (msbetween - mswithin)/(msbetween + [k - 1]mswithin) where msbetween is the mean-square estimate of between-subjects variance, reflecting the normal expectation that different subjects will have true different scores on the rating variable mswithin is the mean-square estimate of within-subjects variance, or error attributed to inter-rater unreliability in rating the same person or target (row).

k is the number of raters/ratings per target (person, neighborhood, etc.) = number of columns. If the number of raters differs per target, an average k is used based on the harmonic mean:

k' =(1/[n-1])(SUMk-[SUMk2]/SUMk). ICC will approach 1.0 when mswithin = 0 -- that is, when there is no variance within targets (ex., subjects, neighbourhoods -- for any target, all raters give the same ratings), indicating total variation in measurements on the Likert scale is due solely to the target (ex., subject, neighbourhood) variable. For instance, one may find all raters rate an item the same way for a given target, indicating total variation in the measure of a variable depends solely on the values of the variable being measured -- that is, there is perfect inter-rater reliability.

The formula above varies depending on whether the judges are all judges of interest or are conceived as a random sample of possible judges, and whether all targets are rated or only a random sample, and whether reliability is to be measured based on individual ratings or mean ratings of all judges. These considerations give rise to six forms of intraclass correlation, described in the classic article by Shrout and Fleiss (1979).

Fig 1. Shrout, P.E., and J. L. Fleiss (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin (86): 420-428.

Ebel, Robert L. (1951). Estimation of the reliability of ratings. Psychometrika 16: 407-424.

NOTE: Problem with intra class correlation is that the ICC is strongly influenced by the variance of the trait in the sample/population in which it is assessed. ICCs measured for different populations might not be comparable.

4. **Kappa**

Kappa is a chance-corrected measure of agreement proposed by Cohen. It is based on the assumption that some agreement between raters is accounted for by chance. There are many who believe that the Kappa statistic is fundamentally floored.

Example

Judge 1

Judge 2	Positive test	Negative test	Total
Positive test	40	9	49
Negative test	6	45	51
Total	46	54	100

The **percentage agreement** is $\frac{40+45}{100}$ = **85%**

The probability of positive test for Judge 1 is 46/100
The probability of positive test for Judge 2 is 49/100
The probability that both pick a positive by chance is 46/100 X49/100=22.5%
The probability that both judges pick a negative is 54/100 X 51/100= 27.5%
So out of 100 people for 27.5 of 100 the judges would agree by chance on a negative and 22.5% of the time they would agree on a positive so overall, in 50% of cases the judges would agree by chance.

Kappa =

$$\frac{\text{sum of the frequency of observed agreement - sum of expected frequency of agreement}}{\text{N- sum of the expected frequencies}}$$

Thus **Kappa**= $\frac{(40+45)-(22.5+27.5)}{100-(22.5+27.5)} = \frac{85-50}{50}$ = 0.7 i.e. **70%** agreement

Notice that this is less than the 85% percentage agreement of raw data; i.e., in the examples above, the trait prevalence has essentially doubled.
See references for a full discussion of this.
HOWEVER, Note:
There are many who believe this reasoning is flawed.
The following example will illustrate one reason why:

Judge 1

Judge 2	Positive	Negative	Total
Positive	80	10	90
Negative	5	5	10
Total	85	15	100

Note the judge **agreement on raw scores** is $\frac{80+5}{100}$ = **85%** (as above)

However the **kappa score** is $\frac{(80+5) -((85/100X90/100) +(10/100X15/100))}{100- -((85/100X90/100) +(10/100X15/100))}$

$$= \frac{85-78}{22} = \mathbf{0.38}$$

Thus even though in both groups the judge agreement on raw scores was 85% the kappa scores were vastly different. Some believe that this is because Kappa is significantly effected by trait prevalence.

Pros
-Kappa statistics are easily calculated and software is readily available
-Kappa statistics are appropriate for testing whether agreement exceeds chance levels for binary and nominal ratings.

Cons
-Kappa is not really a chance-corrected measure of agreement.
-Kappa is an overall index of agreement. It does not make distinctions among various types and sources of disagreement.
-Kappa is influenced by trait prevalence (distribution) and base-rates. Thus not comparable across studies.
-Kappa may be low even though there are high levels of agreement
With ordered category data Kappa must be weighted on opinion
-Kappa requires that two rater/procedures use the same rating categories. It is not suited to different rating systems between judges
Tables that purport to categorize ranges of kappa as "good," "fair," "poor" etc. are inappropriate; do not use them.

Recommended Methods

Data type	2 raters	Multiple raters
Dichotomous	-Log Odds ratio -McNamara's test of homogeneity -tetrachoric correlation co-efficient -raw agreement scores	-latent trait model (continuous trait) -latent class model (discrete) -measure pairs of raters
Ordered-categorical (usually implies continuous trait)	-Polychoric correlation coefficient - McNamara's test of homogeneity - Association models (RC)	As above
Likert-type data	-Pearson's correlation coefficient - McNamara's test of homogeneity - Other tests for likert data	-One factor common factor analysis and measure correlation of each rater with common factor. Describe raters marginal distributions

Nominal	-raw agreement -kappa (to check only if agreement occurs more than by chance). Ignore magnitude -Marginal homogeneity	-Latent modelling -Marginal homogeneity -measure each pair of raters separately

Sensitivity and Specificity

Sensitivity Se

The % of those who are true positives compared to the number of testing positives.

Specificity Sp

The % of those who are negative and test negative compared to the total number of true negatives.

Positive Predictive value +PV

The % that are true positives compared to the total number of test positives.

Negative Predictive value -PV

The % of those that are true negatives compared to the total number of test negatives.

Positive Likelihood Ratio +LR

The ratio of the chance of the test being positive if have condition compared to the chance of testing positive if don't have condition.

Negative Likelihood ratio -LR

The ratio of the chance of the test being negative if have the condition compared to the chance of testing negative in don't have the condition.

Prevalence P

The percentage of the population who have the disease.

Note that :

(1) The gold standard is often far from a gold standard.
(2) The false positive rate is generally easily determined because doctors act on those patients testing positive but often little is known of the false negative rate.
(3) For some conditions there is no hard and fast criteria for diagnosis. Angina is an example. Sometimes the validity of a test is established by comparing results to a clinical impression based on history and examination and then the test is used to *validate* the diagnosis. e.g., bowel manometry to confirm irritable bowel syndrome.
(4) If one choses a gold standard test that is weak and then tries to compare a new test with the standard a potentially improved new test may seem worse!!
(5) Sensitive tests are used if there is a serious consequence of not detecting the disease.
(6) Specific tests are used if trying to confirm a diagnosis when there is other evidence.
(7) Often a trade-off needs to be reached between specificity and sensitivity when clinical data take on a range of values eg blood sugar and diabetes.
(8) Receiver operator characteristic curves (ROC) may be used to find the best values for the cut -off point for a screening test. ROCs plot sensitivity against false positive rate. The curves are really a plot of likelihood ratio. By plotting ROCs of several tests on the same axes, it is possible to determine the best test
(9) The trade-off point between specificity and sensitivity will vary with the stage of a disease

Example

	Disease		
	Present	Absent	Total
Screening test			
Positive	900	4950	5850
Negative	100	94050	94150
Total	1000	99000	100000

Sensitivity 900/1000= 90% P =1000/100000=1/100 or 1%
Specificity 94050/99000=95%
PV+= 900/5850= 15.4%
PV-= 94050/94150=99.9%
LR+= 900/1000: 4950/99000= 0.9/0.05 =18
LR-=100/1000:94050/99000= 0.1/0.9415= 0.105

Note:

- Values for sensitivity and specificity and likelihood ratio are usually estimated on small samples of people who are known to have or not have a condition. Because of random chance in any one sample, particularly a small sample, these values may be misrepresentative. It is thus important to know the 95% confidence interval (CI), which will determine the precision of the estimated value, e.g, if sample size is 10 people, with observed sensitivity of 75%, the true sensitivity could be anywhere from 49% to 100%. If sample size is 50, the sensitivity is anywhere from 60-90% according to the 95% CI.

- The sample chosen for estimating sensitivity and specificity may be different from those that the test is used for. Eg when test is developed those tested will clearly "have" or "not have" the disease. Those tested may be at a less obvious stage of the disease and the specificity and sensitivity may be quite different in this population. Thus even though Se and Sp are thought of as independent to prevalence, in fact several patient characteristics eg stage of disease and duration of disease may change both Se, Sp and prevalence since different kinds of patients are found in high- and low- prevalence populations

- Bias may become relevant in the testing of a sample for the purpose of Se and Sp estimation. For instance the effects of a positive and negative test on physician behaviour is such that a positive test is pursued and a negative test is not. Therefore the true false negative rate may not be clear. Having some added information about a patient may bias the tester eg radiologist. The bias increases the "apparent" agreement between the test and standard validity.

- **Predicitive value** of a test is a useful measure determining the chances of having the disease given the result of the test. It is sometimes referred to as posterior or post-test probability

- **Prevalence** is also called priori or pre-test probability.

$$\textbf{Positive predictive value} = \frac{\textbf{Sensitivity x Prevalence}}{\textbf{Sensitivity x Prevalence + (1-specificity) x (1-prevalence)}}$$

Note predictive value of the test can be increased mainly by:

(1) increasing **specificity** of the test, e.g. if above specificity is increased to 98%, the positive predictive value will increase to 31.3%

(2) increasing **prevalence** in the population being tested.
Note effect of prevalence on positive predictive value with constant specificity and sensitivity.

Prevalence(%)	PV+(%)	Sensitivity(%)	
0.1	1.8	90	95
1.0	15.4	90	95
5	48.6	90	95
50	94.7	90	95

When Se and Sp are high and prevalence is high, the PV is high. When prevalence is low, the PV of the test is very poor.

Estimating prevalence in a particular patient setting

- Medical literature, local database, clinical judgement. Although not precise this value this value is more likely to be accurate than clinical judgement alone. Usually **prevalence is more important than specificity/ sensitivity in determining predictive value.**

Prevalence of condition can be increased by:

Referral process
Selecting demographic groups
Specific clinical situations

5. Presentation and analysis of epidemiological data

Descriptive statistics are used to summarize data to a form that can be used for comparison between study groups or populations. In summarizing data, there is loss of information. The method of summarization chosen depends on the type of variable and how it is measured.

1. Types of variables

1. Discrete
2. Continuous

- **Discrete variables:** -Dichotomous (only 2 alternatives eg dead/aliveivor)
  -Multichotomous (e.g. race)

Scale: Nominal - if no order e.g., race
Ordinal - if order of progression e.g., stage of disease, improvement in mobility: none, mild moderate, considerable
Numeric discrete e.g., parity, episodes of angina

Comparisons will involve **proportions**

- **Continuous variables:** These that can assume any value e.g., height, weight. Limited only by accuracy and precision of measuring instrument

Scale: numeric
Comparisons will often involve **means**

Some variables can be considered in a number of forms by grouping eg age in decades. Thus data can be reduced after collection if appropriate.

2. Data Presentation

1. **Frequency distribution**: for each variable value: the number of times or proportion the observation occurs.

The histogram will show central tendency, distribution shape, percentile rank

2. **Contingency table** representing "r" rows of the number of variable options and "c" columns representing the number in the various groups being compared. For dichotomous data from cases and controls, a 2x2 table would be used

e.g., Peptic ulcer

	cases	controls	total
NSAID use	18	100	118
No NSAID use	12	200	212
Total	30	300	330

3. **Histogram** for grouped values (important for intervals to be the same or misleading)
4. **Frequency polygon** for continuous data

3. Summary statistics

(A) Discrete data

This is most simply represented as the proportion of, or frequency of individuals falling within each category. Because there are no numerical scores it is not possible to calculate mean, variance or standard deviation.

(B) Continuous data

1. Measures of **central tendency**: mean, mode, median. The choice of measure depends on the distribution. Sometime, median may be more informative. Mean is generally used when applying statistical tests.

2. Measures of **Spread or Variability** e.g., range, standard deviation, variance (sum of difference between each observation and mean) which is squared then divided by the number of observations minus one. The standard deviation (SD) is the square root of this.

If distribution is normal the SD can describe the distribution fully. In this case 68% of observations fall within +/-1 SD of the mean, 95% within 2 SD of mean and 99% within 2.5 SD of mean.

NOTE: the distribution of sample means will be normal if either of the following 2 conditions are satisfied:

1. The population from which the samples are selected is normal in respect of the parameter
2. The size of the samples is relatively large (around $n = 30$ or more)

4. Standard error (SE) of a mean and a proportion:
The standard error of an estimate of a population parameter (e.g. the mean or a proportion) is a measure of the 'long-run' variability of that estimate when it is calculated from successive random samples (such as those samples obtained from drawing several samples of say 10 cards drawn from a pack of cards and counting the number of red cards drawn compared to number of black cards drawn).

Statistically, it is the square root of the variance of the estimate calculated from a long run of random samples (i.e. the standard error is the standard deviation of a very large sample of estimates of a population parameter)
However the SE can be estimated for use as a statistic from a single sample (simple random sample) from a formula using,

- variance of the statistic and
- number of cases in the sample

So, for example:
The **standard error of the mean** is square root of $s^2/(n-1)$
where s^2 is the variance of the variable calculated from the sample and
n is the sample size.
Alternatively standard error of mean is:

$$\frac{SD}{\sqrt{n}}$$

The **standard error of a proportion** is square root of $p(1-p))/(n-1)$ where p is the proportion (e.g. proportion answering 'yes' to a question), and n is the sample size.

NOTE: The standard error measures the variability of the sample-mean from the mean of the population from which the sample was taken. It is related to the **variability of the original population** and the **size** of the sample from which it was drawn.

5. *Z* score
The mean of a distribution of sample means is called the expected value *X*
The location of *X* in the distribution of sample mean can be specified by a *z* score:

$$z = \frac{X\text{-mean}}{\text{se}X}$$

The *z* score can then be used to identify the proportion of the of the normal distribution that is either side of the *z* score. i.e, *z* score of 1.65 corresponds to a point along the x axis of a normal distribution at which 95% of the area under the curve is in the larger proportion and 5% beyond it. Note that the standard deviation and mean of the normal population is needed to determine the standard error if *z* scores are used.

6. 95% Confidence Interval
When the sample used to estimate the standard error of the mean is large, the distribution of means tends toward a *normal distribution*. From such a sample we can therefore estimate that the true mean of the population is within 1.96 standard errors of the sample mean with a probability or *p* of 0.95 or 95%. Thus creating the 95% Confidence intervals of the mean. Using confidence limits can be very useful when comparing differences between outcome of 2 or more groups, e.g. of an RCT. Because the confidence limits will give the range through which the true mean or

proportion of each group lies, we can appreciate how large or small the true difference between two or more groups could be. If the 95% confidence interval of the *difference* of 2 samples does not include zero, then the difference is likely to be significant.

7. Comparing proportions and means

(1) Means

If we want to compare means from 2 independent samples and we can assume the samples are large enough to assume the normal distribution, then the calculated **standard errors** are good estimates of the standard distribution of the normal population. We can use this to find the confidence intervals.
E.g., comparison of 2 samples with means x and y with large sample sizes N1 and N2. The expected difference between the sample means is = the difference between population means. The variance of the difference between 2 independent random variables is the sum of their variances. Hence the standard error of the difference is calculated by taking the square root of the 2 standard errors.

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

The confidence interval of the difference of mean can then be calculated (i.e +/- 1.96 multiplied by the Standard error of difference) and if this does not include zero, it is likely that the difference is a true difference that was unlikely to occur by chance.

(2) Proportions:

The same principle can be applied to *differences* in proportions:
The difference in proportions is determined. Then the variances are again added and the standard error of the proportions calculated. A calculation of the standard error (SE) will allow calculation of the 95% confidence interval of the estimation as above:

$$\sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

provided the conditions of a normal distribution are met the confidence interval of the difference is calculated in usual way.

A ratio of proportion may be easier to interpret. The SE is more difficult to calculate because it is a ratio rather than a difference and does not follow a normal distribution. However it can be calculated by taking the log of the ratio.
The final formula for the SE of a *ratio* of proportions is

$$\sqrt{\frac{1 - p_1}{n_1\,p_1} + \frac{1 - p}{n_2\,p_2}}$$

Note:
Increasing the sample size will reduce the standard error.

Improving the sampling design will reduce the variance of the sampling distribution, and will thus also reduce standard error.

Both strategies will increase the precision of the estimates by narrowing the range of the confidence interval.

8. Probabilities and significance levels

An analysis of epidemiological studies will always involve an estimate of the probability that the finding of the study occurred by chance:-

(1) develop a null hypothesis either *Ho: po=p1* or *Ho:* **RR**=1

(2) develop an alternative hypothesis *H1: po≠p1 or Ho* **RR≠1**

(3)Perform a test of statistical significance appropriate to the circumstances, which will lead to a probability *p value* that the observed difference has occurred by chance. The larger the value of the test for statistical significance the smaller the *p* value. The test can be one sided or two-sided depending on the relevance of knowing the direction of the difference. The test size is increased by the magnitude of the difference or the effect size and the sample numbers (decreasing the variability of the estimate), i.e., the significance test size can be seen as depending on the ratio between the effect size in the numerator and the variance in the denominator. Thus, even very small differences can be significant if there is a large enough sample size, and similarly a large difference may not achieve statistical significance if the sample size is small and variability high. The *p* value is thus a guide only. The actual *p* level should be expressed as it helps interpret the findings. The confidence interval is far more informative as it reflects the range in which the true effect is likely to lie with 95% or 99% of confidence. If the interval of a difference does not include zero, the difference is likely to be real *ie P< 0.05.* The confidence interval also gives a guide as to whether a study that shows no difference between groups may be flawed, i.e., a type 2 error has occurred. A large confidence interval suggests that a larger sample may have a different result (i.e., study insufficiently powered to rule out the finding occurring through chance), whereas narrow confidence levels add more weight to the findings being correct.

(4) Note that the level of significance may be calculated in regard to a difference in one direction (one tailed) or both directions (two-tailed)

(5) Note the statistical significance of an association or difference may not be of any biological or clinical significance.

10. Tests of statistical significance

Many statistical tests used in medicine compare population parameters calculated as **means**.

Z scores and T tests are used in comparing means of continuous numeric data about each subject. They are examples of **parametric tests,** because they test a parameter. Other medical studies compare frequencies of events. Because there is no numeric score, there is no mean or variance so **non- parametric** tests need to be used such as Chi-Square.

.

Note that dichotomous data expressed as proportions which follows a binomial distribution which approximates a normal distribution, can be analyzed using the model of the normal distribution.

(1) **Parametric tests**
A *Z* score statistic provides a model for other statistical tests. It is a ratio of the numerator (difference between sample mean and hypothesized population mean) and the denominator (standard error measuring how much difference is caused by chance). A test statistic greater than 1 suggest a difference is more likely than by chance. The alpha level is used to obtain a critical value for the test-statistic. For Z score test, the alpha level of 0.05 produces a critical level of 1.96.

As a rough guide,
An alpha of 0.05 is equivalent to a ratio of approximately 2 (Z score 1.96) (the difference must be twice as big as chance).
An alpha of 0.01 is equivalent to a ratio of approximately 2.5 (Z score 2.58).
An alpha of 0.001 is equivalent to a ratio of approximately 3 (Z score 3.30).

Note that to use Z scores, we require the value of population standard deviation to compute the standard error. The Z scores critical values are based on the normal sample distribution.
The value of the standard error is unchanged by the "treatment or intervention"

If a population standard deviation is not known, we use the T statistic
The T statistic uses the sample standard deviation as a substitute. The only difference between the formulae of Z and T, is the use of the sample standard error rather than population standard error.

Both T test and Z test are examples of tests based on calculating a *critical ratio* of the parameter and standard error of that parameter. The value of the ratio is then looked up in tables to determine the level of significance. If the sample size is larger than around 30, a ratio of greater than 2 usually means the difference is *unlikely* to be due to chance. The tables used adjust the ratio for the sample size by use of *degrees of freedom.* This term refers to the number of observations that are free to vary. Usually it is N-1. The greater the degrees of freedom, the closer the variance of the sample approximates the variance of the population and the closer the T score represents the Z score. The shape of the t-distribution changes with the degrees of freedom.

***z* scores** may be used to **test hypotheses**. In hypothesis testing, all the necessary ingredients of the test are not present. Specifically, we do not know the value for the population mean and need to estimate it.

The most frequently encountered significance tests encountered in the medical literature are the,
-(1)**T test** used for continuous data and can be used with sample size less than 30.
t-test: The normal distribution (z scores) can be used for very large samples or if the variance of the population is known. More commonly, this is not the case and the T test is used for determining the significance of differences in means. The T distribution of values is different to the normal distribution, and the criteria for using the test is less restrictive.
There are several varieties of T tests for different circumstances:

One sample T test: Comparing a mean of a sample to the normal **population**
Independent samples T test: Comparing 2 independent samples. The variance of each of the samples are added together to create pooled variance. Note the assumptions of this test are that, both samples come from normally distributed populations and the variances of the populations are the same (Homogeneity of variance should be close to 1.00)
T test for Paired samples (eg related means eg same sample at time 1 and time 2)

The **ANOVA or analysis of variance test** is used to compare differences in 3 or more independent samples of continuous data. The one way ANOVA is an *F* test ie *F* is determined by $F = \frac{\text{between- groups- variance (or mean square as estimate)}}{\text{within- groups- variance (or mean square as estimate)}}$
The *F* test only tells us if there is a significant difference between groups but does not tell us where the significance lies.

Which tests are used depends on whether we have predicted there would be a difference:

If a hypothesis had been made about any two groups, a T test could be used
If we had not predicted a difference, then it is likely that when we carry out a large number of comparisons, some will be significant by chance. A number of tests are available which take this into account (Post Hoc tests eg Scheffe test).

Note **Levene's** test compares *variances* between groups rather than *means.*

The MANOVA test is used to compare 3 or more related means.

When more than one variable is being compared between groups, multiple factor-design studies can be performed and differences analyzed using multiple-factor analysis of variance.

-(2)Non parametric tests
-Most non-parametric tests make few assumptions about the population distribution.
-Non parametric tests are suited to data measured on ordinal or nominal scales.
-Non parametric tests are not as sensitive as parametric tests in detecting real differences.
Chi -square test used for testing hypotheses about the proportions of discrete data if the sample size for each cell is >5.
If less than 5 per group expected: 2 paired sample: use Exact McNamara test,
2 independent samples : Fisher's exact test
2 independent samples: Cochrane's Q-test

These tests are based on distributions different to the normal distribution and the assumptions of the tests must be met.

For **Chi-Square test**:
(1)Determine difference between frequency of data and the hypothesis frequency e.g., null hypothesis comparing to normal population.
(2) Square difference.
(3) Divide by frequency of expected.
(4) Sum values from all categories (Degrees of freedom are: number-of-categories -1)
(5) To determine significance consult table of Chi distribution and read for appropriate degrees of freedom).

Note, there are **2 types of Chi square tests**:
(1) **Goodness of fit** to determine if there are changes from expected by the null hypothesis.
(2) **Chi square test for independence** to test for a relationship between 2 variables
Note the
Phi-Coefficient is used to determine the degree of relationship between 2 variables.
Other tests:
Binomial test when only 2 categories. Equivalent to Chi-square when 2 categories

Sign test is a special application of binomial test for determining direction of difference.
It may be used to in matched subjects for differences between treatments and classifies difference as an increase or decrease. Used as alternative to related T test or Wilcoxen when data do not satisfy criteria of these tests

Mann Whitney U test converts individual scores into rank order from 2 separate samples to test a hypothesis about differences in the populations or different treatment. It can be used in the same situation as the independent T test where the scores can be ranked and when the samples do not satisfy the criteria of T test.

Wilcoxen T test
Uses data from repeated measures or matched-samples to evaluate differences in two treatment conditions. Used where differences scores can be ranked in order, but do not satisfy requirements of T test.

11. Tests of correlation

A correlation measures the relationship between 2 variables.
The correlation is described by;
The direction,
The form
The degree

The most commonly used correlation is the **Pearson's correlation *r*** which measures the linear correlation,

$$r = \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}}$$

r^2 value is proportional to the variation of one variable explained by the other

he Spearman correlation:
Is used to measure correlation between variables measured on an ordinal scale

Linear regression
This can be used to predict a value of X for a value of Y from correlation data.

12. Survival Curves

Success of a treatment is often measured in terms of survival. Data reporting mortality is not enough, as it does not indicate length of survival or loss to follow-up.
Two methods are used to collect this data,

(1) **Person -time Method**. eg measuring person years. This is useful for outcomes that occur repeatedly eg back pain

(2) Life Table analysis

(i) The **Acturial Method**. Calculates survival during fixed time intervals eg at one year survival is 0.83 and at 2 years it is 0.65.

(ii) The **Kaplan-Meier Method** use principle that the survival rate over a certain time is made up of the product of survival rate over intervals of that time. Each time a death occurs a new survival line is calculated. The curve appears as uneven steps.

Life table methods are subject to bias if two groups being compared have different rates of loss to follow-up.. It is important to also consider the pattern of survival over time as well as the end result.

Tests of significance for survival:
-Significance test of proportion e.g., t tests and z tests
-Logrank test probability for each death that each death would have occurred in the treatment group, and the probability it would have occurred in control group
-Proportional Hazards Models (Cox Models) . Used with Kaplan Meier curves using multiple logistic regression

13. Sample size and Power

There are two broad approaches to estimating sample size

(1)Consider the sample size needed for adequate statistical precision of a single estimate
or a difference between estimates.

(2)Consider the sample size needed for adequate statistical power for hypothesis testing,

NOTE: These two approaches can give quite different estimates of an 'appropriate' sample size,

(1)Taking the 'Precision of a Single Estimate' approach and assuming that the data will be mainly reported as percentages

We need to determine,
-the desired width of the confidence interval (say, ± 2 = 4 percentage points)
-an acceptable probability level for that interval (usually 0.05)
-an estimate of the population percentage of the 'focus' response variable (say, the proportion people with back pain in a town of 50,000).

We Estimate an Appropriate Sample Size when the 'True' Proportion of the 'Target' Variable is Unknown
-From pilot studies (if based on a small simple random sample).
-From previously published surveys.
-From a preliminary sample, that will be augmented when the desired sample size is known (sometimes known as two-phase sampling).
-From prior knowledge of the structure of the population and behaviour of the required statistic (e.g. in working with proportions we know, for absolute accuracy, that the upper bound of the sample size is achieved when p = 0.5

Example
(1) Sample size for estimating frequency of a condition
Formula for calculation of sample required to estimate the frequency of patients with back pain in a town of 50,000 persons. The researchers want an estimate within + or – 2% of the population value: that is with a SE of .01 (i.e. a 95% CI of + or – 2%): 95% of samples lie within 2 standard errors of mean

Assuming that about 0.3 will have back pain (eg based on other studies: 30% of population have had back pain)
n = p * (1-p) divided by SE squared. (see previous section on comparing proportions)
n = (0.3) * (1-0.3) divided by .01 * .01
=0.3*0.7 divided by .01 * .01
= .21 / .0001
= 2100

The previous example focussed on avoiding a type I error and did not consider detecting differences.
The power of a statistical test is the probability that the test will correctly reject the null hypothesis, i.e., if there is a true difference in a treatment group, the statistical test will be powerful enough to detect it. A type 2 error (ß) is an error of not detecting the difference when it truly exists. The power of a test is 1- ß. Note the power of a test will be *reduced* by:
- decreasing the critical cut off to detect a difference (reducing the alpha level)
- decreasing the effect-size,
- two- tailed rather than one tailed test,
- smaller samples

NOTE: In order to determine POWER, we must decide on an EFFECT- SIZE.
It is simplest to think of effect size as the difference in scores between an experimental and control groups. Simply considering the difference between means is unsatisfactory because it,

-fails to account for the unit of measurement: e.g. difference measured in millimetres would look greater than one measured in centimetres.
-fails to account for the variance or dispersal of scores: a small difference between means from two widely dispersed distributions would be less important than the same difference between two distributions with very low variances.

NOTE: One widely used effect-size is the difference in mean scores divided by the mean of the standard deviations of the two scores.

-Conventionally, as suggested by Cohen, effect-sizes may be classified:

- 0.2 = small effect
- 0.4 = moderate effect
- 0.8 = large effect.

There are 2 aspects to be considered in calculation of sample size

-(1)to calculate the number of subjects required to assure a given probability of obtaining a difference of given magnitude if one truly exists **(sample calculation)**

-(2) to calculate the probability of demonstrating a statistically significant effect of a given magnitude amongst a given number of subjects if it truly exits **(power calculation)**

Power analysis and Sample size determination

-Ascertain the **minimum true difference** between the groups that is deemed clinically meaningful

-Estimate the **expected variability** of the data

Chance of making a **type I error** or rejecting the null hypothesis when it is true is = to the probability level set which is commonly set at **5%**

Accepted level for risk of **beta error** is often **20%**. The power of the study is 1- beta or 0.8 ie the study is 20% likely to be making a type 2 error and failing to reject the null hypothesis when H1 is true ie 80% power of detecting a difference.

Sample size determination

Once the above have been determined, the specific formula used depends on the type of data, study design, question being asked, i.e.,

(1) Is the data paired data or unpaired?
(2) Is a beta error (type II or false negative) being considered in addition to an alpha error (type I or false positive)?
(3) Is the variance expected to be large or small?
(4) Is the standard 0.05 p value or 95% CI being used?
(5) What is the size of the difference or treatment effect?

The basic formula for sample size may be derived from the formula for a paired T test

$$t\alpha = \frac{\text{mean difference}}{\text{se (mean)}} \quad \text{se (mean)} = \frac{}{\sqrt{N}}$$

After substituting the formula and substituting *z* for *t* because *z* is independent of sample size

The formula becomes $N = \frac{z\alpha^2 * (s)^2}{\text{difference}^2}$

Note (1)that for a study that was not using the same group as a control and using the student T test for independent samples eg RCT then it would be necessary to calculate N for each group.

(2) the above formula does not consider a type II or beta error. A *z* level for the beta error must also be included.

Note also from the above formula that
(1) the larger the variance, the larger the sample size
(2) to have considerable confidence that a mean difference is real implies a small *p* and thus a large $z\alpha$. *If p=0.05 the* $z\alpha$ *=1.96.* In the formula this will be squared to make 3.84. A change in α to 0.01 would mean a change by a factor of about 75% $(2.58)^2=6.66$ in numbers.
(3) to detect a smaller difference means increasing sample size considerably because the difference is squared and in the denominator
(4) the beta error has not been added

Sample size considering Power, Precision and Effect Size

(1) Choose the appropriate formula (Note: * means multiplication)

Studies using paired T test (eg before after) and considering type I error only	$N = \frac{(z\alpha)^2 * (s)^2}{(\bar{d})^2}$
Studies using student T and considering alpha only eg RCT with one experimental group and one control group(considering type 1 error only)	$N = \frac{(z\alpha)^2 * 2 * (s)^2}{(\bar{d})^2}$
Studies using student T and (considering type 1 and type II error)	$N = \frac{(z\alpha+z\beta)^2 * (s)^2}{(\bar{d})^2}$
Studies using a test difference in proportions (and considering type I and type II errors)	$N = \frac{(z\alpha+z\beta)^2 * 2 * \bar{p}(1-\bar{p})}{(\bar{d})^2}$
N=sample size, $z\alpha$=value for alpha error, $z\beta$=value for beta error, $(s)^2$=variance $(d)^2$=difference to be detected, p= mean proportion of success	

(2) The values must be specified. The variance must be based on some knowledge of existing data, e.g., a pilot study, or based on reported literature. If the outcome is a change in proportion, then the mean proportion is easily calculated.

Note the calculation of power and sample size is often done using a computerized calculation. In addition to these basic formulae, there is the possibility to modify formulae, e.g., it may be easy to find control subject rather than cases or visa versa. In this case it is satisfactory to have unequal groups to maximize statistical power, but the analysis is more complicated.

14. A note on Multi-variable and Multivariate statistics

Multivariable statistics are analyses of the relationship of multiple independent variables to one dependent variable, whereas multivariate analysis refers to methods used in the analysis of multiple independent variables with more than one dependent variable. (The terms are often confused). Multivariate statistics are complex and will not be discussed in this document.
Note the term **Bivariate analysis** refers to the analysis of one independent variable and one dependent variable.

Multivariable statistics:
Statistics for assessing one outcome variable in relationship to multiple independent variables.
These statistics are generally based on a **general linear model**
i.e, $y = a + b1f1 + b2f2 + b3f3 + e$

Where y is the variable of interest, "a" is an anchor point or constant, "e" is error of measurement and the "b1,2,3" are factors that modify and (f1,2,3) are weighting of the factors.

Multivariate statistics are used:
(1) To understand the relative importance of independent variables when either acting alone or together in influencing a variable.
(2) To determine interaction between variables
(3) To develop prediction models.
(4) When it is difficult to conduct studies in which there are multiple factors that are difficult to control.

The type of statistical test used depends on whether the dependent and independent variables are continuous, dichotomous, nominal, ordinal or a combination.

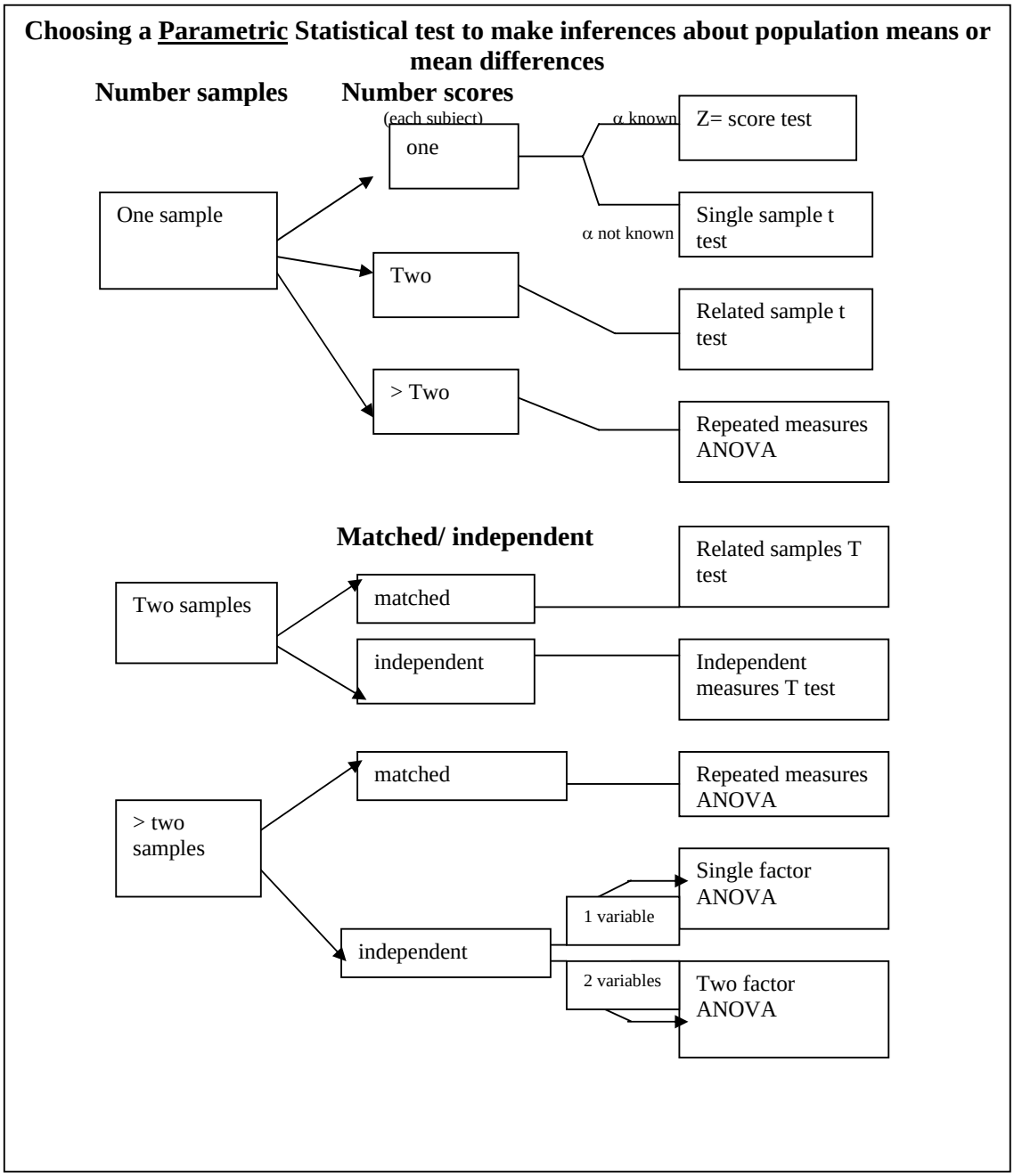
Examples are ANOVA, ANCOVA, multiple linear regression, logistic regression, log linear analysis, discriminant function analysis.

If the dependent variable is time related and dichotomous eg live/die then a commonly used form of logistic regression is the (Cox) proportional hazards method used to test for differences between Kaplan-Meier survival curves.

Choosing statistical tests

(1)Parametric tests

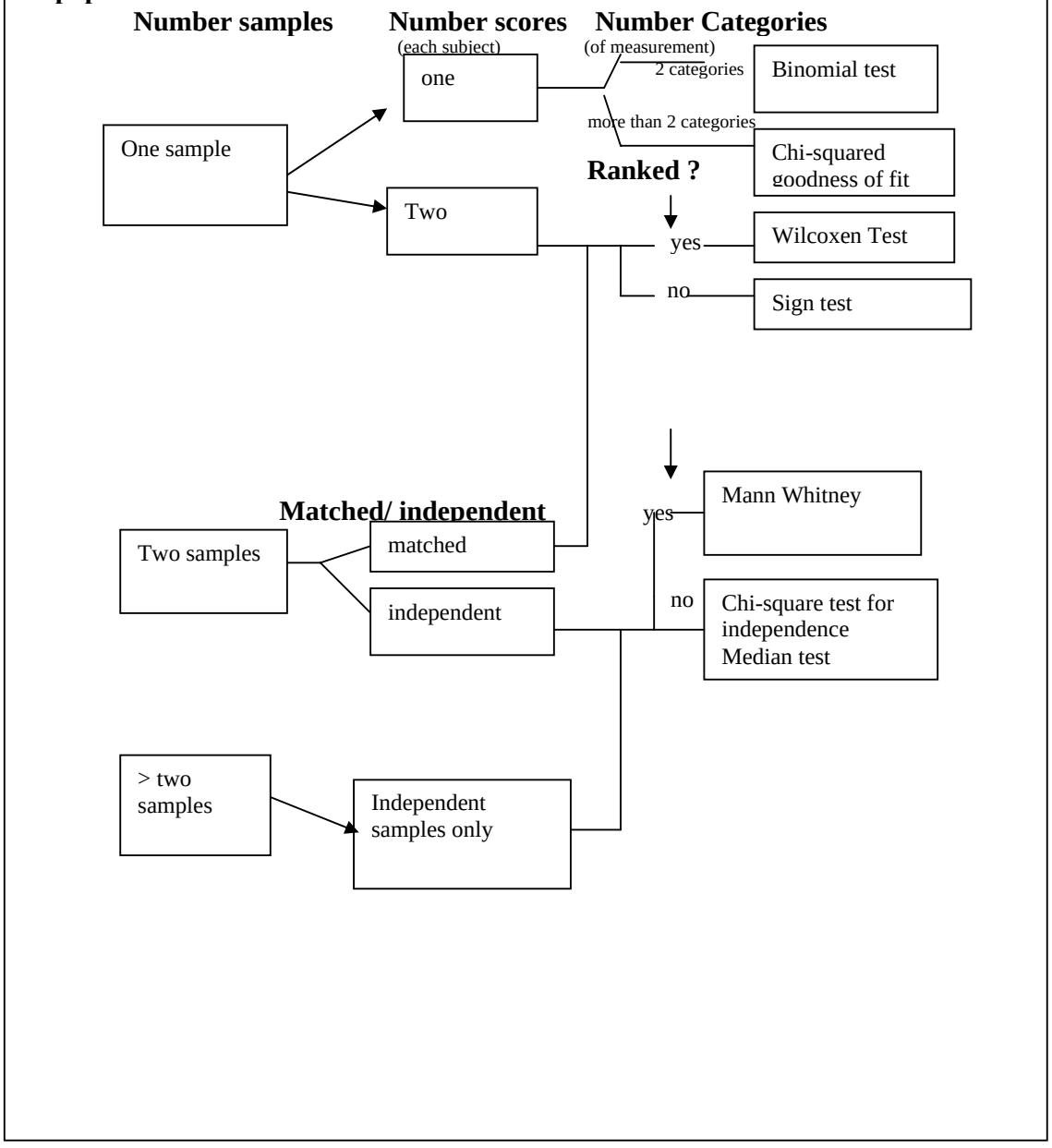
- Measurement on an interval or ratio scale
- Tests use means obtained from sample data as basis for testing hypotheses about population means
- Each test makes assumptions about population distribution and sampling technique



2. Non-Parametric tests

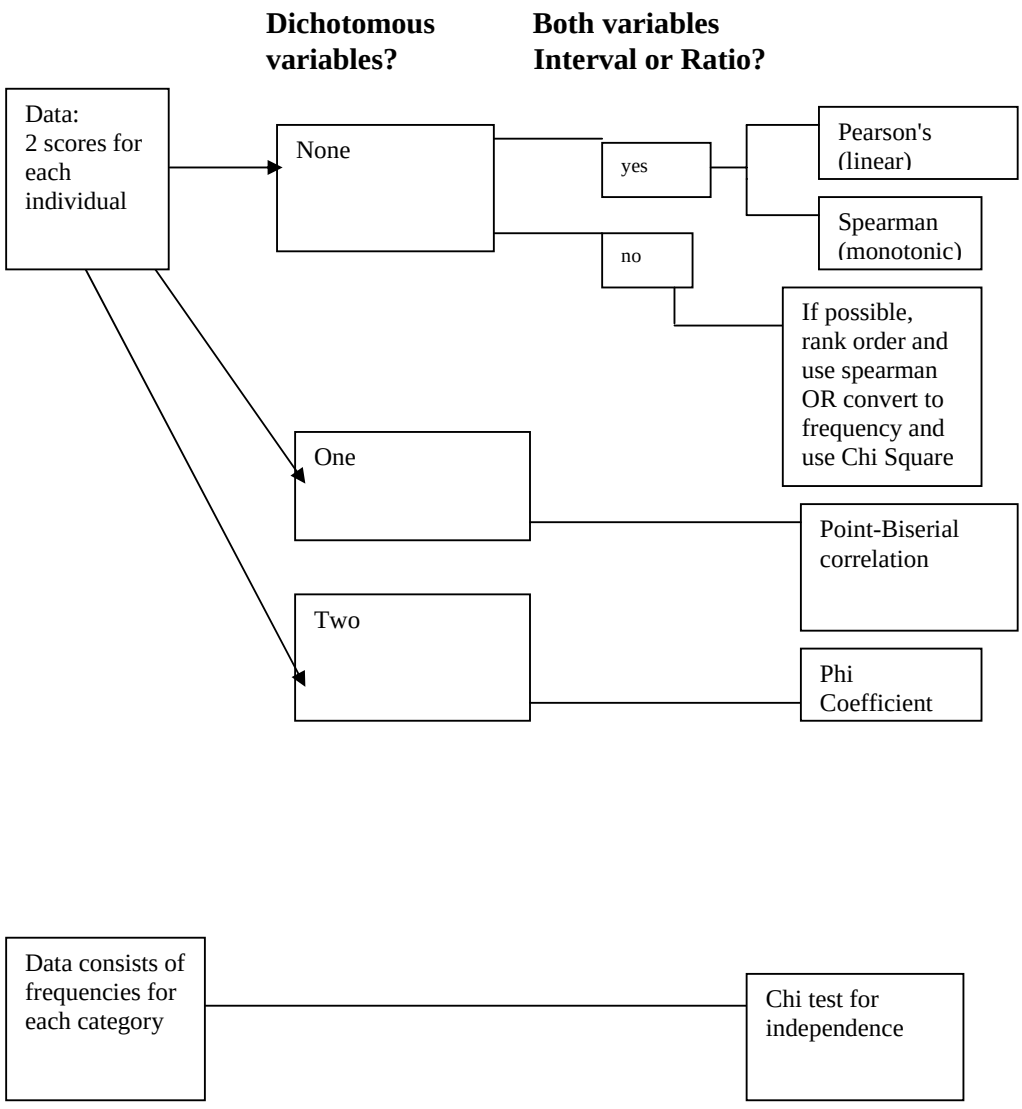
- Data measured on nominal or ordinal scale, therefore cannot compute means and variances required for parametric tests.
- The data doesn't satisfy assumptions of parametric tests such as population distribution assumptions.

Choosing a Non-Parametric Statistical test for making inferences about populations or population differences:



(3) Relationships between variables

Choosing a measure of correlation between 2 variables



6. Evidenced-based Medicine.

The evidence-based medicine is about finding the best available evidence, appraising the evidence, making the evidence and using the evidence. Generally the evidence-based medicine movement has adopted randomized controlled trials (RCTs) and systematic reviews (SR) and meta-analyses as the most powerful tools from which to gain evidence for the benefits of an intervention. It must be remembered, however, that RCTs are specific to the experimental environment and patient characteristics of that trial which may not necessarily mimick the true clinical situation of an individual or patient group. It is necessary to understand the downside of RCTs, as well as their benefits.

Systematic reviews may provide much information which can be very useful, just interesting or misleading. The review needs to be structured to answer the specific question of interest and to be relevant to the individual clinical situation. Many systematic reviews are inconclusive due to low numbers of comparable trials which is secondary to low quality and different methodology of the reviewed trials. In addition SRs are always subject to publication bias, i.e., there is a bias for reporting positive trials. Systematic reviews are of variable quality.

Descriptive techniques such as audit, large cross-sectional studies are sometimes the most practical way of providing evidence of a more general effect on a larger scale. Because of variations within and between populations, the numbers needed for these type of studies needs to be large, and the sampling technique is very important. It is important to understand the limitations and benefits of these types of studies. The following are some useful tools commonly used when analyzing evidence.

(1) When searching for a relevant systematic review for a clinical question it is important to add at least 3 parts to the question.

e.g., Is Dorsal column stimulation effective?

- Add in comparison to what?…
- Add in which patient group?
- Add for what condition?
- Add over what time interval?
- Add at what cost (economic/ morbidity)?

(2) It is also important to define what level of effectiveness is being measured.

Table of evidence classification:

Level I: Strong evidence from at least one systematic review of multiple well designed RCTs
Level II: Strong evidence from at least one well designed RCT
Level III: Evidence from well-designed trials without randomization, single group pre/post, cohort, time series or matched case-control
Level IV: Evidence from well-designed, non-experimental studies from more than one centre of research group
Level V: Opinions of respected authorities, based on clinical evidence, descriptive studies, or reports of expert committees.

A systematic review:
A systematic review identifies all relevant studies about a particular clinical question and analyzes these studies using stringent predetermined methodology which assess quality of the studies so as to reduce bias.

A meta-analysis
A meta-analysis is a review that assesses each trial separately and then summary statistics are then combined to give an overall result. A meta-analysis may not be possible if the data cannot be combined.

Collaborative overview
Also known as a Meta-analysis based on individual patient data is the central collection of data on each and every patient randomized.

Oxman and Guyatt index of scientific quality
This is a popular index to evaluate the quality of research overviews. The following questions are asked and scored:-

1. Were the search methods used to find evidence on the primary question stated?
2. Was the search for evidence reasonably comprehensive?
3. Were the criteria used for deciding which studies to include stated?
4. Was bias in the selection of studies avoided?
5. Were the criteria for assessing validity of studies included?
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?
7. Were the methods used to combine the findings of the relevant studies stated?
8. Were the findings of the relevant studies combined appropriately relevant to the primary question of the overview?
9. Were the conclusions made by the author(s) supported by the data and/or the analysis reported in the overview?
10. How would you rate the scientific quality of the overview?

Search Strategies:
In order to review relevant articles, it is necessary to adopt a search strategy that *includes* relevant terms, e.g. randomized or double blind or both, analgesic interventions with pain or adverse effect as outcome, or interventions using pain as an outcome measurement. The following databases should be searched: MEDLINE, EMBASE, Cochrane library, CINAHL, PSYCHLIT. Note EMBASE provides a large European database and may not be as readily available in Australia as the other databases. Older journal articles in particular are often not found using the electronic searching method. Journals known to contain articles related to the topic are hand-searched. The process has been formally described by Jadad *et al.* 1996
A web based site re searching bibliographic data bases is included in the reference list.

Trial quality and validity:
Scoring systems exist to measure quality of trials. These are largely based on giving points for minimizing bias: Randomization (appropriate),
Double blinding (appropriate),
Description of withdrawls/dropouts
In addition blinding of reviewers to author, journal year of publication, etc, improves consistency of quality stores.

It is important to note that,
(1) Inadequate **randomization** can overestimate treatment effect by **41%**
(2) Inadequate **double-blinding** can overestimate treatment effect by **17%**

Clearly it is not always possible to perform double-blinding or single-blinding studies.

Measuring analgesic effect and combining data:
Many pain trials report mean pain scores, or pain relief scores and their standard deviation, or standard error before and after an intervention compared with placebo. This assumes that the data follows a normal distribution.

Unfortunately, there is variation in placebo response between studies using the same and similar drugs. This has been extensively studied in relationship to acute pain studies. Both pain scores and placebo responses are not normally distributed. Placebo effect is more common when measuring smaller effect size. Thus using mean values to describe treatment response for analgesic trials is inappropriate. It is not generally possible or practical to access individual patient data. Some meta-analytic studies have converted mean scores to a dichotomous measurement, e.g., percentage achieving >50% pain relief, and found the calculation reliable when compared to actual patient data from multiple analgesic trials using placebo and various analgesics.

Two measures commonly used in analgesic research studies and applicable to systematic reviews and meta-analysis related to pain are,
(1) summation of pain and
(2) pain relief over time

i.e., the area under the pain-intensity or pain-relief versus time curve is measured before and after an intervention. This gives rise to the following measures:
Summed Pain Intensity Difference (SPID)
Total Pain Relief (TOTPAR)

These values can be derived from pain scores. Continuous data can be converted to dichotomous data by defining a cut-off level for pain relief eg >50% pain relief at a certain time.

There are formulae available to derive proportion of patients achieving at least 50% pain relief from mean data using the following outcome measures: categorical pain relief, categorical pain intensity, VAS pain relief, VAS pain intensity. When converted to this form, the data from RCT is able to be used in meta-analysis. Note the validity of these formulae have only been established for single dose studies in the acute pain model only.

Combining data:
Combining data will depend on the whether there is quantitative information, the type of data collected (combining continuous data may be difficult), and relative quality of the trials.
Methods of combining:
There are three stages
(1) L'abbé plot (scatter plot of control treatment verses active treatment
(2) Statistical test eg odds ratio or relative risk (combined with their 95% CI)
(3) Clinical significance measure NNT (Numbers needed to treat)

Systematic reviews databases:
Where to find them and what is available re PAIN
Some examples of resources for systematic reviews and meta-analyses
http://www.update-software.com/clibing/cliblogon.htm
The Cochrane library database contains over 1000 reviews related to pain. This may be accessed free of charge
Other useful sites are bandolier
1. http://www.jr2.ox.ac.uk/bandolier/

2. http://www.ahfmr.ab.ca/publications.html
Prevalence of chronic pain systematic review
Evidence for effectiveness of multidisciplinary pain programs for chronic pain

3. http://www.ccohta.ca
vertebroplasty
trigger point injections

4. Translating evidence into Practice
http://www.tripdatabase.com