# MedicalBiostatistics.com

From: Basic Methods of Medical Research, Second Edition
By A. Indrayan
AITBS Publishers, Delhi (+91-11-22054798, +91-11-22549313)

## Glossary of Methodological Terms

In the explanation, the terms included in this Glossary are in italics. They can be referred within this Glossary.

**Absolute risk** — Same as *risk*. Contrast it with *relative risk*.

**Absolute risk reduction** — *Risk* difference between the control group and the intervention group.

**Accuracy** — Truthfulness or correctness of a particular value of a measurement to the reality. Age recorded as 7 years 4 months and 14 days is more accurate than recorded only as 7 years, although this additional accuracy may be redundant.

**Adaptive design** — A research design that builds into itself at the time of protocol that certain specified aspects such as inclusion/ exclusion criteria, allocation strategy and, particularly the sample size can be changed at interim stages when so indicated by the data obtained at those satges.

**Addition rule (of probability)** — The probability of occurrence of one of two or more *mutually exclusive events* is the sum of the probabilities of their individual occurrence.

**Additive model** — A model stating that combined effect of two or more factors is the sum total of their individual effects – valid when there is no *interaction*.

**Adjusted correlation** — Same as *partial correlation*.

**Adjusted odds ratio** — The odds ratio obtained after eliminating the effect of other *concomitant variables* that might be affecting the OR. This is generally obtained by including these concomitants in the *logistic regression*.

**Adjusted rate** — The net rate obtained after eliminating the effect of other *concomitants* that might be affecting the rate. This adjustment increases the comparability between rates in different segments of population.

**Adjustment** — A procedure by which the effect of structural differences in the two or more groups is minimized—thus improving the comparability. Common methods of adjustment are *regression* and *standardisation*.

**Aetiology diagram** — A diagram that depicts the inter-relationship of various aetiological factors leading to the disease.

**Aetiological factor** — The characteristic that contributes to the occurrence of disease or a health condition. It may or may not be a causal factor.

**Age-adjusted death rate** — See *age standardisation*.

**Age standardisation** — A procedure of adjustment to remove the effect of differences in age composition of the groups to increase comparability. This adjustment is required when, for example, one group has older subjects than in the other and the outcome of interest is death. For deaths, this is called age-adjusted death rate. This could be done as *direct standardisation* or *indirect standardisation*.

**Agreement** — When two procedures, two observers, or two sites, etc., tend to give same result in each subject, they are said to be in agreement with one another. For qualitative

data the extent of agreement is statistically measured by *Cohen's kappa* and for qualitative data by *limits of disagreement*.

**Aleatory uncertainties** — Uncertainties arising from variation in the factors internal to the system such as biological, psychological and environmental. Contrast it with *epistemic uncertainties*.

**Allocation bias** — The systematic error in results arising from specific allocation of subjects to the test and control group, such as due to nonrandom allocation.

**Alpha error** — Same as *Type I error*.

**Alpha error rate** — The probability of *Type I error*.

**Alpha level** — Same as *significance level*.

**Alternative hypothesis** — A plausible hypothesis that is accepted by default when the *null hypothesis* is rejected.

**Analysis** — The process of going into the deep of a phenomenon, data-set, thought, etc., and looking at its various components.

**Analysis bias (for data)** — Gearing data analysis to support a particular hypothesis and ignoring aspects that contradict the hypothesis, e.g., using an inflated *P-value* for some statistical tests, using proportions when mean is appropriate, etc.

**Analysis of variance** — Breaking variance into its components such as within groups and between groups. This method is used in *regression analysis* and in various other situations but more commonly for comparing three or more means.

**Analytical study** — A study with the objective to identify the determinants or correlates of an outcome, such as *aetiological factors*, or to delineate their specific contribution to the outcome. See *observational study*, *experiment*.

**Anecdotal evidence** — Incidental observation that lacks scientific scrutiny.

**ANCOVA** — Acronym for analysis of covariance. A statistical procedure used when the dependent (or outcome) variable is quantitative, and the independent (or antecedents) variables are a mixture of quantitative and qualitative variables.

**ANOVA** — Acronym for *analysis of variance*. A method for analysis of quantitative outcome dependent on qualitative characteristics, particularly groups of subjects.

**Antagonism** — The situation where combination of two or more factors depresses the intensity of outcome compared to the sum of their individual effects: a negative *interaction* effect.

**Antecedent** — A characteristic that precedes the outcome. It may or may not be (fully or partially) responsible for the outcome.

**APACHE score** — Acronym for Acute Physiology and Chronic Health Evaluation: A scoring system that helps to assess the prognosis in case of critical condition of a patient. This is based on state of consciousness, reflexes, eye movements, blood pressure, etc.

**Apgar score** — Sum total of (0 to 2) scores assigned to each of heart rate, respiration, muscle tone, skin colour, and response to stimulation in a newborn. A low score such as less than 4 out of possible 10 is an indication of poor prognosis.

**Area sampling** — The method of sampling that uses geographical area as the unit. When the *sampling frame* of the *study units* is not available, area sampling can be used to select some areas by using a map, and then the subjects in those selected areas.

**Area under the curve** — The area from the *ROC curve* to the base: used as an indicator of the *efficacy* of a test in terms of *sensitivity* and *specificity* – can be used to compare performance of various tests. The area under the *concentration curve* is used to assess performance of a regimen and to compare performance of two or more regimens.

**Arithmetic mean** — Same as *mean*.

**Arm of a trial** — The case (or test or intervention) group is one arm of a *trial* and the control group is the other arm. The test arm may have more than one regimen.

**Ascertainment (or assessment) bias** — Bias due to paying more attention to cases than controls in assessment, or giving more attention to specific outcomes of interest. Also when more accurate history is given by diseased then nondiseased, or when they are more cooperative than the others. Also when outcome of interest is morbidity but there are early deaths that no longer can contribute to morbidity. Assessment bias also occurs when the observer is able to establish more rapport with some respondents than others, e.g., due to cultural similarities. Diabetes and gallstone may appear to be associated because diabetics are regularly checked for gall stone and nondiabetics are not checked.

**Association** — The property of change in one qualitative factor being accompanied by the change in the other. This change can be causal, incidental, or due to a third intervening factor. The association can be full or partial of various degrees, and can be negative or positive.

**Attack rate** — New spells during a specified time interval (such as the period of an epidemic) as percentage to the total population at risk in the same interval. See *secondary attack rate*.

**Attributable fraction** — The proportion of risk that can be validly assigned to a particular exposure.

**Attributable risk** — The additional *risk* that can be attributed to the presence of a risk factor. If risk of diabetes in those with both parents diabetics is 0.30 and in those with one parent diabetic is 0.25 (no other difference) then risk attributable to the diabetes in the second parent is 0.30 – 0.25 = 0.05.

**Attribute** — A qualitative characteristic.

**Attrition** — Loss of subjects during the course of the study. This can happen due to temporary unavailability, migration, refusal to participate, severe injury, or unrelated death.

**Attrition bias** — Systematic difference between the groups in the pattern of *attrition* of the subjects.

**Bar diagram** — A diagram appropriate for disjoint categories to show the number of subjects or mean or rates by bars of corresponding height.

**Baseline data** — The data that show the status of the subjects at the initial stage, generally at the time of beginning of the study. The baseline is used for evaluating any subsequent change either due to intervention or otherwise.

**Bayes' rule** — The rule that converts probability of A given B to the inverse probability of B given A. Most common use of this in medicine is in obtaining the probability of disease given signs/symptoms by first obtaining the probability of signs/symptoms given disease, and in obtaining *predictivities* on the basis of *sensitivity* and *specificity*. Both require some additional information.

**Before-after study** — Assessing subjects before an intervention and after the intervention to find changes—thus assessing the utility of the intervention without using a control group; also called a non-controlled trial. Part of the effect could be psychological.

**Bell-shaped distribution** — Same as *Gaussian distribution*.

**Berkson's bias** — Occurs when the study is based on hospital cases but the exposure is such that increases the chance of admission. Thus hospital cases will have more exposure than hospital controls such as fracture in motor vehicle injury cases.

**Beta error** — Same as *Type II error*.

**Beta error rate** — The probability of *Type II error*.

**Bias** — A systematic error that can falsify or distort the results of a study. Contrast it with *random error*. Bias can arise due to a variety of sources:

(i) errors from faulty logic or incorrect premises;

(ii) nonrandom sample such as volunteers;

(iii) small sample that fails to represent the entire spectrum of subjects;

(iv) concomitant medication or concurrent disease that might seem unrelated;

(v) unmatched case and control group;

(vi) unaccounted confounders;

(vii) wrong or blurred definitions that give room to assessor to use subjective interpretations;

(viii) errors in diagnostic or screening criteria;

(ix) insufficient instructions regarding what to do in unusual or unforeseen situations so that observers use their own discretion that can vary from observer to observer;

(x) more care in assessment of cases than controls;

(xi) differential recall of past events by cases and controls;

(xii) not being able to recall events of the past, or selective recall of serious events and not of mild events;

(xiii) differential compliance of regimen and instructions by cases and controls;

(xiv) suppression of information by patients because of stigma or otherwise;

(xv) premeditated response by a group after consulting each other;

(xvi) aberration in response due to unsuspecting event such as sudden illness or death in the family;

(xvii) deteriorated response in repeat testing due to fatigue in either observer or respondent, or improvement due to learning effect;

(xviii) some control subjects or some in treatment group getting unaccounted therapy outside the study;

(xix) detection of some cases in early phase of the disease and some in the late phase;

(xx) prevalent cases could be more of those who survive for longer duration, and thus may be of better health;

(xxi) the study is based on hospital cases but the exposure is such that increases the chance of admission;

(xxii) observer able to establish rapport with some and not with other subjects due to cultural or other differences;

(xxiii) observer being more careful or attentive for specific type of subjects or special responses;

(xxiv) faulty instrument that gives unpredictable readings;

(xxv) improper use of instrument;

(xxvi) systematic error in measurement, which could be either due to the instrument, or due to the observer;

(xxvii) subjects changing response just because they are being observed;

(xxviii) more accurate history by diseased than nondiseased or better cooperation by one group than the other;

(xxix) selective nonresponse;

(xxx) exclusion of some patients during mid-course if they develop an unrelated condition such as injury;

(xxxi) third party or natural introduction of unsuspecting new intervention among the respondents or their habitat that could alter the outcome;

(xxxii) contamination of controls by spill-over effect of cases when the two groups are not isolated;

(xxxiii) selective handling of outliers;

(xxxiv) digit preference;

(xxxv) gearing analysis to support a particular hypothesis;

(xxxvi) using differential *P*-values to support a particular view;

(xxxvii) lack of statistical power to detect medically important difference;

(xxxviii) incorrect interpretation of results;

(xxxix) preparing a report to support a particular view by suppressing some facts;

(xl) selective publication by journals; and

(xli) bias in presentation of results emphasizing one particular aspect and de-emphasizing the other.

The list is still not exhaustive. Some of these are explained in more detail under the term for respective bias.

**Bibliography** — A list of citations of the related literature. This is different from the list of references because references are restricted to the literature actually cited in the text. Bibliography includes references to the other literature as well that are not cited but are related.

**Binary variable** — A characteristics that is assessed only in two categories such as ascites present or absent (or yes/no), or gender as male or female. A qualitative variable is mostly binary (in some cases could be *polytomous* and/or *ordinal*) but a quantitative variable can also be made binary by dividing into two categories such systolic blood pressure <140 and ≥140 mmHg.

**Binomial distribution** — A statistical distribution that describes the probabilities of various numbers of subjects with a characteristics out of total when the chance of one subject being positive is known, and the subjects are independent. If it is known that 30% of diabetics have eye problems, the probability that all five randomly selected diabetics will have eye problems is only $(0.30)^5 = 0.0024$, and the probability that 3 will have eye problems and 2 will not is ${}^5C_3(0.30)^3(0.70)^2 = 0.132$. Such probabilities come from binomial distribution.

**Bioequivalence** — Similar course of the disease process in the two regimens under comparison: also evaluated in terms of comparable bioavailability of drug products, say, within 80% to 125% with respect to area under the *concentration curve* and $C_{max}$. Also see *equivalence* and *clinical equivalence*.

**Biological plausibility** — Consistency with the present biological knowledge, which can be explained.

**BioSIS** — The UK-based citation service that processes articles from a large number of journals, books, monographs, conference proceedings, etc., on all topics of biological sciences.

**Biostatistics** — The science dealing with medical uncertainties in one or more groups of subjects—their identification, measurement, and control—leading to decision with least error.

**Bivariate analysis** — A statistical analysis of data by considering two variables together, such as maternal haemoglobin level and birth-weight category, or alcohol intake and occurrence of liver cirrhosis. One or both variables can be either quantitative or qualitative.

**Black box approach** — Using computer to solve problems without understanding the implications of the underlying procedure.

**Blinding** — Keeping the experimental subject or the observer or both ignorant about which subject is in the case group and which in the control group in a trial.

**Body mass index (BMI)** — Weight in kg divided by square of height in meters: a measure of obesity in adults. A BMI <20.0 kg/m$^2$ is considered lean, 20.0-24.9 is considered 'normal', 25.0-29.9 overweight, and ≥30.0 obese.

**Bonferroni procedure** — When two or more comparisons or other statistical tests of hypothesis are done on the same set of data, the total probability of *alpha error* can increase much beyond the prefixed level such as 5%. In order to keep the error probability within the specified level $\alpha$, the Bonferroni procedure is to do individual comparisons at $\alpha/k$

level of significance where $k$ is the total number of comparisons. If $k = 4$ and $\alpha = 0.05$, each comparison is done at 0.05/4 = 0.0125 level. This is a conservative procedure in the sense that the total level of significance is actually less than $\alpha$.

**Box-and-whiskers plot** — Same as *box plot*.

**Box plot** — A diagram that shows the median, the first and the third *quartile* (the difference between them giving an idea of dispersion, and the distance from median on either side, of skewness), and the lowest and highest value: a very effective method to present so many features of data in one diagram.

**Burden of disease** — A composite measure of premature mortality from a disease and morbidity equated to mortality through a weighting system based on age, discounted duration and severity of disease. Premature mortality is assessed in comparison to the mortality in the population with highest *life expectancy*.

**Capture-recapture method** — Consider counting of fishes in a pond. Capture some (say $m$), mark them and release. They thoroughly mix up in a day or two. Now capture again (say $n$) and find that $k$ of them have the mark. This gives an estimate of $m \times n/k$ fishes in the pond. In health and medicine, this method is used when incomplete count of cases is available from two independent sources such as hospital records, physicians in private practice, and death certificates. The overlap between them can be used to estimate the total number of cases if they are really randomly distributed.

**Case** — A person or unit of interest possessing a specified characteristic, such as a person with the disease or a family living in the conditions of interest.

**Case-control study** — Investigation of the antecedents in a group of cases and equivalent controls without introducing any intervention. The groups are defined on the basis of presence or absence of disease or any other outcome of interest. The logic of the design leads from effect to the cause. All case-control studies are inherently retrospective.

**Case-fatality rate** — The number of cases who die of those who suffer from a particular disease. This measures the *virulence* of the disease. Case-fatality in typhoid is low and high in tetanus. For chronic diseases such as cancers, the case-fatality may be 100% but measures such as 5-year death rate may be better indicators of the 'virulence' of disease.

**Case group** — The group of subjects that already has the disease or the condition under study.

**Case-referent study** — Same as *case-control study* but comparison is with some other disease and not placebo.

**Case series** — The group of patients with the disease of interest; generally consecutive cases reporting in a clinic, or observed in a community. There is no control group in this setup. Case series is one method of a *descriptive study*.

**Case study** — Study of one individual, particularly with regard to the *antecedent* and *outcome* factors as observed in that person.

**Categorical data** — All qualitative data are categorical. In addition, quantitative data are also summarized into categories. For example, if you measure *smoking index* of 300 individuals, the table for reporting these measurements may have categories such as 0, 0.1-4.9, 5.0-9.9 and 10.0+. Some times such categories are used for analysis and inference purposes also.

**Cause-effect relationship** — *Statistically significant* dependence of an outcome on an antecedent so that any change in antecedent makes corresponding changes in the outcome when all confounders are absent. The relationship should also meet other criteria such as temporality, consistency and biological plausibility. See *necessary cause* and *sufficient cause*.

**Cause-specific rate** — The rate obtained when numerator is restricted to a particular cause (e.g., of morbidity or of mortality). Cause-specific death rate is the number of deaths due to a cause per thousand population. Sum of cause-specific death rates for all causes is the same as *crude death rate*.

**Census** — *Survey* of the entire population.

**Centiles** — Same as *percentiles*. See also *deciles, quartiles, tertiles*.

**Central Limit Theorem** — The result stating that the chance of a summative measure such as sample mean following a *Gaussian distribution* rapidly increases in almost all practical situations as the number of individuals in a sample increases (i.e., sample size becomes large). As a result of this theorem, while dealing with sample mean, there is rarely any need to worry about the nonGaussianity of the underlying distribution when the sample size is large.

**Central tendency** — Among variations, there is still a tendency for a set of values to gather around a central value. Mean, median and mode are measures of central tendency.

**Chance** — 1. Colloquial name for factors that are unknown or too complex to comprehend such as collision occurring on road by chance.

2. Colloquial term for probability.

**Chance node** — A point in a *decision tree* where the outcome depends, to a certain degree, on chance such as result (positive or negative) of a test in a suspected case. Contrast it with *judgment node*.

**Chi-square test** — A versatile statistical procedure that is used to test different types of hypothesis on proportions, such as equality, trend and relationship.

**Citation** — Identification data of a document containing the authors' name, title, publication name, volume, publication date, page numbers, etc.

**Classification** — Placing a unit to one of the two or more known classes or categories. Units within each category share some similarity among them, and units in different classes are dissimilar.

**Clinical agreement** — See *agreement*.

**Clinical epidemiology** — Application of principles of epidemiology to individual subjects, particularly to the patients.

**Clinical equipoise** — Genuine uncertainty among the experts about the relative merits of the regimens under trial: thus the research subjects are not particularly disadvantaged.

**Clinical equivalence** — Two regimens giving same *efficacy* under identical clinical conditions.

**Clinical significance** — A result that is capable of modifying the management of a patient.

**Clinical thresholds (of normal range)** — A range of values of a quantitative medical measurement in healthy subjects that has least overlap with the values found in diseased subjects so that the chance of misclassification is minimum. But the chance of error remains. For example, for blood pressure this is 140/90 mmHg.

**Clinical trial** — A medical experiment on human subjects, particularly in a clinic setup, such as to find *efficacy* and safety of a new therapeutic or diagnostic regimen.

**Clinimetrics** — Assigning scores to clinical entities for diagnostic or rating purposes—thus qualities are converted to quantities.

**Close-ended question** — A question for which list of possible answers is already provided.

**Cluster** — A group of subjects with some commonality—generally available in close proximity of time and space.

**Cluster analysis** — The statistical procedure to classify units or individuals into groups such that the units are similar within each group but dissimilar across groups: generally used when the number and nature of the groups are not known.

**Cluster random sampling** — Dividing the target population into *clusters* of specified size and selecting a few clusters by random method.

**Cochrane Collaboration** — An international organisation of producers and consumers of medical research that helps to clarify the research achievements, particularly health care interventions such as drugs, diet alteration and behaviour change. The focus is mostly on systematic reviews or correct *meta-analysis* of the relevant studies.

**Cochrane Review** — A review of trials, mostly based on *meta-analysis*, following specific guidelines of *Cochrane Collaboration*.

**Coding** — Assigning a numeric to a qualitative characteristic, such as code 1 for hypertension, code 2 for diabetes, code 3 for cancer, etc. Codes are not quantities, and care should be exercised that they are not used as quantities at the time of analysis. These are used only for convenience in data entry.

**Coefficient of determination** — The percentage of variation in a variable explained by one or more of the others. In a *simple linear regression* setup, this is square of the correlation coefficient. In a *multiple linear regression* setup, this is square of the *multiple correlation coefficient*. In some other setups, this is explained sum of squares as proportion of total sum of squares.

**Coefficient of variation** — Standard deviation divided by mean. This unit-free measure is used to compare dispersion of one variable with the other, such as dispersion of cholesterol level with dispersion of body temperature in a group of cases.

**Cohen's kappa** — For qualitative data, a measure of agreement in excess of chance between two or more observers, methods, sites, etc.

**Cohort** — A clearly defined person or a group of persons with some common feature, who are followed for an outcome in a specific period beginning from a defined common baseline: not necessarily beginning at the same time. A cohort of women taking contraceptive pill may start from the day of first intake but it may include women starting in any chronological month of the year. See also *inception cohort*.

**Cohort study** — A prospective study of a *cohort* for a specified period, generally to observe the occurrence of an outcome of interest, and thereby determine the *incidence*.

**Collation of data** — The process of rearranging the data into intelligible form so that either the conclusions can be drawn or analysis can be done.

**Community trial** — Same as *field trial*.

**Compliance bias** — Either higher noncompliance by treatment group relative to the control because of discomfort or poor intake of the drug, or better compliance by them because they are improving.

**Concealment of allocation** — The process of allocation of subjects to the groups that is impervious to any influence of the person making the allocation. Among methods of such concealment are centralised randomisation without participation of the observer, coded and identically looking packing of the placebo and drug, and sequentially numbered opaque envelopes.

**Concentration curve** — Plot of quantitative response versus time such as of concentration of drug in the system at different points of time.

**Conceptual bias** — Errors arising from faulty logic or incorrect premises.

**Concomitant variable** — Same as *confounder*.

**Concordance** — Agreement or similarity between two individuals in a paired setup.

**Concurrent validity** — *Consistency* of response to two or more questions in the sense that they reflect the same pattern. For example, high calorie intake and low exercise together should correspond to greater obesity. Response to these three items should be consistent with one another.

**Conditional probability** — The probability of occurrence of an event such as disease when some a-priori information such as sign-symptoms are known: denoted by P(A/B) where after slash (/) sign is what is known a-priori.

**Confidence interval** — The interval within which results of other similar studies is expected to lie with a certain *confidence level*.

**Confidence level** — The degree of assurance that other studies of similar type will have the same results as obtained by the participants in the current study.

**Conflict of interest** — Personal, financial, or other interest of any investigator that could influence the finding or the interpretation.

**Confidence limits** — The upper and lower boundaries of a *confidence interval*.

**Confounder** — An extraneous factor that could be an explanation of the outcome of interest in addition to the factor under study so that its effect can not be differentiated from the other: such as dietary factors when examining relationship between smoking and cervical cancer. Presence of unaccounted confounders decreases the *validity* of a study.

**Confounder bias** — Bias due to presence of one or more unaccounted *confounders*.

**Consecutive sample** — The sample of subjects enrolled on the basis of the sequence of their arrival, and none is excluded if eligible till the desired number is reached.

**Consistency** — Same as *reliability*.

**CONSORT** – Acronym for CONsolidated Standards Of Reporting of Trials that provides guidelines on how a trial results should be reported.

**Construct validity** — The ability of a set of items or questions to assess a given theoretical concept. Suppose positive health is defined as the ability to withstand physical stress, and the intention is to measure it by excess in haemoglobin level, forced vital capacity, and pain bearing capacity. The ability to withstand stress is also separately measured by items such as lack of restriction in daily activity by injury or fever. The correspondence between these two assessments will indicate construct validity of the measurements. It is the agreement between the theoretical concept and the specific device used to measure that concept.

**Content validity** — Sufficiency or adequacy of the items or questions to measure a phenomenon. For example, spinal palpatory test may not have good content validity to identify spinal neuromusculoskeletal dysfunction. Content validity is judged by a panel of experts.

**Contingency table** — A table containing the number of subjects with different characteristics, which should be mutually exclusive and exhaustive, such as number of subjects with and without disease, and each with a positive or negative test. A contingency table is used to test if one characteristic is associated with the other.

**Continuous variable** — A variable that can theoretically have infinite number of possible values within a short range. Age is continuous since within 8 and 12, it can be 8.17, 10.874, 9.756 years, etc. Age can be measured in terms of days, hours and minutes, although practically there is no need to do this. Blood pressure is a continuous variable but measured in integers for convenience. Parity is not a continuous variable because there is no possibility of it being 2.75 or 1.6.

**Control (group)** — Used in two senses:

1. The group of subjects that do not receive the test regimen. They may receive placebo, or the existing standard regimen, or any other regimen that is appropriate for comparison.
2. The group without disease or without any other outcome of interest.

**Control (programme)** — A defined series of steps to reduce or eliminate a disease.

**Control (statistical)** — The statistical process of adjusting for any extraneous influence on the results.

**Control (subject)** — A person or unit of interest but not possessing the specified characteristic, such as a person not having the disease of interest, or a person being treated by regimen other than under test.

**Controlled trial** — A trial that compares intervention group to a control group: when not further qualified this generally indicates a trial with nonrandom or *quasi-random allocation* of subjects to the test and the control group.

**Convenience sample** — The group of subjects that are selected primarily because they were available at a convenient time or place. This is one of the several ways that a *purposive sample* can be drawn.

**Correction for continuity** — Mainly used for a *chi-square test* for comparing two proportions so that discrete data (number of subjects with and without characteristic) can approximate continuous data. This approximation is required since chi-square is a continuous variable whereas the data are for discrete variables.

**Correlates** — Factors that are related in someway to an outcome of interest. They may or may not be contributors to the outcome.

**Correlation** — The degree or strength of relationship between two quantitative variables. Loosely used for qualitative variables also. For *linear relationships*, it is measured by *Pearsonian correlation coefficient* that ranges from –1 to +1. A negative correlation means that increase in the value of one variable is accompanied by linear decrease in the other, and vice-versa. A positive correlation means that the two move together in the same direction. A correlation close to zero means that increase or decrease in one does not linearly affect the other. Correlation coefficient can be close to zero when a strong relationship is present but is nonlinear.

**Correlation coefficient** — See *Pearsonian correlation coefficient*. The other types of correlation are Spearman's, multiple, partial, etc.

**Cost-benefit analysis —** The assessment of benefit per unit of cost, when both are measured in monetary units.

**Cost-effectiveness analysis** — The assessment of effectiveness (life saved, disability restricted, year of life gained, etc.) per unit of cost.

**Covariance** — A measure of how the product of two quantitative variables behave—used in calculating *correlation coefficient*.

**Covariate** — Same as *concomitant variable*: generally used in a restrictive sense for quantitative variables only.

**Cox regression** — A type of *regression* that models logarithm of hazard ratios on the covariates affecting this ratio. Also see *proportional hazards model*.

**Criterion standard** — Same as *gold standard*, but a preferred term.

**Criterion validity** — Ability of a device to provide a measure that correlates or agrees with the criterion known to correctly measure the characteristic of interest.

**Critical value (in the hypothesis testing)** —The threshold value of the statistical test criterion such as $\chi^2$, *t* and *F*, beyond which it is considered statistically significant.

**Cronbach's alpha** — A measure of internal consistency of a tool such as a questionnaire or an index. This utilises the average of the *correlations* between quantitative responses to various components (items) of the tool. Cronbach's alpha ranges from 0 to 1. A higher value indicates a higher consistency. A difficulty with Cronbach's alpha is that it can attain a higher value if there are a large number of items in a test even when they are not consistent.

**Cross-over design** — A design that stipulates that same subjects will get the test and the control regimen after a washout period, but the sequence is randomised. Half the subjects get regimen A followed by B, and the other half B followed by A. Using same subjects for both the regimens reduces variability and thus also controls the level of uncertainty in the results.

**Cross-product ratio** — Same as *odds ratio*.

**Cross-sectional study** — An analytical study with a format that elicits information on the antecedents and the outcomes at the same time. Such a format is poor to investigate *cause-effect* type of relationship but is good to generate *hypothesis*.

**Cross-tabulation** — The process of obtaining the number of subjects of various types when divided by two characteristics simultaneously such as distribution of myocardial infarction cases by their lipoprotein(a) and homocysteine levels. In most cross-tabulations, these levels would be in categories such as 10.0-14.9, 15.0-19.9, etc.

**Crude death rate** — Total deaths in one year per 1000 population at mid-year.

**c-statistic** — A statistic that measures the area under an *ROC curve*.

**Current Contents**—The electronic database of table of contents and bibliographic citations from current issues of more than 7500 research journals in sciences, social sciences, arts, and humanities.

**Curve** — Opposed to straight line, the curve depicts relationship that varies at different values of two variables. For example, one variable may increase for lower values of the other and then decrease for the higher values of the other. Another example is the relationship between glomerular filtration rate (GFR) in chronic renal failure cases and plasma creatinine level. As creatinine level rises, GFR declines steeply in the beginning and then declines slowly.

**Curvilinear relationship** — A relationship that is depicted by a curve opposed to a line, but can be converted to a line through some mathematical transformation.

**DALYs** — Short for *disability-adjusted life years*.

**Data** — A set of observations, generally in numerical format but can be in text format also (plural of datum).

**Database** — A collection of items of data arranged in a predefined format, usually held on a computer.

**Data cleaning** — The process of correcting and deleting the incomplete or apparently wrong observations from the data set. Sometimes this may suggest a relook at the original source for finding the correct values.

**Data dredging** — Initially used for excessive analysis of data in search of new hypotheses, but now used for re-analysing data after deleting some inconvenient values so as to fit them into supporting a particular hypothesis—a very unfair practice that has now become so easy because of wide availability of computers.

**Data and Safety Monitoring Board** — An independent body of experts who are asked and agree to look after a research project for data integrity, safety of subjects, and adherence to protocol.

**Death spectrum** — Variety of causes of death: some people meet death slowly such as by cancer, and some sudden such as by myocardial infarction.

**Deciles**—Nine cut-points of values of a variable that divide the total number of subjects in ten equal parts: obtained after arranging the values in ascending order.

**Decision analysis** — The process of reaching to a decision after considering probabilities of various outcomes and value judgments regarding the *utility* of those outcomes. Decision analysis is considered a very effective method to take a valid decision under conditions of uncertainties.

**Decision node** — See *judgment node*.

**Decision tree** — A diagram that depicts the full path of decision process indicating various possibilities, including the varying probabilities at different *chance nodes* and the choices or decisions at different *judgment nodes*.

**Deduction method** — The method of reaching to a conclusion for a particular individual on the basis of a known generalised result—from general to particular.

**Degrees of freedom** — The number of observations in a dataset that can freely vary once the parameters have been estimated. This concept is used in a *chi-square*, *Student's t* and other statistical procedures since their distribution depends on degrees of freedom.

**Definition bias** — The bias due to

(i) wrong or substandard definition such as of impotence on the basis of erectile dysfunction alone; or

(ii) blurred definition that gives room to assessor to use subjective interpretation such as blood pressure ≥140/90 for hypertension without specifying what to do if systolic level is higher and diastolic is lower. Errors in diagnostic or screening criteria also come under this category.

**Delphi method** — A less scientific but a quick method to arrive at a consensus among experts. In this method, the responses from a panel of experts are iteratively obtained that

are progressively refined by reducing the options in successive rounds depending on what options are preferred in the previous rounds.

**Dependent variable** — A variable that is sought to be explained by one or more of the other variables. The dependent variable is generally the *outcome* of interest whereas independent variables are the *antecedents*.

**Descriptive study** — A study with the objective to delineate the distribution of disease or a health condition in a defined population, or to describe clinical features of specific kind of subjects. It also includes estimation of parameters such as mean and percentiles. This does not examine causality or aetiology. A descriptive study could be a *census*, sample *survey* or a *case series*.

**Design (of research)** — The format of collection, compilation and analysis of observations. See *descriptive study, analytical study.* Also see *sampling design*.

**Design bias** — A design of a study where the selection of subjects is not random, control group is not adequately matched, definitions of subjects of characteristics to be studied are loose, confounders are not properly accounted, etc.

**Design effect** — The effect on the *variance* of an estimate, and thus on efficiency of a study, due to the *design* of the study, such as of *cluster random sampling* relative to the *simple random sampling*.

**Determinant** — A factor that is responsible, fully or partially, for an outcome.

**Diagnostic score** — A numerical score used as an aid in establishing diagnosis, such as thyroid scoring system to distinguish hypo-, eu- and hyperthyroidism on the basis of clinical assessment.

**Diagnostic test** — A criterion used for confirming the presence of disease or a condition. This should have high positive predictivity. The criterion could be a laboratory test, a radiological test, or a clinical observation.

**Dichotomous variable** — Same as *binary variable*.

**Digit preference** — Preference for certain digits such as 0 and 5. Then they become predominant in reporting and recording. This can affect the conclusion.

**Direct standardisation** — A procedure of adjustment to bring the differential structure of two or more groups to a common (standard) base to increase their comparability: most commonly done for differential age structure that can easily affect mortality.

**Disability-adjusted life expectancy —** *Life expectancy* adjusted for equivalent life lost due to disability of varying degrees during the life time. The disability arises from sickness from time to time.

**Disability-adjusted life years lost** — The loss of years of life due to premature mortality compared to the most healthy population, plus loss of equivalent years of life due to various disabilities of different severity and duration from morbidities or otherwise. This measures the *burden of disease* in a population.

**Discrete variable** — A variable that can take only finite, practically small, number of possible values. Parity for a woman is a discrete variable. Deaths in an area with population 200,000 is theoretically discrete but can be considered continuous for statistical purposes because it can take a large number of possible values.

**Discriminant analysis** — The statistical procedure to find a *discriminant function* that classifies subjects into known groups with least misclassification.

**Discriminant function** — A combination of variables that can be used to classify a subject as belonging to one group or the other, such as diseased or nondiseased, with minimum error.

**Disease spectrum** — Among many types of disease spectra, the one of special epidemiological interest is the proportion of population that is susceptible, proportion that has infection (apparent + inapparent), proportion that has disease, the proportion that has serious form of disease, and the proportion that die. Note that the subsequent proportion

is nested in the preceding proportion. Disease spectrum helps to plan the research accordingly, and to take measures to control the disease.

**Disease thresholds (of normal range)** — The range of values of a quantitative measurement beyond which there is a considerable risk of presence of disease or occurrence of disease. The chance of *misclassification* remains in this threshold also as in any other such threshold.

**Dispersion** — The degree of variability or scatteredness of the values of a variable when measured for different subjects or at different times.

**Dissertation** — A detailed discourse or treatise on a particular topic providing it a new perspective: generally the written work submitted by a candidate for the award of a doctoral degree.

**Distal measures of health** — The background characteristics that indirectly affect the health condition under consideration. Socioeconomic factors such as education, income and occupation make an impact on diet, hygiene and exercise that in turn affect the pathophysiological parameters, on the basis of which a condition is assessed. Thus socioeconomic are distal and pathophysiological parameters are *proximal measures*.

**Distribution (statistical)** — The pattern of values when obtained for a large number of subjects. For incubation period of AIDS, the distribution would tell us how many or what percentage of cases have incubation period between 5 and 6 years, what percentage between 6 and 7, 7 and 8, etc. Thus, the distribution tells what has been the dispersion and where has been the concentration of values. See *binomial distribution, Gaussian distribution*, *skewed distribution*.

**Dose-response relationship** — Any kind of relationship between quantity of dose and the degree of response but generally indicating that higher the dose, higher the response, such as between smoking and lung cancer.

**Double-blind trial** — A *trial* in which neither the experimental subject nor the assessor knows that the subject has received the test regimen or the control regimen. This removes the possible bias in response and ascertainment.

**Dropout** — A subject who is initially enrolled for a study but whose subsequent measurements as required under the study protocol could not be obtained.

**DSMB** — Short for *Data and Safety Monitoring Board*.

**Dunnett's test** — A statistical test used to compare means in two or more test groups with mean in the control group.

**Effectiveness** — The extent to which a regimen is effective in meeting its objectives in actual field conditions or routine circumstances—generally measured in percentage. *Pragmatic trials, field trials* or *post-marketing surveillance* is used to evaluate effectiveness of drugs.

**Effect size** — Magnitude of effect of a factor on an outcome. This could be proportion, difference, odds ratio, regression coefficient, measure of association, etc. This could also be transformed effect such as standardised mean difference between the test and control group obtained as the mean difference divided by its *standard error*.

**Efficacy** — The extent to which a regimen is effective in meeting its objectives in ideal conditions—generally measured in percentage. Generally *RCTs* are used to evaluate efficacy of a new drug.

**Efficiency** — The frequency of desired outcome per unit of resource inputs such as time, money and manpower. For example, coronary angiography using 4 French catheters may be more efficient procedure than 6 French catheters because of early ambulation without sacrificing the quality of images. Case-control studies are considered more efficient, particularly for rare outcomes, than prospective studies because of lower cost. Campaign against smoking may be more efficient for avoiding deaths or for increasing life-years than treatment of lung cancer patients.

**Efficiency of an estimate** — Inverse of the *standard error* of the estimate or a derivative of this inverse. Relative efficacy is the ratio of efficiencies of two methods of estimation.

**Electronic resources** — Resources available in electronic format such as on compact disk (CD), diskette, and particularly the world wide web (www). For literature, these include books, journals, reports, other documents, citation databases (e.g., MedLine), etc.

**EMBase** — Europe-based Excerpta Medica electronic dataset of medical and health research literature covering 3500 journals from more than 100 countries.

**Empiricism** — The system based on observations and evidence, opposed to theories.

**Epidemiology** — The study of factors that affect distribution and determinants of disease or a health condition in a human population.

**Epidemiologic consistency** — The correspondence among the incidence, prevalence, duration of disease, mortality, etc., so that their underlying relationship is properly reflected.

**Epistemic uncertainties** — Uncertainties arising from limitation of knowledge and biases Contrast it with *aleatory uncertainties*.

**Equipoise (subject)** — Subjects are such that the outcome of test or control regimen is uncertain, i.e., the subjects are not chosen in a manner that they favour one regimen or the other—an ideal situation for conducting *RCTs* because either of the two regimens can be used without raising ethical issues, and the results would be more valid. Also see *clinical equipoise*.

**Equipoise (clinical)** — See *clinical equipoise*.

**Equivalence** — Difference between two or more groups or regimens by not more than a prespecified clinically irrelevant amount. Two regimens are considered therapeutically equivalent if their efficacy does not differ by more than a prespecified small amount. They are considered bioequivalent if the course of the disease or recovery in the two regimens is nearly the same. Therapeutic equivalence considers only the outcome whereas bioequivalence considers the entire course.

**Equivalence trial** – A trial with the objective to examine if a new regimen is different from another regimen by more than a prespecified medically unimportant margin.

**ERMed** — Short for Electronic Resource for Medicine; A database of more than 1200 medical journals promoted by National Medical Library in India.

**Error** — See *bias*, *false positive, false negative, random error*, *Type I* and *Type II error.*

**Error in research (honest, negligent and deliberate)** — Honest error is absolutely unintentional that arises from limitation of knowledge, such as not taking care of unforeseen bias. Negligent error is gearing up investigations or findings to support a particular view neglecting the other evidence. Deliberate error is misconduct of plagiarism, reporting inflated sample size, cooking up the results, etc.

**Ethics** — See *medical ethics*.

**Etiological factor** — See a*etiological factor*.

**Evidence-based medicine** — Using evidence as available in literature, in records, or in newly generated data, for managing patients after proper accounting of risks and benefits.

**Exclusion criteria** — The set of conditions presence of which will exclude an otherwise eligible subject from the study. Generally these conditions are indicative of severe form of disease or complications that render a subject unsuitable for that research. Exclusion criteria are part of the case definition that delineates the target population. The other part is *inclusion criteria*.

**Expectation of life** — Same as *life expectancy*.

**Experiment** — A study of effect of intentionally introduced intervention or alteration of a factor, under controlled conditions. See *trial*.

**Experimental design** — The structure of an experiment in terms of inclusion and exclusion criteria for subjects, method of allocation of subjects to intervention or other groups, and number of subjects in various groups. Sometimes this also includes details of assessments, reliability and validity of tools and of measurements, etc.

**Expert system (medical)** — An intelligent computer aid to diagnosis, treatment and prognosis. It is capable of storing all the information, and more importantly to selectively retrieve the relevant information as an aid to the physician, and is capable of suggesting a spectrum of diagnosis for the given set of complaints, examination results, laboratory findings, etc., and also suggests a possible line of treatment and the prognostic implications.

**Explanatory trial** — A trial done under near ideal conditions with practically no deviation. Contrast it with *pragmatic trial*.

**Explanatory variable** — Same as *independent variable*.

**Exploratory study** — A small-scale study of relatively short duration, which is carried out to gain baseline knowledge about a problem for which little is known.

**External validity** — The generalisability of a result to those subjects that are not included in the study, i.e., the applicability of results to a larger population and not just to the study participants.

**Face validity** — The property of a measurement or an instrument to apparently look right or reasonable.

**Factor** — A characteristic that can, or suspected to, cause alteration in an outcome.

**Factor analysis** — The statistical procedure to discover a construct out of data that can possibly explain the variation and relationship among different *variables*.

**Factorial design** — A design that includes all possible combinations of the *antecedent* factors under study. If there are three antecedent factors for visual acuity in adults—age in years as <40, 40-49 or 50+, gender as male or female, and haemoglobin level in g/dl as <10.0, 10.0-12.9, 13.0-14.9 or 15.0+ then a total of $3\times2\times4 = 24$ combinations are required to be studied in a factorial experiment, and each combination must be tried on at least some subjects by design.

**False negative** — A person with disease classified as without disease. In place of disease, false negativity can be for any other *attribute*.

**False positive** — A person without disease classified as with disease. In place of disease, false positivity can be for any other *attribute*.

**Field trial** — An *experiment* on human subjects in a community, such as vitamin A supplementation to children less than 3 years for examining improvement in their nutrition status.

**Fisher's exact test** — A statistical test used for $2\times2$ contingency table to find statistically significant difference when the number of subjects is small. When the number is large, this is approximated by *chi-square*.

**Fixed effects model** — A statistical model that stipulates that the levels of factors under study are the ones of interest. Contrast it with *random effects model*.

**Follow-up study** — Same as *prospective study*.

**Force of mortality** — *Hazard* of death at an instantaneous point of time. In an air accident, the force of mortality is exceedingly high. In home it is very low. The force of mortality in leukaemia is high relative to breast cancer.

**Four-fold table** — A *contingency table* with two rows and two columns so that the total number of cells is four (excluding the totals).

**Frequency** — Used in two senses:

1. Frequency of occurrence per unit of time (per month, per year, etc.).
2. Number of subjects with a particular characteristic or with values in a particular interval, such as how many have homocysteine level between 10 and 14 μmol/l.

**Frequency curve** — A graphical representation of the *frequency distribution* by a smooth curve.

**Frequency distribution** — A statistical distribution of subjects that displays the number of subjects with different levels of measurement, e.g., how many have diastolic blood pressure <70 mmHg, how many between 70-74, 75-79, etc.

**Frequency matching** — Same as *group matching*.

**Frequency polygon** — A graphical representation of the *frequency distribution* by a polygon shape.

**Friedman test** — A *nonparametric test* for comparing *central tendency* in two or more groups in a *repeated measures* design.

***F*-test** — A statistical procedure for *test of hypothesis* on equality of three or more means, and for other setups such as *regression*.

**Garbage-in, garbage-out syndrome** — The tendency of getting poor output or poor outcome when the inputs or efforts are poor.

**Gaussian distribution** — Distribution of values of a quantitative variable such that they are symmetric with respect to a middle value with same mean, median and mode, and the frequencies taper-off rapidly in a particular manner on both sides—a bell-shaped distribution.

**General linear model** — A model that describes a *dependent variable* as a linear combination of a set of *independent variables* with two important features –

(i) the variables could be qualitative or quantitative or mixture, and

(ii) the variables could be square, cube, logarithm, or any such function of the original variable.

*ANOVA*, *ANCOVA* and *regression* are special cases of general linear model.

***g*-index** – A measure of quality of research of a person: Computed as *g* if *g* is the number of top cited papers of the researcher received at least $g^2$ citations by others. See also *h-index*.

**Gnatt chart** — A chart that shows *time-line* of a project: the duration and dates of beginning and ending its various phases.

**Gold standard** — In medical assessment, a method, procedure or a measurement with almost 100% sensitivity/specificity or 100% predictivity: an extremely difficult proposition to achieve. As a compromise, the best available is considered “gold”, against which the newer tests are compared. Thus the better term is criterion standard. This need not be a single or simple procedure but could include follow-up of subjects.

**Goodness of fit** — How well the actual observations fit into a specified pattern. The goodness of fit is statistically tested mostly by *chi-square* method.

**Group matching** — See *matching*.

**Grouped data** — Quantitative values in categories such as age into 0-4, 5-9, 10-14, etc.

**Group sequential design** – A design that seeks to sequentially add a prespecified group of subjects in stages depending on the results obtained at interim analysis at previous stage.

**Half-life** — The duration at which time dependent outcome or response is 50%. This could be survival, availability of drug in the body after intake, or any other such outcome. For a drug, this is the time at which the availability of drug in the body is one-half of what was initially injested.

**Haphazard sampling** — A mixture of convenience, volunteer, snowball sampling, etc., that does not follow any specific procedure.

**Hawthorne effect** — The tendency of subjects changing their response when they know that they are being observed. This can happen with the test group as well as the control group.

**Hazard rate** — The force of occurrence of events at a (instantaneous) point of time, such as force of mortality or of morbidity when measured per unit of time. This could exceed one.

**Health expectancy** — Same as *healthy life expectancy*.

**Health-adjusted life expectancy** — Same as *disability-adjusted life expectancy*.

**Healthy life expectancy** — The remaining portion of life at any age that would be spent without any morbidity. Life expectancy at age 40 may be 36 years but healthy life expectancy would be only 30 years if 6 years on average are spent in nonhealthy states as per the current pattern of morbidities.

**Helsinki Declaration** — A set of ethical rules to plan, conduct and report a study on human subjects.

**Herd immunity** — When a large percentage of susceptibles such as more than 90% is immunized, the infection feels strangulated, as it is not able to find susceptibles in its proximity. Thus the infection fails to spread.

**Heterogeneity** — Variability or differentials among measurements, subjects, specimen, results, estimates, etc. Contrast it with *homogeneity*.

***h*-index** — A measure of quality of research of a person: Computed as h if each of at least *h* papers of that person have been cited at least *h* times by others. See also *g-index*.

**Histogram** — A diagram showing the number or percentage of subjects with values in different intervals by means of contiguous bars: used only for quantitative variables.

**Historical cohort** — A *cohort* whose common baseline is in the past, and mostly outcomes too have already occurred. But the format of investigation still is from antecedent to outcome. The past is reconstructed mostly on the basis of records or recall.

**Historical control** — A *control group* or a subject for which information was collected earlier than the group being currently studied. This information can be biased because of changes over time in risk pattern, techniques, concepts, etc.

**Homogeneity** — Similarity among measurements, subjects, specimen, results, estimates, etc. Contrast it with *heterogeneity*.

**Hypothesis** — A statement of belief that is made before the investigation regarding the status of parameters under study, including those that measure relationship. See *null hypothesis, alternative hypothesis*.

**ICD** — Short for *International Classification of Diseases*.

**Iceberg phenomenon** — When there are a large number of undetected cases for each detected case, such as in the case of HIV infection. Also when one clinical case means many infected but inapparent cases such as for AIDS.

**Impact factor (of a journal)** — The average number of times articles from a journal are cited by others in preceding two years—a service of the Institute for Scientific Information for journals covered by Citation Index.

**Imprecise probability** — Probability intervals based on personal belief rather than facts and figures such as a physician telling a seriously suffering patient that the chance of full recovery is between 30 and 40 percent. This may be fuzzy with an element of gamble.

**IMRAD format** — A format for writing research articles—Introduction, Material and Methods, Results, And Discussion.

**Inception cohort** — A cohort assembled at the beginning phase of disease or a condition that is followed up to examine its course.

**Incidence** — The number of cases newly occurring or arising over a defined period of time.

**Incidence density** — Same as *incidence* but now number of new spells are counted instead of persons newly affected. One person can have more than one spell of diseases such as diarrhoea and angina in a defined period.

**Incidence rate** — Incidence per unit of time and per unit of population. If incidence of benign prostatic hyperplasia in a population of 1,00,000 adults in six months is 16 cases, the incidence rate is 0.32 per 1000 adults per year. This can also be calculated per *person-year* or per 100 person-years, etc.

**Incident cases** — Same as *incidence*.

**Inclusion criteria** — The set of characteristics such as age, disease and severity, which are necessary in a subject to be considered eligible for inclusion in the study. Some of these

subjects may become ineligible when *exclusion criteria* are imposed. The inclusion and exclusion criteria together define the study subject and the *target population*.

**Independence** — If occurrence of one event does not affect the occurrence of the other, they are called independent. Body temperature is generally independent of blood pressure levels but not of heart rate. Occurrence of typhoid is independent of colour blindness—none affects the other. Although, in adults, weight does not affect height but they are not independent since height affects weight. Independence is both ways.

**Independent variable** — A variable that is used as an explanation of another variable. Independent variables are mostly the *antecedent factors* that affect or suspected to affect the (dependent) outcome.

**Index** — A composite of two or more indicators, such as body mass index (BMI) for obesity that combines height and weight, and Indrayan's *smoking index* for a combination of age at initiation, duration, quantity and type of smoking, and the time elapsed since quitting by exsmokers.

**Index case** — An affected person who might affect others also. In the case of infection, the infected person is an index case who can spread the disease. When a new person is infected, he can be an index case for other susceptibles.

**Indicator** — A single measurement that indicates the existence or magnitude of a condition. Signs-symptoms are indicators of a disease, smoking an indicator of lung cancer risk, and infant mortality rate (IMR) is an indicator of mortality. Contrast it with an *index*, which is obtained by combining two or more indicators.

**Indirectly standardised death rate** — For age, same age-specific death rates are used on the observed age-structure of two or more groups to recalculate the death rate. These (called, standard) age-specific death rates are chosen by the researcher. Age is the most frequent factor for standardisation but similar standardisation can be done for any factor. Indirect standardization is used when observed specific rates are unreliable due to small numbers, or are not available.

**Induction** — The method of reaching to a generalized conclusion after compiling the individual cases—from particular to general.

**Infectiousness** — The property of a disease to be able to infect *susceptibles* on exposure: but generally used to measure how many can be infected. Measles is a highly infectious disease since an exposed susceptible is very likely to catch the infection.

**Infectivity** — The proportion actually got infected out of those exposed.

**Information bias** — Suppression of some information by some subjects because of stigma or any other reason. Also loosely used for any type of bias in the data.

**Informed consent** — Agreement by a subject to participate in a research or some such endeavour after he is fully explained the favourable and adverse implications of participation.

**Instruction bias** — Use of varying individual discretion to resolve doubtful and unforeseen situations when the instructions are not complete, or not properly understood.

**Instrument bias** — A systematic error in an instrument to consistently give either lower or higher values than actual. Presence of air bubble in mercury column of a sphygmomanometer makes the instrument biased. Improper use of an instrument can also cause this bias.

**Intention-to-treat analysis** — Analysis that includes *dropouts* or other subjects with incomplete data or those who had to be shifted from one regimen to the other due to developing some medical complication to the group they originally belonged. Mostly worst but sometimes average outcome is assumed for them.

**Interaction** — Simultaneous presence of two or more antecedent factors affecting the outcome either negatively or positively so that the net effect is not the same as the sum of their individual effects. It is called *antagonism* when the interaction effect is negative and called *synergism* when this is positive. But there could be other interactions that are not classifiable into any of these two categories.

**Internal consistency** — Mostly, consistency within the set of actual observations. One observation should be consistent with the other, or one variable should be consistent with the other. If systolic blood pressure of a person is 110 mmHg and the diastolic blood pressure is 95 mmHg then they are inconsistent unless there are specific reasons for such disparate readings. At the group level, HIV prevalence should be higher among those practicing multipartner sex. If the prevalence in this group is lower than those with single partner then this is inconsistent, and raises doubts about internal validity of data. If a study shows higher morbidity after exposure but lower mortality then this also may be internally inconsistent.

**Internal validity (of a study)** — When the biases are sufficiently under control so that the difference in the outcomes among groups can be legitimately assigned to the hypothesized factor under investigation. Thus the results hold true at least for the subjects included in the study. An internally valid study may or may not be externally valid. There are situations when a sample provides excellent results for itself but fails when used on another sample from the same *target population*.

**International Classification of Diseases (ICD)** — A system of classification of diseases, injuries and causes of death into relevant groups, and assigning code to each condition, so as to promote uniformity and comparability across health care establishments in various countries. The ICD is revised every 10 years by the World Health Organisation to incorporate new diseases and new understandings.

**Interpretation bias** — Incorrect interpretation of results, either knowingly to support a particular hypothesis, or unknowingly.

**Interval estimation** — The process of assigning a range of values to a parameter within which it is expected to lie in repeated studies of that type.

**Intervention study** — A study of the impact of an intentionally introduced intervention on a predefined outcome. Experiments and trials are intervention studies.

**Interviewer bias** — Greater attention paid by interviewers to certain type of subjects or certain responses, relative to the others—thus introducing bias in the recorded responses.

**Inter-observer variability** — The variation between observers that occurs when the same measurement on the same subject is taken by different observers. A high inter-observer variability indicates poor *reliability* of the measurement.

**Intra-class correlation** — The *correlation* among same quantitative measurements within the same subjects or such other units at different times, by different observers, by different methods, etc.

**Intra-observer variability** — The variation that occurs when a measurement is taken repeatedly by the same observer. High variability indicates poor *reliability* of that observer.

**Inter-rater reliability** — The extent of agreement between the measurements obtained by different raters when they use the same measuring device on the same group of subjects.

**Inverse probability** — The conditional probability P(B/A) compared to P(A/B): often used to find predictivities P(D+/T+) and P(D−/T−) using sensitivity P(T+/D+) and specificity P(T−/D−) of a test, where D and T are for disease and test, respectively.

**Judgment node** — A point in decision tree where the physician has to choose one out of various possible options, such as treat or not to treat, or to advise CAT scan or not. This requires judgment. Also referred to as a decision node.

**Kaplan-Meir method** — The method of *survival analysis* that is used when the exact survival duration is assessed—thus these durations are not fixed time intervals.

**Kappa** — A measure of agreement in excess of chance in qualitative data; used for assessing inter-rater reliability and for other such agreements.

**Keywords** — The set of words that describes the essential features of a study. These words are used for indexing purposes so that the article is quickly retrieved for that category.

**Kuskal-Wallis test** — A *nonparametric test* for comparing *central tendency* in three or more groups.

**Lead-time bias** — Can occur when some subjects under study are enrolled in early phase of the disease and some in late phase of the disease. This may apparently show that early detected cases have higher duration of survival without any real prolongation of life.

**Left-skewed distribution** — See *skewed distribution*.

**Length bias** — The bias due to inclusion of disproportionately more cases with longer survival time in one group than the other: thus cases that show rapid progression of disease are not well represented.

**Level of significance** — The maximum tolerable probability of *Type I error* that is fixed in advance, such as 5%: denoted by $\alpha$. In statistical terms, the agreed threshold of probability of rejecting a true null hypothesis.

**Life expectancy** — The average number of years a person is expected to live in a given community on the basis of current pattern of mortality. It can be calculated 'at birth' or at any other age. A life expectancy of 36 years at age 40 means that the average life span after the age 40 years is 36 years in that community. In this population, life expectancy at birth could be only 71 years. Life expectancy is a mortality indicator measured in terms of survival duration.

**Life table** — A summary of the death and survival pattern of a group of people—generally for the entire population of an area, but can be used for patients of a particular disease also.

**Life table method** — The method of *survival analysis* that is used when the survival is assessed at fixed time intervals, such as weeks, months or years. The time intervals are fixed in advance.

**Likelihood ratio** — Relative odds of the result of interest (such as occurrence of a complication) in patients against the controls. Positive likelihood ratio measures the increase in odds of disease when the test result is positive, and negative likelihood ratio measures the decrease in odds of disease when the test result is negative.

**Limits of disagreement** — A procedure for measuring extent of disagreement between two quantitative measurements obtained by two methods, two laboratories two sites, etc., on the same subjects. These limits are obtained as (mean of differences) ± 2(SD of differences). If these limits are far too wide that can change clinical assessment, the disagreement is considered beyond clinical tolerance.

**Line diagram** — A diagram showing the trend by a line.

**Linear regression** — See *linear relationship*.

**Linear relationship** — A relationship that moves in a line with either positive or negative slope. The essential feature of a linear relationship is that one variable changes exactly by same amount when the other changes by one unit. If systolic blood pressure rises by ½ mmHg per year of age over the entire adult age from 20 to 59 years, the relationship is linear in this age-interval. Linearity can also include many variables. See *multiple linear regression*, also *curvilinear relationship*.

**Logistic regression** — The regression where the dependent variable is the probability of occurrence of an event. The independent variables may be qualitative or quantitative. This regression is based on a specific mathematical form, called logistic model.

**Longitudinal study** — A study where the same set of individuals is periodically assessed for one or more defined outcomes.

**Mann-Whitney test** — A *nonparametric test* for comparing *central tendency* in two groups: analogous to *t-test* for Gaussian data. Gives exactly same result as *Wilcoxon test*.

**MANOVA** – Acronym for multivariate *analysis of variance*. Used for simultaneous analysis of related quantitative outcomes when dependent on qualitative factors. This takes care of interrelationships and adjusts *P-value*s accordingly.

**Mantel-Haenszel procedure** — A statistical procedure for *stratified analysis* of qualitative data that combines evidence from two or more inter-related *contingency tables*.

**Masking** — Whereas the term blinding is used for the subjects and investigators, masking is for the regimen and the procedures. They are wrapped or administered in a manner that they look similar.

**Master chart** — An arrangement of data that prepares one record for each subject: thus data available in several pages of questionnaire/ schedule are converted to one row in, say, Excel software. This helps to get full view of the data in one shot.

**Matching** — Deliberate selection of control subjects that have the same characteristics as the cases except for the disease or the condition under study so as to increase the comparability. In practice only a few characteristics can be matched. Mostly it is one-to-one matching, which is called pair-matching, but sometimes can be group matching also. In the former, each control is matched with one case, and in the latter one group is matched with the other on average or for the pattern on the whole.

**McNemar test** — A chi-square test used for paired qualitative data, e.g., same subjects tested by histology and polymerase chain reaction (PCR) for extrapulmonary tuberculosis. McNemar test will reveal whether histology and PCR significantly disagree or not.

**Mean** — The average.

**Measurement bias** — Systematic error in measurement. This could be either due to faulty instrument, or due to carelessness of the observer.

**Measures of association** — The parameters that quantify the degree of association between two or more qualitative factors. Chi-square based measures are (i) phi coefficient, (ii) Cramer's V, and (iii) contingency coefficient. More useful measures are (i) proportional reduction in error, and (ii) relative risk or odds ratio. For ordinal data, these are Kendall's tau, Somer's d and Goodman-Kruskal gamma.

**Median** — The most middle value obtained after arranging values in increasing or decreasing order. Median seeks to divide the group in two equal halves, each with $n/2$ individuals. Sometimes in practice exactly equal halves are not possible, and they are divided into nearly equal halves.

**Medical decision process** — The process of taking decisions regarding diagnosis, treating or not treating a patient, what treatment to prescribe, when to stop, etc., after considering the chances of success of various alternatives and their respective utility in terms of likely outcome.

**Medical ethics** — The discipline that considers individual patient's welfare above every thing else—thus puts restrictions on how research involving human subjects should be done. Sometimes animal experimentation is also included in its domain. See *Helsinki Declaration*.

**Medically significant** — A result that is capable of modifying the management of any aspect of health or disease.

**Medical uncertainties** — Uncertainties in medical measurements and outcomes due to various sources of variation, and other factors such as lack of knowledge, poor compliance, incomplete information on the patient, etc. These can be diagnostic, treatment, prognostic predictive, or other types of uncertainties. See *aleatory uncertainties, epistemic uncertainties*.

**MedLine** — An electronic database of citations from more than 7500 medical journals published in different languages in different parts of the world. This is the most useful resource of medical research literature.

**Memory lapse** — See *recall bias*.

**MeSH** — Medical Subject Heading: an important resource to search articles in *MedLine*. This reduces problems arising from, e.g., British and American spellings, and has a tree structure that branches off into a series of progressively narrower terms.

**Meta-analysis** — A procedure of combining evidence in different reports on the same aspect. If different trials on the same regimen report varying efficacy, they can be combined to come to a unified conclusion, which may command substantially more confidence than result of any one of the individual trials.

**Metric scale** — Measurement in terms of numerics such as blood glucose and cholesterol level. Contrast it with measurement in terms of attributes such as gender and signs-symptoms. Metric scale gives rise to quantitative data.

**Mid-course bias** — Bias arising from exclusion of some patients who develop unrelated conditions during the course of the study such as injury. Some may have to be excluded because of related but serious condition requiring special care. In a field trial, this bias can occur when a new health facility or a new health problem starts in the study area that was not visualised earlier, and has potential to affect the results.

**Misclassification** — Classifying diseased as healthy (or nondiseased) or nondiseased as diseased. The first could be called *missed diagnosis* and the second as *misdiagnosis*. In place of healthy/diseased this could be any other categorisation.

**Misdiagnosis** — Diagnosing a person as suffering from a particular disease when he does not have that disease (the person can have any other disease).

**Missed diagnosis** — Not being able to detect a particular disease in a person when it is present.

**Mode** — The most commonly occurring value, i.e., a value seen in highest number of subjects.

**Model** — A simplified version of a complex process. A model could be mathematical, graphical, structural, etc.

**Multicentric study** — A study conducted at different locations with a common *protocol*.

**Multicolinearity** — Existence of high correlation between two or more *independent variables* in a *regression analysis* setup.

**Multifactorial aetiology** — Occurrence of disease depending on multiple factors: hypertension is a univariate disease because the diagnosis depends entirely on blood pressure level, but it has multifactorial aetiology since its occurrence depends on heredity factors, life stress, diet, obesity, etc. Malaria has one-factor aetiology.

**Multiple comparisons** — Several comparisons based on the same data. If each comparison is statistically done at 0.05 *level of significance*, the total probability of *Type I error* can be enormously large. To keep this within the specified level, procedures such as *Tukey*, *Bonferroni* and *Dunnett* are used for comparison of group means.

**Multiple controls** — More than one control subject for each case. In a case-control setup, sometimes it is easier to enroll controls than cases. The reliability of the results can be increased in this situation by enrolling 2 or 3 or even 4 controls per case.

**Multiple correlation coefficient** — The degree of *linear relationship* of one *quantitative variable* with two or more simultaneously considered quantitative variables.

**Multiple linear regression** — A *regression* in which a *dependent variable* is sought to be explained by linear combination of more than one *independent variables*: such as regression of systolic blood pressure on age, obesity, and socio-economic status. For one dependent and one independent variable, see *simple linear regression*.

**Multiple regression** — A *regression* that expresses the nature of relationship of one (dependent) variable on two or more of the other (independent) variables. This could be linear or nonlinear.

**Multiple responses** — More than one response to one question or one item, such as two or more complaints of a patient at the same time, or listing of two or more sources of infection when asked about HIV.

**Multiplication rule (of probability)** —The probability of joint occurrence of two or more *independent events* is the multiplication of their individual probabilities.

**Multistage random sampling** — The process of sampling where a subset is chosen at *random* from units at different stages. First stage units can be cities, second stage hospitals (within chosen cities), third stage wards (within chosen hospitals) and fourth stage patients (within chosen wards).

**Multivariate analysis** — A set of statistical procedures that considers several variables together for drawing a conclusion. If the variables are inter-related, as they would in most situations, the results of multivariate analysis could be very different from separate *univariate analyses*.

**Multivariate diagnosis** — The diagnosis that depends on a multitude of measurements. The diagnosis of liver cirrhosis depends on oedema, ascites, spider naevi on the chest, oesophageal varices, gastric ulcer, etc., whereas the diagnosis of diabetes mellitus depends only on blood glucose level. The former is a multivariate diagnosis and the latter is univariate.

**Multivariate setup** — A situation where several variables are considered simultaneously.

**Mutually exclusive events** — The set of events wherein only one can occur at a point of time. Blood group of a patient would either be O, or A, or B or AB. These are mutually exclusive. Signs-symptoms such as pain, diarrhoea and vomiting are not mutually exclusive—they can occur together in a patient.

**Necessary cause** — A cause that must be present to change the outcome. Exposure to an infection is necessary for it to produce that particular disease but it is not sufficient since in some cases infection can remain subclinical and may not produce the disease. Hypertension is not a necessary cause of stroke (neither it is sufficient). See *sufficient cause*.

**Negative association** — Presence of one factor associated with absence of the other, and vice-versa, in more subjects than expected by chance.

**Negative correlation** — Higher values of one variable generally accompanied by lower values of the other, and vice-versa, i.e., the two variables tend to move in quantitatively reverse direction.

**Negative predictivity** — Short for 'predictive value of a negative test'. This is the probability that a person with negative test really turns out to be free from the disease. Since a test is used only on suspected cases, the predictivity should be evaluated on the basis of suspected cases only.

**Negative trial** — A *trial* that reports that difference between the test and control regimens is not statistically significant, i.e., the test regimen is not found effective.

**Nested case-control study** — A *case-control study* where cases are identified through a *prospective study.* Controls may or may not be from the prospective study.

**N-of-1 trial** — A trial on one patient who undergoes repeated pairs of treatment periods such that he gets experimental treatment one period and the control therapy the other period. The sequence can be randomised. The patient and the physician can be blinded regarding the sequence. Treatment periods are replicated until a result one way or the other is obtained.

**Nominal scale** — Assessment of a characteristic in terms of names only. Blood group is on a nominal scale since O, A, B and AB are just names with no order or no grading among them. The other type of scale for qualitative variable is *ordinal* where grading is present such as hypotensive, normotensive, probably hypertensive and definitely hypertensive.

**Nomogram** — A collection of inter-related lines or curves such that the corresponding values can be read by using straight-edged ruler put across those lines or curves.

**Noninferiority trial** – A clinical trial with the objective of examining if a regimen is not worse than the other by more than a prespecified clinically unimportant margin.

**Nonlinear relationship** — A relationship that is characterised by a curve instead of a line. Oestrogen level in a woman has a cyclic variation over menstrual periods, and thus the relationship is nonlinear. Many organisms multiply exponentially as the days after exposure pass, and not linearly.

**Nonparametric test** — A statistical *test of hypothesis* that does not focus on a parameter such as mean. This does not require *Gaussian* or any other specific form of distribution of the variable. The usual tests such as *t* and *F* require Gaussianity but *chi-square* is

nonparametric. Other popular nonparametric tests are *Mann-Whitney* (or *Wilcoxon*), *Kruskal-Wallis*, and *Friedman*.

**Nonrandom sample** — Same as *purposive sample.* Can include volunteers, referred cases, case series, convenience sample, etc.

**Nonrandomised controlled trial** — A trial in which the subjects are assigned to the case group and the control group as per the convenience. Contrast it with *RCT*.

**Nonresponse** — Not being to able to collect full, or partial information on subjects once they are included in a study. This can happen due to unrelated death, injury, moving out of the area, left against medical advice, refusal to cooperate, etc.

**Nonsampling errors** — Opposed to *sampling errors*, these arise mostly due to lack of planning or due to lack of knowledge. The examples are presence of *confounders*, *nonresponse*, partial compliance, biased sample, inadequate measurement, etc.

**Nonsense correlation** — A correlation between two variables that incidentally occurs because each is related to a third irrelevant variable. Such correlation has no biological plausibility. For example, there might be a correlation between births in India and temperature in Boston, both of which rise in the months of August and September each year.

**Normal deviate** — The difference of a value from its mean when expressed in SD units, such as LDL cholesterol in a patient being more than 1.25SD away from the mean in healthy subjects. Normal deviate is 1.25 in this example. Compared to the absolute difference, this deviate gives a more realistic assessment of how far the value is from mean. Generally, same as *z-score*.

**Normal distribution** — Same as *Gaussian distribution*.

**Normalisation (of variables)** — Transformation of variable such that they lie between a standard range such as between 0 and 1, or between –1 and 1.

**Normal level** — A level generally seen in healthy individuals. This is not necessarily ideal or optimal. Normal level may be different for children than for adults, or different for males than for females, etc.

**Normal range** — The range of values (of a quantitative medical measurement), which is generally seen in healthy individuals in a population or its specified segment.

**Nuisance variable** — A variable not of interest but interfering and spoiling the picture.

**Null hypothesis** — A *hypothesis* that says that there is no difference, or that asserts the existing knowledge, and is tested for refutation by the study.

**Number needed to treat (NNT)** — The average number of subjects that must be treated to get one favourable outcome or to prevent one adverse outcome. If improved blood pressure control of 15 subjects is required for 10 years to prevent one death from myocardial infarction, then NNT for this outcome is 15 subjects for 10 years. Mathematically, this is reciprocal of *absolute risk reduction*.

**Observational study** — A study based on observation of the natural occurrences (no intervention). See *case-control study, cross-sectional study, prospective study*,

**Observer bias** — Observer being more careful or attentive to specific type of patients or particular responses.

**Odds** — The chance or frequency of occurrence or presence of a characteristic relative to its nonoccurrence or absence. If the chance of occurrence is 75%, the odds are 3:1. Generally calculated for presence of antecedent factors.

**Odds ratio** — The ratio of *odds* in one group to the other (generally the control group).

**One-sided alternative** — A directional alternative hypothesis in the sense of asserting that the value can be only either more or less than the null value. Contrast it with the *two-sided alternative*.

**One-tailed test** — While testing equality of two groups, it is sometimes not known before hand that which group could be better (or worse). For example, this happens when a test

regimen is being compared with the existing regimen. This requires a two-tailed test. However, while comparing a test regimen with placebo, if there is an assurance that test regimen can not be worse than placebo, one-tailed test is used.

**One-to-one matching** — See *matching*.

**One-way classification —** The division of subjects of interest by only one characteristic, such as dividing cases of bronchial asthma by their smoking status.

**One-way design** — A study that is planned to investigate the effect of levels of only one factor. The levels could be two such as presence and absence, or more than two such as none, mild, moderate, and serious.

**Open-ended question** — A question whose answer is allowed to be recorded in verbatim as given by the respondent. Contrast it with a *close-ended question* that provides a list of possible answers.

**Open trial** — Nonblind and/or nonrandomised trial.

**Ordinal association** — Association between two ordinal characteristics such as severity of disease and socioeconomic status. This is measured by Kendall's tau, Somer's d or Goodman-Kruskal gamma.

**Ordinal scale** — A scale that measures a *polytomous* characteristic in a defined order, such as severity of disease into mild, moderate, serious and critical. The 'distance' between mild and moderate is undefined. Division of blood group into O, A, B, and AB is polytomous but not ordinal since these blood groups do not have any order—none is better or worse than the other.

**Outcome** — A disease or a health condition of interest including any change in health status that may occur after exposure to antecedents or interventions. It may or may not be a result of the antecedents.

**Outlier** — A value that is far away from the other values. If duration of hospital stay after a cholecystectomy is 2, 3 or 4 days for most patients but happens to be 14 days for one patient because of complications, this value 14 days is an outlier.

**Outlier bias** — Differential management of different *outliers*: e.g., considering some extreme value as outlier and not others, or not ignoring outliers if they support a particular hypothesis.

**Pair-matching** — See *matching*.

**Paradigm** — A system, a pattern of thought, or a model regarding a phenomenon.

**Parameter** — A summary measure for any characteristic in the *target population*, such as percentage of cirrhosis patients with high aspartate aminotransferase, or rate of increase of systolic blood pressure in healthy subjects per year of age. The parameter pertains to the entire population of interest and not to the sample.

**Parsimonious model** — A *model* containing small number of explanatory factors yet providing adequate explanation.

**Partial correlation** — The *correlation* between two quantitative measurements when third or other measurements affecting them are considered fixed, and thus their effect is eliminated. Generally considered only for linear relations.

**Pathogenecity** — The ability to produce the clinical manifestation of disease in an infected person. Measles is not only infectious but also highly pathogenic—the disease appears in most susceptibles when infected. Generally measured by the percentage of infected who develop the disease. Tuberculosis is not a highly pathogenic disease.

**Pearsonian correlation coefficient** — A measure of the degree of *linear relationship* between two quantitative variables. This ranges from –1 to +1 with zero in between indicating no linear relationship. A negative correlation means that increase in one is accompanied by decrease in the other or vice-versa, whereas a positive correlation means that both increase or decrease together at least to some extent.

**Peer review** — A refereeing process of a research proposal, article, thesis, etc., by expert colleagues for technical merit.

**Percentiles** — Ninetynine cut-points of a variable that divide a group of subjects into one hundred segments after arranging in ascending order such that each segment has the same number (*n*/100) of subjects.

**Person-years** — The sum total of years observed for different individuals. If one person is followed-up for 3 years, second for 1½ years and third for 2 years, then the person-years of follow-up is 3 + 1½ + 2 = 6½ years. Similarly there could be person-weeks or person-months.

**Phases of a trial** — In phase I, the maximum tolerated dose and pharmacological properties including toxicity and safety are determined by a trial on a group of volunteers, usually without controls. The objectives of phase II are to investigate clinical efficacy, incidence of side-effects, identify a dose schedule, and to collect further pharmacological data. Most phase II trials have a control group. Phase III is an RCT that is done after achieving success in the first two phases to firmly establish efficacy and safety. *Post-marketing surveillance* is sometimes called phase IV.

**Pie diagram** — A diagram showing the proportions of subjects in different groups or with different *mutually exclusive* characteristics by means of segments of a circular pie.

**Pilot study** — A small-scale forerunner study to learn about the situation and the variables.

**Placebo** — An inert substance or a procedure that is neither harmful nor beneficial. The objective of placebo is that the subject gets the perception that he is receiving treatment— thus removing perception bias in trials.

**Placebo effect** — The psychological effect on a patient of the perception that he is receiving a treatment although the treatment is dummy. This is the main reason for conducting placebo-controlled trials. Also see *Hawthorne effect*.

**Plagiarism** – Copying somebody else's ideas, writings, data, etc. and projecting as your own.

**Point estimation** — The process of identifying a single value of a parameter as an estimate based on the study group.

**Polytomous variable** — A characteristics divided into three or more exclusive categories, such as severity of disease into mild, moderate, serious and critical, or liver disease as cirrhosis, hepatitis, and malignancy. A quantitative measurement such as cholesterol level can be made polytomous when divided into small number of categories such as –99, 100-179, 180-249 and 250+ mg/dl.

**Population** — The totality of individuals or units of interest. There could be a 'population' of blood samples collected in a year. If the interest is restricted to only suspected cases of liver diseases, the population comprises blood samples of such cases only. If the interest is further restricted to the cases attending OPD in a group of hospitals, the population is also accordingly restricted.

**Population attributable risk** — The risk in the total population minus the risk in unexposed subjects. In the population, some are exposed but generally most are unexposed. This measures the impact on the population of eliminating that exposure.

**Positive association** — Presence of one factor associated with presence of the other, and absence with absence, in more subjects than expected by chance.

**Positive correlation** — Higher values of one variable generally accompanied by higher values of the other and lower values with lower, i.e., they tend to move in the same direction.

**Positive predictivity** — Short for 'predictive value of a positive test'. This is the probability that a person with positive test really turns out to be suffering from the disease. Depends heavily on the prevalence of the disease. Since a test is used only on suspected cases, the predictivity should be evaluated on the basis of suspected cases only.

**Posterior probability** — See *prior probability*.

**Posthoc comparison** — The comparison of groups with regard to their initial equivalence after collection of data.

**Post-marketing surveillance** — Keeping a tab on outcome and side-effects of a formulation after it is introduced into the market: sometimes called phase IV of a *clinical trial*.

**Post-test probability** — The probability of occurrence or presence of an event such as disease after the test results are available. See *prior probability*.

**Power** — The probability that a study or a trial will be able to detect a specified difference. This is calculated as 1–Probability of *Type II error*—i.e., the probability of correctly concluding that a difference exists when it is indeed present. This measures the ability to demonstrate an association when one really exists, and depends primarily on the number of subjects in a study.

**PowerPoint** — A software of Microsoft Corporation that helps to make slides, which can be directly projected from the electronic format. This software has several features regarding designing the slide. The presentation of research to an audience can be very effective with the help of PowerPoint slides.

**Pragmatic trial** — A trial done under standard clinical practice so that accepted variation such as during drug intake are allowed. Contrast it with explanatory trial done under near-ideal conditions with practically no deviation.

**Precision** — Same as *reliability* but measured statistically as inverse of the *variance*.

**Predictive validity (of a test)** — The average of the *positive predictivity* and *negative predictivity* of a test. This can be used as a combined measure of the two types of predictivities when both are equally important for the outcome of interest.

**Predictivity (of a test)** — See *positive predictivity* and *negative predictivity*.

**Prevention trial** — A human experiment for a preventive strategy such as exercise and diet changes or a regimen involving vitamins, to prevent occurrence or recurrence of a disease, or any other adverse condition.

**Pretest probability** — The probability of occurrence or presence of an event such as disease before the test results are available: generally the same as prevalence rate in the specified group. Contrast it with *post-test probability*. See also *prior probability*.

**Pretesting** — Checking the workability, adequacy, reliability, etc., of a tool before using it for actual study. The tool could be an instrument, a laboratory procedure, a questionnaire, or any other.

**Prevalence** — The number of cases of interest present or existing at any specific time, usually at the time of the survey.

**Prevalence rate** — *Prevalence* per unit of population or per unit of susceptibles, such as percent, per thousand and per million. Note that prevalence rate is not a 'rate' as it does not signify frequency of occurrence—it is only a proportion. Conventionally, but wrongly, it is called a rate.

**Prevalence ratio** — Ratio of *prevalence rate* in one group to the other.

**Prevalent cases** — Same as *prevalence*.

**Primary data** — Data that are directly collected from the respondents. Contrast it with *secondary data* that already exist in databases, records, reports, articles, etc.

**Primordial factors** — Factors that work behind the scene, are precursors, or are those that give rise to risk factors. Life style is a primordial factor that can give rise to risk factors such as obesity and smoking.

**Prior probability** — The chance of occurrence or presence of an event such as disease or death in a patient when nothing is known about the condition of the patient. Once something such as signs-symptoms-measurements are known, the diagnosis becomes substantially more focused and the probability changes. The latter is called the *posterior probability*. When further information becomes available, this posterior becomes prior probability and the new probability based on the fresh information becomes posterior.

**PRISMA** – Acronym for Preferred Reporting Items for Systematic reviews and Meta Analyses: A statement of recommendations of how a systematic review and meta analysis should be reported.

**Probability** — A measure of belief in occurrence of an event or presence of a characteristic. This can be obtained either on the basis of theoretical considerations such as 1/6 for each of the 6 faces of a dice, on the basis of experience, or on the basis of frequency of occurrence when total occurrences are very large. Probability is the degree of certainty of occurrence of an event on a 0 to 1 scale. Probability of death is 1 for all individuals but the probability of death of a pancreatic cancer case within 5 years of detection could be 0.6. See *addition rule, conditional probability, Bayes' rule, multiplication rule.*

**Procite** — The software that manages bibliographic citations; can be used in conjunction with *MedLine*.

**Proforma** — A prototype or a sample of a format on which the observations are to be recorded.

**Prognostic factor** — A characteristic that can predict the eventual development of an outcome such as recovery, complication, and death. This does not necessarily imply a *cause-effect relationship*.

**Prognostic stratification** — A *stratification* done after examining the pattern of observations, such as categorising patients as mild, moderate, serious on the basis of new criteria developed after the patients are seen. In the usual stratification, the criteria are decided before seeing the patients.

**Prophylactic trial** — An experiment on a prophylactic measure such as amnioinfusion for meconium-stained amniotic fluid at the time of child-birth, or in the community such as iron supplementation to adolescent girls.

**Proportion** — The measure of how big is the part of a whole. The whole is considered as one. If 20% of a population have blood group A, the proportion is 0.20 or 1/5.

**Proportional hazards model** — A model that works when logarithm of ratio of hazards in test group to control group remains same all through the period of observation. This is an important prerequisite for the usual Cox model.

**Proportional reduction in error** — The reduction in error in prediction of the outcome when a particular antecedent is used for prediction relative to when it is not used.

**Prospective study** — A study that investigates outcomes for known antecedents. The follow-up of subjects is inherent in this kind of study since the occurrence of outcome can take time.

**Protocol** — A comprehensive statement regarding steps to be taken—the plan of a study. See *research protocol*.

**Proximal measures** — Measurements directly on *outcome*, contrasted with *distal measures*. Impact of vitamin A supplementation to children below three years can be measured proximally by rise in retinol level. In contrast, distally, the impact can be measured by growth pattern. Research results many times depend on appropriate choice of proximal measures for the outcome of interest.

**PubMed** — An extension of *MedLine* database of articles published in selected journals. Most comprehensive database of world medical literature freely available to the users on internet.

**Publication bias** — Publication of one type of views more often at the cost of other or opposite views, such as more frequent publication of positive results than negative results.

**Purposive sampling** — Nonrandom sampling to include subjects that serve the specific purpose, such as volunteers in phase I of *clinical trials*. See *convenience sampling, haphazard sampling, snowball sampling, volunteer studies.*

***P*-value** — The probability of *Type I error*, i.e., the chance that a difference or association is concluded when actually there is none: the chance that the result could have been produced by random sampling fluctuations rather than being actual. This is the probability that the observed data agree with the null hypothesis. Small *P*-value indicates that the chance of null being true is small.

**Qualitative data** — A set of observations on qualitative characteristics of individuals such as signs and symptoms. These can be *nominal* or *ordinal*. The only real summary measure

for qualitative data is proportion of subjects with a specified characteristic, although for some ordinal data, scores can be assigned that can be treated as numerics.

**Qualitative variable** — A characteristic that is assessed in terms of attributes such as gender and degree of severity of disease: a variable that yields *qualitative data*.

**Quantiles** — Cut-points of values of a variable that divide the total number of subjects into desired number of equal groups. Examples are *percentiles*, *deciles*, *quartiles* and *tertiles*.

**Quantitative data** — Collection of observations on characteristics that could be numerically expressed for an individual such as haemoglobin level, blood pressure, and blood glucose level. Most common summary measures for quantitative data are *mean* and *standard deviation (SD)*.

**Quantitative variable** — A characteristic that is measured in terms of numerics, such as creatinine level and parity of a woman: a variable that yields quantitative data. See also *continuous variable, discrete variable*.

**Quartiles** — Three cut-points of values of a variable that divide a group in four equal parts with regard to number of subjects, after the values are arranged in ascending order.

**Quasi-random allocation** — Allocation of subjects to test and control group by following apparently random method such as alternation and based on birth data that are not strictly random.

**Questionnaire** — A survey instrument that contains a predetermined series of questions that are supposed to be put in verbatim to the respondents or can be self-administered. It contains space for recording responses also.

**Quota sampling** — Purposive selection of pre-specified number (quota) of subjects from each segment of population without using random method.

**Random** — Unpredictable, like lottery—generally when each unit has same chance of being picked up, but the chance can be unequal also in some situations.

**Random allocation** — Same as *randomisation*.

**Random effects model** — A statistical model that stipulates that the levels of factors under study are random samples of the possible levels. Contrast it with *fixed effects model*.

**Random error** — An error that has no bias, and which is natural to occur in observations because of biological or other variation beyond control. These errors are small, and some are positive some negative so that the long-term average is close to zero.

**Random sampling** — Sampling in a manner that the selection can not be predicted. The chance of selection of various units can be equal or unequal. The popular methods of random sampling are *simple, systematic, stratified, cluster*, and *multistage*.

**Randomisation** — Allocation of subjects to different groups in a random manner with equal chance. The objective is that unaccounted factors are almost equally distributed among groups, and there is no bias on this count. Randomisation could be open so that the participants or the observers know which subject is in which group, or it could be concealed.

**Randomised clinical trial** — An experiment on human beings where the subjects are randomly allocated to various arms of a trial. These arms may be various dosage groups. When one arm is the control group, this becomes *randomised controlled trial*.

**Randomised controlled trial** — A *trial* where there is a control group (in addition to the test group) and the allocation of subjects to the control and test groups is by random method. This is considered to be the ideal methodology to evaluate efficacy of a new regimen (preventive, therapeutic or diagnostic) particularly when it is *double-blind*.

**Rate** — The frequency with which events occur, such as deaths per year or new cases per month. Time is a necessary ingredient of a rate. Generally measured per unit of population such as percent, per thousand and per million.

**Ratio** — Strength or magnitude or number of one quantity relative to the other, such as male-female ratio and albumin-globulin ratio.

**RCT** — Short for *randomised controlled trial*.

**Recall bias** — Not being able to recall events occurring far away in the past with the same frequency as those occurring recently. This introduces *bias* in favour of recent occurrences. Also occurs when diseased cases are able to recall because of their suffering but controls fail to recall as much. Also when serious episodes are easily recalled and mild episodes tend to be neglected.

**Receiving operating characteristic curve** — Same as *ROC curve*.

**Record linkage** — The process of linking different records of one person to make one comprehensive record.

**Reference population** — Same as *target population*.

**Reference values** — Same as *normal levels*.

**Referred sample** — A group of subjects that are referred for specialised handling. A study can be carried out on a referred sample although the results would be biased.

**Regression analysis** — The statistical procedure to find a *regression equation*.

**Regression equation** — The nature of relationship of one variable with one or more of others, generally expressed as a mathematical equation that best fits the data. Also called regression model.

**Regression coefficient** — The quantity that delineates the change in dependent variable for one unit change in independent variable. If regression coefficient of birthweight (in gm) on maternal haemoglobin level is +60, it means that birthweight increases on average by 60 gm for each 1g/dl increase in maternal haemoglobin level.

**Regression line** — Graphical presentation of *linear regression*.

**Regression model** — Same as *regression equation*.

**Regressor** — Same as *independent variable* in a *regression equation*.

**Relationship** — The property of change in one quantitative variable when the other changes. This change can be causal, incidental, or due to a third intervening variable.

**Relative risk** — *Risk* of occurrence of an outcome in the presence of one factor (exposure) relative to the risk in the presence of another (generally control) factor.

**Relative risk reduction** — Reduction in *absolute risk* after an intervention as percentage of the risk in exposed group before intervention.

**Reliability** — Ability to repeat the performance. The performance could be poor but same performance every time means good reliability. Statistically, this means smaller *variance* in repeated measurements.

**Repeatability** — Same as *reproducibility*.

**Repeated measures** — When the same subject is observed repeatedly after specified time gaps, such as monitoring blood pressure and heart rate at 1, 5, 10, 15 and 30 minutes after a administering anaesthesia.

**Replication** — Trying the same regimen on more than one equivalent group so as to get an idea of the repeatability. Replications, when yielding similar results, increase the reliability of the results.

**Reporting bias** — Highlighting findings in a report that support a particular view at the cost of the other.

**Reproducibility** — The ability to give similar result when conducted in identical conditions.

**Reproductive rate (of infection)** — The rate at which an *index case* infects others in the entire transmission phase. If the reproductive rate is one or more, the infection sustains itself and spreads in the population, like HIV is doing in some countries. If the reproductive rate is less than one, the infection will die down or will stabilize at a low level in course of time.

**Research** — Discovery of new facts, enunciation of new principles, or fresh interpretation of the known facts or principles.

**Research design** — Same as *design*.

**Research protocol** — Statement on planned steps of research, including background information and rationale, objectives and hypotheses, review of literature, methodology, ethics, statistical evaluation, and references.

**Response bias** — (i) Not giving proper history due to stigma such as in STDs or for any other reason, and (ii) selective nonresponse, i.e., persons who are not seriously ill do not fully cooperate.

**Retrospective cohort** — Same as *historical cohort*.

**Retrospective follow-up study** — A study based on a *historical cohort*.

**Retrospective study** — A study that investigates antecedents for known outcomes. The recruitment of cases can be prospective spanning a duration such as all cases reporting in one year period. But the logic is from effect to the cause.

**Right-skewed distribution** — See *skewed distribution*.

**Risk** — The chance of occurrence of an *outcome* of interest, generally calculated per year of exposure. It is not necessary that the outcome is adverse to health. See *attributable risk, population attributable risk,* and *relative risk*.

**Risk difference** — The difference in risk of occurrence of a condition in two setups, such as risk of glaucoma in vegetarians vs. risk in nonvegetarians. Same as *attributable risk*.

**Risk factor** — A characteristic that is suspected to affect the outcome, such as obesity for hypertension.

**Risk ratio** — Same as *relative risk*.

**Robust method** — A method that is not much affected by minor variation in its applicability conditions.

**ROC curve** — A curve that depicts the relationship between sensitivity of a test for different thresholds and the corresponding (1–specificity), such as for different $T_4$ values for diagnosis of hyperthyroidism. This curve helps to evaluate the applicability of a test and helps to compare performance of one test with the other, such as of $T_4$ with $T_3$, or $T_4$ with TSH.

**Sample** — A part of the *target population*, which is actually studied.

**Sample size** — The number of subjects or units in a sample.

**Sampling** — Choosing a part from the whole, such as choosing 300 child births out of 5000 in a hospital in one year for studying the intrauterine growth retardation. See *random sampling, purposive sampling*.

**Sampling bias** — The bias due to (i) *nonrandom sample* such as volunteers that do not represent the *target population*, and (ii) small sample that fails to represent the entire spectrum of subjects.

**Sampling design** — The method of selection of *sample* out of the *population* of subjects.

**Sampling error** — The tendency of one sample giving a different result than the other sample, and neither possibly giving exactly same result as in the population. This is not an error in conventional sense.

**Sampling fluctuation** — The tendency of different samples giving different results, even if drawn from the same *target population*. This happens because different samples contain individuals who are different from those in other samples.

**Sampling fraction** — The extent of sampling such as one out of eight, or one out of twenty.

**Sampling frame** — A list of units in the *target population* from which sample is drawn.

**Sampling method** — The procedure to choose or select a fraction out of the *target population*.

**Sampling unit** — The unit used for sampling. In a *multistage sampling*, there is a separate sampling unit for each stage.

**Sampling variation** — Same as *sampling fluctuation*.

**Scales of measurement** — The system of differentiating one type of observation from the other. When names are used (opposed to grades) for differentiation, the scale is called nominal. Signs and symptoms are generally measured on *nominal scale*. Textual grades such as mild, moderate, serious, are measurements on *ordinal scale*. Numeric measurements such as body mass index are on *metric scale*.

**Scatter diagram** — A diagram displaying values of one quantitative variable for different values of the other variable by plotting points. This is also called the (*x, y*) plot.

**Schedule** — A form that contains a set of items on which information is to be obtained. It contains space also to record the responses.

**SciSearch** — The software of the Institute for Scientific Information for searching citations in Science Citation Index and Current Contents.

**Score** — Quantification of a set of (mostly) qualitative measurements, such as APACHE score. Scores in medicine are used either to grade severity of a condition, as an aid to reach to a diagnosis, or to assess prognosis.

**Screening test** — A criterion used to locate possible positives that can be later confirmed by stricter criterion. A screening criterion should have high negative predictivity. The criterion could be a laboratory test, a radiological test, or a clinical observation.

**SD** — Short for *standard deviation*.

**Secondary attack rate** — The number of people who get sick within the incubation period after exposure to an infective person. This infective person is called the *index case*.

**Secondary data** — The data that are already lying somewhere as in records or literature. Contrast it with *primary data*.

**Selection bias** — The bias occurring due to selection of nonrepresentative group of subjects—thus affecting the generalisability of the findings. This can occur either because the selection is done with a purpose, such as of volunteers, or unwittingly due to selection of surviving subjects (if the disease under study is rapidly fatal), due to selection of younger subjects who have survived (many of the older ones may have died), etc. Lack of matching in case and control groups can also be called selection bias. All these can affect the *validity* of results.

**Sensitivity** — Ability to identify known positives as positives. If CPK ≥120 IU is present in 60% known cases of myocardial infarction, then sensitivity of this cut-off is 60%. Sensitivity is not a *valid* indicator of diagnostic value of the test. It only measures the intrinsic ability of a test to detect a disease when it is already known to be present.

**Sensitivity analysis** — The analysis to assess the impact of changes in assumptions on the outcome. Because of limitation of knowledge, various assumptions are made in a study. For example, it is generally assumed that cure rate depends linearly on treatment regimen, host characteristics and compliance. Two kinds of *epistemic uncertainties* arise—one what happens if the dependence is not linear but is quadratic or any other type, and second, what happens if this is assumed to depend also on physical strength and mental toughness of the patients. Answers to such questions are obtained by sensitivity analysis.

**Sequential sampling** — Serial sampling of subjects, one by one, to be stopped when scientifically acceptable result either way is available.

**Severity score** — A *score* used to grade severity of a condition such as *Apgar score* and *APACHE score*.

**Sign test** — A *nonparametric test* based on sign (negative or positive) of differences between two sets of observations. This test is used to compare two groups for their *central tendency*.

**Significance** — See *statistical significance* and *medical significance*.

**Significance level** — Same as *level of significance*.

**Simple linear regression** — A *regression* in which a dependent variable is sought to be explained linearly by only one independent variable, such as regression of systolic blood pressure on age. Other correlates are ignored in this setup, and the relationship format is limited to linear.

**Simple random sampling** — Sampling in a manner that all individuals have same chance of being selected.

**Simple regression** — Relationship between one independent and one dependent variable. It could be *linear* or *nonlinear*.

**Simpson's paradox** — In some cases the tendency of aggregated data showing a result very different from the results from disaggregated (stratified) data. This can happen due to *interaction* effect of the stratifying variable.

**Single blind** — No subject is informed that he is receiving test regimen or the control regimen in a *trial*. But the investigator is aware of the allocation.

**Skewed distribution** — Opposed to symmetric such as *Gaussian*, generally a distribution is considered skewed when values on one side of *mode* vary much more than on the other side. It is right-skewed when values more than mode have more variation, and left-skewed when values less than mode have more variation.

**Smoking index** — A measure of life-long burden of smoking with weightage for age at initiation, quantity of smoking, duration of smoking, type of smoking (filter/ nonfilter/ bidi/ cigar), and duration elapsed since quitting by exsmokers.

**Snowball sampling** — One eligible person, such as a client of sex-worker, is asked to list others known to him and eligible that can be included in the sample, and those in turn are asked to identify others, so on.

**Spearman's correlation** — A measure of strength of relationship between two quantitative variables when they are converted to ranks.

**Specificity** — Ability to identify known negatives as negatives. If AFB is negative in 95% of the healthy subjects, its specificity is 95%. In practice, a test is rarely used on healthy subjects. It is used on suspected cases. In them, the specificity may be very different.

**Spectrum of disease**—The distribution of subjects by affected and not affected, and among affected by severity of affliction.

**Spurious association** — A false *association* that could arise due to presence of *confounders*, *bias,* or merely a *chance*.

**Spurious correlation** — A false *correlation* that could arise due to presence of *confounders*, *bias,* or merely a *chance*.

**Standard deviation (SD)** — Most common and generally most appropriate measure of dispersion obtained as positive square root of *variance*.

**Standard error (SE)** — The measure for sample-to-sample variability in a summary measure such as *mean* and *median*. Just as the measurements such as haemoglobin level differ from person to person, so do the sample summaries from sample to sample. Mean haemoglobin level in one sample would be different from another sample even when both are drawn from the same *target population*. The extent of variability in such summaries is measured by their respective standard errors. Actually, this is the *standard deviation* of mean or median or any other summary measure as the case may be.

**Standardisation (of groups)** — The procedure that makes two different groups comparable by bringing them on to a common base or a common standard, such as *age-standardisation*.

**Standardisation (of variables)** — Subtracting mean from the variable and dividing by the standard deviation (SD): thus standardisation makes mean = 0 and SD = 1.

**Standardised death rate (directly standardised)** — Recalculated death rate when one or more factors affecting deaths (such as age structure) are brought at par with the comparison group, or both brought to a common base. This substantially increases

comparability that crude death rate lacks when one group is young and the other is old. Common base structure (called, standard) of population is chosen by the researcher.

**Standardised mortality ratio** — Ratio of actually observed deaths to expected deaths based on standard death rates. If this ratio is more than one, the *force of mortality* is higher in the study group compared to the standard group.

**Statistic** — A summary measure for any characteristic in the sample or the group actually studied, such as *mean*, *median* or *standard deviation* of a sample, or proportion of subjects found affected in a sample.

**Statistical analysis** — Subjecting data to the rigours of statistical methods so that the uncertainty levels are either quantified or minimized, or both.

**Statistical fallacies** — Many fallacies can occur in the results because of improper use of statistical methods. Some of these are due to (i) use of improper denominator for computing rate or percentage, (ii) not accounting for variable periods of exposure that could affect the rate of outcome, (iii) considering mixture of two groups as one, (iv) misuse of percentages such as 2 out 4 being stated as 50%, (v) using means for emphasising a point without considering the standard deviation, (vi) inappropriate scales in the graphs, (vii) looking at linearity when in fact the relationship is *nonlinear*, (viii) ignoring important prerequisites such as randomness, independence, equality of variances and *Gaussian* form of distribution, (ix) using means where proportions are adequate, or vice-versa, (x) ignoring baseline values, (xi) using too many *statistical tests* on the same set of data without adjusting *P-values* (xii) quantitative analysis of codes, (xiii) jumping to *cause-effect relationship* without sufficient examination of data, and (xiv) multivariate conclusions on the basis of several *univariate analyses*.

**Statistical power** — Same as *power*.

**Statistical significance** — A result is statistically significant if the chance of wrongly rejecting *null hypothesis* is less than the prefixed level such as 5%. The implication is that chance differences in samples would produce that kind of result in less than 5 times out of 100— thus chance is not an explanation for that result, and it is most likely real.

**Statistical test** — A procedure to find *P-value* corresponding to a *null hypothesis* on the basis of the given data. Depending upon the type of data and the type of hypothesis, a large number of statistical tests are available. Most popular of these are *Student's t-test* and *chi-square test*.

**Statistical threshold (of normal range)** — Mean $\pm$ 2SD where mean and standard deviation (SD) are obtained from measurements in a large number of healthy individuals. There is a chance of classifying nearly 5% healthy individuals with extreme values as sick when this range is used as normal.

**Statistics** — A science that helps to manage uncertainties in the data. Also as plural of *statistic*.

**Strata** — The divisions obtained after *stratification* (the singular is stratum).

**Stratification** — Division of subjects into relevant groups.

**Stratified analysis** — A statistical procedure to adjust for the effect of *confounders* or other correlates without estimating their individual effects.

**Stratified random sampling** — Separate *simple random sampling* from each group after *stratification*.

**Strength of relationship** — The consistency of change in one factor when the other changes. If one unit change in one factor results in nearly the same change in the other in all the subjects, the strength is high. Note that the strength is not related to the magnitude of change. Small change, if consistent, would mean high degree of relationship.

**STROBE statement** – Acronym for STrenghthening the Reporting of OBservational studies in Epidemiology. A check list is available at http://www.strobe-statement.org that guides on how to report results of an observational study.

**Structured questionnaire (or schedule)** — A form that provides a list of possible responses and the respondent or the investigator is required only to mark his response among the choices given. One of the choices could be "Any other (specify)" to make it exhaustive.

**Student's *t*-test** — Same as *t-test*.

**Study design** — Same as *design*.

**Study setting** — The environment in which the study is conducted. The setting could be a general hospital, a referral centre, private practice, ambulatory care, community, etc.

**Sufficient cause** — A cause that by itself is sufficient to change the outcome. Measles virus exposure to a susceptible is sufficient to produce the disease. In fact this cause is both necessary and sufficient. Hanging by rope is sufficient to cause death but is not necessary as a cause for death.

**Surrogate outcome** — Outcomes that are not of direct interest but reflect the outcome of direct interest, such as pallor for nutritional level. Surrogates are generally easy to assess and do not require long follow-up.

**Surrogate variable** — A makeshift substitute variable that is used when the actually required variable can not be measured. If the respondents are shy of revealing income, surrogates such as size of own house, car, telephone and television can be used to assess the level of income. Since they are surrogate, they may or may not reveal the true status.

**Survey** — A *descriptive study* done generally on scientifically selected subjects. Another descriptive study methodology is *case series*.

**Survival analysis** — Analysis of survival durations. Survival duration is generic—it can be duration between any specified events such as between end of operation and beginning of consciousness, between beginning of treatment and time at complete recovery, etc. Two popular methods of survival analysis are *life table method* and *Kaplan-Meir method*.

**Survival curve** — A graph that depicts survival pattern of the subjects, over the observed period of time. It begins at 100% (all alive) and shows gradual or rapid decline as the time passes depending on the disease under study.

**Survival function** — A mathematical expression for *survival curve*.

**Susceptibility** — Proneness to catch an infection or a disease. A person who is effectively immunised against tetanus is not susceptible to tetanus, and a child is not susceptible to the usual sexually transmitted diseases.

**Synergism** — A situation where two factors when present together accentuate the outcome more than their individual capacities. Iron and folic acid are synergistic for increasing haemoglobin level.

**Synthesis (of research)** — The process of combining diverse evidence from different researches to come to a holistic conclusion.

**Systematic error** — Same as *bias*.

**Systematic random sampling** — Selection of first individual at random and others automatically at regular spacing depending on *sampling fraction*.

**Target population** — Same as *population*. Also called reference population.

**Tertiles** — Two cut-points of values of a variable that divide a group into three equal parts (each part with same number of subjects) after arranging in ascending order.

**Test group** — The group of subjects that is receiving or has received the regimen under test.

**Test of hypothesis** — The procedure used to test whether or not sufficient evidence exists against a *null hypothesis*: mostly a statistical procedure.

**Test of significance** — A statistical procedure to test whether or not the observations fall into a specified pattern, such as equal means of two or more groups, or following a linear trend. If they do not, the result is called statistically significant. This requires prior fixing of the *level of significance* that specifies the maximum tolerable probability of *Type I error*.

**Test-retest reliability** — Agreement between responses when the same instrument such as a questionnaire is administered to the same set of people again; stability of the responses in repeated use of the instrument.

**Therapeutic equivalence** — Comparable safety and efficacy of two or more treatment modalities when administered under the conditions specified for each modality. These conditions could be different for different modalities. Also see *equivalence*.

**Therapeutic trial** — A *clinical trial* on a therapeutic agent or regimen to evaluate its *efficacy* and safety.

**Thesis** — A proposal or hypothesis forwarded after a careful investigation accompanied by full details: generally the written work submitted by a candidate in fulfillment of partial requirement for the award of Master's degree.

**Time-line** — A chart or a table that states the durations and dates of beginning and ending various phases of a project, some of which can overlap.

**Training sample** – A sample of subjects that is utilized to build a model. This model is subsequently tested on *validation sample*.

**Translational research** — A research that seeks to bridge the gap between research findings and their practical utilization for the benefit of mankind.

**Transmissibility** — Used in two senses:

1. The ability to transmit (infection or disease) from one to the other. HIV is transmissible through blood transfusion. Injury is not transmissible.
2. The force of transmission: when exposed what percentage of susceptibles get the infection or disease. HIV is more transmissible from male to female than from female to male.

**Trial** — An experiment on human subjects. See c*linical trial, field trial, prophylactic trial, therapeutic trial* and *vaccine trial*.

***t*-test** — A statistical procedure to test hypothesis of equality of two *means*, and for certain other hypotheses relating to parameters such as *regression coefficient*.

**Tukey's test** — A statistical test for pairwise comparisons of means when there are three or more groups. This test keeps the probability of the total *Type I error* within the specified *level of significance*, but this level could actually be lower than specified.

**Two-sided alternative** — An *alternative hypothesis* that stipulates that the *parameter* value can be higher or lower than the null value. Both directions are admissible. This type of alternative is setup when it is not known that the value is going to be higher or lower than the specified value.

**Two-tailed test** — See *One-tailed test*.

**Two-way classification —** Division of subjects by two characteristics simultaneously, such as dividing cases of diabetes mellitus by their obesity category and smoking category.

**Two-way design** — A study that is planned to investigate the effect of levels of two antecedent factors, such as effect of obesity (BMI<20.0, 20.0-24.9, 25.0-29.9 or 30.0+) and smoking (none, mild, moderate or severe) on blood glucose level. Their *interaction* can also be investigated in this kind of *design*.

**Type I error** — The error of rejecting a true *null hypothesis*, i.e., concluding that there is a difference when actually there is none. The sample or data might be such that this lead to such a wrong conclusion. This error leads to false positive result.

**Type II error** — The error of wrongly concluding that there is no difference when actually some difference is present. This error leads to false negative result.

**Uncertainty analysis** — The analysis that delineates the effect of change in the value of the *parameter* under assessment on the conclusion. One component of this is the *sampling fluctuation* in the estimate of the parameters, and the other is the plausible change in the values themselves that can affect the outcome.

**Uncertainty principle (in a clinical trial)** — The principle that says that the outcome of various arms of trial should be a-priori uncertain. One specific uncertainty situation is that

the a-priori chances of each regimen being successful are nearly equal—called *clinical equipoise*.

**Unit of study —** The unit (individual, patient, blood sample, biopsy, etc.) that is used to obtain the required information.

**Univariate analysis** — The analysis regarding one variable at a time. The combined result of several univariate analyses may be very different from the result obtained by simultaneous consideration of these variables. See also *multivariate analysis*.

**Universe** — The broad group of subjects for which the findings could be generalized or implicated. Statistically this is broader than the *target population* and can include future subjects.

**Utility (of an outcome)** — The value judgment assigned to relative worth of an outcome: generally on 0 to 1 scale. It is easy to say that death as an outcome has a zero utility and full recovery has one utility. But it is difficult to assign the utility to adding 3 years of life, particularly if it is to be accompanied by restriction in movements that affects quality of life.

**Vaccine trial** — An experiment on human population to evaluate potency, efficacy and safety of a vaccine.

**Validation sample** — A sample used to validate the results of the study. It could be a subsample of the original sample that is kept aside for this purpose, or a new sample of subjects.

**Validity (of a measurement)** — Ability to hit the target (or around it): ability to assess what is really intended to be assessed. Weight by itself is not a valid indicator of obesity in adults but body mass index that uses height also is. For a medical test, validity is measured by *sensitivity, specificity* and *predictivities* (positive and negative).

**Validity (of a study)** — The ability of a study to provide correct conclusion, considering the representativeness and size of sample, *validity of measurements*, and the soundness of methods. See also *internal validity* and *external validity*.

**Validity (of a survey instrument or a test)** — Ability of an instrument (such as a questionnaire or a schedule) or of a test to provide the information that matches with the objectives of the study. For its varieties, see *concurrent validity*, *content validity*, *construct validity*, *criterion validity*, *face validity*, and *predictive validity*.

**Vancouver style (or format)** — A style of writing research papers agreed by editors of more than 500 medical journals published around the world. It includes instructions on who could and should be authors, what help must be acknowledged, how to structure the text, and how to cite references. The system of citing references matches closely with the system followed by *PubMed*.

**Variable** — A characteristic that varies from person to person, or from situation to situation. Platelet count in different persons is variable but number of eyes or number of fingers is not a variable. See *quantitative variable, qualitative variable, discrete variable, continuous variable, dependent variable,* and *independent variable.*

**Variance** — A measure of dispersion or scatteredness of *quantitative data* obtained as average of the squared deviations from mean.

**Verbal autopsy** — The process of finding a cause (generally of death) through enquiry.

**Virulence** — The ability to produce severe form of disease that can threaten life. Rabies is a highly virulent disease and cholera is not—measured as percentage of cases who go into severe form. See *infectivity, pathogenecity*.

**Volunteer studies** — A study done on volunteers, opposed to randomly selected subjects. Results from such studies can be used to identify some side-effects of a test regimen but can not be used to estimate efficacy.

**Waist-hip ratio** — The ratio of waist to hip measurement: sometimes considered a more valid measure of obesity than body mass index. Waist-hip ratio measures only central obesity, whereas body mass index is for overall obesity including central obesity.

**Washout period** — In a *cross-over* trial, the period elapsed between withdrawal of first treatment, and start of the second treatment. This period allows time for any effect of the first treatment to vanish before the second is started.

**Wilcoxon test** — Another method to do *Mann-Whitney test* for comparing central tendency of two groups. Both give exactly same result, and they are algebraically equivalent.

***z*-score** — Same as *normal deviate* but some times measured from median instead of mean, particularly in assessing growth: difference of a value from group mean in terms of how many times of SD.

***z*-test** — A statistical *test of hypothesis* based on Gaussian distribution, generally used to compare two means or two proportions.