**Review Article**

# Overview of methods in biostatistics useful for clinical studies

Sarmukaddam Sanjeev[1]

[1]Biostatistics Consultant, Maharashtra Institute of Mental Health, B. J. Medical College and Sassoon Hospital Campus, Pune – 411 001
sanjeev.sarmukaddam@gmail.com

**Abstract:**

"Looking for clinical significance even when the results are statistically significant is very important" as there are situations where a result could be clinically important but is not statistically significant or vice-versa. In this article, these two concepts are contrasted, logic of statistical testing is explained, interpretation of confidence interval is discussed with examples, and concept of tolerance interval is introduced. To understand the difference between the standard deviation and standard error of the mean and why one ought to summarize data using the standard deviation, few appropriate examples are discussed at length. Important terms like design effect, bias, allocation concealment, etc. are discussed.

**Keywords:** Clinical significance, Statistical significance, Sampling distribution, Null hypothesis, P-values, Confidence intervals, Bias.

## Introduction:

In this assay it is tried to cover few essential concepts related with role of "Biostatistics" in "Research in Medical Sciences", however one can choose to ignore/omit few or do the 'rapid reading' as most of you are likely to be familiar with these very basic things. Though of primary level, these basic things are discussed because it is important to be very clear about them. It is true and often said that "clinicians are more sophisticated consumers of research information and moreover they have a far better understanding of how to find the best information and to judge its validity and generalizability for themselves". Fruitful use of the rapid, almost daily, advancement in medical sciences can be made up-to-date only by enabling one to critically read and interpret the findings reported in medical literature such as journals & the electronic media. For this, at least some knowledge of biostatistics is vitally necessary as it helps develop such ability so that one can apply research findings judiciously in practice. Many clinicians are not actually producers of studies involving statistics/biostatistics, but almost all will be consumers. They must read the medical literature to keep up-to-date and they generally find many papers contain a good deal of statistical analysis. A good analysis must be preceded by a good design. It is well recognised that design deserves as much consideration as analysis. A well designed study poorly analysed can be rescued by a reanalysis but a poorly designed study is beyond the redemption by even sophisticated statistics. Many experimenters consult the biostatistician only at the end of the study when the data have been collected. They believe that the job of biostatistician is simply to analyse the data and produce the final 'P' value. However, analysis is only part of a biostatistician's job. A far more important task is to ensure that the results are valid and generalizable, and so the study design and execution are appropriate.

Before we overview few important methods in biostatistics for utility in clinical studies, it may be pertinent to throw light on one very basic but vital fact that "looking for clinical significance even when the results are statistically significant is very important". There are situations where a result could be clinically important but is not statistically significant. Consideration of these two possibilities leads to two very useful yardsticks for interpreting an article(s) on clinical studies/trials. These yardsticks are - (i) If the difference is statistically significant, is it clinically significant as well? and (ii) If the difference is not statistically significant, was the trial big enough to show a clinically important difference if it had occurred? (1). It is possible to determine ahead of time, how big the study should be. But most trials that reach negative conclusions either could not or would not put enough patients in their trials to detect clinically significant differences. That is, the errors (i.e. type II errors) of such trials are very large and their power (= (1- ) 100%) or sensitivity is very low. In one review with a long list of trials that had reached "negative" conclusions, it is found that most of them had too few patients to show risk reductions of 25% or even 50%. In book on fundamentals of 'Biostatistics'(2) tables to find out the sample size, adequate to detect 25% or 50% risk reduction are given. Few other important aspects of quantitative reasoning are also discussed in that book.

The word significant in common parlance is understood to mean noteworthy, or important. Statistical significance too may have the same connotation but it can sometimes be at variance with medical significance. A statistically significant result can be of no consequence in the practice of medicine and a medically significant finding may occasionally fail test of statistical significance. A small and clinically unimportant difference can become statistically significant if the sample size is large. For example, suppose it is known that 70% of those with sore throat are automatically relieved within a week without treatment due to self-regulating mechanism in the body. A drug was tried on 800 patients and 584 (73%) cured in a week's time. Since P (Probability of type I error) is very small the null hypothesis

is extremely unlikely to be true and is rejected. Statistical significance is achieved and the conclusion of 73% cure rate observed in the sample being really more than 70% seen otherwise is reached. But, is this difference of 3% worth pursuing the drug? Is it medically important to increase the chances of relief from 70% to 73%? Perhaps not. Thus a statistically significant result can be medically not significant.

Some cautions are required in interpreting statistical non-significance as well. Consider the following results (1) of a trial in which patients on regular tranquilizer were randomly assigned to continued conventional management and a tranquilizer support group.

| | Tranquilizer support group | Conventional management group |
|---|---|---|
| Still taking tranquilizer After 16 weeks | 5 | 10 |
| Stopped taking tranquilizer by 16 weeks | 10 | 5 |
| Total | 15 | 15 |

Though the number of patients who stopped taking tranquilizer is double in the support group than in the conventional group yet the difference in not statistically significant ($^2$ (with Yate's correction) = 2.13, P = 0.1441, Fisher's exact P = 0.1431). There is a clear case of a trial on an enlarged n. If the same type of result is found on n=30 in each group then the difference would become statistically significant. The conclusion that the evidence is not enough to conclude presence of difference remains scientifically valid so long as n remains 15 in each group. Following example (3) will hopefully make it crystal clear.

Consider the example of a hypothetical intervention that aims to improve children's IQ. Suppose a population of children has a mean IQ of 100 with a standard deviation of 15. An intervention is introduced to improve their IQ. Suppose 4 students undergo the intervention and 4 do not. Then, it can be calculated that the intervention will be considered statistically significant if the intervention produces at least a 20.6-point increase in the IQ (assuming a constant SD of 15). Similarly, if 9 children are studied, the intervention should produce a 12.3-point increase in IQ, if 100 children are studied, the intervention should produce only 3.5-point increase in IQ and if 900 children are studied, the intervention should produce only 1-point increase in IQ. This example illustrates the limitation of relying only on statistical significance in making clinical decisions. Statistical tests in inferential statistics are, in general, designed to answer the question "How likely is the difference found in a sample due to chance (when actually no such difference exists in the population, the null-hypothesis)?"

**Methods:**

Let us discuss 'How biostatistics works?' in brief (in hypothesis testing situation only). This (may appear simple, but nevertheless should be absolutely clear i.e. there should not be a slightest confusion about it and so) is illustrated with one simple example. Consider a trial of "tossing of a fair coin". Since the coin is fair, probability of head coming up is one-half. Suppose we perform 20 trials and keep a count of heads coming up. About 10 heads are expected. Although 10 heads are expected, the actual number could be different. If the number is 9, 8 or 7, (or large as 11, 12, 13) we are not bothered. However, if the number is 6 or smaller (or larger than 13), we doubt about 'fairness' of the coin. If the number is very small like 0 or 1 (or very large like 19 or 20) we are almost sure about 'un-fairness'. We can find the actual probability of all possible events. These are calculated using 'binomial' distribution (3) and are displayed in table below

**Table I: Coin tossing experiment with n=20**

| Number of Heads | Probability (Percentage) |
|---|---|
| 0 | 0.00009 |
| 1 | 0.0002 |
| 2 | 0.018 |
| 3 | 0.109 |
| 4 | 0.462 |
| 5 | 1.479 |
| 6 | 3.696 |
| 7 | 7.393 |
| 8 | 12.013 |
| 9 | 16.018 |
| 10 | 17.620 |
| 11 | 16.018 |
| 12 | 12.013 |
| 13 | 7.393 |
| 14 | 3.696 |
| 15 | 1.479 |
| 16 | 0.462 |
| 17 | 0.109 |
| 18 | 0.018 |
| 19 | 0.0002 |
| 20 | 0.00009 |

This is called as 'sampling distribution'. Statistics/Biostatistics helps to estimate such probabilities easily in various situations (assuming that the variable follows some relevant particular theoretical distribution).

Inter-sample variability has another type of implication. If the mean decrease in cholesterol level after a therapy in a sample of 60 subjects of age 40-49 years is 9 mg/dL and in a second sample of 25 subjects of age 50-59 years 13 mg/dL, can it be safely concluded that the average decrease in the two groups is really different? Or is this difference just occurred by chance in these samples? Statistical methods help to take a decision one way or the other on the basis of the probability of occurrence of such a difference. When the conclusion is that the difference is very likely to be real then the difference is called statistically significant. In order to describe the concept of statistical significance more fully, we briefly visit the methodology followed in all empirical conclusions. This will also help in understanding the concepts of null hypothesis and of P-values which are so vital to the concept of statistical significance. These concepts are intimately related to Confidence Interval.

The concepts are best understood with the help of an example (2) of a court decision in a crime case. Consider the possibilities mentioned in following table. When a case is presented before a court by prosecutor, the judge is supposed to start with the presumption of innocence.

**Table II: Errors in various settings - Court setting**

| Decision | Assumption of innocence | |
|---|---|---|
| | True | False |
| Pronounced guilty | Serious error | √ |
| Pronounced not guilty | √ | Error |

In a court of law, it is up to the prosecutor to put-up sufficient evidence against the innocence of the person and changes the initial opinion of the judge. Guilt should prove beyond reasonable doubt. If the evidence is not sufficient, the person is acquitted whether the crime was committed or not. Sometimes the circumstantial evidence is strong and an innocent person is wrongly pronounced guilty. This is considered a very serious error. Special caution is exercised to guard against this type of error even at the cost of acquitting some criminals!

Diagnostic Journey (2): In the process of diagnosis a healthy individual may be wrongly classified as ill (false positive - misdiagnosis) and some really ill person may fail the detection procedure (false negative - missed diagnosis). Diagnostic journey always starts with the presumption of "no diagnosis" with respect to a particular disease. To rule out or to confirm the presence, a thorough clinical examination and/or some diagnostic test is used. But since the whole procedure is not full proof, the above two types of errors (misdiagnosis and missed-diagnosis) are possible.

**Table III: Errors in various settings - Diagnosis setting**

| Diagnosis | Actual condition | |
|---|---|---|
| | Disease absent | Disease present |
| Disease present | Misdiagnosed | √ |
| Disease absent | √ | Missed diagnosed |

In the process of hypothesis testing, two types of errors are possible to be committed. The error committed when a true null hypothesis is rejected (i.e. conclusion of significant difference where in fact there is no real difference) is called "Type I error". Probability of committing type I error is generally denoted by ". The error committed when a false null hypothesis is not rejected (i.e. conclusion of not significant difference in presence of true difference) is called "Type II error". Probability of committing type II error is generally denoted by ". The complement of type II error, '1-' (i.e. rejecting $H_o$ whenever $H_o$ is false) is called as Power.

In case of empirical decisions, the initial assumption is that there is no difference between the groups. This is equivalent to the presumption of innocence in the court setting and is called the null hypothesis. The notation used for this is $H_o$.

**Table IV: Errors in various settings - Empirical setting**

| Decision | Null hypothesis | |
|---|---|---|
| | True | False |
| Rejected | Type I error | √ |
| Not rejected | √ | Type II error |

The sample observations serve as evidence. Depending upon this evidence, the $H_o$ is either rejected or not rejected. In empirical set-up, the $H_o$ is never accepted. The conclusion reached is that the evidence is not enough to reject $H_o$. This may mean two things, - (i) carry out further investigations and collect more evidence, and (ii) continue to accept the present knowledge as though, this investigation was never done. The 'truth' remains unchanged.

Consider the claim of a manufacturer that his drug is superior to the existing ACE inhibitors in improving insulin sensitivity in diabetic hypertensives. In a trial (3) on matched cases, the improvement was seen in suppose 4 out of 10 patients on the new drug compared to 3 out of 10 on the existing drug. The sample size n = 10 in each group is too small and the difference is small to provide sufficient evidence to reject the $H_o$ of no difference between the drugs. If so, the claim of superiority is not tenable. The manufacturer needs to withdraw the claim forever or till such time that more evidence is available for scrutiny. The hypothesis of equality of two drugs is called as null hypothesis because it nullifies the effect which we want to prove. It is not always that null hypothesis is the statement, which nullifies the effect but it is the statement under which there exits only one condition. That's why for a given null hypothesis there is only one " value but number of " values because under the alternative hypothesis there exist many possibilities, and for each possibility there is one " value. In test of hypothesis we make " value (type I error which is in fact significance level) small but we generally exercise no control over " value (type II error).

The claim made is called the 'Alternative hypothesis' or 'Research hypothesis'. This is denoted by $H_1$. In the above example, the claim is that of superiority of the new drug. This is the alternative hypothesis, $H_1$, in this case. This is a one-sided alternative because only superiority is claimed and inferiority is ruled out. One sided alternative can be considered as saying that one group is 'at least as good as' or 'worse' than the other, and two-sided as saying that one group is 'either better or worse' than the other group. Sometimes it is not possible to claim that one group is better than the other but the only claim is that they are different. In the case of peak expiratory flow rate (PEFR) in factory workers exposed to different pollutants, there may not be any a priori reason to assert that it would be more in one than the other. Then the alternative is that the PEFRs are unequal. This is called two-sided alternative. The null is that they are equal in the two groups. The values observed in the sample serve as evidence against $H_o$. But these values are subject to sampling fluctuations and may or may not lead to a correct conclusion. The error of rejecting a true null hypothesis is similar to punishing an innocent. This is more serious and is called Type I error. This is popularly referred to as P-value. Thus P-value is the probability that a true null hypothesis is wrongly rejected. This is the probability that the conclusion of presence of difference is reached when actually there is no difference. In a clinical trial setup, this is the probability that the drug is declared effective or better when actually it is not. This wrong conclusion can allow an ineffective drug to be marketed as being effective. This clearly is unacceptable and need to be guarded against. For this reason, P-value is kept at a low level, mostly less than 5%, or $P<0.05$. The maximum P-value allowed in a problem is called the 'level of significance' or sometimes as -level. When P-value is this small or smaller, it is generally considered safe to conclude that the groups are indeed different.

The second type of error is failing to reject a $H_o$ when it is false. The probability of this error is denoted by ". In a clinical trial setup, this is equivalent to declaring a drug ineffective when it actually is effective. A drug which could possibly provide better relief to hundreds of patients is denied entry into the market. If the manufacturer believes that the drug is really effective, the company will carry out further trials and collect further evidence. Thus, the introduction of the drug is delayed but not denied. Let us suppose (2) that treatment 'A' produces 30% cure and treatment 'B' produces 55% cure. Now let us conduct a study, taking a pair of samples of size 30 each and administering treatment 'A' to one group and treatment 'B' to another group. Use the test of significance with the Type I error of 5% and judge whether there is any significant difference between the treatments based on these samples. Suppose that we have repeated such studies a large number of times and have judged that there is significant difference between the treatments only in 40 per cent of the studies. This means that in 60% of the pairs of samples we have failed to detect a difference that is large enough to reject the null hypothesis. We call this 60% as Type II error. The magnitude of risk of this error is related to the actual difference between the populations.

Every investigator is anxious to keep both Type I and Type II errors at the lowest but it is not possible to reduce both the errors simultaneously. For a given sample size, if one is reduced, the other automatically increases. Usually the Type I error is fixed at a tolerable limit and the Type II error is minimized. After fixing the Type I error, Type II error can be decreased by increasing the size of the sample. There is another useful concept closely associated with Type II error. If the Type II error is 60%, its complement i.e., 40% is known as the 'power' of the test. The power is a numerical value indicating the sensitivity of a test. Thus, power of a test is the probability of rejecting a $H_o$ when it is false. This depends on the magnitude of the difference between the observed and the real value present in the target population. The power of a test is high if it is able to detect small difference and reject $H_o$. Suppose the mean PEFR in workers of tyre manufacturing industry is 296 liters per minute and that in workers of paint-varnish industry 307 liters per minute. The mean difference is 11 liters per minute. This difference seems small relative to the PEFR values. A test with high power is needed to detect this difference and to call it significant. A low power test will not be able to reject $H_o$ of equality and will give conclusion that the difference is likely to have arisen due to chance in the samples studied.

Power becomes especially important consideration when the investigator does not want to miss a specified difference.

For example, a hypotensive drug may be considered useful if it reduces diastolic BP by an average at least of 5 mmHg after use for one week. A sufficiently powerful statistical test would be needed to detect this kind of difference. (1-) is an important consideration in this setup. However, one would like that the difference (5 mmHg) in this case is chosen with some objective basis. Increasing the size of the sample, beside by choosing an appropriate design of the study can increase power of a test. When the observed probability of Type I error, P, is less than a low threshold such as 0.05, the null hypothesis is rejected and the result is said to be statistically significant. The exact form of test criterion for obtaining P-value depends mostly on (i) the nature of the data (qualitative or quantitative), (ii) the form of the distribution if the data are quantitative (Gaussian or non-Gaussian), (iii) the number of groups to be compared (two or more than two), (iv) the parameter to be compared (can be mean, median, correlation coefficient, etc., in case of quantitative data; it always is proportion or a ratio or frequencies in case of quantitative data), (v) the size of sample (small or large), and (vi) the number of variables considered together (one, two or more).

Results from a single sample are subject to statistical uncertainty, which is strongly related to the size of the sample. These quantities (value of any measure) will be imprecise estimates of the values in the overall population, but fortunately the imprecision can itself be estimated (role of 'statistics' science!) and incorporated into the findings. When an average value or a proportion (or any other quantity such as ratio) is calculated from the sample drawn by the method of random sampling, we can estimate the range within which the corresponding population parameter is expected to lie with a given degree of probability. This probability is called confidence and the range so obtained is called a confidence interval (CI). The standard deviation and standard error of the mean measures two very different things. The standard error of the mean tells not about variability in the original population, as the standard deviation does, but about the certainty with which a sample mean estimates the true population mean. Since the certainty with which we can estimate the mean increases as the sample size increases, the standard error of the mean decreases as the sample size increases. Conversely, the more variability in the original population, the more variability will appears in possible mean values of samples. Therefore, the standard error of the mean increases as the population standard deviation increases (as $S.E._{Mean} = (SD/n)$ where n is the sample size).

Most medical investigators summarize their data with the standard error of the mean because it is always smaller than the standards deviation. It makes their data look better. However, unlike the standard deviation, which quantifies the variability in the population, the standard error of the mean quantifies uncertainty in the estimate of the mean. Since readers are generally interested in knowing about the population, data should never be summarized with the standard error of the mean. To understand the difference between the standard deviation and standard error of the mean and why one ought to summarize data using the standard deviation, consider the following example (4).

Suppose that: Average duration of gestation period in 100 women was found to be 280 days with standard deviation of 5 days. Because the sample size is 100, the standard error is 0.5 and the 95 per cent confidence interval for average gestation period of the entire population is 279 to 281. These values describe the range, which, with about 95 per cent confidence, contains the average gestation period of the entire population from which the sample of 100 women was drawn. This is not the interval that contains gestation period of 95% of the women. If we want that interval, then we should use standard deviation and not the standard error. So the interval which contains gestation period of 95 per cent of the women (assuming 'normal' distribution) is 280 2 (5) = 270 to 290. Such interval is called "tolerance interval" and the end points of such interval are called "tolerance limits."

Consider one more example: suppose that in a sample of 25 patients an investigator reports that the mean cardiac output was 5L/min with a standard deviation of 1 L/min. Since about 95 per cent of all population members fall within about 2 standard deviations of the mean, this report will tell you that (assuming that the population of interest followed a normal distribution) it would be unusual to observe a cardiac output below about 3 or above about 7 L/min. Thus, you have a quick summary of the population described in the paper and a range against which to compare specific patients you examine. Unfortunately, it is unlikely that these numbers would be reported, the investigator being more likely to say that the cardiac output was 5.0 ± 0.20 L/Min. If you confuse the standard error of the mean with the standard deviation, you would believe the range of most of the population was narrow indeed (4.6 to 5.4 L/min). These values describe the range, which, with about 95 per cent confidence, contains the mean cardiac output of the entire population from which the sample of 25 patients was drawn. In practice, one generally wants to compare specific patient's cardiac output not only with the population mean but with the spread in the population taken as a whole.

There are generally many assumptions made while constructing a test (deriving mathematically the sampling distribution of test statistic / estimation method). We either do not know (study it to that extent) or do not bother to verify that they are fulfilled in given situation. However, they are underlying and one should be aware of them (to apply appropriate/most applicable method because otherwise it is likely produce lot of bias in the study/results). For example, many sample size ('n')formulas assume 'simple random sampling' and when any other sampling scheme is used we have to multiply this sample size by "design effect" (5). In this

context it is essential to be aware of other assumptions (6) made while applying any sample size formulas/estimation methods. When the required conditions are not fulfilled, usual methods may fail. To illustrate failure of usual procedure (condition: extreme proportion) consider an example (7). Suppose a particular surgeon has done 10 operations without a single complication. His observed complication rate p is 0/10 = 0 percent for the 10 specific patients he operated on. This is impressive but it is unlikely that the surgeon will continue operating forever without a complication. Therefore the fact that p = 0 probably reflects good luck in the randomly selected patients who happened to be operated on during the period in question. To obtain a better estimate of p, the surgeon's true complication rate, we will compute the 95 per cent confidence interval for 'p'. Usual procedure yields SE as zero and so the CI is from 0 to 0. This result does not make sense. Obviously, the approximation breaks down. Exact confidence intervals for proportions corresponding to the observed 'p' is indicated and for 95% confidence which is from 0% to 31%. In other words, we can be 95 percent confident that his true complication rate, based on the 10 cases we happened to observe, is somewhere between 0 and 31 percent.

Nearly all information in medicine is empirical in nature and is gathered from samples of subjects studied from time to time. Besides all other sources of uncertainty, the samples themselves tend to differ from one another. For instance, there is no reason that the 10 year survival rate of cases of carcinoma breast in two groups of women of 100 each the first group born on odd days of any month and the second group on even days of any month, is different, but there is a high likelihood that this would be different. This happens because of sampling error or sampling fluctuation. This depends on two things - i)the sample size 'n', and ii) the intrinsic inter-individual variability in the subjects. The former is fully under control of the investigator. The latter is not under human control, yet its influence on medical decisions can be minimized by choosing an appropriate design and by using appropriate methods of sampling. It must be clearly kept in mind that tests of statistical significance and confidence intervals evaluate only the role of chance as an alternative explanation of an observed association between an exposure and disease. While an examination of the P value and or confidence interval may lead to the conclusion that chance is an unlikely explanation for the findings, this provides absolutely no information concerning the possibility that the observed association is due to the effects of uncontrolled bias or confounding. All three possible alternative explanations (chance, bias, confounding) must always be considered in the interpretation of the results of every study (8). Any study has two main aspects – generalizability (sometimes called as External Validity) & validity (or sometimes prefix internal is used). By using a big sample, only generalizability aspect is insured but by no means the important validity aspect. Therefore, sample size is not everything because if the study (and so the results) is/are less valid, what is the use of generalizability? It is well known that increasing sample size decreases the standard error as it is inversely proportional to sample size. However, reduction in sampling error can be achieved by using the appropriate sampling (or study) design instead.

It is all most impossible to even overview all the aspects of methods in biostatistics useful for clinical studies, however, let us discuss one more very important aspect namely 'bias' A dictionary definition of 'bias' is 'a one-sided inclination of the mind. In statistics, 'bias' is 'systematic error' that can produce results that depart from the true values. That is, bias is a trend in the design, collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth. Medical research results many times become inconclusive because some bias is detected after the results are available. Therefore, it is important that all sources of bias are considered at the time of planning a study, and all efforts are made to control them. Other type of error is 'random error'. Random variation can never be eliminated totally; however, one can reduce the role of chance by proper design, adequate sample size, and appropriate analyses. Chance should always be considered when assessing the results of clinical observations. But it is very important to note that these two sources of error – bias and chance – are not mutually exclusive. It is well known that 'bias' can produce dramatic change in study results. Few such dramatic effects of bias are shown in an excellent article by Sackett (9) on 'biases in analytic research'.

Randomization in randomized controlled trials (RCTs) involves more than generation of a random sequence by which to assign subjects. For randomization to be successfully implemented, the randomization sequence must be adequately protected (concealed) so that investigators, involved health care providers, and subjects are not aware of the upcoming assignment. The absence of adequate allocation concealment can lead to selection bias, one of the very problems that randomization was supposed to eliminate. Authors of reports of randomized trials should provide enough details on how allocation concealment was achieved so the reader can determine the likelihood of success. Fortunately, a plan of allocation concealment can always be incorporated into the design of a randomized trial. Keeping knowledge of subjects' assignment after allocation from subjects, investigators/health care providers, or those assessing outcomes is referred to as masking (also known as blinding). The goal of masking is to prevent ascertainment bias. In contrast to allocation concealment, masking cannot always be incorporated into a randomized controlled trial.

Both allocation concealment and masking add to the elimination of bias in randomized controlled trials.

Sackett (9) identified 56 possible biases that may arise in any analytic research of which over two-thirds are related to aspects of study design and execution. Methodologically inferior trials might produce bias in both directions, thereby causing greater variability in estimates of treatment effects. Bias may also lead to fallacious interpretation of study/trial results. In few books (2) a complete chapter devoted to 'statistical fallacies' enumerate indirectly the effect of many such biases. Wrong choice may produce lot of bias. Though many instances can be quoted of wrong choice of methodology of data analyses, quoting one common example should suffice to highlight the relevance. In several types of studies we may want to examine the consistency of an observed relation across two or more subgroups of the individuals studied. For example, in a clinical trial we might want to know if the observed treatment difference is the same for young and old patients or for different stages of disease at presentation. In such cases we are interested in examining whether one effect is modified by the value of another variable. This may be viewed as the examination of the heterogeneity of an observed effect such as treatment benefit across subsets of individuals. The statistical term for heterogeneity of this type is "interaction"(8). The medical concept of "synergy" is the same thing (opposition in physiological action is "antagonism"). The statistical term interaction relates to the non-independence of the effects of two variables on the outcome of interest. It is advised very strongly (with reasoning) in the literature that to conclude presence of interaction always "compare effect sizes and not the P values", comparing 'P' values alone can be misleading. Comparing confidence intervals is less likely to mislead. However, the best approach is to compare directly the effect sizes using "test of interaction" (8). Still one can often see the practice of comparing 'P' values alone in such situations.

It may be noted that most of the 'biases' fall into one of three broad categories:

1. Selection Bias (occurs when comparisons are made between groups of patients that differ in determinants of outcome other than the one under study).
2. Measurement Bias (occurs when the methods of measurements are dissimilar among groups of patients).
3. Confounding Bias (occurs when two factors are associated i.e. travel together and the effect of one is confused with or distorted by the effect of the other).

Some steps are suggested for minimizing bias after giving a list of newer biases elsewhere (8). Note that if a study is planned, designed, executed, analyzed, interpreted, etc., properly then occurrences of any type of 'biases' are less likely. Remember that statistical significance and non-significance are equally important. Further ask 'can there be non-causal explanations of the results? Are there any confounding factors that have been missed? Whether chance or sampling error could be an explanation?'. Such consideration will help you to develop proper design, and to conduct the study in an upright manner. Discussion regarding specific issues like newer designs including equivalence/non-inferiority trials, adaptive allocations, new indexes of safety, clinical agreement/disagreement, evaluation of diagnostic/screening tests, interpretation of diagnostic test results, measures of clinical significance, comparison of paired proportions, RIDIT analysis, and time-to-event or survival analysis - clinical life table are not covered for requirement of lot of space (but are described well elsewhere (8).

**References:**

1. Indrayan Abhay and Sarmukaddam S.B. 'Medical Biostatistics'. Marcel Dekker Inc., New York, 2001.
2. Sarmukaddam SB. 'Fundamentals of Biostatistics'. Jaypee Brothers Pvt. Ltd., New Delhi, 2006.
3. Sarmukaddam SB. 'Clinical Biostatistics'. New Age International Publishers, New Delhi, In-press.
4. Sarmukaddam SB. 'Biostatistics Simplified'. Jaypee Brothers Pvt. Ltd., New Delhi, 2010.
5. Sarmukaddam SB. 'Sample size versus choice of appropriate sampling design' Medical Teacher 2001; 23:102-103.
6. Sarmukaddam SB, Garad SG. 'On Validity of assumptions while determining sample size' Indian Journal of Community Medicine 2004, XXIX: 87-91.
7. Sarmukaddam SB. 'Interpreting "statistical hypothesis testing" results in clinical research'. Journal of Ayurveda & Integrative Medicine 2012; 3:65-69.
8. Sarmukaddam SB. 'Methods & Controversies in Clinical Trials'. Centre for Behavioral Medicine, Pune, In Press.
9. Sackett DL. 'Bias in analytical research'. J Chronic Dis 1979; 32:51-63.