Interpreting "statistical hypothesis testing" results in clinical research

Sanjeev B. Sarmukaddam

Maharashtra Institute of Mental Health, B.J. Medical College and Sassoon Hospital Campus, Pune, Maharashtra, India

ABSTRACT

Difference between "Clinical Significance and Statistical Significance" should be kept in mind while interpreting "statistical hypothesis testing" results in clinical research. This fact is already known to many but again pointed out here as philosophy of "statistical hypothesis testing" is sometimes unnecessarily criticized mainly due to failure in considering such distinction. Randomized controlled trials are also wrongly criticized similarly. Some scientific method may not be applicable in some peculiar/particular situation does not mean that the method is useless. Also remember that "statistical hypothesis testing" is not for decision making and the field of "decision analysis" is very much an integral part of science of statistics. It is not correct to say that "confidence intervals have nothing to do with confidence" unless one understands meaning of the word "confidence" as used in context of confidence interval. Interpretation of the results of every study should always consider all possible alternative explanations like chance, bias, and confounding. Statistical tests in inferential statistics are, in general, designed to answer the question "How likely is the difference found in random sample(s) is due to chance" and therefore limitation of relying only on statistical significance in making clinical decisions should be avoided.

Key words: Clinical research, clinical significance, hypothesis testing, statistical significance

INTRODUCTION

The science of statistics should not be condemned because it can be abused (fallacies designed to mislead) or misused (fallacies committed unintentionally) for fault lies not with statistics as such but with the user of the subject. For instance, (though most of these are "truistic/self-evident" examples and so are too obvious for mention), if a person takes a wrong medicine or excessive dose of medicine and dies, we cannot blame the medicine as such. Even if statistical hypothesis testing (significant "*P*" value) shows that the risk of dying from an illness while in the hospital is many times greater than the risk of dying from an illness while at home (though most of the people many not be able to precisely identify and explain the logical fallacy implied by this type of statistical reasoning), the nonsense of it is sufficiently apparent so that no one is likely to avoid beds or hospitals in order to prolong his life. Yet the same sort of statistics too often finds its way into the general health literature and mass media where the errors of logic are less apparent and so pass unrecognized.

It is very rightly said somewhere that "He who accepts statistics indiscriminately will often be duped unnecessarily. But he who distrusts statistics indiscriminately will often be ignorant unnecessarily." Every day we see misuses of statistics which affect the outcomes of elections, change public policy, win arguments, get readers for newspapers, impress readers, support prejudices, inflame hatreds, provoke fears, sell products, etc. It is common to sneer at the subject and say that, "You can make statistics say anything," (Lies, Dam lies and statistics!). It is only through abuse that you make statistics say, "anything." Good statistics tells only the truth.

While interpreting "statistical hypothesis testing" results in clinical research, it is very important to keep in mind the difference between "Clinical Significance versus Statistical

Address for correspondence:
Dr. Sanjeev B. Sarmukaddam, 25,Sangeet Sadhana,
Krishna Colony, 11th Lane, Paramhans Nagar,
Paud Road, Pune, Maharashtra, India.
E-mail: sanjeev.sarmukaddam@gmail.com

Received: 13-Feb-2012
Revised: 16-Apr-2012
Accepted: 23-Apr-2012

Access this article online

Quick Response Code:

Website: www.jaim.in

DOI: 10.4103/0975-9476.96518

Significance." Philosophy of "statistical hypothesis testing" is sometimes unnecessarily criticized based on such wrong premise (failing to understand the difference). Randomized controlled trials are also wrongly criticized similarly.[1] Such wrong arguments appeared even in the past but died their own early death because there were "no takers" to such meaningless arguments. For any statistical test to be applicable, we assume a "population model" (i.e. existence of population from which a random sample(s) is/are drawn) but in RCTs (there is/are no random sample(s), we apply only random allocation) we evoke "population model" and apply statistical test. This is just an example, which points that we often use statistical methods to suit our convenience and mold them to make applicable. It is better to bear in mind that "statistical hypothesis testing" is not for decision making always (occasionally done keeping limitations in mind). Also, note that field of "decision analysis" (a philosophy of decision making that helps get to so called clarity of action) is an integral part of (or topic in) statistics as a branch of science.

On making sense of data and/or results

There are generally many assumptions made while constructing a test (deriving mathematically the sampling distribution of test statistic). 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. For example, many sample size formulas assume "simple random sampling" and when any other sampling scheme is used, we have to multiply this sample size by "design effect."[2] When the required conditions are not fulfilled usual methods may fail. To illustrate failure of the usual procedure (condition: extreme proportion/percentage) consider an example. Suppose a particular surgeon has done 10 operations without a single complication. His observed complication rate "*P*" is 0% for the 10 specific patients he operated on (or success rate is 100%). 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% confidence interval (CI) for "*P*." Usual procedure yields SE as zero and so CI is from zero to zero. This result does not make sense. Obviously, the approximation breaks down. Exact CIs (based on binomial distribution) for proportions corresponding to the observed '*P*' is indicated which for 95% confidence level is from 0% to 31%.

In other words, we can be 95% confident that his true complication rate, based on the 10 cases we happened to observe, is somewhere between 0% and 31%.[3] The specific 95% CI we obtain depends on the specific random sample we happen to select for observation. The CI gives a range that is computed in the hope that it will include the parameter of interest. A particular interval (associated with a given set of data) will or will not actually include the true value of the parameter. The confidence level associated with the interval (say 90%, 95%, or 99%) gives the percentage of all such possible intervals that will actually include the true value of the parameter. Unfortunately, you can never know whether or not that interval does. All you can say is that the chance of selecting an interval that does not include the true value is small (10%, 5%, or 1%). Therefore, a specific 95% CI associated with a given set of data may or may not include the true size of the treatment effect, but in the long run 95% of all possible CIs will include the true value associated with the treatment. So it describes not only the size of the effect but quantifies the certainty with which one can estimate the size of the treatment effect. Therefore, saying that "CIs have nothing to do with confidence" is not correct.

95% CI is {sample value of parameter ± 1.96×SE}. Note that when the variable under consideration is "qualitative/dichotomous" there is no Standard Deviation (SD), but when the variable under consideration is "quantitative" SD exists and is very much important. Most medical investigators summarize their data with the standard error (SE) of the mean because it is always smaller than the standards deviation [SE= (SD/square root of sample size)]. It makes their data look better. However, unlike the standard deviation, which quantifies the variability in the population/sample, the standard error of the mean quantifies uncertainty in the estimate of the mean. Since readers are generally interested in knowing about the study 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 this example. Suppose that: Average duration of gestation period in 100 women was found to be 280 days with standard deviation of 5 days. As the sample size is 100, the standard error is 0.5 and the 95% CI for average gestation period of the entire population is 279 to 281. These values describe the range, which, with about 95% confidence, contains the average gestation period of the entire population from which the random 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% of the women is $280 \pm 1.96 \times 5 \approx$ 270 to 290. Such interval is called "tolerance interval" and the end points of such interval are called "tolerance limits."

As pointed out in recent Nature's editorial[4], though clinical trials are the best way to assess efficacy of treatment, trials in their present form may not be suitable for Complementary and Alternative Medicine (CAM). Therefore, it has become essential to modify or identify methods/techniques suitable for CAM including Ayurveda. In one such attempt[5] conducting trials according to "Equivalence Trial Design" is recommended and one possible "Safety Index" is proposed. A case is made for the appropriate use and relevance of pragmatic trials in the evaluation of alternative and complementary medicine in article by Hugh MacPheron.[6] The main strength of pragmatic trials (more detailed discussion can found in the reference) is that they can evaluate a therapy as it is used in normal practice. Pragmatic trial could be used to test an overall "package" of care (similar to WHO's "black-box" design) and it is easier to grant the practitioners the freedom to treat the patients normally, allowing them to use individual approaches for different patients. It may be specifically noted that pragmatic trial philosophy goes well with the equivalence or non-inferiority trial.

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 CIs 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 CI 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.

Sample size *n* plays a dominant role in statistical inference. The standard error (SE) can be substantially reduced by increasing *n*. This helps to increase the reliability of the results. A narrow CI is then obtained that can really help in drawing a focused conclusion. At the same time, a side effect of large *n* is that a very small difference can become statistically significant. This may or may not be clinically/medically significant. Any study has two main aspects – generalizability (sometimes called as External Validity) and 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. If the study (and so the results) is 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.[7]

Clinical significance goes beyond arithmetic and is determined by clinical judgment. Nevertheless, measures such as number needed to treat (NNT) could be of help to sort out whether the benefits of a particular treatment are big enough. The results of most clinical trials are presented as relative risk reduction or odds ratios, but these ignore the role of event rate on overall clinical benefit. Therefore, in clinical trials a better quantification of overall clinical benefit is provided by presenting results as number needed to treat which is defined as the number of people that needed for a given duration to prevent one death or one adverse event. Method of calculation of NNT and its CI are given in many text books[8] on the subject. 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 on a clinical trial. These yardsticks are ¾ (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?

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 β errors of such trials are very large and their power (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 above quoted book,[8] 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 this book.

Not being able to reject null hypothesis (Ho) is analogous

to pronouncing in a court that the person is "not proven guilty." This is different from saying that the person is "not guilty." The other way that this could be understood is that a null hypothesis is "conceded" but not accepted. Distinction must also be made between "not significant" and "insignificant." Statistical tests are for the former and not for the latter. A statistically "not significant" difference is not necessarily "insignificant." With statistical inference, the results can seldom, if ever, be absolutely conclusive, as the *P*-value never becomes zero. There is always a possibility, however small, that the observed difference arose by chance alone. Whenever statistical significance is not reached, the evidence is not considered in favor of Ho—it is only not sufficiently against it. Samples provide evidence against Ho and in favor of alternative hypothesis (H1), but never in favor of Ho and against H1.

The word significant in common parlance is understood to mean noteworthy, or important. Statistical significance too has 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. The SE depends heavily on the sample size. A result based on a large sample is much more reliable than a similar result based on a small sample. This reflects on the width of CI on one hand and on *P*-value on the other. A small and clinically unimportant difference can become statistically significant if the sample size is large. For example,[9] 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 caution is required in interpreting statistical non-significance as well. Consider the following results[9] of a trial in which patients on regular tranquilizer were randomly assigned to continued conventional management and a tranquilizer support group [Table 1].

Although 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 [χ^2 (with Yate's correction) = 2.13, $P = 0.1441$, Fisher's exact *P* (two tailed) = 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.

EPILOGUE

An important point to be emphasized is "A statistically significant result can be of no consequence in the practice of medicine as it depends heavily on the sample size." One more example will help 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 four students undergo the intervention and four do not. Then, it can be calculated that the intervention will be considered statistically significant (at $P \leq 0.05$ i.e. 5% level) if the intervention produces at least (approximately) a 26.5-point increase in the IQ (assuming a constant SD of 15). Similarly, if 9 children are studied (in each group), the intervention should produce (approximately) a 14.99-point increase in IQ, if 100 children are studied, the intervention should produce only (approximately) 4.24-point increase in IQ and if 900 children are studied, the intervention should produce only (approximately) 1.38-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 random sample(s) due to chance (when actually no such difference exists in the population, the null-hypothesis)?" This fact (limitation of relying only on statistical significance in making clinical decisions) is illustrated with example(s) in many text books[10] sometimes may be in different context like correlation/association versus cause–effect relationship.

Baye's theorem of conditional probabilities and its application in estimating positive/negative predictive values

Table 1: Result of a trial of 'tranquilizer' – an example showing clinically important difference which is not statistically significant[9]

	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

for any given "prior probability" and sensitivity as well as specificity or likelihood ratio [for positive test result which is LR+= {(Sensitivity) / (1 – Specificity)} and for negative test result which is LR_ = {(1 – Sensitivity) / (Specificity)}] is given in many text books.[8-11] The fact that "posterior probability" changes according to "prior probability" is well known in the field of statistics. For excellent discussion on other "quantitative aspects of clinical reasoning" readers may refer to the book by Sackett DL *et al*.[12]

REFERENCES

1. Raha S. A critique of statistical hypothesis testing in clinical research. J Ayurveda Integr Med 2011;2:105-14.
2. Sarmukaddam S, Garad S. On validity of assumptions while determining sample size. Indian J Community Med 2004;29:87-91.
3. Sarmukaddam S. Biostatistics Simplified. New Delhi, India: Jaypee Brothers Pvt. Ltd.; 2010.
4. Grayson M. Traditional Asian medicine. Nature 2011;480(7378):S81.
5. Sarmukaddam S, Chopra A, Tillu G. Efficacy and safety of Ayurvedic medicines: Recommending equivalence trial design and proposing safety index. Int J Ayurveda Res 2010;1:175-80.
6. MacPheron H. Pragmatic clinical trials. Complement Ther Med 2004;12:136-40.
7. Sarmukaddam S. Sample size versus choice of appropriate sampling design. Med Teach 2001;23:102-3.
8. Sarmukaddam S. Fundamentals of biostatistics. New Delhi, India: Jaypee Brothers Pvt. Ltd.; 2006.
9. Indrayan A, Sarmukaddam S. Medical Biostatistics. New York, USA: Marcel Dekker; 2001.
10. Sarmukaddam S. Clinical Biostatistics. New Delhi, India: New Age International Publishers Ltd.; [In Press].
11. Fletcher R, Fletcher S. Clinical Epidemiology – The essentials. 4th ed. Philadelphia, USA: Lippincott Williams and Wilkins; 2005.
12. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology – A Basic Science for Clinical Medicine. 2nd ed. Boston, USA: Little Brown; 1991.

How to cite this article: Sarmukaddam SB. Interpreting "statistical hypothesis testing" results in clinical research. J Ayurveda Integr Med 2012;3:65-9.

Source of Support: Nil, **Conflict of Interest:** None declared.