

Finding Your Way Through the Forest – A TCM Practitioner’s Guide to Evaluating Research: Part 2

Tony Reid


Abstract

Evidence based medicine is the prevailing paradigm of modern healthcare. However, practitioners of traditional Chinese medicine (TCM) vary significantly in their ability to appraise and understand modern research. Part 2 of this series elaborates on basic statistical and methodological concepts in medical studies. Some of the key axioms that underlie statistical science and clinical trial design are explored and discussed.

Keywords

Evidence based medicine, research, clinical trial, statistical methodology, normal distribution, standard deviation, probability, confounding factors

Introduction

 In order to fully appreciate the strengths and weaknesses of contemporary clinical trial literature, in relation both to statistics and trial design, it is important to carefully review some of the fundamental concepts and the nature of the information that they provide to us. Then we will be better equipped to understand and appreciate their limitations and potential for misuse, many of which have been raised within the scientific community.¹ Our purpose at this point is to clearly understand the nature and scope of statistical studies: what they can do and what they cannot do.^{2,3}

The scope of a statistical study

Statistical studies are not designed to provide proof; this is the domain of mathematics and logic. A statistical study of a medical treatment provides information about a specific relationship, ie the strength of association between an intervention and the outcomes that have been observed. Such studies are only able to demonstrate association, but

not causation.

Acceptable proof of causation in medicine may be obtained through a complex series of steps, which go beyond the demonstration of a strong association using statistical analysis. Originally developed by Hill, the nine ‘Bradford Hill Criteria’, provide the basis for ongoing refinements in scientific discussions on causal inference.^{4,5} Moreover, in any statistical study, because of the limited sample size (ie the number of subjects with a specific disorder within a clinical trial versus the total population with that disorder), the results will always carry some degree of uncertainty. This is the reason why a clinical trial is usually repeated several times in different locations and with different members of the population of interest (eg children, pregnant women and the elderly), before the results can be accepted into mainstream practice. The critical point is this: statistical studies aim to minimise uncertainty and inaccuracy - they do not, and cannot, entirely remove them.

Thus, the results of a clinical trial are never entirely true nor, for that matter, entirely false; even though

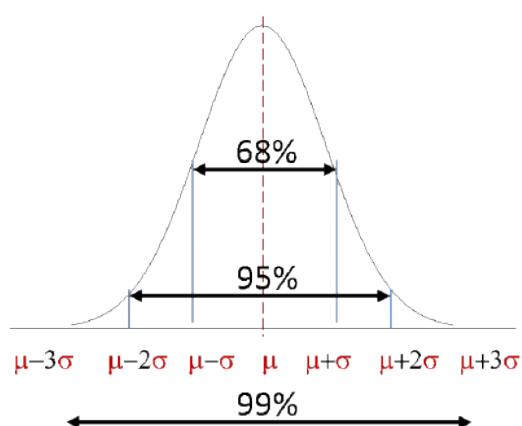


Figure 1: Normal distribution – the bell-shaped curve

μ = the average or mean value, generally referred to simply as ‘the mean’

σ = the standard deviation (SD)

Approximately 68% of values in the distribution are within one SD of the mean, ie, they lie between one SD above or one SD below the mean value

Approximately 95% of values in the distribution are within 2 SD of the mean

Approximately 99% of values in the distribution are within 3 SD of the mean

some experts in the field seem to imply the contrary.⁶ The best that can be provided by a statistical study is a likely range of values, such as effect size and percentage of positive responders to treatment, together with the associated degree of uncertainty. This is generally expressed in terms of the 95 per cent confidence interval (CI) and the mean (average) outcome, around which we expect most of clinical results within the target population to be clustered. In other words, although a clinical trial may start out with a ‘research question’, we never end up with a yes or no answer. If clinical studies are carried out properly, we can be reasonably confident that the outcomes will be reflected in our clinical practice. Thus, while inaccuracies and uncertainties are always part of a statistical study, the aim should always be to minimise them.

Normal distribution

The most fundamental statistical concept in studies of populations is the ‘normal distribution’ curve. Please take a few moments to familiarise yourself with Figure 1. In general, if we measure a particular characteristic (eg height) in every member of a particular population (eg adult males), and we plot the resulting measurements on a graph, with number of people on the vertical (left hand) axis and height measurement on the horizontal (bottom) axis, the results will conform to the pattern of distribution in Figure 1. This is referred to as the normal distribution curve (or simply ‘normal distribution’). In this scenario, we find that the measured values (height) will cluster around a ‘mean’ (the average height) and be distributed in such a way that approximately 95 per cent of the population have a height

that is within two ‘standard deviations’ (SD) of the mean (ie within the range of two SD below the mean and two SD above the mean). The value of the standard deviation is calculated mathematically from the data that have been collected. In the example just given, our measurements and analysis are highly accurate because we have, at least theoretically, measured every member within the population. There is no uncertainty here because we have measured everyone and the degree of accuracy is only limited by the accuracy of our measuring.

Critical assumptions

Now, what if we do not have the resources or the time to measure every member of our designated population, and we only measure a small portion of subjects? How accurate would our figures be when applied back to the entire population of interest? Here we have our basic ‘leap of faith’. At this point, the assumption of statistical science is that we can quantify both the range of measurements and the degree of uncertainty, based on data obtained from a small portion of the total population. However, this assumption may or may not be true. Imagine a situation where we only measured the height of adult males in a nursing home, or alternatively, of an elite basketball team. The average height would likely be quite different in each scenario; neither would be representative of the entire population. This is an example of a ‘confounding factor’, which has a significant effect on the accuracy, and hence veracity, of the results obtained in our research.

If we consider how this may apply in a clinical trial, one very important assumption is that the participants taking the active treatment are evenly matched with those who are taking the placebo. Factors such as age, severity of illness and psychological state may profoundly influence a patient’s response to an intervention. Therefore, we must always remember that a critically important difference between the two groups may have been overlooked.

In conclusion, no matter how well designed or carefully analysed, there is always a chance, albeit a small one, that the results of any clinical trial may not be applicable to the total population of interest. One or more of our assumptions, conscious or not, may have been erroneous: the participants in the study may not truly represent the total population with the disorder. Moreover, the diagnosis of the disorder may be invalid, and therefore the subjects within the trial are falsely assumed to be a homogenous group - for example, studies of depressed patients may include people who are feeling sad due to a loss of some kind (and who tend to ‘get over it’ within a few months) along with those who are feeling sad for no reason (and who tend to become chronically depressed).⁷ Thus, while every effort

may have been made to ensure accuracy and reliability in a clinical trial, there always exists the finite possibility that there are hidden errors and that the results may not apply in a real-life clinical setting. Nevertheless, we must also acknowledge that a carefully designed clinical trial, analysed correctly and reported transparently, can provide valuable clinical information, and also that statistics do have a rightful place in our clinical decision making. Our goal is to discover the best ways to evaluate this kind of information.

Mean and standard deviation

Returning to our example of height measurement in adult males, if we were to pick only one member of this population at random and measure his height, the result will lie somewhere between that of the shortest and the tallest member of this group. Moreover, in the absence of data collected from the whole population, including the mean and standard deviation, our ability to predict his height, or the likely range in which his height may fall, is at its weakest. It should be noted here that this is also what happens in the clinic: we are seeing an individual whose response to a treatment can be anywhere from the poorest to the best response as per the relevant clinical trials. The patient in question has nothing to do with any of these trials (unless, of course, he participated in one of them). If the trial data leads to the conclusion that seven out of ten patients with his condition will benefit in a clinically significant manner from a particular treatment and three will not, we have no way of knowing whether our patient belongs to the three or the seven, nor to what degree this person will respond if positive results are indeed forthcoming.

Returning to the example of height, when we have data from either the whole population, or a large enough representative portion, we are then able to make a judgement about whether an individual’s height is ‘normal’ or ‘abnormal’. Intuitively we know that a person’s height would be considered normal when it is close to the mean (average) height, and abnormal when a person is an outlier, outside of the majority. Statistical science quantifies this in the following way: normal height = the range of height measurements that are clustered around the mean (ie within the range of plus or minus two SDs – where we will find 95 per cent of the subjects), and abnormal height would be any measurement that is outside of this range. The more people we examine, the more likely we are to find that their heights fall within the ‘normal’ range. Turning

Any study dealing with a limited number of subjects will generate a mean that is likely to be different from the true mean of the whole population of interest.

this around, we can say that with fewer subjects, the SD is larger, indicating less accuracy (in terms of the true range of normal); with more subjects the SD becomes smaller, more closely approximating the true SD for the entire population. If we apply this idea to a clinical trial, the most accurate results will be obtained if we are able to enrol every member of a target population (ie everyone with a particular

disease); the fewer the subjects, the less accurate the results will become. Therefore, we need to find a practical compromise – an appropriate number of subjects that will provide meaningful results. This is calculated mathematically, based upon the size of the

effect (that we are hoping to achieve or avoid) and the cut-off point that we choose for deciding whether the association between intervention and effect is significant.^{8,9}

Confidence intervals: two additional bell-shaped curves

The discussion above illustrates the fact that any study dealing with a limited number of subjects will generate a mean that is likely to be different from the true mean of the whole population of interest (remembering that the mean is the average, above and below which the actual measured values will be clustered). This degree of uncertainty can be quantified and is provided in clinical studies by the ‘confidence interval’ (CI), which represents the range within which the true mean for the entire population is likely to occur. This is a separate calculation from the SD, and is meant to provide an estimation of both the best and worse case scenarios (ie outcomes) of a clinical trial when the results are applied within the general population. Thus, a more realistic interpretation of results in a clinical study could be visually represented by adding two additional bell-shaped curves (with the same SD as the original), one to the left and one to the right of the original in Figure 1 above. One of our new bell-shaped curves will be centred on the lowest value of the CI and the other on highest value. In this way we can see a more realistic range of outcomes for a clinical trial. The ramifications of this idea will be further explored in part 3.

Probability and statistical significance

Another important component of medical statistics is probability theory, the use of which may be illustrated with a simple example: calculating the probability (denoted by the symbol, ‘p’) that a coin tossed five times will come

down with the same side up every time. The first toss determines which side we are looking for; and in each of the four subsequent tosses the probability of getting this particular side (say, heads) is one in two (ie 0.5 or 50 per cent) for each toss. To calculate the probability of the four subsequent tosses showing heads we multiply these probabilities together: $0.5 \times 0.5 \times 0.5 \times 0.5 (= 0.0625)$. This gives us a probability of 6.25 per cent ($p = 0.0625$).

The precision of the maths belies the fact that we are not actually being provided with a true measurement here. The above calculation implies that if we were to repeat the example 200 times, we would find 13 sequences of five with identical faces (or 6.5 in 100). Unfortunately, this is unlikely to be correct. What the original p value (ie $p = 0.0625$ or 6.25 per cent) really means is that if we repeated the second experiment (200 lots of 5 throws) many times, the number of times we get five identical faces will get closer and closer to 13 the more times we repeat the experiment. Unfortunately, this does not really tell us very much in a practical sense. When making real-world decisions, the values and expectations of the observers play a major role in how this information is to be used. Returning to our original example, after the five identical tosses, we might be suspicious that this coin is weighted (or biased), but as p is not less than 0.05 (the generally accepted cut-off for significance) we may be swayed by the statistical paradigm and be inclined to accept that the coin could indeed be normal. However, if we added an additional throw and got another head ($p = 0.03125$), then we can start to become suspicious. Of course, this whole process is now beginning to look somewhat ridiculous. Personally, I would be examining the coin very closely after the third or fourth head, especially if I were gambling and had placed my money on tails!

The above example illustrates several important points about the application of statistics in research. We, or rather the statisticians, can only calculate the probability that the difference between the results in the two study groups is for all intents and purposes due to random chance. This is the p value, and in clinical studies the level of statistical significance is generally set at five per cent ($p < 0.05$). It is important to note that the p value is calculated as a measure of the likelihood that the trial results occurred due to random chance: it is a measurement of non-association. By convention we say that the results only become statistically significant when this probability is less than five per cent (expressed as $p < 0.05$). In other words, this means that when p is less than five per cent the results are not insignificant: it is deemed unlikely that they are not unrelated. It should be noted that we have, in reality, only measured insignificance, not significance. We are assuming that if the results are not

shown to be insignificant, then they must be significant. Unfortunately, this assumption is not logically sound. Equally unfortunate is the fact that research literature uses the positive language of 'association', avoiding the clumsy but truer language of 'not unassociated'.

The cut-off point for significance at five per cent is much like an 'industry standard' in research. The widespread acceptance of this standard means that researchers who use it find it easier to apply for funding, get their paper published in a journal, and gain approval from their peers. But we need to remember that it is not an expression of objective truth, nor a demonstration of proof (and these are common misinterpretations); a study could just as well use 10 per cent or 1 per cent ($p < 0.1$ and $p < 0.01$ respectively); the appropriateness of the p value selected depends on whether the benefits and risks of the treatment being researched are major or trivial. If there is a major risk involved in a treatment (ie death or severe and disabling side effects) we may be willing to accept a lower probability of association between the intervention and the negative outcome, and hence a larger p value.

The finding of statistical significance is often used inappropriately, especially in trials where there is only a very small difference between the placebo and treatment groups. In a clinical trial where the placebo group has very few positive responders and the treatment group has a majority of positive responders, the results are obvious, and we do not need a statistical analysis to tell us whether or not the treatment is working. However, when the placebo group has around 40 per cent of subjects responding (measured as remission or significant improvement), and the active treatment group does only marginally better, as in most published trials on antidepressants^{10,11} then the finding of a 'statistically significant' difference between the two groups may be misleading. The p value may easily be nudged over the line of significance by (unethically) adding more subjects to each group and extending the duration of the trial, as for mathematical reasons this will cause p to decrease.^{8,9} Apart from avoiding the issue of clinical significance (ie whether or not we can expect to see real benefits for patients in the clinic), attention is diverted away from issues that may need to be examined more closely, such as the validity of the diagnosis and the normal course of the illness (eg whether or not many or most patients tend to get better anyway with time).

Application of statistics within clinical trials

Clinical trials are designed to assess the likelihood that a particular health outcome will occur within a given population when a particular therapeutic intervention is

applied. There will always be some degree of uncertainty, but by applying statistical methods we endeavour to minimise this uncertainty to acceptable levels. There are three important variables in this process, each having a critical influence on the others: the size of the effect (that we are hoping to achieve or avoid), the size of our sample (ie the number of participants in a trial) and the cut-off point that we choose for deciding whether the relationship is significant.^{8,9} In this way, statistical studies can never be completely objective. Preconceived ideas (biases) are incorporated into them in the form of assumptions that are used to set the statistical parameters (eg how many people are needed in the study, how the treatment effects are measured or what size effect is clinically relevant). These assumptions are derived from decisions that have been made at each critical step in the design of the study, and are based on the researchers’ values, ideas about an illness and what level of risk or benefit is deemed to be acceptable or desirable. This is why a good study report should include a discussion of the known assumptions and how the trial results may have been influenced by them.

Now, what exactly does the outcome of a clinical trial tell us? The common misconception is that the outcome of a trial quantifies a treatment’s effects – therapeutic and/or adverse. When the outcome of a clinical trial is found to be statistically significant, the effects of the treatment (ie the ‘active intervention’) are deemed not likely due to random chance. A low p value (below five per cent or another nominated cut-off point for significance) tell us that we have sufficient evidence to reject the ‘null hypothesis’ - the proposition that the intervention is doing nothing. This process is referred to as ‘null hypothesis significance testing’ (NHST). Trial results with a low p value (ie below 0.05) indicate that:

- The observed differences between the two study groups are not due to random chance.
- We are more than 95 per cent sure of the above statement.
- There is less than a 5 per cent chance that the original statement is wrong.

This is what the ‘evidence’ gained from a clinical trial is telling us: it is highly likely that the active treatment is not doing nothing.

The why and the whence of null hypothesis significance testing

Why are clinical trials conducted in this way? Historically, clinical trial methodology was developed from the methods used in epidemiology, where the most practical way to test for a significant factor in a disease outbreak is to first analyse the data to see whether or not the factor under consideration has effects that are not simply due to chance. Obviously, there are an almost unlimited number of factors at play within any given scenario, and therefore it is more likely that an incorrect one rather than a correct one will be chosen. Hence, the need for an efficient, low-cost method that does not require large resources while repeated tests are conducted to find something that may, in fact, be having an influence (or rather ‘not having no influence’) on the outbreak, spread and severity of the disease being studied. Moreover, this methodology is most suitable for assessing scenarios in which there are a number of different

factors (‘variables’) at play, some of which may only be having a small, but significant, effect. Thus, epidemiological methods are designed to detect variables with different degrees of influence, ranging from quite small

The common misconception is that the outcome of a trial quantifies a treatment’s effects – therapeutic and/or adverse.

to quite large.^{12,13}

In this way, an epidemiological study begins with the proposal, or ‘hypothesis’ that a particular factor is having no influence on a disease – the null hypothesis. The data are collected and analysed in order to accept or reject the null hypothesis according to the value of p. The hypothesis is rejected when p is less than 0.05 and accepted when greater than 0.05. This is how the p value was originally used; it is the outcome of NHST. When applying this methodology to assess the effects of a single medical intervention, there are several critical areas where errors can occur, and these will be explored in Part 3 of this article series.

Ground zero: the placebo group

In a trial that compares an active treatment with a placebo, when the p value is less than 0.05 the difference between the effects of the two interventions being compared (eg between active and placebo) is deemed not to be due to random chance and that the treatment is not doing nothing. The placebo arm of a study provides the reference that defines what ‘doing nothing’ means in measurable terms. It is important to bear this in mind when we come across examples such as the following:

In a study comparing St John’s Wort and citalopram (a selective serotonin reuptake inhibitor) in the treatment of ‘minor depressive disorder’, (ie mild depression), both treatments at first sight appeared to be moderately effective, with just under 50 per cent mean improvement in both treatment groups, measured in terms of reduction in symptoms, improvement in quality of life and improvement in psychological state. However, a third placebo group ended up showing the best response, which was a little over 50 per cent improvement. By way of explanation, the authors reported that ‘these findings were clearly due to the consistently high placebo response rate on all outcome measures’.¹⁴

Through their explanation, the authors above appear to be trying to justify the anomalous results for citalopram, which has been shown (in other exclusively drug company funded trials) to be ‘effective’. This line of reasoning contains two major flaws. One is that by definition the placebo response is ground zero: the clinical response in the placebo group represents the zero setting that is to be used in order to accurately measure the results (if any) of the active treatment. This means that whatever response is found in the placebo group, the only way it can be legitimately used is to subtract it from the response of the active treatment group, and thus obtain a measure of the actual response to the active treatment. The second error is that clinical trial protocol demands that you ignore the ‘within group’ responses, ie the difference between the measures taken at the beginning of the trial and those taken at the end of the trial within a particular group. The reason for this is that in a self-limiting disease, where patients tend to get better without any treatment, both groups will improve over the course of the trial. If the results of the placebo group are ignored, the ‘treatment’ may falsely appear to be effective.


Confounding factors

The assessment of confounding factors in a clinical trial is a critical design component, and constitutes an important step in removing potential sources of bias. In a well-designed clinical trial the potential confounding factors are taken into account, so that the study groups are equally matched, or ‘controlled’. A well-reported clinical trial

should discuss how potential confounding factors were prevented from influencing the trial outcomes, together with a brief discussion of other possible factors. Important confounding factors include age, gender, severity of illness, duration of illness, previous treatments (and how long ago they were stopped before entering the trial), current medications, socio-economic factors, education level, patient expectations, attitudes to the illness (eg perceived benefits from being ill), and the validity of the diagnosis (ie do the subjects all have the same disease?).^{15,16}

As a previous US Secretary of Defense once publicly explained: ‘There are known knowns. There are things we know we know. We also know there are known unknowns. That is to say, we know there are some things we do not know. But there are also unknown unknowns, the ones we don’t know we don’t know.’ This speaks to perhaps the most important consideration of all - one which is the foundation upon which scientific knowledge is built: acknowledgement of ignorance. In spite of careful assessment of the potential confounding factors in a clinical trial, in which every effort has been made to ensure that these factors are evenly matched between groups, it is always possible that some other yet-to-be-discovered factors have played a decisive role in the trial outcomes. This is yet another reason to be cautious about accepting the results of a single clinical trial.

Concluding remarks

The above discussion outlines some of the important critical issues related to statistical methodology and clinical trial design. Further elaboration on these issues is required before we can realistically determine whether statistics are being used appropriately, whether a study is well or poorly designed and how best to interpret the information that is provided. This will be covered in Part 3, the final article in this series, which also contains a summary checklist that can be used as an aid to assessing information quality when reading a clinical trial report. 

Tony Reid is a graduate of the Sydney Institute of Traditional Chinese Medicine and holds master’s degrees in acupuncture and TCM from the University of Western Sydney. He has contributed to TCM as a clinician, lecturer, administrator, course designer and industry consultant since the early 1980s.

References

1. Evans, S. (2010). Common Statistical Concerns in Clinical Trials, *J Exp Stroke Transl Med*, 3(1)1-7.
2. Krousel-Wood, M., Chambers, R., Muntner, P. (2006). Clinicians’ guide to statistics for medical practice and research: part I, *The Ochsner J*, 6(2), 68–83.
3. Krousel-Wood, M., Chambers, R., Muntner, P. (2006). Clinicians’ guide to statistics for medical practice and research: part II, *The Ochsner J*, 7(1), 3–7.
4. Hill, A. B. (1965). The Environment and Disease. Association of Causation?, *Proceedings of the Royal Society of Medicine*, 58, 295–300.
5. edak, K., Bernal, A., Capshaw, Z., et al. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology, *Emerg Themes Epidemiol*, 12:14.
6. Ioannidis, JP. (2005). Why most published research findings are false, *PLoS Med*, 2(8):e124.
7. Horowitz, A., Wakefield, J. (2007). *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*. New York: Oxford University Press.
8. Thiese, M., Ronna, B., Ott, U., (2016). P value interpretations and considerations, *J Thorac Dis*, 8(9):E928-E931.
9. Sullivan, G., Feinn, R., (2012). Using Effect Size-or Why the P Value Is Not Enough, *J Grad Med Educ*, 4(3):279-82.
10. ic Fallacies: Implications for the Evaluation of Antidepressants’ Efficacy and Harm, *Front Psychiatry*, 8:275.
11. Jakobsen, J., Katakam, K., Schou, A. et al. (2017). Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and Trial Sequential Analysis, *BMC psychiatry*, 17(1), 58.
12. Sur, R., Dahm, P., (2011). History of evidence-based medicine, *Indian J Urol*, 27(4):487-9.
13. Zimmerman, A., (2013). Evidence-Based Medicine: A Short History of a Modern Medical Movement, *Virtual Mentor*, 15(1):71-76.
14. Rapaport, M., Nierenberg, A., Howland, R. et al. (2011). The treatment of minor depression with St. John’s Wort or citalopram: failure to show benefit over placebo, *J Psychiatr Res*, 45(7), 931–941.
15. Skelly, A., Dettori, J., Brodt, E. (2012). Assessing bias: the importance of considering confounding, *Evid Based Spine Care J*, 3(1):9-12.
16. Lambert J (2011). Statistics in brief: how to assess bias in clinical studies?, *Clin Orthop Relat Res*, 469(6):1794-6.

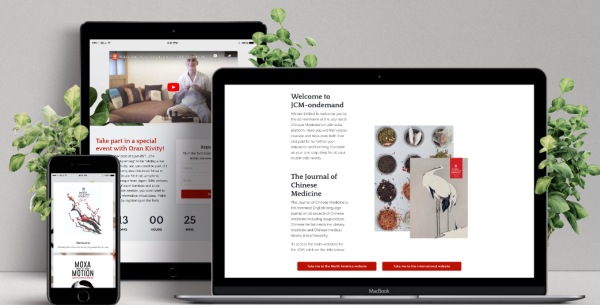


Contribute to The Journal of Chinese Medicine

Do you have special experience or perspective in treating specific conditions that you would like to share with the profession? We are always looking for clinically-based articles - especially for commonly-encountered everyday problems.

Articles don't have to be long, and our highly skilled editorial team can help you best convey your knowledge to our worldwide community of practitioners and students.

See our author guidelines here:
www.jcm.co.uk/news/author-guidelines



The JCM-ondemand multimedia platform
The home of all our multimedia content

COURSES, RESOURCES,
EDUCATION AND
EVENTS

JCM-ONDEMAND.COM