Finding Your Way Through the Forest – A TCM Practitioner's Guide to Evaluating Research: Part 3

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 3 of this series reviews the most important criteria for assessing research papers, so that practitioners of traditional Chinese medicine and other related disciplines are able to identify errors in medical research in order to inform treatment and assist their patients. A summary checklist is included to guide readers when assessing clinical research papers.

Keywords
Evidence based medicine, research, clinical trials, TCM, Chinese medicine, RCT, bias, statistical methodology, confidence interval

Introduction

The application of statistical methodology to medical studies was first championed by Sir Austin Bradford Hill, who in the early 1960s introduced randomised controlled trials into clinical medicine.[1] He also repeatedly warned of their limitations and potential for misuse: problems that are still current some 50 years down the track.[2] Hill cautioned against the overemphasis of statistical significance as well as neglecting to attend to the possibility of undetected errors. Unfortunately, this advice continues to be pertinent today when there are still too many unnecessary 'false positive and non-replicable results' in clinical research.[3]

A renowned professor of statistics published a seminal paper in 2005, entitled 'Why Most Published Research Findings are False'.[4] More than ten years on, several researchers have noted that statistical errors are still 'all too common'.[3,5,6] Mills' famous statement from 1993 still rings true: 'If you torture the data long enough, they will tell you what you want to hear'.[7] You do not need to be well trained in the minutiae of statistics to spot the major problem areas if you know where to look. The following is a summary of the important criteria for assessing statistical accuracy, or otherwise, in a medical paper.[8,9]

Lack of homogeneity between study groups

In a clinical trial the effects seen in a group of subjects receiving a treatment are compared with those seen in a similar group who are given a placebo and/or a no treatment group. You should not trust the 'random allocation' of subjects, even when the authors use p-values that appear to confirm that there are no significant differences in confounding factors between the study groups (e.g. age, severity of illness or previous treatments). The logic behind null hypothesis significance testing (NHST - see Part 2 of this series of articles) dictates that this test is neither accurate nor appropriate for assessing whether or not there are significant differences between groups in a clinical trial. We should remember that 'p' only tells us the probability of obtaining this result if the null hypothesis (that there is no difference between the subgroups) were true; it does not tell us that the null hypothesis is true, given the data. Moreover, there are rarely sufficient people in each of the various subgroups (with a particular confounding factor) to provide sufficient statistical power to generate a meaningful p-value.[6] Thus, although the subjects may have been allocated to their respective groups by a computer-generated random sequence, it is still possible that the groups are unevenly matched in terms of confounding factors. This should have been checked by the researchers before commencing the trial and any discrepancies noted, together with the strategies adopted to mitigate them. If the authors have provided p-values when comparing the potentially confounding characteristics of two (or more) study groups, this is a meaningless measure, and you will need to investigate further.

Good reporting requires that all relevant characteristics of people within each trial group should be comparable; the best way to show this is in the form of a table. If this information is provided, you should check that the subgroups are, in fact, equal and homogenous, and that no important factors are omitted. Factors that could potentially affect a trial's outcomes include age, gender, severity of illness, duration of illness, previous treatments (and how long since they were stopped before the trial), current medications, socio-economic factors, education level, patient expectations, attitudes to the illness (e.g. perceived benefits from being ill), and the validity of the diagnosis (i.e. do the subjects all have the same disease?). In the discussion section towards the end of a trial report, the influence of known potential confounding factors should be acknowledged, and other possible confounders briefly explored.

Good reporting requires that all relevant characteristics of people within each trial group should be comparable.

A critical issue when assessing the homogeneity between groups is to look at pre-trial measurements of the condition (generally relating to the degree of severity) to ensure that not only is the mean of these measurements similar in each of the study groups, but that the way in which the individual measurements are spread around the mean is also similar. If we take a trial of a treatment for depression as an example, we need to look at the measure of severity of depression (the 'main outcome variable') to ensure that the study groups are, in fact, comparable. Even when the mean values are the same (or very close) for each group at the beginning of the trial, if there is a markedly different spread of values on either side of the mean in one group, this can have a profound influence on the trial outcomes. A group of subjects who are either severely depressed or only mildly depressed (i.e. measurements of severity are spread very widely around the mean) may show markedly different responses to a treatment than a more homogenous group with scores that cluster very closely around the same mean.

Let us say we have found a successful trial on a non-drug treatment for mild-to-moderate depression of relatively short duration in which all subjects were carefully screened to exclude those with severe and/or long-term depression (in whom this treatment is known to be ineffective). The positive result may not be reproducible if a subsequent trial were to contain a number of long-term severely depressed subjects in the treatment group. In our 'doomed-to-failure' trial, the placebo arm may still consist of exclusively short-term mild-to-moderate cases, but the treatment arm may contain a mix of severe and very mild cases such that the mean values for severity are comparable. The increased number of failures in the treatment arm of the second trial may readily produce a negative outcome. Therefore, in addition to ensuring that the mean pre-trial measurements of the main outcome variable (and any key confounding factor, such as duration of illness in this example) are comparable in each arm of the trial, we need to look at the standard deviations (SD) as well, because this is a measure of the spread of these measurements. Both the mean and the SD values should be similar in each study group at the beginning of a clinical trial.

Inappropriate statistical tests to analyse the data

There are several different statistical tests for significance that are used in biomedical research, such as the z-test, the Student's t-test and the chi-square test, each one suited to a particular type of study and the nature of the variables that are being tested. We need to be certain that the appropriate one has been used, as in some cases the statistical test may be an incorrect match for the type of data that has been collected. While this requires a level of expertise that is beyond most of us, a study should at least report the particular statistical test and provide intelligible reasons for the choice. If this is not the case, we should be cautious in accepting the conclusions, especially when the size of the treatment effect is very small and the statistical analyses appear to be very complex. A basic rule of thumb is that if a treatment really works in clinical practice (i.e. provides clinically significant results), or if one treatment really is better than another, it should be obvious when looking at the data. The statistical analysis provides a likely range of values for the results and may also be used to compare different subgroups (e.g. older versus younger patients, males versus females).

Placing too much importance on the p-value

As discussed in Part 2 of this article series, the p-value, which is an expression of statistical significance, does not tell us what we really want to know. It only tells us the probability of obtaining these particular results if the null hypothesis were true; a low p-value tells us that the null hypothesis is unlikely to be true.[10] Thus, it begins with the assumption that there is no significant difference in a particular factor (i.e. the expected outcome of an intervention) between the two groups being compared. When p is below the generally accepted level, i.e. less than 0.05, or below five per cent, then we reject the null hypothesis. In doing this we can infer that the therapy is most likely producing effects that are different to those occurring in the placebo group - this is the only conclusion that can be drawn from a low p-value, no matter how low. We should be extremely cautious about accepting any other type of comments made by the trial authors in light of p being less than 0.05.

The p-value is mathematically derived from formulas that are based upon three critical factors: the size of the treatment effect, the sample size and the variability within the sample. Thus, 'p', the measure of statistical significance, is a function of these same factors, and it varies (either directly or inversely) in proportion to variations in these factors. The p-value alone does not provide information about any one of these factors. In particular, the p-value does not give an indication of the size of the treatment effect.[11,12] As discussed above, the lower the p-value, the more confident we can be that the treatment is not doing nothing. We may then infer that it is likely that our treatment is largely responsible for the observed clinical effects (all else being equal), but we still need some way to quantify these effects. This information is provided by the confidence interval (CI).

There is a critical distinction between statistical significance (we strongly suspect that our treatment is having some sort of effect) and clinical significance (the effect is not trivial and will make a real difference to the health and well-being of patients - and therefore also their caregivers). Here again, the p-value does not provide the necessary information. The effects of the treatment need to be quantified, so that we can know that they are not trivial. This information should always accompany the p-values in a trial report and is given in the form of the confidence interval.[12]

Misinterpreting the confidence interval

The CI relating to the size of the treatment effect should always be given along with the p-values in a trial report. We need to remember that the 'effect size' that we are talking about is the mean (or average) effect size. The individual effect sizes of the participants in the trial are generally clustered around this mean in a normal distribution pattern. The CI, generally expressed as the '95 per cent CI', is the range of values within which there is a 95 per cent possibility that the true mean of the treatment effect lies when applied to the whole population of interest. The 95 per cent CI is generally a fairly narrow range of values. However, the significance of the 95 per cent CI is often misunderstood, as it is often described as the range in which the 'true value', is most likely to be. Unfortunately, this 'true value' does not refer to the actual size of the treatment effect that we are most likely to see in this population. The 'true value' refers to the mean treatment effect size that we would expect to see in the general population of patients with the condition for which the treatment protocol is being tested. This is a critical distinction: the CI does not signify that if we gave this treatment to 20 people, at least 19 of them will have a clinical response that lies within the CI range. The CI only tells us the range within which it is most likely that the population mean (or 'true mean') may occur - with most of the real values (i.e. the actual size of the treatment effect in an individual patient) falling on either side of this mean value.[12,13]

There is a critical distinction between statistical significance (we strongly suspect that our treatment is having some sort of effect) and clinical significance (the effect is not trivial and will make a real difference to the health and well-being of patients).

This concept raises some important issues when applied to clinical practice. Say, for example, that we are reviewing a clinical trial in which the outcomes of a treatment (measured as 'effect size') above and including the mean were clinically significant, while outcomes below the mean were measurable but not clinically significant. If we based our clinical expectations on the mean effect size of the treatment group in the trial, we would be confident of achieving clinically significant outcomes in more than 50 per cent of patients. However, if the 95 per cent CI range is large, i.e. the mean treatment effect for the entire population falls within a fairly wide range, things do not look quite so good. If we take the worst-case scenario (i.e. that the lowest value of the CI, as all values are equally likely), we find that less than 50 per cent would have a favourable outcome and more than 50 per cent would fail to experience clinically meaningful results. These considerations may have a major influence on whether or not we choose to give this treatment to our patients. However, this information would be concealed from us if we rely solely on the mean treatment effect that was found in the trial by misinterpreting the CI in the manner described above.

It may not be easy to make the necessary calculations, as studies in which the 95 per cent CI shows the active treatment in a less favourable light, may not provide the relevant data, especially not in the abstract. Another rule of thumb: the authors of a study should always clearly delineate the response level above which the clinical results are meaningful, as well as provide standard deviation values and confidence intervals; if not, it is likely that the authors wish to conceal something. This is an area in which those who report research findings are able to 'creatively' present the data. Obviously, if researchers choose to report trial results as if the upper CI were the true mean for the population being studied, things will look very much better than the scenario based on the lower CI (as the true mean). Therefore, in a clinical trial where only the treatment outcomes above and including the mean were clinically significant, as in the example above, the best that can be said is that 'further research is warranted'.

Poor handling of dropouts and outliers

Inevitably, some subjects will not correctly follow the designated protocol in a clinical trial, or fail to continue up until the end of the trial (e.g. because of intolerable side effects or impatience for results) – just as some patients that we see in clinic fail to continue with a course of treatment or never come back after the first consultation. Additionally, some subjects in a trial, as in our clinics, do not follow instructions and fail to take their medicine regularly. Subjects like this are referred to as 'dropouts' and researchers are often tempted to exclude them from the analysis of the trial results.

Another critical sub-group of subjects are those - both in the treatment and placebo arms of the trial - who experience effects that are considerably outside the usual responses, ranging from no effect at all to a dramatic and rapid effect. How are these people to be treated in the trial results? Do they represent random 'freak' events that crop up from time to time within the general population, and therefore should be excluded from the results? Do they belong to the five per cent of outliers that we would expect to find in any normally distributed variable, and therefore must be included? If these subjects are part of that outlying five per cent, we can expect that within the entire population there will be an equal balance of extreme clinical results on either side of the mean. Additionally, extreme treatment effects may occur in considerably more than five per cent of the general population for reasons related either to the intervention itself or the person receiving the treatment.

In light of these considerations, some researchers may be inclined to completely exclude outliers and dropouts from the final analysis of the trial results. However, in real world clinical scenarios where practitioners see only a small segment of the total population, the anomalous outcomes seen in a trial may indeed reflect the possible outcomes seen in an individual clinical practice. Therefore, a good trial should include all dropouts and outliers in the final analysis, as this reflects real-life and helps provide a realistic assessment of the intervention being studied. This is referred to as intention-to-treat (ITT) analysis. Always check that the numbers of subjects analysed at the end of a trial are the same as the numbers enrolled at the beginning. Generally, dropouts should be counted as 'treatment failed'. If a large number of dropouts occur in the treatment arm, the treatment may be causing unpleasant side effects and/or is ineffective. For the same reasons, subjects with effects that are extreme and unusual should also be included in the end of trial analysis, or the researchers should provide valid reasons why they were excluded.

Within-group comparisons

When researchers report on a comparison between the baseline (beginning of trial) measurements and the end of trial measurements within the one group, this is called a 'within-group paired test'. Despite the technical jargon, this is not a valid statistical test for clinical trials. Even when this comparison shows a clinically significant improvement, it is completely irrelevant. There could have been other factors, both known and unknown, that caused a similar improvement within the placebo group, thus neutralising the apparent effects of the active treatment. The only valid comparison that can be made in a clinical trial is between the placebo and the treatment groups, as this is the only way to gauge the true effect of the treatment. Within-group comparisons, if given or discussed, should always raise suspicion, and should not influence assessment of the trial results. Comparison between the active treatment and placebo group is a fundamental principle of clinical trial methodology. Unfortunately, this is sometimes ignored when researchers or sponsors want to hide the facts and give a positive spin to the trial results.

P-value hacking

The main way in which data are 'tortured until they tell you what you want to hear'[7] is through p-value 'hacking'. This involves using the same set of data to test out one or more new hypotheses, especially when evidence in support of the original one has failed to reach statistical significance. Let us suppose that we have a clinical trial on a treatment for depression, in which the average response of the active treatment group is only marginally better than that of the placebo group. However, there are quite a few subjects in the active treatment group with very good clinical outcomes, far exceeding the best ones in the placebo group; unfortunately, there are also a number of subjects in the active group with minimal or no improvement - hence the low mean response within the active group. The logical next step would be to look for common characteristics in the subgroup with a good response and compare them with similar patients in the placebo group. This is a sub-group analysis and, strictly speaking, should not be part of the legitimate results of the original trial. Every time a different subgroup analysis is conducted, the p-value becomes further and further diluted, such that the 'statistical significance' of successive analyses becomes less and less meaningful. The only valid use for this observation is to develop a new hypothesis and conduct another trial to test it; in the above example the new hypothesis could be that certain characteristics in patients with depression lead to consistently good outcomes when using the treatment protocol tested in the original trial. Therefore, any new hypotheses that are formulated after the trial data have been gathered and analysed should not be given much weight. The main reasons for this are:

- There are usually too few subjects with the specific sub-group characteristics in the treatment and placebo groups for a meaningful comparison
- The placebo and treatment group subjects may not be matched in terms of other important characteristics
- Mathematically, the more hypotheses you try to prove with a single set of data, the more likely you are to have erroneous findings.

'P-value hacking' is an attempt by researchers to find something that is statistically significant in the face of a non-significant finding as the main trial outcome...

Often used as a means to get a research paper published, 'p-value hacking' is an attempt by researchers to find something that is statistically significant in the face of a non-significant finding as the main trial outcome. The data that have been collected are analysed in different ways, looking at various subsets and (inevitably) finding one or more that provide a statistically significant result, often without any real clinical significance. Sadly, this bogus statistically significant result is sometimes reported as if it were the main finding of the study, possibly appearing in the title or at least in the abstract. The converse may also occur, where a subgroup that was part of the initial trial protocol is conveniently omitted when the results do not suit the interests of the researchers. These practices (or rather malpractices) of post hoc hypothesising are also known as 'HARKing' (hypothesising after results are known). [12,13]

Of course, an important part of analysing the data at the end of a trial is to look for patterns, both in terms of the desired effects of a treatment as well as the unwanted effects. If this leads to a new hypothesis being developed (e.g. side effects are more common in subjects who are over 60) that is a good thing. However, this new hypothesis should not be applied back to the original trial – it should only be used as the basis for future trials.

Mistakenly inferring effect size from the p-value

As noted above, the p-value is a function of the sample size: as the sample size increases, the p-value automatically decreases. This means that if the p-value is too high to give statistical significance to your test results, you just have to continue the trial, adding more and more subjects until you get to the point where the p-value is less than 0.05. In this way you can produce a 'significant' result, even though this result is exactly the same as that of the original smaller scale trial – with the same mean value, same spread of outcome measures around the mean and the same difference between the two groups.[14] Therefore, a low or very low p-value should never be interpreted as an indication of a favourable effect size. Moreover, 'statistical significance' should never be taken to mean 'clinical significance'. While different methods of measuring clinical outcomes may show a small difference in favour of the active treatment group, we always need to be sure that the net effect of the active treatment does, in fact, make an appreciable positive difference in the life of patients and carers.

Hidden sources of bias and scientific errors

In addition to mishandling of statistical methodology, there are a number of other common sources of error in clinical trials. Although CONSORT and PRISMA guidelines have been widely promulgated, inadequate or improper reporting of clinical trials, along with failures to adhere to best practice in methodology, are still common. Additionally, there are several weaknesses within the accepted clinical trial methodology that critically impact the quality of the results. The following list outlines some of the more readily detectable ones.[3,8]

Use of casual and imprecise language

The use of casual, imprecise or highly emotive language, especially in the abstract, should be a red flag for a 'spin alert'. Authors should use precise language and clearly summarise the results of a trial, giving the key numerical findings. A recent example appeared during the development of the novel mRNA vaccine by Pfizer. The title of the review paper included the descriptive word 'miracle' and the abstract contained the phrase, 'gave the world a sense of light at the end of a tunnel' while continuing on to quote the '95% efficacy' slogan (as discussed in Part 1). Including emotive messaging, such as 'hope spread worldwide' and 'a return to normality became a tantalising possibility', the paper appeared to be more concerned with the fact that 'stock markets rallied' than establishing any objective scientific data.[14]

Conclusions drawn from insufficient data.

Authors should provide sufficient data and the right kind of data to justify their conclusions. The precise values of the standard deviation (SD) and the confidence intervals (CI) should be given along with the p-value. We should also bear in mind that other comparative data, such as the odds ratio and relative risk are potentially misleading, as discussed in Part 1.

There have been instances where the abstract has contained conclusions that are at odds with the actual findings of the trial. This is fraud. Most of the large pharmaceutical companies have been found guilty of this type of crime, often on several occasions.[20,21] This is a good reason why clinical trials should not be run by those with vested interests, and when they are, all of the raw data should be made available for scrutiny.

Poor description of methods and results

When reading a description of a clinical trial, you should ensure that the methods and results are described accurately and in sufficient detail for a clear understanding of how the researchers conducted the trial and why they chose to adopt their chosen methodology. Similarly, the results should be presented in a realistic way and include a discussion of their potential application in clinical practice and any possible limitations, cautions and caveats.

Specification of main and secondary outcomes

The main outcomes should be described in sufficient detail to be unambiguous. There should also be a description of any secondary outcomes. Moreover, secondary outcomes should be proposed at the beginning of the trial and included in the trial design. If they have been added in at the end, after the results have been collected and analysed, they are to be regarded as speculation, separate from the results of the trial.

Description of adverse events

There will always be some adverse events in any trial – both in the treatment group as well as the placebo group. These should all be accurately described and recorded for comparison. Trials should always report both the benefits and the risks; the results of a trial should always include the frequency of all adverse events. Clinicians need to know the benefits as well as the risks of any treatment, so that these can be weighed against each other.

Publication bias and submission bias

It is 'well known' (but difficult to discover the full extent) that studies with negative findings rarely get published,particularly if the trial is sponsored by a large pharmaceutical company.[15,16] These trials are generally not submitted for publication and are often stopped before completion. Such trials only get to see the light of day due to freedom of information (FOI) requests, e.g. sequestered unpublished trials on SSRI 'antidepressants' have been included by researchers in updated meta-analyses, showing much smaller efficacy than is generally accepted.[17,18] One may assume - noting that the vast majority of published clinical trials report positive findings - that this practice is widespread.

It follows, therefore, that we should be suspicious if we are only able to find a single study with a positive result on a particular treatment that was published several years ago, with no other studies reported since then. Generally, we would expect other studies, perhaps larger, or with a different segment of the population to have been conducted subsequently, in the hope of gaining more positive results. If we cannot find any of these, we may be justified in presuming that the treatment has since not been found to work. Therefore, we should beware of treatments that are supported by only one study. We would expect to find that there are several studies supporting a particular treatment, conducted within a short time after publication of the initial positive one. Of course, this may not always be the case. If a treatment being tested is a non-pharmaceutical intervention that has the potential to supplant a widely used drug treatment, it may be difficult, if not impossible, for researchers to obtain the necessary funding for larger trials, e.g. lifestyle modification to manage gastric reflux.[19]

Validity of the diagnosis

The elephant in the room, particularly regarding trials on conditions such as depression or IBS, is the validity of the diagnosis. Are we looking at a single disease with the same cause in each case? Or are we taking several different disorders with different aetiologies and lumping them together because they share a major symptom, which may only be defined quite loosely? In the case of 'depression', the definition of 'major depressive disorder' has become so elastic that it now includes people who are experiencing sadness due to a loss of some kind, and who tend to get over it within a few months. This would explain the relatively high rates of 'remission' or improvement seen in the placebo groups in trials on treatments for depression. This is an area that is readily exploited by vested interests. To continue the above example, industry sponsored trials, in which the raw data have been sequestered (i.e. not published along with the trial report), may have a severe mismatch between subjects in the placebo and active treatment groups. It is possible that subjects whose depressed mood is long term may predominate in the placebo group, while those with more recent onset depressive symptoms may predominate in the active treatment group. This type of placement is likely to give an advantage to the drug treatment. Moreover, given the current definition of major depression, this type of arrangement is completely undetectable. Thus, even in the absence of any deliberate 'stacking' of the two groups, such a mismatch could also occur by chance. The same kind of thing may happen in crossover trials when patients on active treatment are changed to placebo (inevitably suffering withdrawal symptoms, which are classed as 'depression relapse') and the placebo patients who remitted are excluded from this part of the trial.[20,21]

We should be suspicious if we are only able to find a single study with a positive result for a particular treatment that was published several years ago, with no other studies reported since...

The issues surrounding diagnosis are especially pertinent for TCM-based research. Most diseases as defined by biomedicine are broken down into subgroups with different syndrome patterns in TCM. Thus, a single biomedically defined disease may arise due to different aetiologies in different patients. Because this is an essential component of diagnosis and treatment, a different approach to clinical research is required for TCM. While a detailed discussion of the issues associated with this paradigm shift is beyond the scope of this article, we should be aware of this limitation when looking at clinical research on TCM that uses a predominantly Western style research study paradigm.

Measurement of effect size: accuracy and validity

In many trials, both complete remission as well as significant improvement are bundled up together and reported as a 'positive result'. There are several issues here that require additional scrutiny. How is 'clinical remission' defined? What sort of follow up procedures are in place to provide information about whether or not the remission is maintained for a certain period, and whether or not 'remission' needs to be maintained by continuing with the therapeutic intervention, possibly indefinitely? Does the trial provide data about numbers of subjects experiencing complete remission as well as numbers of subjects with significant improvement? Is the level of 'significant improvement', as defined in the trial, the same as 'clinical improvement'? How are these measured, and what is the margin for error in these measurements?

If we look at clinical trials on treatments for depression, we find that many of these trials use the Hamilton Depression Rating Scale (HDRS). However, serious issues regarding its validity have been raised and it has been described by critics as psychometrically and conceptually flawed.[22] Moreover, when it is used, it should be administered and interpreted by a qualified and experienced psychiatrist, otherwise the likelihood of error is much greater. However, given these limitations, the appropriate definition of clinically significant improvement using the HDRS should be a 50 per cent or more decrease from the baseline score, equivalent to a 7 to 11 points reduction from baseline.[23] Unfortunately, the commonly accepted criterion in American and European trials is a reduction by 3 points from the baseline reading; even then, most trials on SSRI's fail to achieve this.[24]

The Abstract	The main content is clearly described. Purpose of the research outlined. The relevance or importance of the work clearly stated. Main outcomes given with sufficient data to support the conclusions. The language is precise and scientific, not emotive.
Methodology	Methods described in sufficient detail. Rationale behind the choice of methodology given.
Homogeneity between study groups – pre-trial assessments	Compare age, gender, severity of illness, duration of illness, previous treatments (and how long since stopping them before entering the trial), current medications, socio-economic factors and education level; if p-values are given, ignore them. The SD value for each group (reflecting how widely the pre-trial measurements are spread) should be similar, as should the two mean values.
Unknown confounding factors	Are you able to think of any potential confounding factors that have not been acknowledged by the researchers?
Statistical tests - appropriate or not	The reason for applying a particular significance test should be clearly given. The data should be self-explanatory (i.e., the treatment is obviously more effective than placebo or other treatment). If the data need to be put through a complex series of statistical tests to reveal the 'true' results of the trial, this should raise suspicion.
Null hypothesis significance testing: P-values, confidence intervals (CI) and standard deviations (SD).	The size of 'p' should not be linked in any way to the size of the treatment effect. CI and SD should always be given along with p-values. The minimum effect size for clinical significance should be clearly stated. If a study does not clearly define the response level, above which you have clinically meaningful results, does not provide SD values, or omits the CI – the authors may be trying to conceal something.
Dropouts and outliers	The study should provide an intention-to-treat (ITT) analysis. Always check that the numbers of subjects analysed at the end are the same as the numbers enrolled at the beginning of the trial. All dropouts should be counted as 'treatment failed'. Subjects with effects that are extreme and unusual should also be included in the end of trial analysis, or the researchers should provide valid reasons why they were excluded. Large numbers of dropouts in the treatment arm may mean that the treatment causes unpleasant side effects and/or is ineffective.
Within-group comparisons	If the study gives 'within-group paired test', this is invalid and generally reflects a bias or vested interest.
Specification of main and secondary outcomes	The main and secondary outcomes should be specified at the beginning of the trial. If there are secondary outcomes that have been added after the trial results have been collected, they should not be taken as results of the trial; they may only be used as the basis for a new hypothesis that is yet to be verified.
P-value hacking	Post hoc hypothesising is only useful as a rationale for having future trials involving the subgroup/s in question, not for generating additional results.
Mistakenly inferring effect size from the p-value.	P-value is a function of the sample size; as you increase your sample size, the p-value automatically decreases. 'Statistical significance' should not be taken to mean 'clinical significance'. If the trial was deliberately continued by enrolling additional subjects (so that the results could reach statistical significance), the treatment is most likely ineffective.
Use of casual and imprecise language	Methods and results should be reported in precise and scientific language. The methods should be described clearly and in sufficient detail to be critically evaluated.
Conclusions drawn from insufficient data.	You should check that the data collected during the trial actually support the conclusions given in the report or the abstract. Sometimes additional unjustifiable conclusions are given along with the correct ones, or the conclusions may be completely false. We need to be especially cautious of trials run by those with vested interests.
Description of adverse events	The adverse events in all trial groups should be clearly described, preferably in tabular form.
Publication bias and submission bias	Ideally there should exist other studies that confirm the results of a particular trial. If you can't find any, the treatment may not be effective; alternatively, it may not serve vested interests.
Validity of the diagnosis	One should always be aware of the possibility that the diagnosis is not valid (e.g. depression, irritable bowel syndrome).
Measurement or detection of the disorder being studied. Measurement of severity of the illness.	We should consider the accuracy and validity of the measurement or grading system for the disease being studied (e.g., the Hamilton rating scale for depression, especially when not administered by a psychiatrist)

Table 1: Checklist for assessing a randomised controlled trial (RCT)

Epilogue

The opening paragraph of Leo Tolstoy's novel, Anna Karenina, begins with a bold statement that has spawned several iterations of the 'Anna Karenina principle'.[25] This grand generalisation, laying claim to universal truth and placing Murphy's law into its proper context, speaks to the notion that there are only a few ways to get something right – and a seemingly unlimited number of ways to get it wrong: 'Happy families are all alike. Every unhappy family is unhappy in its own way.[26]' On reflection, it appears that the number of ways to 'get it right' or achieve a desired outcome are strictly limited, while the number of different ways to err is several orders of magnitude greater. The comforting fact is that, as we are living in a finite world, the number of mistakes that can be made should also be finite.

This review and summary of the 'popular' errors in contemporary medical research is current at the time of writing. Optimistically, the scientific community will correct them where possible or learn to make allowances for them where unavoidable. Realistically, however, we should expect to find new errors cropping up on a regular basis. Hopefully, with concerted efforts to overcome them, we will reach the end of our finite number of mistakes in the not-too-distant future.

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. Armitage, P. (1991). Obituary: Sir Austin Bradford Hill, 1897-1991, *Journal of the Royal Statistical Society*, 154(3), 482–484
2. Hill, A. B. (1966). Reflections on the Controlled Trial, *Ann. Rheum. Dis*, 25, 107-113.
3. George, B., Beasley, T., Brown, A., et al. (2016). Common scientific and statistical errors in obesity research. *Obesity (Silver Spring, Md)*, 24(4), 781–790.
4. Ioannidis JP. (2005). Why most published research findings are false. *PLoS Med*. 2(8):e124.
5. Szucs, D., Ioannidis, J. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front Hum Neurosci*, 11: 390.
6. Choi, S., Cheung, C. (2016). Don't judge a book by its cover, don't judge a study by its abstract. Common statistical errors seen in medical papers. *Anaesthesia*. 71. 10.1111/anae.13506.
7. Mills, J. (1993). Data Torturing. *NEJM*, 329:1196-99.
8. Evans, S. (2010). Common Statistical Concerns in Clinical Trials. *J Exp Stroke Transl Med*, 3(1)1-7.
9. Strasak, A., Zaman, Q., Pfeiffer, K., Göbe,l G., Ulmer, H., (2007). Statistical errors in medical research - a review of common pitfalls. *Swiss Med Wkly*, 137(3-4):44-9.
10. Panagiotakos, D.(2008). Value of p-value in biomedical research. *Open Cardiovasc Med J*, 2:97-9.
11. 1Gliner, J., Leech, N., Morgan, G. (2002) Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say?, *J Exp Educ*, 71:1, 83-92.
12. Motulsky, H. (2014). Common Misconceptions about Data Analysis and Statistics. *J Pharmacol Exp Ther*, 351:200-205.
13. Sullivan, G., Feinn, R., (2012). Using Effect Size-or Why the P-value Is Not Enough. *J Grad Med Educ*, 4(3):279-82.
14. Badiani, A., Patel,J., Ziolkowski, K., Nielsen, F. (2020). Pfizer: The miracle vaccine for COVID-19?. *Public Health Pract (Oxf)*, 1:100061.
15. Dickersin, K., Chan, S., Chalmers, T., Sacks, H., Smith, H. (1987). Publication bias and clinical trials. *Control Clin Trials*, 8(4):343-53.
16. Mitra-Majumdar, M., Kesselheim, A. (2022). Reporting bias in clinical trials: Progress toward transparency and next steps, *PLoS Med* 19(1): e1003894.
17. Joober, R., Schmitz, N., Annable, L., Boksa, P., (2012). Publication bias: what are the challenges and can they be overcome? *J Psychiatry Neurosci*, 37(3):149-52.
18. Healy, D., Le Noury, J., Wood, J. (2020). *Children of the Cure. Missing Data, Lost Lives and Antidepressants*. Samizdat Health Writer's Co-operative Inc: Toronto.
19. Randhawa, M., Mahfouz, S., Selim, N., Yar, T., Gillessen, A. (2015). An old dietary regimen as a new lifestyle change for Gastro esophageal reflux disease: A pilot study. *Pak J Pharm Sci*, 2015;28(5):1583-1586.
20. Healy, D., Healy, D. (2012). *Pharmageddon*. Berkeley: University of California Press, pp.96-128.
21. Gøtzsche, P. (2013). *Deadly Medicines and Organised Crime. How Big Pharma Has Corrupted Healthcare*, London, UK: Radcliffe Publishing, pp.47-69, 264-267, 281-287.
22. Bagby, R., Ryder, A., Schuller, D., Marshall, M. (2004). The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry*,161(12):2163-77.
23. Bobo, W., Angleró, G., Jenkins, G., et al. (2016). Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. *Hum Psychopharmacol*, 31(3):185-192.
24. 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.
25. Bornmann, L., Marx, W. (2011). The Anna Karenina Principle: A Concept for the Explanation of Success in Science. *Cornell University* (Computer Science, Digital Libraries), from <https://arxiv.org/abs/1104.0807v2> [retrieved 04.02.2022].
26. Tolstoy, L. (1878). *Anna Karenina*. Barnes & Noble Digital Library: New York.