9–11

Probability, proof, and clinical significance

Author Andrea C Skelly
Institution Spectrum Research Inc, Tacoma, WA, USA

A study reports that treatment A "provided significantly better pain relief" than treatment B. How do you know if the effect is real or due to chance? Assuming that the difference between treatment groups on the outcome is statistically significant, does this mean that your patients will have a clinically significant improvement? This article explores these questions as well as provides some additional points to consider when critically appraising and conducting research.

When a difference in an outcome (eg, pain) between exposures (eg, treatment groups) is observed, one needs to consider whether the effect is truly due to the exposure *or* if alternate explanations are possible. In other words, in order to evaluate the validity of a research study, factors that might distort the true association and/or influence its interpretation need to be carefully considered. This means evaluating the role of bias and considering the study's statistical precision.

Bias relates to *systematic* sources of error which need to be considered. (Bias will be discussed in more detail in future issues). By contrast, evaluation of statistical precision involves consideration of *random error* within the study, random error being the part of the study that cannot be predicted, ie, that part attributable to chance. In addition, one needs to consider whether a clinically meaningful improvement is represented.

Precision: power, statistics, and chance

One major consideration regarding precision of a study is the size of the study population. In general, as the number of study participants increases, random error and variability are decreased and the ability to detect a statistically significant difference between study groups (ie, the study's power) is enhanced. All research is performed on samples of subjects, which means there is always a possibility, at least in theory, that the results observed are due to chance *only* and that no true differences exist between the compared treatment groups. Statistical analysis assesses the role of "chance." Misuse, misinterpretation, and over interpretation of statistics are common in the medical literature. Thus, it is worth discussing some of the issues.

Analytical statistics assess the effects of treatment and risk factors on specific outcomes and the probability that any effects are due to chance. This evaluation/assessment relies on the testing of statistical hypotheses. The primary hypothesis that is tested, termed the null hypothesis, is that there is "no effect" or "no difference" other than that which may be expected by chance. Statistical significance depends on three inter-related factors:

1. Sample size—with larger sample sizes, statistical significance is more likely to be seen.
2. Variability in patient response or characteristics, either by chance or by nonrandom factors. The smaller the variability, the easier to demonstrate statistical significance.
3. Effect size or the magnitude of the observed effect between groups. The greater the size of the effect, the easier to demonstrate statistical significance.

Formal hypothesis testing usually involves examining the observed value for some factor compared with an expected value and includes consideration of the standard error (ie, the inherent variability) of the estimate. The process generates a "test statistic" value that is then used to determine the probability of having obtained the result by chance (often sampling error) alone. The *P* value is the *probability* that the observed difference would happen if the null hypothesis of no association was true.

Statistical tests help sort out how likely it is that the observed difference is due to chance only.

- The smaller the *P* value, the less likely that the result obtained could be due to chance if the null hypothesis was true.
- The *P* value is usually compared with an arbitrary value to evaluate statistical significance. By convention, this level of statistical significance is usually .05 and corresponds to the probability of rejecting the null hypothesis when no association really exits (called the alpha or type I error).

Failure to reject the null hypothesis does not necessarily mean that no association is present. Likewise, rejection of the null hypothesis *does not necessarily "prove"* that the association exists, nor does it mean that the relationship is causal. Consider the following:

- First of all, at the .05 level, the sampling will be off 5% of the time and by chance we may/may not observe a difference. In other words, we will be wrong 1/20 times.
- There are also times when we support the null hypothesis when it is false (called the type II or beta error). In this case we do not find a statically significant finding when there really *is* a difference.

The results are statistically significant: why isn't that enough?

While evaluation of the *P* value against this arbitrary number of .05 may provide some guidance with respect to the role of sampling variability and random error, *it should not be the sole criterion on which the value of a study or set of decisions is based*. Just because a statistical test declares the results "significant," it does *not* mean that the differences are meaningful. Why not? Simply stated;

- Statistical significance relates to how likely the observed effect is due to chance (ie, random error due to sampling) instead of a "true" difference between treatments or groups. Random error is the part of the study that cannot be predicted, ie, that part attributable to chance.
- The role of bias (eg, confounding) and its potential impact on the results need to be considered. In a poor-quality study, bias may be the primary reason the results are or are not "significant" statistically. Bias may preclude finding a true effect. This is a topic for future articles.
- Clinical significance relates to *the magnitude of the observed effect* and whether the magnitude or "effect size" is big enough to consider changes to clinical care.

To expand on this last point: An association that is statistically significant may not be biologically or clinically significant. For example, if the difference in blood pressure in one group of patients was statistically significant but the difference in terms of real numbers was 1 mm Hg; does it represent a biologically (or clinically) significant difference given the limitations and variability inherent blood pressure determination? No. Increasingly, reviewers and policy makers are looking beyond statistical significance for evidence of a "minimal clinically important difference" for commonly used outcomes measures (eg, pain, the Oswestry disability index). Factoring clinical significance into study design and power calculations and reporting results in terms of clinical as well as statistical significance will become increasingly important for evaluating efficacy.

Summary

Statistical tests help distinguish true differences (associations) from chance and result in a *P* value which is an estimation of probability that the results are due to chance. An arbitrary test threshold value (eg, usually alpha = .05) is used to distinguish results that are assumed to be due to chance from the results that are due to other factors. The bottom line:

- If the probability that the results are due to chance is less than the threshold value ($P < .05$), it is assumed the differences are due to these other factors (eg, true differences in treatment effects). However, we may be wrong 5% of the time.
- Significance testing in itself does not take into account factors which may bias study results. The possible exception to this is multivariate analysis, which is a subject for future articles.
- Sample size and random variation play an important role in whether a result is statistically significant or not and, together with expected effect size, whether the study is "powered" to detect statistical differences.
- A statistically significant result does *not* "prove" anything and *does not* establish a causal relationship between the exposure and outcome. The finding of an "association" does not mean that the association is causal in most instances. This is a topic for future articles.
- Although a result may be statistically significant, the effect size (ie, magnitude of the effect) may not be biologically or clinically important. In critiquing, designing, and reporting studies, the minimal clinically important difference is important to consider.

While statistical testing is an important part of research analysis, its limitations, uses, and misuses need to be considered in order to put results of a study in the proper context. In the next issue, we tackle the topic of confounding, an important source of potential bias which needs to be considered in all studies regardless of design.