EDUTORIAL

Significance and Limitations of the *p* Value

Keywords: Methodology, Interpretation of science, Statistics

In research authors, journals, and readers all aim to produce, publish, and consume significant (worthwhile) content. Thereby, an assumed measure of "significance" (i.e. $p < .05$) is easily mistaken as a handy indicator. However, the .05 threshold is completely arbitrary and *p* values as such are inappropriate to guide clinical or scientific decision making. *p* stands for (statistical) probability and not for (clinical) certainty; thus, it characterizes individual comparisons statistically but without clinical interpretation. *Significance* is thereby defined very narrowly and must not be confused with clinical relevance, generalisability or even meaning of findings.

Two biological observations are never identical, but will always show a natural degree of variation even if the same sample was evaluated twice under identical conditions. The main challenge lies in differentiating whether the observed difference reflects such "background noise" or a real difference (attributable, for instance, to an intervention). The *p* value is a direct measure of the *probability* that the observed difference is a simple chance finding (i.e. unreal). If this probability is very small, for instance less than 5%, then the assumption of a true difference (or treatment effect) may seem justified.

Importantly, the 5% threshold is not absolute but relies on convention only. There is no critical difference between $p = .045$ and $p = .055$: the likelihood of a chance finding differs just by 1%. At other times, a remaining uncertainty of 5% may be unacceptably high: a car with a known 5% risk of brake failure would probably not be licensed! A *p* value provides at best a crude orientation regarding the probable realness of specific group differences, but is too simplistic to explain the "big (clinical) picture".

Specifically, *p* values must *not* be mistaken as a substitute for critical appraisal in many crucial aspects.

(1) A *p* value does *not* indicate whether the described comparison was justified (e.g. whether the compared groups were comparable to begin with). This fundamental precondition must be ascertained within the study design. For instance, randomised controlled experiments approach the ideal of unbiased study group comparability best, but to a certain extent this can be emulated by stratified or confounder adjusted observational studies.

(2) A *p* value *ignores* whether the selected statistical test was appropriate. The correct choice depends on the data to be assessed, the sample size, the comparative concept, and the outcome format, all of which must be checked during critical appraisal.

(3) As elaborated, the threshold at .05 leaves significant *uncertainty*, whether the assumption of a difference (or treatment effect) is, in fact, correct (*alpha* error). The need for additional safety margins (i.e. a lower degree of uncertainty, for instance $p < .01$) depends on the clinical context. Conversely, as *p* values refer to specific samples only, a *"non-significant"* $p > .05$ does *not* exclude relevant effects of an intervention in clinical reality (so called *beta* error): *absence of proof is not proof of absence*!

(4) A *p* value depends on the sample size: the larger the sample, the smaller the associated *p* value and the higher the risk of "accidental" significance at the 5% threshold. Remember, *p* values do *not* reflect the clinical relevance of a finding, even if the underlying difference is real. A clinically modest treatment effect may appear "significant" when tested in a large enough ("overpowered") sample. If, for instance, a trial reported a real antihypertensive drug effect ($p < .001$),[1] the clinical decision whether to expose your patient to any potential adverse effects (for the benefit of a diastolic pressure reduction by 4.4 mmHg at 8 weeks) should not be driven by the *p* value. Clinical relevance must be appraised by appropriate measures such as *effect size* (e.g. relative risk, absolute difference or number needed to treat) with estimated precision (i.e. confidence intervals). The latter represents an important alternative for the assessment of statistical (and clinical) significance. Conversely, small ("underpowered") study samples must not be used to dismiss treatment effects (see (3)): a power calculation is always required for adequate appraisal.

(5) A *p* value does *not* indicate whether the study design was predefined or the analysis plan adopted before data inspection. Therefore, *p* values ignore biased selection of patients or study periods just as they ignore, statistical fishing expeditions (i.e. multiple hypothesis testing). Consequently, explorations should always be validated in hypothesis driven investigations in different study samples.

(6) A *p* value, as such, *never* indicates causality. Other criteria including chronological sequence, biological plausibility, and exclusion of confounding effects must be met before a causal relationship may be assumed.

And (7), a *p* value refers to summary statistics of specific study samples *only*. The application of study findings to individual patients is only justified after appraisal of their external validity (i.e. generalisability).

Clearly, *p* values represent a precious first aid for orientation. However, they must be carefully interpreted against study design, sample size, comparability of study groups, and appropriateness of statistical tests, and be pondered against clinical significance. Categorisation (e.g. $p < .05$, $p =$ n.s., etc.) obscures the interpretations of this continuous measure and is unacceptable. At any rate, comprehensive appraisal of scientific information must go beyond a single indicator. It is the responsibility of anyone dealing with summary statistics to assure that study question and design, statistical approach, and presentation of results are sound before accepting or dismissing reported findings. A *p* value $< .05$ may be significant *statistically*, but never proves clinical significance.

REFERENCE

1 Giles TD, Weber MA, Basile J, Gradman AH, Bharucha DB, Chen W, et al. NAC-MD-01 Study Investigators. Efficacy and safety of nebivolol and valsartan as fixed-dose combination in hypertension: a randomised, multicentre study. *Lancet* 2014;**383**:1889–98.

F. Dick*
Department of Vascular Surgery, Kantonsspital St. Gallen, 9007 St. Gallen, Switzerland
Department of Cardiovascular Surgery, Swiss Cardiovascular Centre, University Hospital of Bern and University of Bern, 3010 Bern, Switzerland

H. Tevaearai
Department of Cardiovascular Surgery, Swiss Cardiovascular Centre, University Hospital of Bern and University of Bern, 3010 Bern, Switzerland

*Corresponding author. Kantonsspital St. Gallen, Rorschacher Strasse 95, CH-9007 St. Gallen, Switzerland.
Email-address: florian.dick@kssg.ch (F. Dick)

http://dx.doi.org/10.1016/j.ejvs.2015.07.026