

In recent years there has been debate on the merits and pitfalls of reporting p-values, and more generally the hypothesis testing framework that supports published research (1-10). The unrest reflects longstanding issues at the root of statistics and a response to widespread concerns about the lack of reproducibility of many scientific findings (11-14). How should those of us actively engaged in clinical research react to all this? A recent editorial in *Nature* suggesting that we should go as far as to *retire* statistical significance altogether (15).

The Research Methods Group and the Editorial Board of the Journal have come together and present a discussion of the issues and the expectations of authors submitting to our Journal.

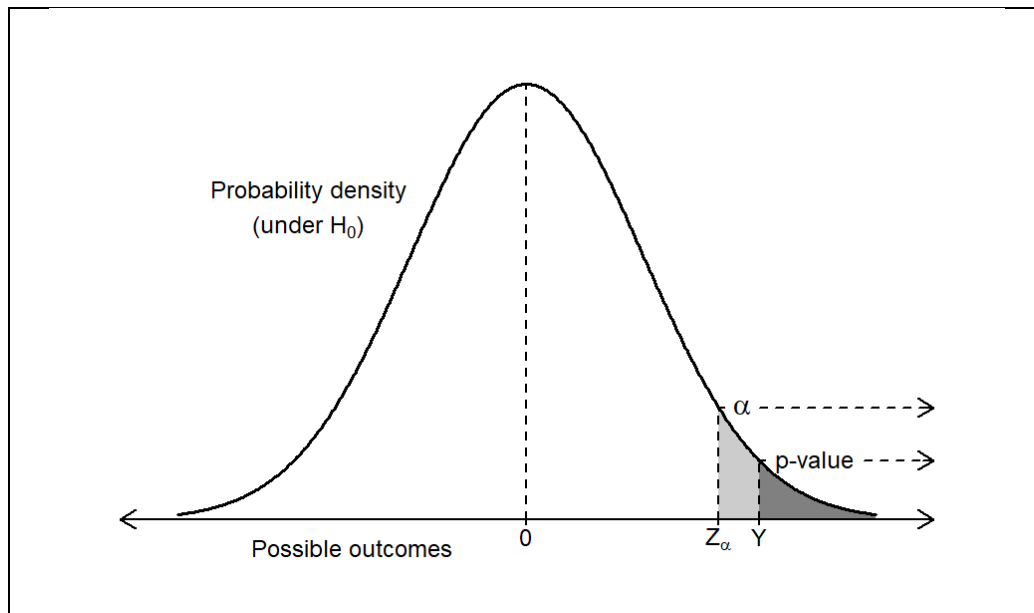
The issues are deceptively simple. The methodology that we now refer to as ‘null hypothesis significance testing’ originated in the early 20th century with the ideas of Fisher and Neyman and Pearson (16, 17). What we now think of as null hypothesis significance testing is a rather inconsistent mixture of ideas from both Fisherian and frequentist (Neyman-Pearson) approaches, that was much disputed between the parties at the time and for many years after on philosophical grounds (18, 19).

What is a p-value? The technical bit

The cornerstone of null hypothesis significance testing is the *p-value*. The *p-value* is the probability, under the *null hypothesis* (i.e. no effect or no difference), of obtaining a result as extreme or more extreme than that observed. The *p-value* could be described informally as an objective measure of the discrepancy between the observed data and the *null hypothesis*.

Consider a study that aims to compare two groups. Our *null hypothesis*, which we state at the start of the study, might be that there is no difference in the group means. The study outcome (*test statistic*) measures the difference in means between the groups. Figure 1 shows the *probability density curve* of every possible outcome of the study, under the *null hypothesis* (H_0); i.e. assuming the *null hypothesis* is true.

Figure 1. Probability density of study outcomes under the hypothesis (H_0), with significance level α , critical value Z_α , observed test statistic Y and p-value indicated



The area under the *probability density curve* in Figure 1 is such that it sums to one; i.e. every possible outcome must lie somewhere in the distribution. The larger the value of the *probability density* the more likely the observation. Under the *null hypothesis* the most likely outcome is zero (no difference between groups), and the least likely outcomes are those at the extreme ends of the density curve.

The *type I* or *false positive error rate* (α), which is set at the start of the study, is the probability (*alpha level*) at which a result is declared to be significant; the shaded area of the curve to the right of the *critical value* (Z_α). This is often called the *rejection region* (the *null hypothesis* is rejected if the *test statistic* falls in this region) and is conventionally set to be 0.05 (5% level). The *p-value* is the probability of getting the observed test statistic (Y), or more extreme, from the study data if the *null hypothesis* were true; i.e. the dark shaded tail error of the curve. If the *p-value* is less than α then the result is declared significant and the *null hypothesis* is *rejected*, at level α .

Guidance on understanding and using p-values

In response to ongoing debate between and amongst scientists and statisticians, the Board of Directors of the American Statistical Association (ASA) published the following six principles in 2016, accompanied by statements from leading statisticians, to help non-statisticians understand statistical significance and p-values (20, 21):

- P1. *p-values can indicate how incompatible the data are with a specified statistical model*
- P2. *p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone*
- P3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold*
- P4. *Proper inference requires full reporting and transparency*
- P5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result*

P6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis*

Some of these statements follow immediately from the definition of the p-value, while others are part of the bedrock of statistical textbooks and teaching. It is worth emphasising the importance of P3 – “*Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold*” - and recall that the p-value was by derivation (and definition) proposed as an informal index to quantify the discrepancy between observed data and the null hypothesis. It was not intended, and should not be used in place of sound scientific reasoning. The ASA statement concludes with an elegant sentence summarising this, that we would all do well to remember; “*no single index should substitute for scientific reasoning*”. However, despite highlighting numerous issues with misinterpretation and misuse of p-values, the ASA statement failed to make explicit recommendations about alternative methodologies. This was largely because there is no consensus amongst statisticians regarding an appropriate replacement. It is now three years since publication, and it is not clear if the ASA’s statement has had any effect on reporting or practice (22).

Multiple testing

One of the main issues of concern, which was not addressed by the ASA statement, is that of *multiplicity* or the ‘*multiple testing problem*’. Stated simply, if we undertake multiple (independent) statistical tests (with each providing a p-value), then it becomes increasingly likely that at least one of the tests will lead us to reject the null hypothesis. If the null hypothesis is true (i.e. there is no true effect) and we are testing at the conventional 5% level, then 5 incorrect rejections (false positives) would be expected for every 100 tests. Analysing multiple outcome variables but only deciding which to report after the analysis has been completed has major consequences for interpretation of medical research findings, even when researchers do not set out to deliberately deceive their audience (23). There are many widely-used multiple comparison procedures that aim to control the overall false positive rate (24); i.e. many will be familiar with the Bonferroni procedure. These procedures typically adjust the *alpha level* of individual tests, as a means to control the overall false positive rate across all the tests, thus making it less likely that any individual null hypothesis is rejected. However, there is a price to pay for this kind of adjustment: reducing the false positive rate for null associations inevitably means reducing the power to detect true associations (if they are present), and so increases the chance of getting false negative results. As with much else concerning null hypothesis significance testing, there is no consensus amongst statisticians as to when corrections for multiple testing are necessary. Some have even argued that corrections for multiple testing should never be used (25). However, in highly regulated areas of clinical research, such as drug development, procedures to adjust for multiple comparisons procedures are strongly advocated and widely used (26, 27).

Those of us actively involved in orthopaedic research should consider the following important issues regarding multiplicity, when designing, undertaking and reporting studies:

- (a) The simplest approach to the *multiple testing problem* is to avoid it by prospectively selecting a single *primary outcome* measure, and to explicitly identify all other outcomes as

secondary. This ensures that there is a single primary significance test for inferential and decision-making purposes, with claims for significance for the secondary outcomes being made only after the primary objective of the study has been achieved. This approach is routine in clinical trials, but can be used more generally in other research designs. However, often it makes no sense to identify a single important measure prospectively. For instance, in exploratory studies where the aim is to test many variables as a means to identify possible significant associations.

(b) If a study makes a formal argument that there are multiple primary endpoints (e.g. co-primary endpoints for survival and quality of life), then it should be clearly stated how the results of the multiple tests are to be interpreted. For instance, if for statistical significance to be declared it needs to be established for *all* primary endpoints at the stated significance level, then no formal adjustment of the significance level is required for the individual hypothesis tests. Another common scenario is where a hierarchical series of tests are undertaken in a *pre-defined* order (*fixed-sequence tests* (27)). For instance, we might test for differences between groups, in a fixed order based on a clinical assessment of importance, in (i) survival, then (ii) function and then (iii) quality of life. In this setting, we do not have to make adjustments for multiple testing, if we *only* reject for an individual hypothesis test at the chosen level of significance if *all* the hypothesis tests higher-up the hierarchy are *also* rejected at the same level. For instance, we only reject at the set significance level for quality of life, provided we have already rejected at this same level for both survival and function.

(c) Multiplicity issues can often arise from claims of significance after tests in many subgroups of the full study population. Again, this provides a number of opportunities to claim significance, and thus is likely to increase the chance of false positive results. Reliable conclusions from subgroup analyses, in a confirmatory setting, require pre-specification and appropriate methods of analysis. In an exploratory or observational study setting, many subgroup analyses can legitimately be undertaken but should be motivated by scientific reasoning, ideally limited in number and reported with due caution.

(d) If a number of defined (intervention or treatment) groups (A, B, C...) are compared to each other or with a single fixed control group then a number of tests follow naturally; e.g. A *versus* B, B *versus* C, A *versus* control, B *versus* control etc. In settings such as this multiple comparison procedures (e.g. *Tukey's HSD*, *Dunnett's test* (24)) should be used; they are well established and available routinely in most statistical software.

(e) Another commonly observed setting is where groups are compared at many occasions over time, without any clear definition of the primary endpoint (as one would typically do in a clinical trial). Multiple cross-sectional analyses at many time-points give us many opportunities to undertake tests to identify significant differences between groups. A much better approach than this is to use a single *longitudinal model* to assess the pattern of between group differences over time (e.g. a mixed-effects model). In the fully longitudinal analysis, a single test is used to assess the significance of the temporal group effects, and thus the false positive rate is controlled. Where such time courses of data are reported, appropriate longitudinal modelling methods should be used in preference to multiple cross-sectional analyses.

(f) In general, additional analyses undertaken on subsets of the main study data, with the purpose of exploring the robustness (sensitivity) of the conclusions from the primary analysis, do not require adjustment to p-values to control overall false positive error rates.

(g) In settings where many significance tests are reported, and none of the above options offer a method of adjustment, the most common method of controlling the false positive rate is the Bonferroni procedure (each individual hypothesis is tested at level α/m , where α is the overall test

level and m is the number of hypotheses tested). This procedure is known to be very conservative, so should *not* be used routinely without careful consideration of other options. For instance, the Holm-Bonferroni procedure (24) is known to be uniformly 'better' than the Bonferroni procedure.

(h) Multiplicity cannot simply be ignored. However, in many circumstances it is perfectly acceptable to proceed and report results of many analyses without adjustment of p-values, in what we might call a 'conventional' manner. Provided that it is made clear that this is a deliberate choice and that this is acknowledged as a limitation of any reported analyses.

Redefining significance thresholds

Rather than proposing that p-values be dropped or replaced, others have suggested thresholds for significance should be re-defined (28). They argue that the ubiquitous 5% level was itself arbitrary when it was originally suggested by Fisher (17), and that a new threshold of 0.5% ($p < 0.005$, rather than $p < 0.05$) would be more appropriate for claims of discovery for new effects. Although the proponents of this new threshold acknowledge that any p-value threshold represents a trade-off between false positive and false negative errors, they believe the current balance is such that it produces too many false positive results. The main disadvantage to changing the conventional p-value threshold to 0.5% is that this would require considerable increases in sample size to maintain study power: to achieve 80% power, at the 5% significance level, sample sizes would need to be increased by about 70% (28). Fewer, but larger and more rigorous, studies may please many of us, but would not be universally popular and would certainly be more expensive. A more practical way forward may be to select different thresholds dependent on the research specialty (e.g. genome-wide association studies currently use a threshold of $p < 5 \times 10^{-8}$ (29)) and the type of study (e.g. confirmatory or exploratory analyses (30)). No one-size-fits-all approach to re-setting thresholds is likely to be universally adopted and accepted.

Clinical research and p-values

Within the wider scientific community, clinical trials are often singled out as paragons of good practice, when contrasted with much basic, biomedical, pre-clinical, and observational research (11). The requirement to make available protocols that clearly identify the primary outcome and other important details of the study design and analysis plan (31, 32), do not in themselves prevent abuses of the null hypothesis significance testing methodology. However, it would be difficult to imagine a funded clinical trial proceeding now without a clear statement of the hypotheses under test, a clear reasoned scientific rationale as to why the study is taking place and *a priori* justification of both the significance level and how outcomes (e.g. effect estimates) will be interpreted.

In the clinical sciences, we have obvious advantages compared with for instance the behavioural sciences (e.g. psychology) where there are well-established minimum clinically important differences (e.g. Oxford hip and knee scores (32)). For example, many clinical studies report outcomes for which we have a natural sense of whether a reported difference is important (e.g. revision rates for an implant), or (patient-reported) outcomes where there are well-established minimum clinically important differences (e.g. Oxford hip and knee scores (35)). An intuitive understanding of such outcomes help us to make decisions about the importance of the results of any statistical analysis.

Of course we need to ask “*what is the p-value for my chosen statistical test?*” This is important, and size matters, but not in isolation. We also need to consider the effect estimate (36), confidence interval (37), and whether the size of the effect is both sensible and scientifically credible. If the effect is small, it may not be clinically meaningful (38). All these things matter, and cannot simply be ignored purely on the basis that a p-value passes a pre-determined threshold.

An intuitive understanding of such outcomes help us to make decisions about the importance of the results of any statistical analysis. Of course, we need to ask “*what is the p-value for my chosen statistical test?*” This is important, and size matters, but not in isolation. Importantly, we also need to consider the effect estimate (33), confidence interval (34), and whether the size of the effect is both sensible and scientifically credible. If the effect is small, it may not be clinically meaningful (35). All these things matter, and cannot simply be ignored purely on the basis that a p-value passes a pre-determined threshold.

Reporting guidelines

The following ten guidelines have been accepted by the Editorial Board of the BJJ and are intended to help authors to address many of the issues we discuss here when submitting manuscripts. For many widely used study types specific guidelines for reporting are available from the EQUATOR network (39)(31, 36).

- (1) Null hypothesis significance testing has become an integral part of the scientific method for orthopaedic research, and is likely to remain the standard inferential tool for the foreseeable future. However, there are alternative methods available (e.g. Bayesian analysis) and these can be reported in BJJ, provided sufficient detail and justification are provided.
- (2) If null hypothesis significance testing is reported, then authors should make a clear statement of the scientific reasoning and basis for undertaking such tests. With, where appropriate, clear statements of the hypotheses under test.
- (3) The study design and objectives should be explicitly stated at an early point in all manuscripts, as a basis for how the results of statistical tests are reported and interpreted. For example exploratory and confirmatory research have very different aims, and may legitimately use different approaches to significance testing and interpretation of p-values.
- (4) Any distinctions between primary and secondary study outcomes including timepoints, or similar data hierarchies, that may affect how results of significance tests are interpreted should be made clear in the methods section of the manuscript. Also, if subgroup analyses are reported, then they should be justified on a clinical or scientific basis and identified as either planned (pre-specified) or *post hoc* by the authors.
- (5) The default level (threshold) for significance tests should be 5%; i.e. the p-value is equal to or below 0.05. Authors can decide to set lower levels if they wish. However, lower levels should not be arbitrarily selected without good scientific reason, and the level used for testing should be clearly stated before any analysis is reported. Ideally, dependent on the type of research reported, the significance level used for testing should be stated in a published protocol or study registration document.
- (6) Unless there is a legitimate scientific reason not to do so, p-values should be reported from all statistical tests undertaken in a reported analysis, not just the subset of values deemed to be

significant. And they should be reported to 3 decimal places, and not simply stated as crossing specific thresholds (e.g. $p < 0.05$ or $p < 0.01$).

- (7) In addition to p-values, authors should also *always* report effect estimates (e.g. odds ratios, differences of means, hazard ratios) and error estimates (e.g. confidence intervals, credible intervals).
- (8) Multiplicity issues (i.e. decisions about whether to modify the significance level when undertaking more than one test) are left to the discretion of the authors of manuscripts. However, there are a number of well understood settings where there is an expectation that methods should routinely be implemented (see above notes, Multiple Testing (a) – (g)). If formal multiple comparison methods are reported then authors should justify the choice of method, describe briefly and reference appropriately in the statistical methods section of the manuscript.
- (9) When interpreting p-values, and making statements about the significance of reported effect estimates, authors should be mindful of the distinction between *statistical significance* and *clinical significance*. The former is in large part determined by the study sample size, and thereby the capacity to detect small effects with relatively high power. Whereas *clinical significance* relates to a difference, or change, in outcome that is relevant (meaningful) to a patient, study participant or may cause a clinician to consider a change in practice.
- (10) Authors should comment in the discussion section of the manuscript on the number of significance tests reported in the manuscript. Where this is perceived to be an issue, authors should acknowledge that reporting many p-values can have consequences for how the results of statistical tests are interpreted, clearly state this as a limitation of the analysis and if appropriate make explicit reference to the exploratory nature of the work.

Summary

Null hypothesis significance testing and the interpretation of p-values is a contentious issue, and is likely to remain so for the foreseeable future. The truth is that many statisticians and scientists like p-values, including the past president of the Royal Statistical Society (RSS), Professor Sir David Spiegelhalter (22). He presented the views of many in an address to the RSS in 2017 (40) He presented the views of many in an address to the RSS in 2017 (37) by saying that “[p-values] are *good and useful measures of compatibility between data and hypotheses, but insufficient distinction is made between their informal use in exploratory analysis and their more formal use in confirmatory analyses*”. Going on to say that “*there is too strong a tendency to use p-values to jump from selected data, to a claim about the strength of evidence, and to conclusions about the practical importance of the research*”. Sir David went on to summarise the core of the problem elegantly by saying: “*p-values ‘do what they say on the tin’, but people do not read the tin*”.

References

1. Ioannidis J. Why most published research findings are false. *Plos Medicine*. 2005;2(8):696-701.
2. Ioannidis J. Why most published research findings are false: Author's reply to Goodman and Greenland. *Plos Medicine*. 2007;4(6):1132-3.
3. Goodman S, Greenland S. Why most published research findings are false: Problems in the analysis. *Plos Medicine*. 2007;4(4):773-.
4. Berry D. A p-Value to Die For. *Journal of the American Statistical Association*. 2017;112(519):895-7.
5. Briggs W. The Substitute for p-Values. *Journal of the American Statistical Association*. 2017;112(519):897-8.
6. Gelman A, Carlin J. Some Natural Solutions to the p-Value Communication Problem-and Why They Won't Work. *Journal of the American Statistical Association*. 2017;112(519):899-901.
7. Laber E, Shedden K. Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-Values for Statisticians. *Journal of the American Statistical Association*. 2017;112(519):902-4.
8. McShane B, Gal D. Rejoinder: Statistical Significance and the Dichotomization of Evidence. *Journal of the American Statistical Association*. 2017;112(519):904-8.
9. McShane B, Gal D. Statistical Significance and the Dichotomization of Evidence. *Journal of the American Statistical Association*. 2017;112(519):885-95.
10. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond " $p < 0.05$ ". *American Statistician*. 2019;73:1-19.
11. Begley C, Ioannidis J. Reproducibility in Science Improving the Standard for Basic and Preclinical Research. *Circulation Research*. 2015;116(1):116-26.
12. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*. 2017;4(12).
13. Colquhoun D. The reproducibility of research and the misinterpretation of p-values (vol 4, 171085, 2017). *Royal Society Open Science*. 2018;5(3).
14. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Science Translational Medicine*. 2016;8(341).
15. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature*. 2019;567(7748):305-7.
16. Fisher RAS, Bennett JH, Fisher RASSmfrw, Fisher RASDoe, Fisher RASSm, inference s. *Statistical methods, inference and experimental design : a re-issue of Statistical methods for Research workers, The design of experiments, and Statistical methods and scientific inference*. Oxford: Oxford University Press; 1990.
17. Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. 1933;231:289-337.
18. LEHMANN E. THE FISHER, NEYMAN-PEARSON THEORIES OF TESTING HYPOTHESES - ONE THEORY OR 2. *Journal of the American Statistical Association*. 1993;88(424):1242-9.
19. Lenhard J. Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal For the Philosophy of Science*. 2006;57(1):69-91.
20. Wasserstein R, Lazar N. The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician*. 2016;70(2):129-31.
21. Ionides E, Giessing A, Ritov Y, Page S. Response to the ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician*. 2017;71(1):88-9.
22. Matthews R, Wasserstein R, Spiegelhalter D. The ASA's p -value statement, one year on. *Significance*. 2017;14:38-41.
23. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. *PLoS Biol*. 2015;13(3):e1002106.

24. Bretz F, Hothorn T, Westfall PH. Multiple comparisons using R. Boca Raton, FL: CRC Press; 2011. xvii, 187 p. p.
25. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43-6.
26. Dmitrienko A. Special Issue: Multiplicity Issues in Clinical Trials. *J Biopharm Stat*. 2018;28(1):1-2.
27. Dmitrienko A, D'Agostino RB, Sr. Multiplicity Considerations in Clinical Trials. *N Engl J Med*. 2018;378(22):2115-22.
28. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10.
29. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet*. 2016;24(8):1202-5.
30. Tukey JW. WE NEED BOTH EXPLORATORY AND CONFIRMATORY. *American Statistician*. 1980;34(1):23-5.
31. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ*. 1996;313(7057):570-1.
32. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol*. 2015;68(1):73-9.
33. Parsons N, Griffin XL, Stengel D, Carey Smith R, Perry DC, Costa ML. Standardised effect sizes in clinical research: how to compare shoulder surgeons with hip surgeons. *Bone Joint J*. 2014;96-B(7):853-4.
34. Perry DC, Griffin XL, Dritsaki M, Costa ML, Parsons N, Bone obot, et al. Becoming confident about confidence intervals. *Bone Joint J*. 2017;99-B(5):563-5.
35. Parsons NR, Perry DC, Costa ML. All interventions differ, although some are more different than others. *Bone Joint J*. 2017;99-B(9):1123-4.
36. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Dore C, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*. 2017;318(23):2337-43.
37. Spiegelhalter D. Trust in numbers. *Journal of the Royal Statistical Society Series a-Statistics in Society*. 2017;180(4):949-65.