Clin Orthop Relat Res (2011) 469:3259–3261
DOI 10.1007/s11999-011-2053-1

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

IN BRIEF (BY INVITATION ONLY)

Statistics in Brief

Interpretation and Use of p Values

All p Values Are Not Equal

Frederick Dorey PhD

Published online: 15 September 2011
© The Association of Bone and Joint Surgeons® 2011

Background

In a formal hypothesis testing situation, a question is frequently asked about differences between groups, and based on that question an experiment is designed, data are collected, and a statistical test is performed, usually resulting in one or more p values. The p value resulting from a hypothesis test is heuristically defined as a probability measure of how much evidence there is against the null hypothesis of the test, that is, no difference exists [1]. When the p value is small (however defined), then a decision might be made to reject the null hypothesis and accept the alternative hypothesis that a difference exists. However, in many (if not most) situations today, the reader of a medical journal has made no such prior definition of what is small, or exactly what use should be made of any given p value. Thus, despite the exact definition of what a p value means, how p values in general should be interpreted or how they should influence the readers of medical journals is not clear. Although the definitions involving hypothesis testing and p values are precise, the interpretation and use of the resulting p values are much more subjective and individual processes.

Question

Some insight into these processes might be obtained by answering the following question: What factors should play a role in the understanding, interpretation, and use of p values reported in medical journals? For example, if two treatments are being compared in some medical articles resulting in the same p values, do the two p values together represent twice the amount of evidence against the null hypothesis? If the p values have different interpretations, what factors might cause a reader to place more emphasis on one rather than the other? These questions are especially important as the p value is used routinely as a summary result for a clinical study.

Discussion

Interpretation

Before any published p value resulting from a clinical study can be interpreted (that is, deciding how much evidence it presents against the null hypothesis), readers must make some assessment of the validity and quality of the study.

Validity

All p values are not equal. The believability of any p value depends on the validity of the associated clinical study. If a p value is based on a biased comparison, its particular value likely will have little meaning. In addition, issues

The author certifies that he has no commercial associations that might pose a conflict of interest.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.

Electronic supplementary material The online version of this article (doi:10.1007/s11999-011-2053-1) contains supplementary material, which is available to authorized users.

F. Dorey (✉)
Department of Pediatrics At Children's Hospital Los Angeles, University of Southern California, Keck School of Medicine, 4600 Sunset Blvd., Los Angeles CA 90027, USA
e-mail: fdorey@chla.usc.edu

such as missing data, incorrect statistics, and confounding might call into question whether any valid use can be made of a study and its associated p values. In addition, a comparison of p values should be made only if they are based on studies with the same level of evidence as used by this journal or defined somewhat differently [8]. A p value from a double-blind randomized clinical trial (highest level, and least subject to bias) should carry much more weight than one from a retrospective observational study (one of the lower levels more subject to bias). Some publications address these issues in more detail [1–6]. (A supplemental video is available with the online version of CORR.)

Quality

Even if the validity of two p values is not in question and the publications are based on the same level of influence, the quality of the involved studies should play a role in determining what weight should be given to them. For example, the quality of a randomized study can be greatly affected by issues such as poor patient flow in the study, protocol violations such as failure to adhere to the randomization scheme, small sample size and low power, and failure to adjust properly for confounding variables or interactions in the statistical analysis [4–7].

Use of a p Value

A p value is only one tool to be used in deciding whether a clinical treatment should be changed or modified in some way. Statistical issues such as confidence intervals, statistical power, and especially, the issue of statistical significance versus clinical importance, must play essential parts in assessing the results of a clinical study. For example, studies with very large sample sizes will have several "statistically significant" comparisons that have very little biological or clinical importance. Nevertheless, p values frequently are used as a summary for the results of a clinical study and the numerical value alone can have an influence on a reader's response to the results of a clinical study. The proper use of a p value will depend, to some extent, on the clinical importance and seriousness of the medical issue, the current literature on the subject, and the personal experience and beliefs of the reader.

Clinical Importance and Seriousness of the Medical Situation

Decisions regarding whether a patient should receive limb-salvage versus amputation clearly fall into a different category than decisions involving aspects of postoperative protocols that likely have minimal clinical implications. Thus it might take several p values from high-quality studies to modify a surgeon's decision regarding what surgery to perform, whereas only one or two significant p values might lead to modifications of lesser aspects of the surgery that can be changed with little possible harm to the patient.

Existing Literature and Prior Beliefs of the Physician

When there are a large number of quality published studies on a subject, the results might be formally analyzed using a meta-analysis. In these cases the p value results of the meta-analysis might be sufficient to change a physician's practice in some way. However, even in these cases, the physician's prior beliefs would play a role. In general, for most p values presented in a paper, not only the existing literature information but also the physician's prior beliefs will play large parts in his or her subjective evaluation of the results from a clinical study. A physician who already had a strong belief against a null hypothesis might consider a small, but not quite statistically significant p value to be further evidence supporting that belief, whereas a reader with no such prior belief might view the same p value as minimal evidence at best.

Personal Clinical Experience

A physician's response to a given p value will depend in large part on his or her personal clinical experience. A p value suggesting that one approach is significantly better than another might be totally ignored by a physician whose personal experience was different. In situations like this, one approach would be for the physician involved to investigate why such an extreme difference might exist.

In a few situations only one p value might be enough to decide an issue. For example a young surgeon might have learned how to perform a surgery one way during residency and a different way during a fellowship. This situation might result in the physician performing a randomized clinical trial that might, depending on the results, be the only information necessary for that surgeon to decide the issue.

The way most p values are used in practice is similar to the situation where a physician is diagnosing a patient and is unsure about the true diagnosis. The physician might have a prior belief (probability) that a particular diagnosis is correct, but being unsure, requests an additional test. If the results from the additional test are more likely to have come from a patient with the diagnosis in question rather

than some alternative, the physician's prior belief would be increased, possibly to the point where the diagnosis would be made.

If this reasoning process was done in a more formal statistical manner, it would be a particular example of the statistical methodology referred to as Bayesian statistics [7]. In Bayesian statistics a prior probability belief in a hypothesis is modified by some new data, resulting in a change in that probability. This is in contrast to the p value setting, where if a specific threshold is past and statistical significance is reached, there is an implied decision to be made concerning the hypothesis in question. As suggested earlier, this formal use of a p value is seldom made.

Myths and Misconceptions

(1) If two p values address the same hypothesis, the smaller one provides more evidence against the null hypothesis. As noted above, there are many other important issues that must be addressed before two p values can even be compared.

(2) The notion of statistical significance today (p less than 5% or not) contains all the important information relating to a p value. As noted above, if the study is flawed, the p value is likely flawed. Choosing artificial cut points for the p value does not address the issue that the p value itself might be incorrect.

(3) Including a p value for almost every statement made in a publication strengthens the scientific validity of a paper. In fact, the further the hypothesis related to a p value is from the primary hypothesis that generated the clinical study, the further the p value itself becomes from having a scientific interpretation.

(4) If a comparison of relevant baseline variables in two groups of patients results in no 'statistically significant differences', then a simple statistical comparison between the treatments is sufficient for the analysis. A p value resulting from a multivariate statistical model indicating the need to adjust for confounding variables and interactions related to the treatments is necessary and has much more scientific validity than a simple direct comparison of the treatment results, even if all baseline comparisons were not statistically significant.

Conclusion

This brief discussion has not directly or comprehensively answered the question of how p values should affect clinical practice. I am hopeful it has helped clarify some of the issues that should be part of the decision process. Although the interpretation of p values can be greatly clarified by using good principles of experimental design and statistical analysis, the decision regarding how to use those results in the clinic is much more complex and subjective. Exactly how a p value in a published study might influence a physician's clinical practice will likely vary from issue to issue and study to study. However, this is certain: all p values are not the same in their interpretation or how they will be used.

References

1. Dorey F. The p value: what is it and what does it tell you? *Clin Orthop Relat Res*. 2010;468:2297–2298.
2. Dorey FJ. In brief: statistics in brief: confidence intervals: what is the real result in the target population? *Clin Orthop Relat Res.* 2010;468:3137–3138.
3. Dorey FJ. Statistics in brief: statistical power: what is it and when should it be used? *Clin Orthop Relat Res.* 2011;469:619–620.
4. Greenhalgh T. *How to Read a Paper: The Basics of Evidence-Based Medicine*. Chichester, UK: BMJ Books; 2007.
5. Jolles BM, Martin E. In brief: statistics in brief: study designs in orthopaedic clinical research. *Clin Orthop Relat Res.* 2011;469: 909–913.
6. Lambert J. Statistics in brief: how to assess bias in clinical studies? *Clin Orthop Relat Res*. 2011;469:1794–1796.
7. Motulsky H. *Intuitive Biostatistics*. New York, NY: Oxford University Press; 1995.
8. Oxford Center for Evidence-Based Medicine. Available at: http://www.cebm.net/index.aspx?o=5653. Accessed August 9, 2011.