GREGORY W. FROEHLICH, MD
VA Outcomes Group
Department of Veterans Affairs Medical Center
White River Junction, Vt

Effective Clinical Practice. 1999;2:234-239.

What Is the Chance That This Study Is Clinically Significant?

A Proposal for Q Values

CONTEXT. Clinicians who use the medical literature to guide their practice need to make judgments about the clinical significance of medical interventions.

GENERAL QUESTION. How likely is an intervention to be clinically worthwhile?

SPECIFIC RESEARCH CHALLENGE. Given the results of a study, determining the probability that the true effect of an intervention is at least as great as some minimum worthwhile effect.

CURRENT APPROACH. *P* values are widely used to convey the probability of observed effects arising by chance if there truly is no effect. By convention, *P* values less than 0.05 are interpreted as being "statistically significant."

POTENTIAL DIFFICULTIES. Statistical significance is often confused with clinical significance.

ALTERNATE APPROACH. A different probability could be reported, a probability I call a Q *value*. A *Q* value is the probability that the true effect of an intervention is at least as great as some minimum worthwhile effect. *Q* values are calculated in a manner analogous to that used for *P* values, except that the null hypothesis becomes a minimum worthwhile effect instead of no effect. *Q* values encourage researchers and clinicians to be explicit about what they think a worthwhile effect is and could help shift the focus of study interpretation away from arbitrary statistical conventions.

Clinicians who use the medical literature to guide their practice face the following four basic questions when determining the applicability of a study's result: Is the study result valid?[1] Can it be generalized to the patient in question?[2–5] Is the intervention feasible, given the available resources? What are the chances that the intervention will produce a clinically meaningful effect? The final question, which concerns assessment of clinical significance, can be especially challenging.

Clinicians are often tempted to equate statistical significance with clinical significance. Because statistical significance is almost always quantified and is evaluated against a conventional standard (i.e., $P<0.05$), it has considerable appeal for clinicians. When *P* values are emphasized, however, decisions about significance and importance are relegated to mathematical models.[6] *P* values only convey the probability of the measured effect arising by chance if there truly is no effect. Confidence intervals (CIs), while emphasizing the precision of the measured effect, are often interpreted in the same manner as *P* values. Consequently, it is not unusual to find examples where statistically significant results are of trivial importance or, conversely, where non–statistically significant results are clinically important.[7,8]

To quantify clinical significance, I propose that a different measure of probability be reported after a study, a probability I call a Q *value*. In practical terms, a *Q* value is the probability that the true effect of an intervention is at least as great as

The abstract of this paper is available at ecp.acponline.org.

some minimum worthwhile effect. *Q* values differ from P values in that they are derived from testing a hypothesis of some minimum effect rather than from testing a hypothesis of no effect. Whereas CIs can provide qualitative information about clinical significance, *Q* values provide quantitative information about the probability of clinical significance. To make the concept of *Q* values more tangible, I begin with an example that shows the limitations of existing practices and demonstrates the calculation of *Q* values.

Judging Clinical Significance

The Standard Approach

Consider a recent randomized, multicenter trial of cyclosporine versus placebo in patients with severe rheumatoid arthritis treated with methotrexate.[9] (I chose this example to simplify the statistical calculations because the primary outcomes were expressed as unadjusted means.) All patients had to have six or more actively inflamed, tender, or swollen joints, and the main outcome measure was the number of joints that were no longer tender after 6 months of treatment. The primary finding was that cyclosporine-treated patients had an average of 4.8 fewer tender joints than the placebo group. **Table 1** shows the basic statistics relevant to this finding.

TABLE 1
Basic Statistics on the Primary Outcome in a Trial of Cyclosporine in Patients with Severe Rheumatoid Arthritis*

MEASURE	RESULT
Mean number of joints no longer tender at 6 months	
Cyclosporine group ($n = 75$)	7.5
Placebo group ($n = 73$)	2.7
Effect size (difference in means, or $\Delta\bar{x}$)	4.8
SE of mean difference	2.06
Test statistic	
$t = \frac{\Delta\bar{x} - 0}{SE}$	2.33
P value	
P = probability ($\|T\| \geq \|t\|$)†	0.02
95% CI	
$\Delta\bar{x} \pm t_{0.95} \cdot SE$‡	0.7–8.9

**SE = standard error.*
†Degrees of freedom (DF) = 146.
‡$t_{0.95}$ (given DF = 146) = 1.98

P Values

The *P* value for the effect of cyclosporine is 0.02. This is the probability of a type I error—that is, the probability of finding a difference of at least 4.8 fewer tender joints, by chance alone, if cyclosporine truly has no effect. *P* values are judged against an arbitrary standard for statistical significance: Those that are less than 0.05 are significant, and those that are above 0.05 are not significant. Because the *P* value in the example is less than 0.05, we reject the null hypothesis and conclude that cyclosporine has a statistically significant effect.

Rejecting the null hypothesis, however, is not the same as demonstrating clinical significance. The *P* value of 0.02 tells us that there was an effect, not whether the effect was clinically worthwhile. Furthermore, the *P* value obscures the magnitude of the effect. In other words, the larger the sample, the more significant the *P* value, regardless of the actual size of the effect. Because of these limitations, many journals now encourage authors to provide CIs for their results.[10]

Confidence Intervals

The 95% CI for the effect of cyclosporine compared with placebo is 0.7 to 8.9 fewer tender joints. In this case, the range is wide but does not include zero. The CI provides a range of plausible values for the true effect. In familiar terms, there is a 95% chance that the true or population effect of adding cyclosporine ranges from 0.7 to 8.9 fewer tender joints. CIs emphasize the size and precision of the measured effect as an estimate of the true effect.

In practice, however, CIs are often used as nothing more than tests of statistical significance. If no effect is outside the 95% CI, the result is statistically significant. CIs do not require readers to explicitly consider what is clinically worthwhile, although readers can determine if the CI includes or excludes clinically worthwhile effects. If an improvement in five or more joints is determined to be clinically worthwhile, then five joints is the minimum worthwhile effect. If the CI includes this value, a clinically worthwhile effect is plausible.

Even if readers define a minimum worthwhile effect, however, CIs do not readily convey the probability of that effect, except in the following three cases. If the minimum worthwhile effect is at the lower bound of the CI, the probability that the true effect is clinically worthwhile is 97.5%; if the minimum worthwhile effect is at the upper bound of the CI, the probability is 2.5%; if the minimum worthwhile effect equals the measured effect,

the probability is 50%. In most cases, the minimum worthwhile effect lies within the CI; thus, obtaining this quantitative information from the CI is not possible.[11]

An Alternate Approach: The *Q* Value

The purpose in proposing *Q* values is to provide a general method for quantifying the probability of a clinically worthwhile effect when the minimum worthwhile effect lies within the CI. Determining *Q* values follows three steps.

First, decide what the minimum worthwhile effect is. This is not a trivial exercise, although researchers must give some thought to it in order to calculate a study's sample size. To calculate the sample size for the example, the authors asserted that the minimum worthwhile difference was five fewer tender joints.

Second, calculate a test statistic. Because the effect in this example is a difference between two means, this is a *t* statistic. It is calculated as the difference between the minimum worthwhile effect and the measured effect, divided by the standard error of the effect size: (5−4.8)/2.06 = 0.097.

Third, determine the one-tailed probability for the test statistic, using the appropriate probability distribution. This probability is the *Q* value. In the example, the *Q* value is 0.46. In other words, there is a 46% chance that the true effect of adding cyclosporine to methotrexate is at least five fewer tender joints. A graphic representation of this *Q* value is shown in **Figure 1**. Of course, not everyone will agree that five fewer tender joints is the minimum worthwhile effect when cyclosporine is added. **Figure 2** shows the effect on the *Q* value of specifying different values for the minimum worthwhile effect. Note that the *Q* value, like all probabilities, ranges from 0 to 1. If we believe that the risks of adding cyclosporine are low, we might assert that a difference of even a single joint is worthwhile. In that case, *Q* = 0.97 and the intervention has a high probability of being worthwhile. If, as the authors asserted, the minimum worthwhile effect is five fewer tender joints, *Q* = 0.46 and the intervention is about as likely to be worthwhile as not. Finally, if the benefits of fewer tender joints are outweighed by the side effects of cyclosporine, the minimum worthwhile effect might be higher, perhaps eight joints. In that case, *Q* = 0.06 and there is only a small chance that the true effect of the intervention is worthwhile. The same intervention, with the same *P* value and 95% CI, can have very different *Q* values depending on what effect is believed to be clinically worthwhile.

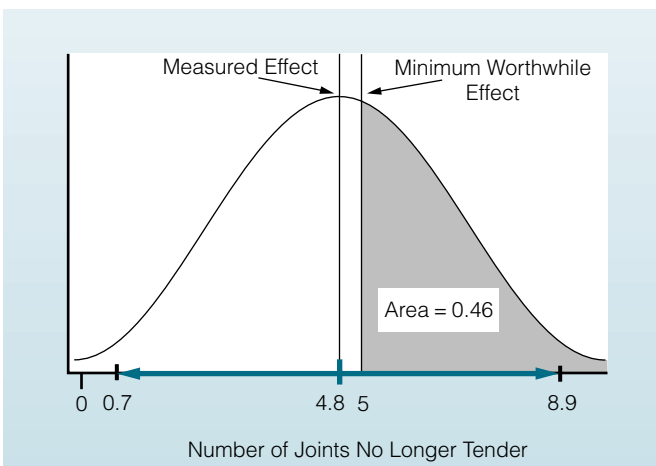

FIGURE 1. *Q* value for the trial of cyclosporine in patients with severe rheumatoid arthritis. The curve for probability density is centered on the measured effect, and the 95% CI is superimposed on the horizontal axis. The *Q* value is the area under the curve to the right of the minimum worthwhile effect.

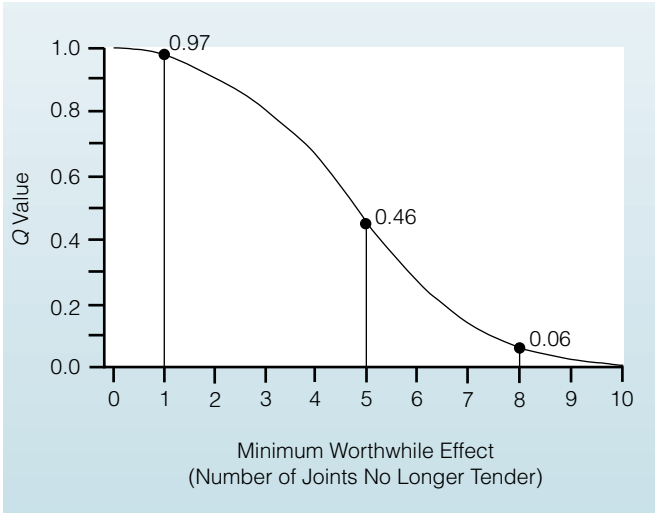

FIGURE 2. Varying the minimum worthwhile effect. The curve shows the relation between different values of the minimum worthwhile effect and the *Q* value, given the study results (measured effect = 4.8 fewer tender joints, SE = 2.06)

General Description of the *Q* Value

A *Q* value is determined as a *P* value would be, with two exceptions. First, the null hypothesis is for a minimum worthwhile effect rather than no effect. **Table 2** provides guidance on how to determine test statistics for *Q* values when comparing means or proportions. Second, *Q* values are always one-sided probabilities. They inform us whether an intervention is good enough or not good enough. From the standpoint of a clinician deciding whether to use an intervention, this is a more relevant issue than whether the intervention is beneficial, harmful, or has no effect. When the result suggests a harmful effect, further investigation and analysis may be worthwhile.[12]

TABLE 2
Definitions and Calculations for *Q* values*

Definition of *Q* value	**The probability that the true effect of an intervention is at least as great as some minimum worthwhile effect given the results of the study.**
Definition of minimum worthwhile effect (δ)	**The smallest true effect of the intervention that would yield a net benefit for the patient.**
Determination of *Q* value when comparing means $(\bar{x}_1 - \bar{x}_2)$	$t_Q = \frac{\delta - (\bar{x}_1 - \bar{x}_2)}{SE}$, Q = probability that $T > t_Q$
Determination of *Q* value when comparing proportions $(p_1 - p_2)$	$Z_Q = \frac{\delta - (p_1 - p_2)}{SE}$, Q = probability that $Z > Z_Q$

**SE = standard error. t_Q and z_Q = t and z statistics for Q value; T and Z = random variables for Student t test, normal distribution.*

Figure 3 highlights how *Q* values can help quantify clinical significance and distinguish it from statistical significance. In the figure, the CI for study A is wide, reflecting the imprecision of the result. In addition, the CI includes no effect; hence, the result is not statistically significant. However, the probability of a worthwhile effect (the *Q* value) is large (well over 50%) because the minimum worthwhile effect is smaller than the measured effect. Study B has a precise result and is statistically significant. Nevertheless, the probability of a worthwhile effect is small (less than 2.5%).

Q values differ from other probabilities that incorporate subjective values. Others have presented probability functions that can indirectly yield the probabilities of a clinically worthwhile effect.[13, 14] However, these approaches do not emphasize the importance of determining a minimum worthwhile effect. Another type of probability, statistical power, is a measure of a study's ability to detect a statistically significant result if the true effect is clinically significant. Power is determined by a study's design and sample size, not by the results; when the results of a study are being analyzed, the precision of the results—not the study's power—is the issue.[15] Power also differs from a *Q* value in that it is predicated on the ability of a study to find statistically significant results, whereas a *Q* value does not incorporate the concept of statistical significance. On the other hand, formulas to calculate power can be readily modified for use in determining *Q* values.

Limitations to the Approach

Like *P* values and CIs, *Q* values do not provide information about the validity of a study, whether the intervention is feasible, or whether the results can be generalized to other patients. Unlike *P* values, *Q* values are not helpful in gaining information about causal relationships between variables. *Q* values should therefore be used to complement *P* values, not to replace them.

Q values have two special limitations—limitations that help guide the process of interpreting a study's results. First, there is no consensus on what constitutes a minimum worthwhile effect. Being able to assert this value is a necessary first step in determining the *Q* value for the intervention. However, the time and effort necessary to determine a minimum worthwhile effect are well spent.

Table 3 provides some guidance for determining the minimum worthwhile effect. Clinicians might first look to others for help. Investigators often assert a value in their calculation of sample size.[15] Expert consensus is another possible source. In rheumatoid arthritis, for

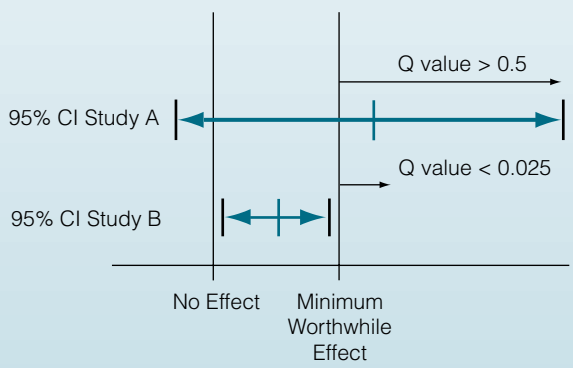

FIGURE 3. Use of *Q* values to discriminate between clinical significance and statistical significance.The results of study A are not statistically significant but have a substantial probability of being clinically significant. The results of study B are statistically significant but have a low probability of being clinically significant.

example, a formal exercise was used to reach the consensus that a difference of five fewer tender joints was worthwhile.[16] Finally, decision analyses may be used to determine the minimum worthwhile effect.[17]

On the other hand, clinicians may have to determine the minimum worthwhile effect on their own. As shown in **Table 3**, the answer to the question of when to require a large effect (or when to accept a small effect) hinges on the cost of the intervention, the risks of the intervention, and the importance of the outcome. In addition, individual clinicians should use their knowledge of the patient faced with the treatment decision to adjust the minimum worthwhile effect.[18]

The second limitation is that using a *Q* value in clinical decision making is more complex than using a *P* value. No convention exists for the level at which a *Q* value should be accepted or rejected, but the absence of a standard for evaluation may help focus attention on the other elements of a study that are equally important and to pay a reduced (but appropriate) amount of attention to issues of probability. Clinical decisions incorporate much more than the statistical significance of a single study's results.[14] For example, if the possibility of bias was high, if we were concerned about generalizability, or if previous studies were conflicting, we might demand a higher *Q* value before using an intervention.

Q values near 0 or 1 are most helpful in making decisions because they represent greater certainty about whether the intervention is clinically worthwhile. *Q* values near 0.50 reflect increasing uncertainty, related either to small sample size or to a study effect that is near the minimum worthwhile effect. In some situations, a *Q* value of 0.50 might support use of an intervention. However, *Q* values in this range may also serve as reminders that small studies generally do not provide strong evidence for or against worthwhile effects.

TABLE 3
Guidance for Determining the Minimum Worthwhile Effect

QUESTION	ANSWER
How can I learn what others consider a minimum worthwhile effect?	• Effect used in sample size calculation • Expert consensus • Cost-effectiveness/decision analysis
If other evidence is absent…	
when should I require that the minimum worthwhile effect be large?	• Costly intervention (in terms of time, money, or other resources) • High-risk intervention • Unimportant outcome, or intermediate outcome with uncertain patient benefit • Risk-adverse patient
when should I accept that the minimum worthwhile effect be small?	• Low-cost intervention • Low-risk intervention • Important and unambiguous outcome (e.g., death) • Risk-taking patient

Future Steps

Q values provide important information about clinical significance, information that *P* values and CIs do not provide. Several steps can be taken to encourage the use of *Q* values in interpreting study results. Researchers may wish to provide *Q* values along with *P* values and CIs, either by determining a minimum worthwhile effect and reporting the *Q* value for that effect or by providing readers with a *Q* value versus a minimum worthwhile effect plot (e.g., **Figure 2**). Readers can use a nomogram (available from the author) to approximate *Q* values for previously published studies, incorporating their own values for minimum worthwhile effects, the standard errors for the results, and the *P* values for the interventions. Because more complex statistical methods are used with increasing frequency, adaptation of these statistical methods to *Q* values is an area for future development.

In summary, *Q* values provide a way to quantify uncertainty about clinical significance, thereby giving clinicians more insight into the role that a study's results should play in decisions about care. *Q* values use a familiar numerical format, but require a conceptual shift, from testing a hypothesis of no effect to testing a hypothesis of a minimum worthwhile effect. Finally, *Q* values help place the burden of interpretation where it belongs. Rather than relying on arbitrary statistical conventions to say when a result is significant, researchers and clinicians need to be explicit about what they think a worthwhile effect is and how the probability of such an effect contributes to decisions.

Take-Home Points

- **Clinicians who use the medical literature to guide their practice need to make judgments about the clinical significance of medical interventions.**
- **Although the standard approach is to use *P* values, this often confuses statistical significance with clinical significance.**
- **I propose reporting *Q* values, which reflect the probability tha t the true effect of an intervention is at least as great as some minimum worthwhile effect.**
- ***Q* values would encourage researchers and clinicians to be explicit about what they think a worthwhile effect is and could help shift the focus of study interpretation away from arbitrary statistical conventions.**

References

1. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993;270:2598-601.
2. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1994;272:59-63.
3. Bailey KR. Generalizing the results of randomized clinical trials. Control Clin Trials. 1994;15:15-23.
4. Davis CE. Generalizing from clinical trials. Control Clin Trials. 1994;15:11-4.
5. Rubins HB. From clinical trials to clinical practice: generalizing from participant to patient. Control Clin Trials. 1994;15:7-10.
6. Feinstein AR. Invidious comparisons and unmet clinical challenges [Editorial]. Am J Med. 1992;92:117-20.
7. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med. 1978;299:690-4.
8. Moher D, Dulbert CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994;272:122-4.
9. Tugwell P, Pincus T, Yocum D, et al. Combination therapy with cyclosporine and methotrexate in severe rheumatoid arthritis. The Methotrexate–Cyclosporine Combination Study Group. N Engl J Med. 1995;333:137-41.
10. Gardner MJ, Altman DG. Estimating with confidence [Editorial]. Br Med J (Clin Res Ed). 1988;296:1210-1.
11. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med. 1986;105:429-35.
12. Bland JM, Altman DG. One and two sided tests of significance. BMJ. 1994;309:248.
13. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA. 1995; 273:871-5.
14. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195-9.
15. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200-6.
16. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. J Rheumatol. 1993;20:561-5.
17. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109-17.
18. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? J Clin Epidemiol. 1994;47:787-95.

Acknowledgment

The author thanks Dr. H. Gilbert Welch for his contributions to this work.

Grant Support

Dr. Froehlich was supported by a VA training grant while this work was being done.

Presentation

An abstract of this work was presented at the Society of General Internal Medicine Meeting, May 2 through May 4, 1996, Washington, D.C.

Correspondence

Gregory W. Froehlich, MD, VA Outcomes Group (111B), Department of Veterans Affairs Medical Center, White River Junction, VT 05009; telephone: 802-296-5178; e-mail: gregory.froelich@hitchcock.org.