Size of Treatment Effects and Their Importance to Clinical Research and Practice

Helena Chmura Kraemer and David J. Kupfer

In randomized clinical trails (RCTs), effect sizes seen in earlier studies guide both the choice of the effect size that sets the appropriate threshold of clinical significance and the rationale to believe that the true effect size is above that threshold worth pursuing in an RCT. That threshold is used to determine the necessary sample size for the proposed RCT. Once the RCT is done, the data generated are used to estimate the true effect size and its confidence interval. Clinical significance is assessed by comparing the true effect size to the threshold effect size. In subsequent meta-analysis, this effect size is combined with others, ultimately to determine whether treatment (T) is clinically significantly better than control (C). Thus, effect sizes play an important role both in designing RCTs and in interpreting their results; but specifically which effect size? We review the principles of statistical significance, power, and meta-analysis, and commonly used effect sizes. The commonly used effect sizes are limited in conveying clinical significance. We recommend three equivalent effect sizes: number needed to treat, area under the receiver operating characteristic curve comparing T and C responses, and success rate difference, chosen specifically to convey clinical significance.

Key Words: Effect size, significance, hypothesis testing, power, confidence intervals, meta-analysis

Suppose a well-done randomized clinical trial (RCT) reports a statistically significant difference between treatment (T) and control (C) groups, with $p = .05$, $p = .01$, even $p = 10^{-6}$. Should these results be automatically considered of clinical significance, the basis of recommending that clinicians use T rather than C for patients like those studied? No. What would be needed in addition to infer clinical significance is the subject of this review.

P values alone have problems. First, *p* values are often miscomputed, the result of misapplication of a test or of post hoc testing or multiple testing leading to exaggeration of significance. The two most obvious sources of miscomputation are possible but least likely: 1) errors in arithmetic; or 2) an intent to deceive; however, even in absence of errors, "naked" *p* values are not sufficient for an inference of clinical significance.

Moreover, many who report correct *p* values misinterpret their message (Cohen 1995; Dar et al 1994; Hunter 1997; Krantz 1999; Nickerson 2000; Shrout 1997; Thompson 1999; Wilkinson and The Task Force on Statistical Inference 1999). Some incorrectly interpret the *p* value as an effect size, as the probability that the null hypothesis is true or of future nonreplication. Some also misinterpret a lack of statistical significance as an indication that the null hypothesis is true. Consequently, many methodologists advocate "banning the *p* value" (Shrout 1997). However, statistical significance testing remains a useful tool, a shame to lose only because it can be misused. Others propose that researchers be better educated on the meaning of *p* values and that every *p* value reported in a research journal (statistically significant or not) be accompanied by an effect size and its confidence interval (Borenstein 1994, 1997, 1998; Grissom and Kim 2005; Jacobson and Truax 1991; Kraemer 1993; Rosenthal et al 2000).

Such a recommendation is currently difficult to implement. Which of the many available effect sizes should be reported? The growing literature on the subject is largely directed to effect sizes that satisfy statistical, not clinical, needs (Grissom and Kim 2005). Even among statisticians there is controversy regarding some of the effect sizes proposed (Hsu 2004; Kraemer, in press). For the present purpose, the effect size required must accomplish three purposes. It must serve 1) to communicate information useful to assessing the clinical significance of any result found in an RCT; 2) as the basis of power calculations in the process of designing the RCT; and 3) to represent the results of that RCT in meta-analysis.

Here, we will first briefly review relationships among statistical significance, clinical significance, power, and meta-analysis. Then we will discuss the two most commonly used effect sizes, Cohen's *d* and the odds ratio, showing their limitations. Then we will recommend three mathematically equivalent effect sizes: NNT (number needed to treat), AUC (area under the receiver operating characteristic [ROC] curve comparing responses to T and C), and SRD (success rate difference), selected for clinical interpretability. We discuss the computation of standard errors and confidence intervals. We will conclude by introducing the important unsolved problem: the determination of the threshold of clinical significance for power computations and interpretation of RCT results.

Statistical and Clinical Significance, Power, and Meta-Analysis

As statistical hypothesis testing is typically performed, a "statistically significant" result with $p < .05$ means that the data indicate that something nonrandom is going on. When $p < .01$, the evidence is more convincing, and $p = 10^{-6}$ very convincing indeed. However, the *p* value is a comment on how convincing the data are against the null hypothesis of randomness; the conclusion is always "something nonrandom is going on." Such a conclusion gives no clue as to the size or importance of the nonrandom effect. To judge the clinical significance of a statistically significant finding, an effect size is needed.

Conversely, a "nonsignificant" result means that the data are not sufficient to support a contention of nonrandomness—a comment on the quality of the data, not the quality of T versus C. It is a serious misinterpretation to suggest that a "nonsignificant" test comparing T and C indicates equivalent effects or to use terms such as "marginally significant," "a trend toward signifi-

From the Department of Psychiatry and Behavioral Sciences (HCK), Stanford University, Stanford, California; and the Department of Psychiatry (DJK), Western Psychiatric Institute and Clinic, University of Pittsburgh, Pittsburgh, Pennsylvania.

Address reprint requests to Helena Chmura Kraemer, Stanford University, Department of Psychiatry and Behavioral Sciences, 401 Quarry Road MC5717, Stanford, CA 94305. E-mail: hck@stanford.edu

Received June 21, 2005; revised August 15, 2005; accepted September 21, 2005.

0006-3223/06/$32.00
doi:10.1016/j.biopsych.2005.09.014

cance," and so on in reporting a nonstatistically significant result. These are merely different ways of saying that the study was not designed quite well enough to be able to establish a nonrandom difference between T and C by conventional standards ($p < .05$). Asking for "post hoc" power calculations, too, is troublesome (Levine and Ensome 2001; Tukey 1993). Preferable would be an effect size whose possible clinical significance could be evaluated and thus a judgment made as to whether pursuing the effect in future, better-designed RCTs remains warranted.

In planning an RCT, for a legitimate 5% statistical test, there must be "*a priori*" assurance that whenever T is no different from C (two-tailed test) or T is no better than C (one-tailed test), with data collected and analyzed according to the proposal, the probability of declaring a result significant at the 5% level is no more than 5%. To demonstrate adequate power for their proposal, investigators need to declare a threshold of clinical significance, an effect size below which the T versus C difference would not be considered clinically significant (Cohen 1988; Kraemer and Thiemann 1987), and to show that the probability of declaring a result significant at the 5% level is greater than, say, 80% whenever the true effect size is above that threshold. The true effect size is, of course, unknown at the time of the proposal—if it were known, why do the study? The rationale and justification for the proposal should, however, indicate that the true effect size is likely to be greater than the threshold of clinical significance. (For illustration we hereafter focus on a 5% one-tailed test, with 80% power.)

These criteria are graphically summarized in Figure 1. Here, we use an as-yet-undefined effect size that is zero when the response to T is the same as that to C; positive when T is better than C, topping out at +1, when every single patient given T has a better response than every single patient given C; negative when C is better than T, bottoming out at −1, when every single patient given C has a better response than every single patient given T. The larger the difference between T and C, the larger the magnitude of the effect size. Whenever the null hypothesis is true (here a zero or negative effect size), the probability of a significant result is 5% or less; whenever the true effect size is greater than the threshold value, the probability of declaring it "statistically significant" is greater than 80%.

One common mistake in RCT design is guaranteeing adequate power, not at or above the threshold of clinical significance, but at or above the desired or hoped-for effect size or one based on very optimistic, underpowered, pilot studies. Then the RCT will typically be underpowered. In such studies, clinically significant results might well *not* be found statistically significant. In contrast, in an RCT powered correctly, it will seldom happen (less than 5% of the time) that T will be declared statistically significantly better than C when it is not, and it will seldom happen (far less than 20% [100%–80%] of the time) that a T that is clinically significantly better than C will not be declared statistically significantly better as well.

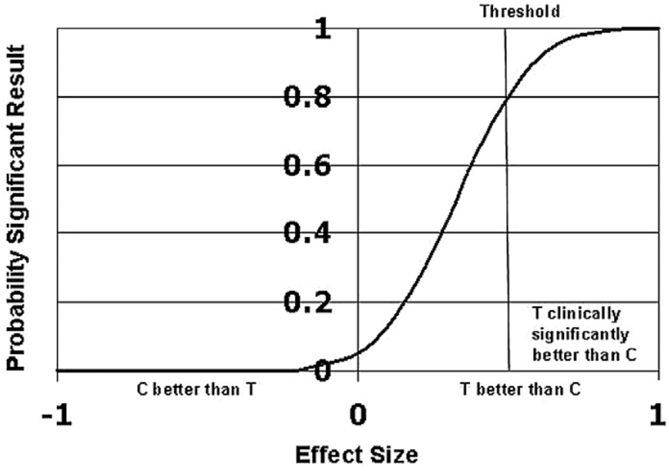

Figure 1. The operating characteristic of a 5% (one-tailed) test of statistical significance, with greater than 80% power to detect any effect above the threshold of clinical significance. T, treatment; C, control.

Finally, one RCT seldom definitively determines that T is clinically significantly better than C—independent replication is necessary. In recent years, the preferred method to seek the consensus of multiple RCTs comparing T and C is meta-analysis (Cooper and Hedges 1994; Cooper and Rosenthal 1980; Hedges and Olkin 1985). In a meta-analysis, each RCT yields an effect size comparing T and C. The homogeneity of effect sizes is examined, and in absence of heterogeneity, the effect sizes are pooled and the confidence interval computed. If that confidence interval lies completely above the threshold of clinical significance, it is established that T is clinically better than C. If the confidence interval lies below that threshold (even if above the null value), then it is established that T is not clinically significantly better than C. If the confidence interval straddles the threshold of clinical significance, more RCTs are needed to resolve the issue.

To summarize, the effect sizes seen in previous related studies guide both the choice of the effect size that sets the threshold of clinical significance appropriate in the medical context and the rationale and justification to believe that the true effect size is above that threshold and thus worth pursuing in an RCT. Then the effect size that determines the threshold of clinical significance is used to determine the necessary sample size for the proposed RCT. When the RCT is done, the data generated are used to estimate the true effect size and its confidence interval. Clinical significance is assessed by comparison of the estimated true effect size to the effect size set as the threshold of clinical significance. In any subsequent meta-analysis, this effect size might then be combined with other effect sizes from replications, ultimately to determine whether T is clinically significantly better than C. Thus, effect sizes play an important role both in designing RCTs and in interpreting their results; but specifically which effect size?

The Most Common Effect Sizes: Cohen's *d* and Odds Ratio

Cohen's *d*

When an RCT outcome measure is scaled, the most common effect size is Cohen's *d* (Cooper and Hedges 1994; Hedges and Olkin 1985), the difference between the T and C group means, divided by the within-group standard deviation. This effect size was designed for the situation in which the responses in T and C have normal distributions with equal standard deviations.

The population parameter estimated by Cohen's *d* ranges across the real line, with zero indicating no difference between T and C, clearly a different scale from the hypothetical one in Figure 1. Clinicians, consumers, and researchers often find interpretation of the magnitude of an effect size in which "large" values can theoretically approach infinity daunting. For that reason, the scaling between −1 and +1 of the hypothetical effect size in Figure 1 is preferable; however, scaling is never a serious problem. Here one might rescale d to $r = d/(d^2+ 4)^{1/2}$

(Rosenthal 1994; Rosenthal and Rubin 2003) or to $2\Phi(d/\sqrt{2}) - 1$ (to be discussed below), where $\Phi()$ is the cumulative distribution function of the standard normal distribution. Both would correspond to the scaling in Figure 1 but would not generally equal each other (see Table 1).

In this context, Cohen suggested that $d = .2$, .5, and .8 are "small," "medium," and "large" on the basis of his experience as a statistician, but he also warned that these were only "rules of thumb" (Cohen 1988). An effect size of .2 could be "large" in some contexts (e.g., Salk vaccine to prevent polio in the general pediatric population) or an effect size of .8 "small" in others (e.g., a drug treatment for depression that substantially increases the risk of suicide).

For statisticians accustomed to dealing with normal distributions, Cohen's *d* is informative because it creates a mental image of overlapping standard normal curves, the degree of overlap determined by *d*; however, for clinicians and consumers, Cohen's *d* carries no clinically interpretable message. Nevertheless, because power computations for the common two-sample *t* tests are completely determined by *d*, any effect size we recommend must be easily converted to Cohen's *d* in those circumstances (normal distributions, equal variances) for which *d* is appropriate.

Odds Ratio

For binary outcome measures in RCTs (success/failure), the most common effect size in current use is the odds ratio (OR) (Grissom and Kim 2005). If the success rate in T is s_T and that in C is s_C, then $OR = [s_T/(1 - s_T)]/[s_C/(1 - s_C)]$. Odds ratio is not scaled like the hypothetical effect size of Figure 1, but again that is of little consequence. Often used instead are the γ coefficient $[\gamma = (OR - 1)/(OR + 1)]$ or Yule's Index $[Y = (OR^{1/2} - 1)/(OR^{1/2} + 1)]$, each of which differently rescales the OR to correspond to an effect size like that in Figure 1.

Table 1. Cohen's *d* and Its Rescaling *r* for Outcome Data Having Normal Distributions with Equal Variances in the Treatment and Control Groups, Translated to the Equivalent Values of AUC, SRD, and NNT

Cohen's *d*	*r*	AUC	SRD	NNT	*n* for 80% Power
−∞	−1.000	.000	−1.000	−1.000	
−1.0	−.447	.240	−.521	−1.921	
−.9	−.410	.262	−.475	−2.103	
−.8[a]	−.371	.286	−.428	−2.334	
−.7	−.330	.310	−.379	−2.636	
−.6	−.287	.336	−.329	−3.043	
−.5[a]	−.243	.362	−.276	−3.619	
−.4	−.196	.389	−.223	−4.490	
−.3	−.148	.416	−.168	−5.953	
−.2[a]	−.100	.444	−.112	−8.892	
−.1	−.050	.472	−.056	−17.739	
0[a]	.000	.500	.000	∞	∞
.1	.050	.528	.056	17.739	1,220
.2[a]	.100	.556	.112	8.892	306
.3	.148	.584	.168	5.953	139
.4	.196	.611	.223	4.490	107
.5[a]	.243	.638	.276	3.619	61
.6	.287	.664	.329	3.043	39
.7	.330	.690	.379	2.636	26
.8[a]	.371	.714	.428	2.334	20
.9	.410	.738	.475	2.103	16
1.0	.447	.760	.521	1.921	13
∞	1.000	1.000	1.000	1.000	∞

A sample size necessary to achieve 80% or more power with a 5% one-tailed test is also presented. AUC, area under the receiver operating characteristic curve; SRD, success rate difference; NNT, number needed to treat.

[a] Cohen's suggestions for "small," "medium," and "large" effect sizes (positive when treatment is better than control, negative otherwise).

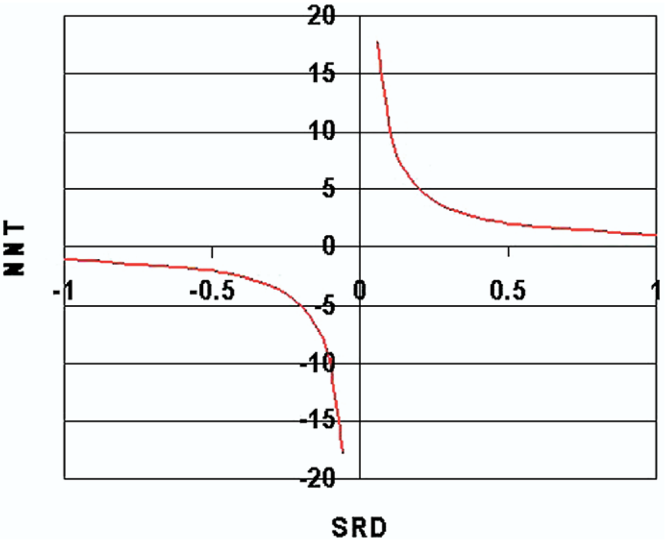

Figure 2. The relationship between NNT (number needed to treat) and SRD (success rate difference).

The situation with OR is quite different from that with Cohen's *d*. It is never possible to propose a sample size large enough so that one would have 80% or more power to detect all ORs exceeding any fixed threshold. The OR was originally introduced to test the null hypothesis of randomness and still remains the most sensitive indicator of whether the success rates differ nonrandomly (e.g., in logistic regression analyses). Odds ratio gained popularity in an effort to salvage retrospective case–control studies (Cornfield 1951, 1956), not in RCTs. There are many reasons to recommend against its use as an effect size (Fleiss 1970; Kraemer 2003; Kraemer et al 1999; Sackett 1996). Consequently, although any effect size we recommend should bear some relationship to Cohen's *d* in those situations in which Cohen's *d* is appropriate, we would not necessarily expect any relationship to the OR in those situations in which OR is appropriate.

Recommended Equivalent Effect Sizes: NNT, AUC, SRD

Number Needed to Treat

The effect size proposed that seems to best reflect clinical significance is one proposed in the context of evidence-based medicine for binary (success/failure) outcomes: NNT (Altman and Andersen 1999; Cook and Sackett 1995). Number needed to treat is defined as the number of patients one would expect to treat with T to have one more success (or one less failure) than if the same number were treated with C. For a binary outcome (success/failure), the success rate difference (SRD) is defined as $(s_T - s_C)$, and $NNT = 1/(s_T - s_C)$. For T better than C, NNT ranges from the ideal value of 1 to infinity; for T worse than C, NNT ranges from −1 to minus infinity.

In Table 1 and Figure 2, we see the major problem with NNT. There is a major discontinuity in NNT when the difference in success rates nears zero. When s_T and s_C are similar, the difference in success rates in different studies might wobble back and forth across zero and thus NNT between positive and negative infinity. Because of this instability, to obtain confidence intervals or to use NNT in meta-analysis might produce peculiar results and is not advisable. Thus, for computational purposes, one might prefer to use SRD, even though for interpretation purposes one might translate SRD to its equivalent NNT.

Another minor problem is that NNT was proposed only for binary outcomes (success/failure). Yet many RCTs have scaled outcome measures (e.g., time to remission, decrease in symptoms). To solve that problem, we turn first to AUC.

Area Under the ROC Curve

In any situation in which one can compare the clinical consequences experienced by two patients—in short, with any response on which an RCT can be based—one can use an effect size we call AUC (label to be explained below). If one sampled a T patient and a C patient, AUC is the probability that the T patient has a treatment outcome preferable to the C patient (where we toss a coin to break any ties) symbolically:

$$\text{AUC} = \text{probability } (\text{T} > \text{C}) + .5 \text{ probability } (\text{T} = \text{C}).$$

Thus, if AUC = .50, the T patient outcome is as likely as not to be better than that for the C patient (i.e., no effect), and AUC = 1.0 means that every T patient has an outcome better than that for every C patient. AUC has been called "The Common Language Effect Size" (McGraw and Wong 1992) or an "intuitive" effect size (Acion et al, in press), suggesting its relevance to interpreting clinical significance. Because AUC ranges from 0 to 1, to get the scaling of Figure 1, we can use 2AUC − 1.

"AUC" was originally generated as the area under the curve of the receiver operating characteristic curve (ROC) comparing T and C scaled responses, which provides graphic insight into what is going on. To generate that ROC, we take every value on the outcome scale and compute the proportion of T patients and the proportion of C patients with response clinically better than that value, again with ties randomly broken. For each possible value, this pair of points is located on a graph, called the ROC plane (see Figure 3). Connecting these points and the two corner points creates a "curve," the ROC curve. When the outcome measure is continuous, the ROC curve smoothly connects the two corner points of the ROC plane; when the outcome measure takes on discrete values (e.g., 3-, 4-, 5- ... point scale), the ROC "curve" is a series of connected straight lines.

"AUC" is the area under this curve. If there were no difference between T and C, the ROC curve would coincide with the main diagonal line, the random ROC (AUC = .5). The higher the ROC curve above the random ROC, the greater the advantage of T over C.

In Figure 3, for example, we show the situation in which the responses are normally distributed in both the T and C groups, but with a much larger variance in the T than in the C group (ROC1). That "bulge" of the ROC1 curve that crosses below the random ROC in Figure 2 results from the unequal variances in the two groups.

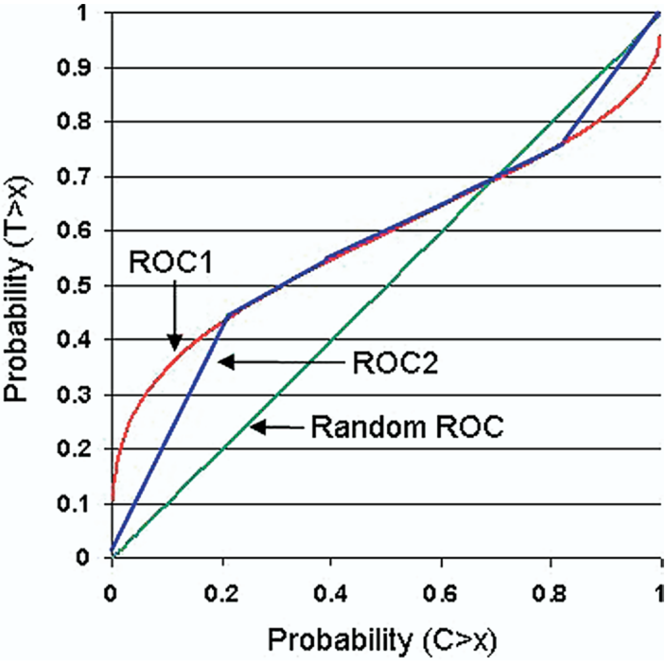

Figure 3. The ROC (receiver operating characteristic) plane with Probability (Treatment [T] Response > x) graphed against Probability (Control [C] Response > x) for all possible values of x in the range of the outcome measure. The ROC1 curve connects all the points for a continuous outcome measure with the two corner points of the ROC plane. The Random ROC is the diagonal line, which results when there is no difference in the T and C distributions. Shown also is the curve (ROC2) resulting from grouping of the original measure into a 5-point scale at the quintiles of the original distribution.

The computation of the AUC does not actually require drawing the ROC curve. When, as is typical, the RCT outcome measures are on some ordinal scale (e.g., decrease in symptoms, time to remission), one can compute the Mann-Whitney *U* statistic available in most statistical computer programs. Then AUC = *U*/mn, where m and n are the sample sizes in the two groups (Acion et al, in press). Even simpler, when the outcome is binary, success rates s_T and s_C in the treatment and control groups, then

$$\text{AUC} = .5(s_T - s_C + 1) = .5(\text{SRD} + 1).$$

If, instead of using a scaled response (continuous or not), we choose to use its logarithm, its square root, or any other rescaling, the ROC curve would not change. Area under the ROC curve is invariant under all rescaling of the outcome measure. This makes any effort to find some transformation that "normalizes" response irrelevant and presents a distinct mathematical advantage over the use of Cohen's *d*, which is invariant only under linear rescaling.

If we grouped the continuous responses, creating, say, a 5-point scale by cutting at the quintiles of the C-group distribution, the ROC "curve" of the grouped responses would be a sequence of straight lines connecting the two corners and the four points on the original ROC curve (ROC2 in Figure 3) corresponding to cut-points. Now the AUC would change, because any curved parts of the original ROC curve that lie above or below the straight lines connecting the points would be lopped off. In most cases, the effect size would become smaller. Because power depends on effect size, this means that larger sample sizes are usually necessary to detect treatment effects using 3-, 4-, 5- ... point response measures rather than the continuum.

Finally, if we chose to dichotomize the continuous response, there would be only one point in the ROC plane, and the "curve" would be a triangle perched above or below the random ROC, with its apex at the point of dichotomization on the original ROC curve. If, in Figure 3, that point lay at the point where ROC1 crosses the random ROC, an effective T would now seem to be no different from C, and if the point of dichotomization fell in the "bulge" of the ROC1 curve below the random ROC, the overall positive effect of T over C would turn negative!

There has long been a struggle between the use of categorical (binary) and dimensional (scaled) measures both in diagnosis (Kraemer and O'Hara 2004) and in clinical research. Warnings of the costs of dichotomization have long been expressed in the research literature (Cohen 1983; Kraemer and Thiemann 1987; MacCallum et al 2002; Veiel 1988), but dichotomization remains common in RCTs. The choice is often depicted as a struggle between statistical considerations favoring dimensional outcomes and clinical considerations favoring categorical outcomes. The ROC suggests that clinical considerations should also favor scaled RCT outcome measures. The problem with dichotomization lies in the compromise in ability to discern response differences between patients, a clinical as well as statistical loss.

SRD Expanded

An effect size closely related to AUC is an expanded version of the SRD (Hsu 2004), originally defined only for binary outcome: the difference between the probability that a T patient has a treatment outcome preferable to a C patient and the probability that a C patient has a treatment outcome preferable to a T patient, symbolically:

$$\mathrm{SRD} = \mathrm{probability}(\mathrm{T} > \mathrm{C}) - \mathrm{probability}(\mathrm{T} < \mathrm{C}).$$

Once again, this effect size has appeared in the literature under several different names (Cliff 1993; Grissom and Kim 2005). The scale of SRD matches that in Figure 1, and SRD = 2AUC − 1.

When the response measure is binary, SRD is still $s_T - s_C = 1/\mathrm{NNT} = 2\mathrm{AUC} - 1$. Thus, with binary outcome, NNT can easily be converted either to SRD or to AUC, and vice versa. With scaled outcome measures, each T patient can be declared a "success" if s/he has a response better than a randomly selected C patient; each C patient can be declared a "success" if s/he has a response better than a randomly selected T patient. Then even with a scaled outcome, NNT is the number of patients to whom one would have to give T, to expect one more success or one less failure than if the same number had been given C. That resolves the one remaining problem with NNT, its limitation to binary outcomes. With a scaled outcome NNT = 1/SRD = 1/(2AUC − 1), exactly as with binary outcome. To develop insights and for computation from data, AUC is preferred; for confidence interval estimation or meta-analysis, SRD is preferred; to interpret clinical significance, NNT is preferred. The different equivalent forms simply facilitate different tasks and insights.

When Cohen's *d* is appropriate (normal distributions, equal variances), AUC estimates the same parameter as does $\Phi(d/\sqrt{2})$. Thus, conversion from SRD or NNT to Cohen's *d*, or vice versa, when appropriate, is readily done. Thus Cohen's "small," "medium," and "large" effect sizes would correspond to NNT = 8.9, 3.6, and 2.3 (Table 1).

Although we would neither expect nor demand it, it is reasonable to ask whether there is any association at all between the commonly used OR, on the one hand, and NNT on the other. The answer is both surprising and informative. Could OR somehow translate to NNT? If we consider all possible pairs of success rates, s_T and s_C, having a fixed nonzero value of OR (e.g., OR = 4.0) then the best (lowest) possible value of NNT is $(\mathrm{OR}^{1/2} + 1)/(\mathrm{OR}^{1/2} - 1)$ [(2 + 1)/(2 − 1) = 3], but NNT could have any value worse (larger) than that. In Figure 4, we present the best possible value of NNT for a range of possible values of OR. The only value of the OR meaningful for clinical significance is OR = 1, which unequivocally means that NNT is infinite. Any other value of the OR, no matter how large, might correspond to a very poor (large) NNT. Unfortunately, the fact that clinical effects that otherwise evaluated seem of trivial importance often have large ORs might be one possible reason for the continued popularity and wide usage of the OR.

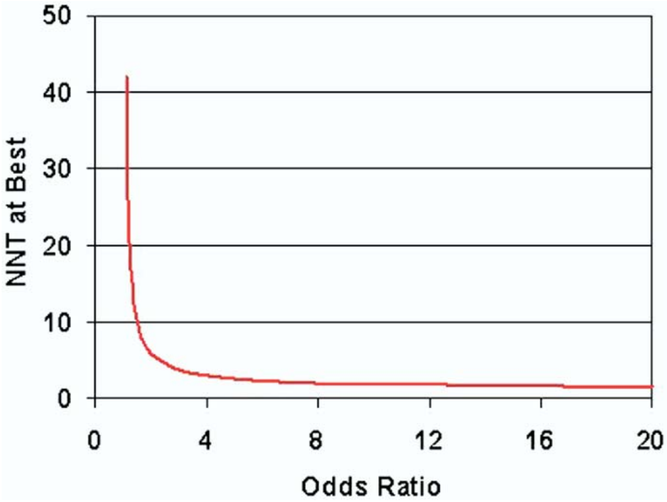

Figure 4. The best possible value NNT (number needed to treat) that corresponds to any particular odds Ratio. Any value of NNT above the curve is a possible value given that odds ratio.

Confidence Intervals and Effect Sizes

In every report of an RCT, we recommend that each *p* value be accompanied by NNT (for interpretability) and SRD with its standard error and confidence interval (for computations). The difficulty is that the correct computation of the confidence interval and the standard error of SRD depends on the distribution of the data underlying that effect size.

In those circumstances in which Cohen's *d* is appropriate (normal distributions, equal variances), the exact distribution of Cohen's *d* is known (Hedges and Olkin 1985). Because SRD is a known function of Cohen's *d*, it is possible to derive exact or approximate confidence intervals or standard errors. With binary outcome, the observed s_T and s_C are independent binomial proportions. The exact distribution of SRD is thus known, and exact or approximate confidence intervals or standard errors can be derived; however, these two situations cover very few of the distributions of outcome measures that might be encountered in RCTs.

In general, we would recommend use of the bootstrap method (Efron and Gong 1983; Efron and Tibshirani 1995) to generate both standard errors and confidence intervals. This is a computer-intensive method that would not have been feasible years ago but today is readily available. Were there consensus regarding the use of SRD as the effect size for use in RCTs, developing a computer program that would rapidly yield the standard errors and confidence intervals in any RCT would be a simple matter.

Discussion: The Threshold of Clinical Significance

To summarize, we propose that for any RCT, along with reporting the *p* value comparing T with C, researchers report NNT and SRD, as well as the standard error and a confidence interval for SRD. If effect sizes were so reported, they could then be used to facilitate consideration of what the threshold of clinical significance might be for design of subsequent related studies.

Here we have attempted to take the first major step, recommending an effect size that is clinically interpretable and statistically justifiable. One objection to this proposal is that this effect size, as do all others, reflects effect on a population level, not what is truly desired, the expected effect for the individual patient. However, because we can never observe any single individual both with treatment and without, it is never possible to estimate a truly individual effect size. At best, we can estimate the effect of T versus C on patients very like an individual patient. The recent emphasis on moderators of treatment response (Kraemer et al 2002, 2005)—that is, the search for pre-RCT patient characteristics (to identify "very like") associated with different effect sizes—reflects the attempt to satisfy this need. The different NNTs, in the moderated subgroups so defined, approach the individual effect sizes both clinicians and consumers need and want.

The most challenging and urgent task remains unsolved: developing the principles that underlie the thresholds of clinical significance in different clinical contexts. This is a task with which biostatisticians have long struggled in designing RCTs, because they need such a threshold to do power calculations, but, in absence of clinical knowledge of the field, they can only arbitrarily set one. On the other hand, clinical researchers have the clinical knowledge of the field necessary to set the threshold but have not to date had the language to translate this knowledge to a form usable for power computations. We have attempted to provide such a language.

In Table 1, we present the sample size necessary to achieve at least 80% power to detect the associated effect size whether the test is the two-sample *t* test (normal distributions, equal variances) or the 2 × 2 χ^2 tests (binary outcome). If one considers these sample sizes in comparison with sample sizes reported for RCTs for treatment of mental health disorders, it seems that, in general, RCT researchers have set their thresholds at NNT approximately 2–4, suggesting that unless NNT is this good, the treatment is of no clinical significance. Is this always true?

Rosenthal et al (2000) ask the crucial question: "How big an effect size is 'important'?" They compare the effect sizes in several important medical contexts. From their results, the NNT associated with the regular use of aspirin to reduce the risk of heart attack was 130. The NNT associated with the use of cyclosporine in the prevention of organ rejection was 6.3. The NNT associated with effectiveness of psychotherapy was 3.1. The regular use of aspirin has become common practice. Cyclosporine to prevent organ rejection is considered a medical breakthrough of considerable practical importance. Yet the effectiveness of psychotherapy, here with the "best" NNT, is often considered "modest." If we accept these effect sizes as true indications of clinical significance, what sense does this make?

The result of ineffective T in aspirin use and cyclosporine use was disability or death, whereas the result of ineffective T in psychotherapy was "less benefit." It is very likely that the more serious the clinical consequences of ineffective treatment, the higher (less stringent) the threshold NNT is likely to be for clinical significance. Moreover, daily aspirin is cheap, relatively low risk, and does not place too much of a burden on patients. Cyclosporine is more costly and risky but still not very burdensome. On the other hand, psychotherapy is costly, although low risk, but burdensome to patients. The less invasive (in terms of cost, risk, burden) is T relative to C, the higher the threshold NNT is likely to be for clinical significance.

Consequently, it does make sense that the threshold of clinical significance for a low-cost, low-risk vaccine or other prevention intervention to protect against a possibly disabling or fatal disease in a low-risk population (i.e., polio vaccine) might have a threshold NNT of 100 or even 1000. However, surgical, or radiation treatment to reduce the likelihood of the same type of outcome of disability or death in a high-risk population (as in the cyclosporine case) might have a threshold value of NNT closer to 10, given the costs and risks of the treatment. If the outcome of psychotherapy research had been to induce early remission from major depressive disorder, it might have had a threshold value of NNT closer to 5, but for the vaguer outcome of "less benefit," the threshold NNT might indeed be closer to 2.

This is far from a complete answer, but it represents the kind of discussion that has yet to take place to determine the factors that influence the threshold of clinical significance and to provide guidelines for researchers proposing RCTs, reviewers evaluating those proposals, and clinicians and consumers hoping to apply the results of RCTs to guide clinical decision making. The result of such discussions would be RCTs that are better designed and reported, and that quite possibly might be more influential in improving clinical practice.

This work was supported in part by grant MH30915 from the National Institute of Mental Health.

HCK is a consultant for Eli Lilly and Company. DJK is on the Advisory Boards of Pfizer, Forest Pharmaceuticals, and Solvay/Wyeth Pharmaceuticals, and serves as a Consultant to Servier Amerique.

Acion L, Peterson JJ, Temple S, Anrndt S (in press): Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat Med*.

Altman DG, Andersen K (1999): Calculating the number needed to treat for trials where the outcome is time to an event. *Br Med J* 319:1492–1495.

Borenstein M (1994): The case for confidence intervals in controlled clinical trials. *Control Clin Trials* 15:411–428.

Borenstein M (1997): Hypothesis testing and effect size estimation in clinical trials. *Ann Allergy Asthma Immunol* 78:5–16.

Borenstein M (1998): The shift from significance testing to effect size estimation. In: Bellak AS, Hersen M, editors. *Research and Methods, Comprehensive Clinical Psychology,* volume 3. Burlington, Maryland: Elsevier Science, 319–349.

Cliff N (1993): Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol Bull* 114:494–509.

Cohen J (1983): The cost of dichotomization. *Appl Psychol Measurement* 7:249–253.

Cohen J (1988): *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen J (1995): The earth is round (p<.05). *Am Psychol* 49:997–1003.

Cook RJ, Sackett DL (1995): The number needed to treat: A clinically useful measure of treatment effect. *Br Med J* 310:452–454.

Cooper H, Hedges LV (1994): *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Cooper HM, Rosenthal R (1980): Statistical versus traditional procedures for summarizing research. *Psychol Bull* 87:442–449.

Cornfield J (1951): A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J Natl Cancer Inst* 11:1269–1275.

Cornfield J (1956): A statistical problem arising from retrospective studies. In: Neyman J, editor. *Proceedings of the Third Berkeley Symposium,* volume IV. Berekely, California: University of California Press, 135.

Dar R, Serlin RC, Omer H (1994): Misuse of statistical tests in three decades of psychotherapy research. *J Consult Clin Res* 62:75–82.

Efron B, Gong G (1983): A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Statistician* 37:36–48.

Efron B, Tibshirani R (1995): *Computer-Intensive Statistical Methods (Technical Report 174)*. Stanford, California: Division of Biostatistics, Stanford University.

Fleiss JL (1970): On the asserted invariance of the odds ratio. *Br J Prev Soc Med* 24:45–46.

Grissom RJ, Kim JJ (2005): *Effect Sizes for Research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hedges LV, Olkin I (1985): *Statistical Methods for Meta-Analysis*. Orlando: Academic Press.

Hsu LM (2004): Biases of success rate differences shown in binomial effect size displays. *Psychol Bull* 9:183–197.

Hunter JE (1997): Needed: A ban on the significance test. *Psychol Sci* 8:3–7.

Jacobson NS, Truax P (1991): Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 59:12–19.

Kraemer HC (1993): Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoneuroendocrinology* 17:527–536.

Kraemer HC (2003): Reconsidering the odds ratio as a measure of 2×2 association in a population. *Stat Med* 23:257–270.

Kraemer HC (in press): A simple effect size indicator for two-group comparisons? A comment on $r_{equivalent}$. *Psychol Methods*.

Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ (1999): Measuring the potency of a risk factor for clinical or policy significance. *Psychol Methods* 4:257–271.

Kraemer HC, Lowe KK, Kupfer DJ (2005): *To Your Health: How to Understand What Research Tells Us About Risk*. Oxford, United Kingdom: Oxford University Press.

Kraemer HC, O'Hara R (2004): Categorical versus dimensional approaches to diagnosis: Methodological challenges. *J Psychiatr Res* 38:17–25.

Kraemer HC, Thiemann S (1987): *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, California: Sage Publications.

Kraemer HC, Wilson GT, Fairburn CG, Agras WS (2002): Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry* 59:877–883.

Krantz DH (1999): The null hypothesis testing controversy in psychology. *J Am Stat Assoc* 44:1372–1381.

Levine M, Ensome M (2001): Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy* 21:405–409.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002): On the practice of dichotomization of quantitative variables. *Psychol Methods* 7:19–40.

McGraw KO, Wong SP (1992): A common language effect size statistic. *Psychol Bull* 111:361k–365k.

Nickerson RS (2000): Null hypothesis significance testing: A review of an old and continuing controversy. *Psychol Methods* 5:241–301.

Rosenthal R (1994). Parametric measures of effect size. In: Cooper H, Hedges LV, editors. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 231–244.

Rosenthal R, Rosnow RL, Rubin DB (2000): *Contrasts and Effect Sizes in Behavioral Research*. Cambridge, United Kingdom: Cambridge University Press.

Rosenthal R, Rubin DB (2003): $r_{equivalent}$: A simple effect size indicator. *Psychol Methods* 8:492–496.

Sackett DL (1996): Down with the odds ratio! *Evidence Based Med* 1:164–166.

Shrout PE (1997): Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychol Sci* 8:1–2.

Thompson B (1999): Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educ Psychol Rev* 11:157–169.

Tukey JW (1993): Tightening the clinical trial: Nonrelevancy of power calculations after the fact (Appendix 1). *Control Clin Trials* 14:266–285.

Veiel HOF (1988): Base-rates, cut-points, and interaction effects: The problem with dichotomized continuous variables. *Psychol Med* 18: 703–710.

Wilkinson L, The Task Force on Statistical Inference (1999): Statistical methods in psychology journals: Guidelines and explanations. *Am Psychol* 54:594–604.