Measures of Clinical Significance

HELENA CHMURA KRAEMER, PH.D., GEORGE A. MORGAN, PH.D., NANCY L. LEECH, PH.D., JEFFREY A. GLINER, PH.D., JERRY J. VASKE, PH.D., AND ROBERT J. HARMON, M.D.

Behavioral scientists are interested in answering three basic questions when examining the relationships between variables (Kirk, 2001). First, is an observed result real or should it be attributed to chance (i.e., statistical significance)? Second, if the result is real, how large is it (i.e., effect size)? Third, is the result large enough to be meaningful and useful (i.e., clinical or practical significance)? In this last column in the series, we treat clinical significance as equivalent to practical significance.

Judgments by the researcher and the consumers (e.g., clinicians and patients) regarding clinical significance consider factors such as clinical benefit, cost, and side effects. Although there is no formal statistical test of clinical significance, researchers suggest using one of three types of effect size measures to assist in interpreting clinical significance. These include the strength of association between variables (r family effect size measures), the magnitude of the difference between treatment and comparison groups (d family effect size measures), and measures of risk potency. In this paper, we review the d and r effect size measures and five measures of risk potency: odds ratio, risk ratio, relative risk reduction, risk difference, and number needed to treat. Finally, we review a relatively new effect size, AUC (which for historical reasons irrelevant to the current discussion stands for area under the receiver operating characteristic [ROC] curve), that integrates many of the others and is directly related to clinical significance. Each of these measures, however, has limitations that require the clinician to be cautious about interpretation. Guidelines are offered to facilitate the interpretation and understanding of clinical significance.

Problems With Statistical Significance

A statistically significant outcome indicates only that there is likely to be at least some relationship between the variables. In other words, the p value indicates the probability that an outcome this extreme could happen, if the null hypothesis were true. Statistical significance, however, does not provide information about the strength of the relationship (effect size) or whether the relationship is meaningful (clinical significance).

Sometimes researchers misinterpret statistically significant results as showing clinical significance. However, it is quite possible, with a large sample, to have a statistically significant result from a weak relationship between variables (i.e., a small effect size). Outcomes with lower p values are sometimes misinterpreted as having stronger effects than those with higher p values. For example, an outcome of p < .01 is implied to have a stronger effect than an outcome of p < .05. In part, this misconception stems from the fact that often smaller p values, given a constant sample size, are correlated with larger effect sizes. In addition, sometimes researchers misinterpret non-statistically significant results as “proving” the null hypothesis, for example as showing that two treatments have equivalent effects. However, a non-significant result may not be due to a small effect size but to unreliable measures, inadequate designs, insensitive analyses, or other determinants of low power. Because the presence or absence of statistical significance does not give information about the size or importance of the outcome, it is critical to know the effect size (i.e., the strength of the relationship between the independent variable and the dependent variable).

Effect Size Measures

Statisticians have proposed many effect size measures. They fall mainly into three types or families: the r family, the d family, and measures of risk potency.

The r Family. One method of expressing effect sizes is in terms of strength of association, with statistics such as the Pearson product moment correlation coefficient, r, used when both the independent and the dependent measures are ordered. Such effect sizes vary between −1.0 and +1.0, with

Accepted July 9, 2003.

Dr. Kraemer is Professor of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA. Dr. Morgan is Professor Emeritus of Education and Human Development, Colorado State University, Fort Collins, and Clinical Professor of Psychiatry, University of Colorado School of Medicine, Denver. Dr. Leech is Assistant Professor of Research, Evaluation and Measurement, School of Education, University of Colorado, Denver. Dr. Gliner is Professor of Occupational Therapy and Education, Colorado State University, Fort Collins. Dr. Vaske is Professor, Human Dimensions in Natural Resource Unit, Colorado State University, Fort Collins. Dr. Harmon is Professor of Psychiatry and Pediatrics and Director of The Irving Harris Program in Child Development and Infant Mental Health, University of Colorado School of Medicine, Denver.

The authors thank Nancy Plummer for preparing this and each of the previous manuscripts in the series; her help was invaluable.

Correspondence to Dr. Harmon, CPH Room 2K04, UCHSC Box C268-52, 4200 East Ninth Avenue, Denver, CO 80262; e-mail: robert.harmon@uchsc.edu.

DOI: 10.1097/01.chi.0000091507.46853.d1
0 representing no effect. This family of effect sizes also includes associational statistics such as the Spearman or Kendall rank correlation coefficients, and the multiple correlation coefficient ($R$).

The $d$ Family. These effect sizes are used when the independent variable is binary (dichotomous) and the dependent variable is ordered. When comparing two groups, the effect size ($d$) can be computed by subtracting the mean of the second group from the mean of the first group and dividing by the pooled standard deviation of both groups. Other $d$ family effect sizes use somewhat different formulas, but they all express the mean difference in standard deviation units. This effect size ranges from minus to plus infinity, with zero indicating no effect; however, it is unusual to find $d$ values in the applied behavioral sciences much greater than 1.

Measures of Risk Potency. These effect sizes are used when both the independent and the dependent variable are binary. There are many such effect sizes, but here we discuss five common ones: odds ratio, risk ratio, relative risk reduction, risk difference, and number needed to treat (NNT). Odds ratios and risk ratios vary from 0 to infinity, with 1 indicating no effect. Relative risk reduction and risk difference range from −1 to 1, with zero indicating no effect. NNT ranges from 1 to plus infinity, with very large values indicating no treatment effect.

$AUC$. Finally we discuss one index that can be used when the independent variable is binary but the dependent variable can be either binary or ordered (Grissem, 1994). This measure ranges from 0% to 100%, with 50% indicating no effect. $AUC$ has a unique status in that it was originally proposed to address clinical significance.

We discuss the advantages and disadvantages of each of the above measures as indicators of clinical significance later in this column. In this discussion, we focus on positive association only; that is, effect sizes ranging from whatever value indicates no effect to that indicating maximal effect. (If association were negative, one need only change the sign of one variable, or switch two categories to make a negative association positive.)

**Issues About Effect Size Measures**

Unfortunately, there is little agreement about which effect size to use for each situation. The most commonly discussed effect size in the behavioral sciences, especially for experiments, is $d$, but the correlation coefficient, $r$, and other measures of the strength of association are common in survey research. In medical journals, an odds ratio is most common.

Wilkinson and the American Psychological Association Task Force on Statistical Inference (1999) stated that effect sizes should always be reported for primary results. In the future, it is likely that results in most behavioral science journal articles will include effect sizes for significant results. In addition, because the results of a single study might be used later in a meta-analysis, we suggest providing effect sizes or the information necessary to compute them (e.g., mean, SD) even if the relationship is not significant statistically.

**Interpreting $d$ and $r$ Effect Sizes**

Table 1 provides general guidelines for interpreting the size of the effect for five measures discussed in this column. Cohen (1988) provided research examples of what he labeled small, medium, and large effects suggested by $d$ and $r$ values. Most researchers would not consider a correlation ($r$) of 0.5 to be very strong because only 25% of the variance in the dependent variable is predicted. However, Cohen argued that when the two variables measure different constructs, an $r$ of 0.3 is typical and 0.5 is about as large as correlations in applied behavioral sciences get. When, as in test–retest reliability measures, the two variables measure the same construct, the standards increase.

Cohen pointed out that effects with a $d$ of 0.8 are “grossly perceptible and therefore large differences, as (for example) the mean difference in height between 13- and 18-year-old girls” (p. 27). Cohen’s medium size effect is “visible to the naked eye. That is, in the course of normal experiences, one would become aware of an average difference in IQ between clerical and semi-skilled workers” (p. 26). Kazdin and Bass (1989), based on a review of psychotherapy research, found

**TABLE 1**

<table>
<thead>
<tr>
<th>Strength of a Relationship</th>
<th>$d$</th>
<th>$r$</th>
<th>$2 \times 2$ Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Interpretation of the Strength (Effect Size) of a Positive Relationship</td>
<td></td>
<td></td>
<td>AUC (%)</td>
</tr>
<tr>
<td>Much larger than typical</td>
<td>$\geq 1.00$</td>
<td>$\geq 0.70$</td>
<td>$\geq 76$</td>
</tr>
<tr>
<td>Large or larger than typical</td>
<td>0.80</td>
<td>0.50</td>
<td>71</td>
</tr>
<tr>
<td>Medium or typical</td>
<td>0.50</td>
<td>0.30</td>
<td>64</td>
</tr>
<tr>
<td>Small or smaller than typical</td>
<td>0.20</td>
<td>0.10</td>
<td>56</td>
</tr>
</tbody>
</table>

*Note: We interpret the numbers in this table as a range of values. For example, a $d$ greater than 0.90 (or less than −0.90) would be described as much “larger than typical” in the applied behavioral sciences, a $d$ between say 0.70 and 0.90 would be called “larger than typical,” and a $d$ between say 0.60 and 0.70 would be “typical to larger than typical.”*
that \( d \) was approximately 0.8 when comparing a new active treatment against an inactive (treatment withheld) placebo, and they found that \( d \) was about 0.5 when comparing one active treatment against another active treatment.

The \( d \) and \( r \) guidelines in Table 1 are based on the effect sizes commonly found in studies in the applied behavioral sciences. They do not have absolute meaning; Cohen’s “large,” “medium,” and “small” were meant to be relative to typical findings in behavioral research in general. For that reason, we suggest using “larger than typical” instead of “large,” “typical” instead of “medium,” and “smaller than typical” instead of “small.” However, as suggested by the Kazdin and Bass (1989) results, it is advisable to examine the research literature to see if there is information about typical effect sizes in that context. The standards then expressed in Table 1 would need to be adjusted accordingly.

There are disadvantages of the \( d \) and \( r \) effect sizes as measures of clinical significance. First, they are relatively abstract and consequently may not be meaningful to patients and clinicians or even to researchers. They were not originally intended to be indexes of clinical significance and are not readily interpretable in terms of how much individuals are affected by treatment.

**Clinical Significance**

The clinical significance of a treatment is based on external standards provided by clinicians, patients, and/or researchers. Unfortunately, to date there is little consensus about the criteria for these efficacy standards. Several such criteria are a lower percentage of treated clients than comparisons with negative outcomes or at risk, elimination of the problem, or normative levels of function (meeting or exceeding the cut-score) at the end of treatment. Jacobson et al. (1999) defined clinical significance as a change to normal functioning due to therapy, and they suggested approaches for identifying patients who made statistically reliable changes that were clinically significant according to their definition.

**Interpreting Measures of Risk Potency**

Clinicians must make categorical decisions about whether or not to use a treatment (medication, therapy, hospitalization), and the outcomes also are often binary. For example, a child is classified as having ADHD or not, or is at risk for some negative outcome or not. In comparing two treatments, a positive outcome might indicate that the patient is sufficiently improved (or not) to meet the criteria for a clinically significant change. These binary decisions and outcomes provide data in a \( 2 \times 2 \) contingency table like those in Tables 2 and 3.

The phi coefficient, which applies the formula for the Pearson or Spearman correlation coefficient to \( 2 \times 2 \) data, is sometimes used here. For example, phi was computed from the outcome of giving the Salk vaccine to half of several hundred thousand children to see if it prevented polio. The effect of receiving the vaccine seemed by the \( r \) standards in Table 1 to be small (phi = 0.01). However, the more different the distributions of the two binary variables, the more restricted the range of phi. Thus, in the Salk vaccine case, much fewer than 1% in either group developed polio, while approximately half of the study sample (50%) got the vaccine. The maximal possible value of phi was considerably less than 1, actually 0.02. Thus, the observed phi was about halfway between random (0) and its maximal value (0.02), rather than 1% of the way if judged on the usual \( r \) scale. For this reason it is difficult, if not impossible, to extract clinical meaning from phi.

All the measures described below are used when researchers and clinicians have a \( 2 \times 2 \) contingency table to express the risk of clinical level outcomes (i.e., the success or failure of a treatment). In some cases, a \( 2 \times 2 \) table results when initially continuous outcome data are dichotomized (e.g., when responses on an ordered outcome measure in a clinical trial are reclassified as “success” and “failure”). Such dichotomization not only results in a loss of information, but also, as Kraemer (1992) pointed out, dichotomizing can result in inconsistent and arbitrary effect size indexes due to different choices of the cut-point or threshold for failure, whatever effect size is used.

To illustrate these effect sizes and the associated interpretation problems, we present the examples in Tables 2 and 3. Although these are hypothetical examples, they are similar to

<p>| TABLE 2 |
|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>Failure (Not Improved and Somewhat) (%)</th>
<th>Success (Very Improved) (%)</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison group</td>
<td>84</td>
<td>16</td>
<td>5.25</td>
</tr>
<tr>
<td>Treatment group</td>
<td>62</td>
<td>38</td>
<td>1.63</td>
</tr>
</tbody>
</table>

*Note: Odds ratio = 3.22; risk ratio = 1.35; relative risk reduction = 26%; risk difference = 22%; number needed to treat = 4.55; area under the receiver operating characteristic curve = 61%.*

<p>| TABLE 3 |
|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>Failure (Not Improved) (%)</th>
<th>Success (Somewhat and Very Improved) (%)</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison group</td>
<td>68</td>
<td>32</td>
<td>2.13</td>
</tr>
<tr>
<td>Treatment group</td>
<td>10</td>
<td>90</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*Note: Odds ratio = 19.12; risk ratio = 6.80; relative risk reduction = 85%; risk difference = 58%; number needed to treat = 1.72; area under the receiver operating characteristic curve = 79%.*
The tables show the percentage of subjects in the treatment recently published data from a randomized controlled trial and comparison groups for each of two outcomes, failure and success. Table 2 shows results when “failure” includes both “not improved” and “somewhat improved” responses, while Table 3 includes only “not improved” as “failures.” Any response not classified “failure” is considered a “success.”

Odds Ratio. The odds ratio, the most commonly reported of these measures, is determined by first computing the odds, the ratio of the percentage judged to fail (failure rate) to the percentage judged as successes (success rate) within both the comparison and intervention groups. In Table 2, these two ratios are 5.25 (84%/16%) for the comparison group and 1.63 (62%/38%) for the intervention group. The odds ratio is then obtained by dividing the comparison group odds of failure to those of the intervention group. The odds ratio is 3.22 (5.25/1.63), indicating that the odds of failing to improve for a comparison group member are 3.22 times as high as those from the treatment group. If one compared the odds of success in the treatment group to that in the comparison group, one would get the same odds ratio. Easier yet, one could compute the cross-product ratio: (84% × 38%)/(16% × 62%) = 3.22.

A major limitation of the odds ratio as an effect size index is that the magnitude of the odds ratio may approach infinity if the outcome is rare or very common, even when the association is near random. The magnitude of the odds ratio varies strongly with the choice of cut-point. For example, the only difference between Table 2 and Table 3 is the choice of cut-point, but the odds ratio in Table 3 is 19.12, while that in Table 2 is 3.22. Thus, there are no agreed-upon standards for what represents a large odds ratio because some very large odds ratios are obtained for situations very close to random association. Consequently, odds ratios can be quite misleading as an effect size indicating clinical significance.

Risk Ratio. Risk ratio is determined by dividing the failure rate of the comparison group by the failure rate of the treatment group, or by dividing the success rate of the treatment group by that of the comparison group. In Table 2, if one compares failure rates, the relative risk is 1.35 or 84%/62%; if one compares success rates, the relative risk is 2.38 or 38%/16%. Risk ratios are always less than odds ratios. In fact, the odds ratio is the product of these two risk ratios (1.35*2.38 = 3.22). In Table 3, the risk ratio of failure is 6.80, and the risk ratio of success is 2.81. (Again, the odds ratio is equal to 19.12 or 6.80*2.81.) The choice of cut-point and which risk ratio is chosen change the magnitude of the risk ratio, making it hard to interpret. Once again, because the risk ratio may approach infinity when the risk in the denominator approaches zero, there can be no agreed-upon standards for assessing the magnitude of risk ratio.

Relative Risk Reduction. Relative risk reduction is computed by subtracting the treatment group failure rate from the comparison group failure rate, and dividing by the latter, or by subtracting the comparison group success rate from the treatment group success rate and dividing by the former. Relative risk reduction can vary between 0 and 1.0. In Table 2 the failure relative risk reduction is 26%, but in Table 3 it is a larger 85%. Because the “failure” relative risk reduction may be very small when the “success” relative risk reduction is large, relative risk reduction is difficult to interpret in terms of clinical significance, and there are no agreed-upon standards for judging its magnitude.

NNT. NNT is a relatively new measure (Laupacis et al., 1988) that has been recommended for improving the reporting of effect sizes, but it has not yet been widely used (Nuovo et al., 2002). NNT is the number of patients who must be treated to generate one more success or one less failure than would have resulted had all persons been given the comparison treatment. Alternatively, in risk studies, it is the number who would need to be exposed to the risk factor to generate one more case than if none had been so exposed.

Mathematically, NNT is the reciprocal of the risk difference. A result of 1.0 means the treatment is perfect, that every treatment subject succeeds and every comparison subject fails. The larger the NNT, the less effective the treatment relative to the comparison. In Table 2, NTT is 1/0.22, or 4.55. This means that out of every 4.55 patients treated, 3.55 would get the same results as they would have had they been in the comparison group, and there would be one “excess” success due to treatment. In other words, 3.55 subjects would be exposed to the costs or risks of the treatment with no benefit, and only 1 subject out of 4.55 would gain whatever benefit the treatment confers.

Clinicians seem to understand the meaning of NNT and to be able to make the appropriate judgments. In the Salk data, the NNT was 2,439. Since the benefit was the prevention of serious disability or death and the cost and risks were very low, subjecting 2,438 subjects to the cost and risk of vaccination who did not benefit from it was considered ac-
ceptable to save one person from polio. If the outcome had been the common cold, and the treatment involved a drug with both cost and side effects, a NNT of 2,439 would have been clinically unacceptable.

AUC. Another relatively new effect size that might substitute for either family measures or measures of risk potency is AUC, which represents the probability that a randomly selected subject in the treatment group has a better result than one in the comparison group.

In situations where d might be used, AUC = Φ (d/√2), where Φ ( ) is the cumulative standard normal distribution function; that is, d/√2 is the z value and Φ (d/√2) is the area under the normal curve up to that z value. As shown in Table 1, one can define standards for AUC that correspond to those for d. For example, a medium or typical effect size of d = 0.5 corresponds to AUC = 64%. Thus, when comparing a treatment subject against a comparison subject, 64% of the time the treatment subject would have a better response.

In situations where measures of risk potency are used, risk difference = 1/NNT = 2 AUC − 1. Thus again, as in Table 1, one can present standards for risk difference and NNT that correspond to those for d. For example, a medium or typical effect size of d = 0.5, which corresponds to AUC = 64%, corresponds also to a risk difference of 28% and to NNT = 3.6 subjects.

AUC is of special interest because it can be computed based on clinical judgments alone. One could randomly select pairs of subjects, one of each pair in the treatment and one in the comparison group, and submit their clinical records to experts blinded to group membership. The experts would then be asked which of the two had a better outcome. The proportion of the pairs for which the experts selected the treatment group subject as better off is an estimate of AUC. For this reason, AUC has special appeal as a measure of clinical significance.

Finally, AUC helps us understand the problem of cut-points. When one imposes dichotomization on an ordered response, there is a tacit declaration that all variation of response above the cut-point and all below the cut-point have no clinical relevance. All that matters in comparing the response of one subject to another is whether their response lies above or below the cut-point. Thus, in a cancer trial, where survival past 5 years is a “success,” someone who survives 5 years + 1 day is considered as having an equivalent response to someone who survives 50 years, and someone who survives 5 years – 1 day as having an equivalent response to someone who survives 1 day. If this reflects good clinical judgment, then a treatment patient who survived 50 years is considered a tied response with a comparison group patient who survived 5 years + 1 day, and the tie is randomly broken in computing the AUC. The more such ties there are, the more the AUC is reduced from its original value computed from the ordered scores to the AUC computed from the binary success/failure outcome, and the variation of the AUC (and thus risk difference and NNT) from one cut-point to another is related to how many tied responses are so induced.

Summary

Which of the above measures of effect size presented here is the most appropriate for reporting clinical significance? Clearly family effect sizes address a problem that none of the other effect sizes address—association between two ordered variables. Within that family, whether one selected the Pearson, the Spearman, or the Kendall correlation depends on the distributions of the variables. Otherwise, when one of the variables is binary and the other ordered, the d family effect sizes are familiar to researchers but not designed for and probably not very helpful to patients and clinicians. However, d can easily be converted to AUC, which is very helpful to patients and clinicians.

When both of the variables are binary, despite their popularity with researchers, odds ratios, risk ratios, and relative risk reduction are best avoided. One could easily convert risk difference or number needed to treat to AUC, and thus have an effect size comparable to that derived from d.

Nuovo et al. (2002) pointed out that the Consolidated Standards on Reporting Trials (CONSORT) statement (Altman et al., 2001) recommends reporting the NNT or the risk difference. However, often risk difference can seem unimpressively small, and NNT may seem very large, suggesting very little effect of treatment. In many such cases, one (not both) of the risk ratios and one (not both) of the relative risk reduction measures and, most of all, the odds ratio can give an inflated impression of the size of the effect, thus exaggerating clinical significance. For this reason, the preferred effect size at this time would tend to be AUC or d or NNT or risk difference, all of which are mathematically equivalent.

We have provided some general guidelines for interpreting measures of clinical significance. It is not possible, however, to present more than a tentative recommendation for which effect size to use, or to provide any fixed standards for any such effect size that a clinician could universally use to conclude that an effect size was clinically significant. It makes a difference whether the treatment is for a deadly disease like polio, or the common cold, and whether the treatment is risky and costly or perfectly safe and free. The context in which an effect size is used matters in interpreting the size of the effect; the choice of effect size is only to facilitate consideration of the effect in the context of its use.

REFERENCES


