II-4

American Journal of Epidemiology
Vol. 135, No. 5
Printed in U.S.A.

A BRIEF ORIGINAL CONTRIBUTION

Acceptable Values of Kappa for Comparison of Two Groups

Daniel G. Seigel, Marvin J. Podgor, and Nancy A. Remaley

A model was developed for a simple clinical trial in which graders had defined probabilities of misclassifying pathologic material to disease present or absent. The authors compared Kappa between graders, and efficiency and bias in the clinical trial in the presence of misclassification. Though related to bias and efficiency, Kappa did not predict these two statistics well. These results pertain generally to evaluation of systems for encoding medical information, and the relevance of Kappa in determining whether such systems are ready for use in comparative studies. The authors conclude that, by itself, Kappa is not informative enough to evaluate the appropriateness of a grading scheme for comparative studies. Additional, and perhaps difficult, questions must be addressed for such evaluation. *Am J Epidemiol* 1992;135:571–8.

bias; clinical trials; epidemiologic methods; models, statistical; odds ratio; relative risk; statistics

Considerable effort is often expended in clinical research to develop grading schemes for pathology. For example, several papers have been published on the grading of lens opacities (1–3). In these publications, the extent to which agreement exists between graders is described. A grading scheme with low agreement is regarded as unsatisfactory. The statistic Kappa is often computed to quantify the extent to which agreement exists, beyond what might be expected by chance. It is uncertain, however, how to interpret any level of Kappa. Figure 1 presents two sets of adjectives that Fleiss (4), on the one hand, and Landis and Koch (5), on the other, propose for characterization of Kappa. For values of Kappa above 0.4, the strength of agreement is similarly described. Below 0.4, however, Landis and Koch tend to be more charitable with their adjectives. The evidence supporting these two scales is

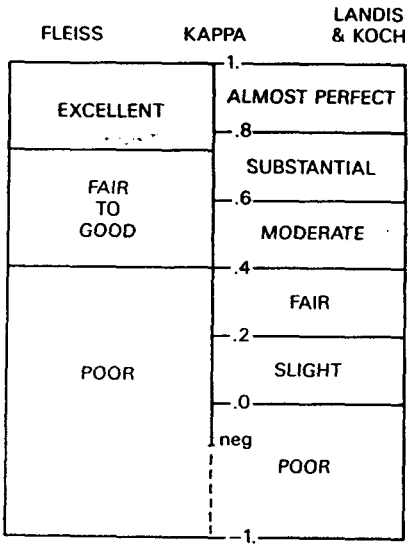

FIGURE 1. Scales for strength of agreement for Kappa, as proposed by Fleiss (4) and Landis and Koch (5).

Received for publication May 20, 1991, and in final form September 20, 1991.

From the Biometry and Epidemiology Program, National Eye Institute, National Institutes of Health, Bethesda, MD 20892.

Reprint requests to Dr. Daniel G. Seigel, Building 31, Room 6A10, Biometry and Epidemiology Program, National Eye Institute, NIH, Bethesda, MD 20892.

not given.

The purpose of this paper is to evaluate Kappa, not in the abstract, but insofar as it predicts the qualities of a system for encoding information observed on images such as photographs, x-ray films, or pathology slides. We will call such systems "grading schemes," and refer to persons who encode the information as "graders." Can Kappa tell us whether the grading scheme is "ready"? We did this by developing a model of a clinical trial in which a grading scheme is used to determine whether an event has occurred for each patient followed. Examples of such events are common in ophthalmologic research and include visual field loss, lens opacification, and development of new vessels in the retina. In particular, we ask whether Kappa offers useful predictions for two statistics: bias in estimates of the treatment effect and the efficiency of the trial.

METHODS

Computations were done for the following clinical trial model. A drug is to be compared with a placebo. At the end of the trial, the sample proportions observed with disease (as determined by the scores provided by the grading scheme) in the drug and placebo groups are compared. A two-tailed test of significance with alpha equal to 0.05 is computed. The sample size is selected so that a power of 0.8 is available for contrasts of the drug and placebo groups. Figure 2 shows how outcomes in the drug group are subjected to misclassification by the graders within this model. A comparable figure could be shown for the placebo-treated group. The key element is that some proportion of those with or without the event have images that are difficult to interpret. Among those, some are classified incorrectly.

In our computations, we explored:

1. Event rates of 0.1, 0.25, and 0.5 in the placebo group.
2. A relative risk of 0.6 in the drug-treated group.
3. Of those with event, proportions difficult to classify of 0.0, 0.1, 0.2, 0.3, and 0.5.
4. Of those with no event, proportions difficult to classify of 0.0, 0.1, 0.2, 0.3, and 0.5.
5. Proportions misclassified (among the difficult to classify, with event) of 0.1, 0.25, and 0.5.
6. Proportions misclassified (among the difficult to classify, with no event) of 0.1, 0.25, and 0.5.

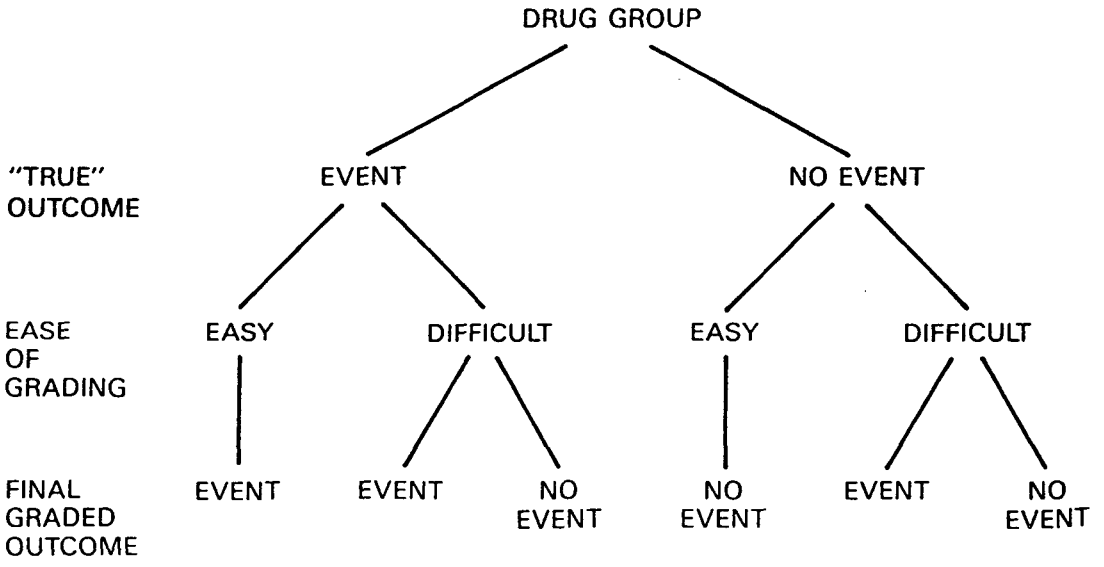

FIGURE 2. Model of grading misclassification.

We computed from the model bias, efficiency, and Kappa, each defined as follows:

1. Bias = $R(m)/R$ where $R(m)$ = relative risk in population, with misclassification and R = relative risk in population, without misclassification.
2. Efficiency = $n/n(m)$ = ratio of sample sizes without and with misclassification for power of 0.8 and $\alpha = 0.05$.
3. Kappa, for the 2 × 2 table of event classification, based on the event rate in the placebo series, for two independent graders who have the same misclassification probabilities.

We illustrate the calculations for the following combination of parameters:

Event rate of 0.5 in the placebo group.
A relative risk of 0.6 in the drug-treated group.
Of those with event, proportions difficult to classify of 0.2.
Of those with no event, proportions difficult to classify of 0.3.
Proportion misclassified (among the difficult to classify, with event) of 0.25.
Proportion misclassified (among the difficult to classify, with no event) of 0.1.

For such statistics, the bottom tier of figure 2, for the placebo group, would have values of 0.5 × 0.8, 0.5 × 0.2 × 0.75, 0.5 × 0.2 × 0.25, 0.5 × 0.7, 0.5 × 0.3 × 0.1, 0.5 × 0.3 × 0.9. The proportion that are finally graded with the event would be 0.49. The proportion that are finally graded without the event would be 0.51.

For the drug group, the bottom tier of figure 2 would have values of 0.3 × 0.8, 0.3 × 0.2 × 0.75, 0.3 × 0.2 × 0.25, 0.7 × 0.7, 0.7 × 0.3 × 0.1, 0.7 × 0.3 × 0.9. The proportion that are finally graded with the event would be 0.306. The proportion graded without the event would be 0.694.

The relative risk with misclassification would be 0.306/0.49, or 0.6245.

The bias would be 0.6245/0.6 = 1.04.

Sample sizes with and without misclassification are computed from the following formula (6):

$$n = \left(\frac{z_\alpha \sqrt{2\left(\frac{p_c + p_t}{2}\right)\left(1 - \frac{p_c + p_t}{2}\right)} + z_\beta \sqrt{p_c(1 - p_c) + p_t(1 - p_t)}}{p_c - p_t} \right)^2$$

Without misclassification:

p_c = proportion with the event in the control group (0.5 in this illustration).
p_t = proportion with the event in the treated group = relative risk × p_c (0.6 × 0.5 = 0.3).

With misclassification:

p_c and p_t are calculated as in the illustration above (0.49, 0.306).

Sample sizes required with and without misclassification are 110 and 93, yielding an efficiency of 0.85.

The 2 × 2 table for two independent graders who have these same misclassification probabilities is computed from the placebo series, and would have cells in which they both declare the event to have occurred (0.5 × 0.8 + 0.5 × 0.2 × 0.75 × 0.75 + 0.5 × 0.3 × 0.1 × 0.1), both declare no event (0.5 × 0.2 × 0.25 × 0.25 + 0.5 × 0.7 + 0.5 × 0.3 × 0.9 × 0.9) and off-diagonal cells in which they disagreed, each with probabilities (0.5 × 0.2 × 0.75 × 0.25 + 0.5 × 0.3 × 0.1 × 0.9). The four cells, respectively, would have values of 0.4578, 0.4778, 0.0323, and 0.0323.

For such a table, Kappa would be computed as 0.87.

RESULTS

Figure 3 presents the relation between Kappa and bias. Each point represents the computed Kappa and bias for each of the 675 combinations of event rates and misclassification statistics. The bias is clearly associated with Kappa. Low values of Kappa imply greater bias. As Kappa approaches one, bias in estimation of the relative risk diminishes. Prediction of the amount of bias has considerable error, however, since the range in the bias is fairly sizable for any fixed value of Kappa. Even with values of Kappa described as good or substantial in figure 1, the relative risk of 0.6 may be inflated enough to render a clinical trial uninformative of treatment benefits. The direction of the bias is as usually seen with misclassification; the relative risk tends to be closer to one.

We have set up the computations for a relative risk less than one. For the constants explored in our model, the bias is never qualitative. That is, for this model, the expected value of the observed relative risk is never greater than one.

We have also computed the bias in estimation of the odds ratio, since it is often of interest in clinical trials and in epidemiologic research. Holding the relative risk at 0.6, as before, figure 4 presents the relation between Kappa and bias in estimation of the odds ratio. Except for a somewhat reduced variability in bias for any fixed value of Kappa, the pattern is similar to that seen for the relative risk.

In figure 5, efficiency is shown to be more predictably related to Kappa than is bias. As Kappa approaches unity, so does efficiency. For the model we explored, efficiency is at best equal to Kappa. More typically, efficiency is the square of Kappa. At worst, at the lower boundary of the plot, efficiency can be as low as the cube of Kappa.

All of the results presented have been for a relative risk of 0.6. Though figures are not presented, we performed all the computations and generated figures for relative risks of 0.9 and 0.3 as well. For both of these, bias is associated with Kappa, with low values of Kappa implying greater bias, as was noted for a relative risk of 0.6. Prediction of the amount of bias for any value of Kappa again showed considerable error, since the range in bias is large, especially for a relative risk of 0.3. For the relative risk of 0.9, the bias is relatively constrained, never exceeding 1.1. For the relative risk of 0.3, on the other hand, the limit of bias is 3.3, and values of 2 or more (a doubling in the estimate of the relative risk) are seen for Kappas as high as 0.75. The graphs of efficiency and Kappa for relative risks of 0.3, 0.6, and 0.9 are all nearly the same.

DISCUSSION

The analysis of the relation between Kappa and the variables efficiency and bias indicates that, for this model, conventional guidelines for the ranges of acceptability are not adequately informative. Can the relations found in this paper be extrapolated to clinical trials with other designs, which are often more complex? They cannot, nor can the relations in those studies be readily predicted. At present, we know of no way, other than the exploration by statistical models, to assess potential cost in bias and in efficiency that will result from weaknesses in a grading scheme.

This is not to say that the statistic Kappa is without value in the development of a scale of measurement for possible use in clinical research. For example, training of personnel in implementing a grading system should be rewarded by increasing values of Kappa. Values of Kappa close to one, moreover, are likely to assure little loss in efficiency because of grading problems, assuming validity is secure.

Bias seems even less well predicted from Kappa than efficiency, since the range of bias is considerable for any level of Kappa. Moreover, we have not discussed ways in which Kappa may be insensitive to yet other sources of bias associated with grading. The graders may have a common misunderstanding in the interpretation of a lesion. Kappa rewards consistency, even when the grading is incorrect. Emerson might have

FIGURE 3. Plot of bias versus Kappa with relative risk = 0.6 where bias = (relative risk observed)/(relative risk expected). Points in figures 3–5 are jittered so that overlapping points are separated slightly.

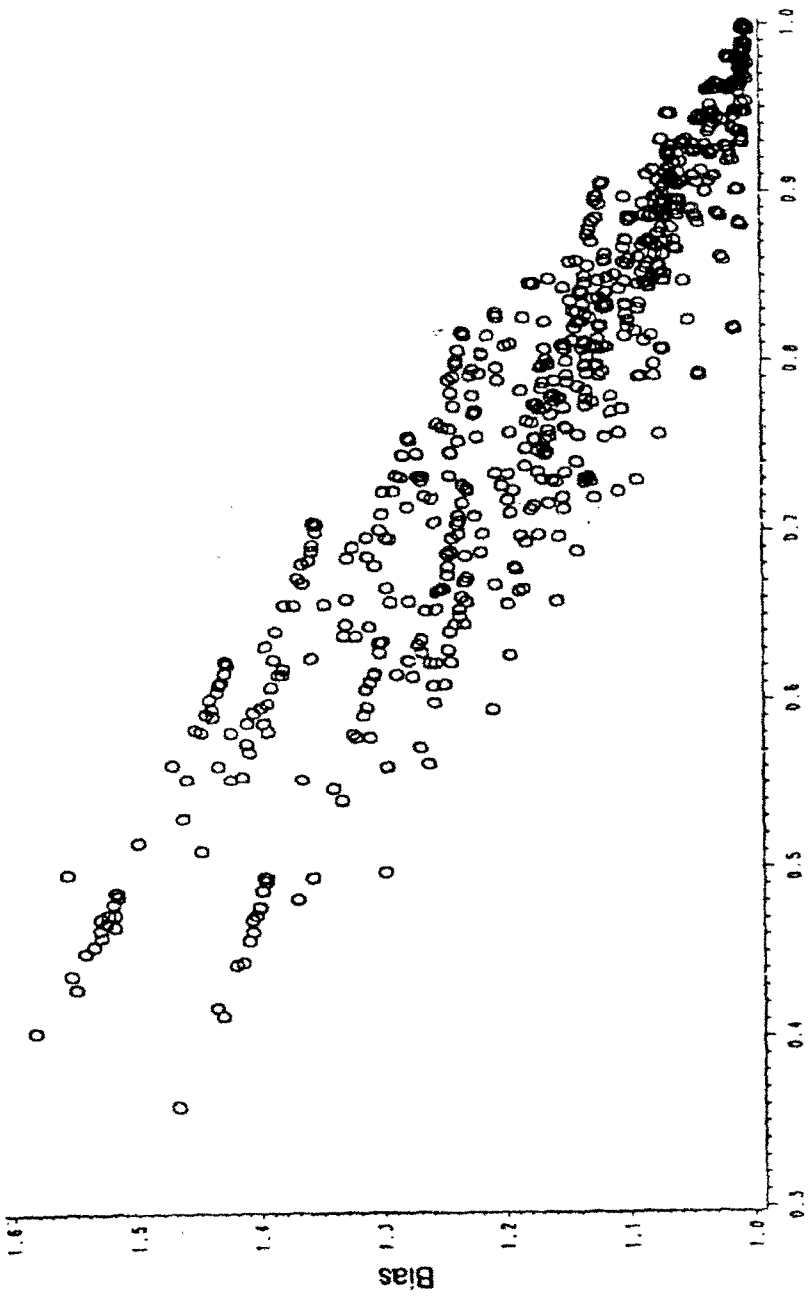

FIGURE 4. Plot of bias versus Kappa with relative risk = 0.6 where bias = (odds ratio observed)/(odds ratio expected).

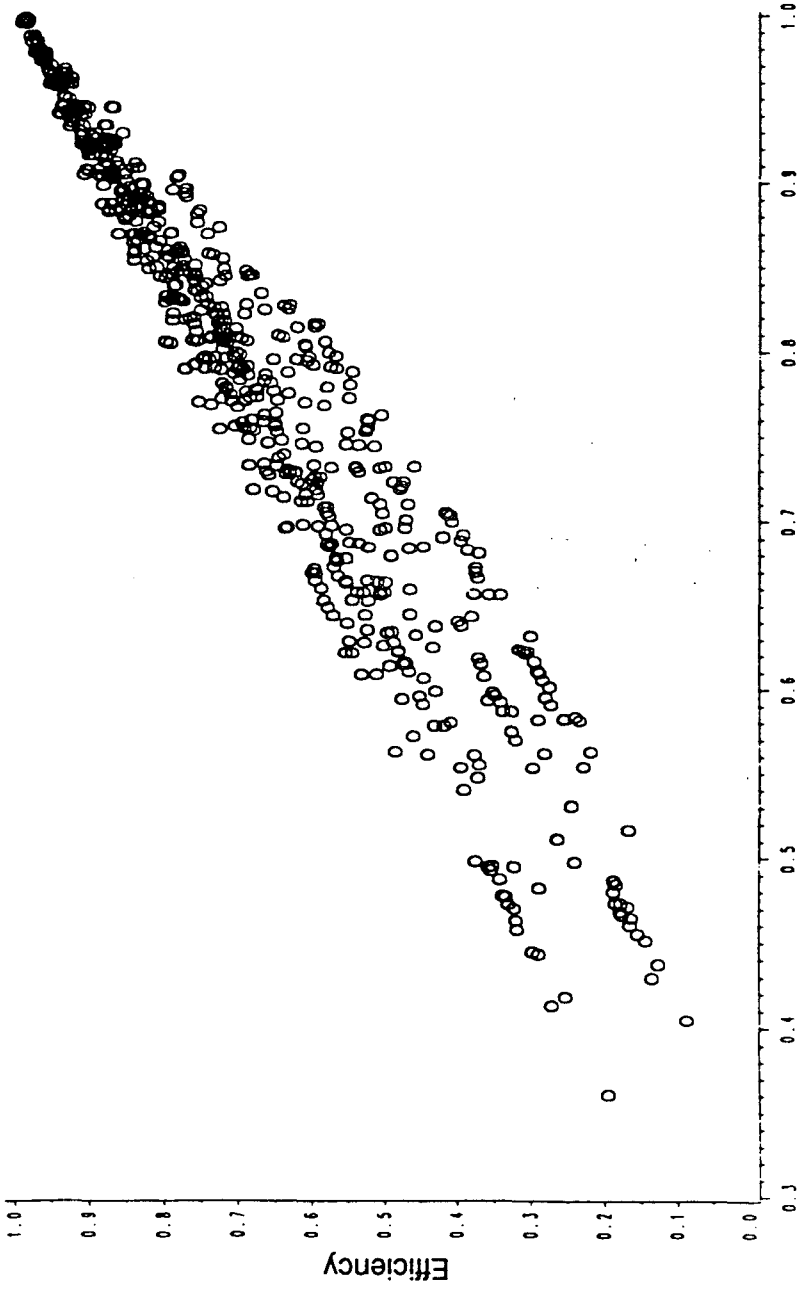

FIGURE 5. Plot of efficiency versus Kappa with relative risk = 0.6.

said that foolish consistency is the hobgoblin of the Kappa statistic.

Kappa shares the limitations of other single statistics and parameters. The complexity of research settings prevents complete description by any such single quantity.

Studies in which morbidity is the primary dependent variable depend heavily on the quality of the scales of measurement. Two recent papers in ophthalmology point the way toward use of statistical models to evaluate potential measurement scales. Seigel and Milton (7) found that side by side evaluation of before and after treatment photos may often yield less power in a clinical trial than independently graded photos. Datiles et al. (8) determined the contribution of measurement errors for lens opacities to overall sample size in clinical studies. Both of these papers illustrate the use of modeling in the selection of appropriate scales of measurement for clinical research. The computation of Kappa, by itself, does not offer comparable information.

We wish that we could offer an easy to follow recipe for evaluating grading scales for use in comparative studies, and in particular in clinical research. We have cautioned that Kappa is not sufficient information. As we imagine consultations with investigators who wish to model, as we have, to determine bias and efficiency, we find ourselves asking for information not easily provided. Is there a "standard" against which the grading can be compared? What are the misclassification rates against such a standard? What will be the distribution of pathology in the population? How often will measurements be taken during the clinical research? How much of a change is anticipated? How much is clinically meaningful? All of these were easy to answer in our model; we simply invented them. They are less easy to answer in the actual research setting. This is clearly an area where methods and experience are needed.

REFERENCES

1. Maraini G, Pasquini P, Tomba MC, et al. An independent evaluation of the Lens Opacities Classification System II (LOCSII). Ophthalmology 1989;96:611–15.
2. Chylack LT, Leske MC, McCarthy D, et al. Lens Opacities Classification System II (LOCSII). Arch Ophthalmol 1989;107:991–7.
3. Taylor HR, West SK. The clinical grading of lens opacities. Aust N Z J Ophthalmol 1989;17:81–6.
4. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons. 1981:218.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
6. Armitage P, Berry G. Statistical methods in medical research. 2nd ed. Oxford: Blackwell Scientific Publications, 1987:183.
7. Seigel D, Milton R. Grading of images in a clinical trial. Stat Med 1989;8:1433–8.
8. Datiles MB, Podgor MJ, Sperduto RD, et al. Measurement error in assessing the size of posterior subcapsular cataracts from retroillumination photographs. Invest Ophthalmol Vis Sci 1989;30: 1848–54.