Public Health Rev 1988; 16: 153–162

A new statistical test for summarizing the results of independent epidemiologic studies

William R. Gaffey, PhD

Epidemiology Director, Department of Medicine and Environmental Health, Monsanto Company, 800 N. Lindbergh Blvd., St. Louis, MO 63167, USA

D-3

INTRODUCTION

Suppose (1) that we have a number of independent epidemiologic studies of the relationship between some exposure and an outcome variable, (2) that each study investigates the hypothesis of no effect versus the alternative that exposure increases risk, and (3) that each of the studies produces a statistic that is a reasonable measure of the relationship. An example is the estimated relative risk, after an appropriate latent period, for a particular cancer in exposed persons compared with unexposed persons. More often than not the results of the studies will not be identical, in the sense that the individual statistics will vary in the degree to which they are statistically significant. If we are interested in testing the hypothesis of no effect at, say, the 5% level of significance, and if every study is statistically significant, there is no problem of interpretation. If a few studies are significant but most are not, we clearly need some way to assess the overall significance of the studies. If none of the studies is significant at the 5% level but most or all of them are nearly significant, the need for some formal overall assessment is even more apparent because intuition tells us that the results *in toto* could be significant.

The studies may be of different designs, and may generate test statistics with different distributions, which complicates the situation if one thinks of pooling the data. In some cases it may even be that nothing is known about the test statistics other than their *P* values, which are given in published reports.

Combining the results of separate studies by using their *P* values is therefore not only expedient, but in some situations is the only thing that can be done. R.A. Fisher (1) proposed taking the product of all of the *P* values and rejecting the hypothesis of no effect (the null hypothesis) when the product is small. More specifically, Fisher found that, under the null hypothesis, the natural logarithm of the product of the *P* values multiplied by minus two has a chi-square distribution with degrees of freedom equal to twice the number of studies. Therefore, the test can be carried out using a commonly available table, and the test using this statistic is usually referred to as Fisher's test of *P* values or as the Fisher algorithm.

However, since epidemiologic studies are observational and not experimental, the null hypothesis may be false for reasons having nothing to do with the exposure in question. For example, it is quite

possible in occupational studies that the exposure in question is innocuous but that there are other exposures in each plant that are carcinogenic. In such a case the results of a collection of studies may suggest that the exposure under investigation is protective because the persons with the longest exposure to that substance will in general have the shortest exposure to the carcinogens present in other parts of the plant. In such a situation we would like to conclude that the exposure under investigation is not carcinogenic, but if the *P* values are generally small, Fisher's test of *P* values will reject the null hypothesis. In other words, we want to test the hypothesis of no effect against the alternative that there is an effect that increases with increased exposure, but Fisher's test of *P* values rejects the null hypothesis in favor of any alternative that creates small *P* values. We propose a modification of Fisher's test of *P* values that is sensitive to the alternatives that are of concern to us but tends to ignore alternatives that are not of interest. At the same time, the test can be carried out by using tables of known distributions.

THE ALTERNATIVES OF INTEREST

If chance is the only factor involved, the *P* value for a particular test statistic has a uniform distribution between 0 and 1. If exposure (or some other factor) has an effect, the *P* value will not be uniformly distributed, but will tend either to be large or to be small, depending on whether the effect is favorable or unfavorable. The latter case is the alternative of interest that we want to detect.

In such a case, if we have a number of studies that are identical except for exposure level, the studies with the greatest exposures will tend to have the smallest *P* values.

On the other hand, if in such a case we have a number of studies that are identical except for sample size, the *P* values will tend to be smallest in the largest studies.

In real life, studies differ both in their powers against a particular alternative, that is, in their sample sizes, and in the alternatives against which it is appropriate to test them, that is, in the level and duration of exposure. Conceivably, a small study with heavy exposure could be more likely to detect an effect than a large study with light exposure, that is, the *P* value for the first study could be generally smaller than

the *P* value for the second study if the exposure actually has a harmful effect.

Under the alternative of interest, therefore, one would expect the *P* values of a set of studies to be ranked in a way that depends on their sample sizes and their relative exposures. In theory, one cannot predict this ranking without knowing the dose–response relationship for the exposure effect. In practice, it is usually possible to predict a ranking that is not very dependent on the dose–response relationship that one is willing to assume. Consequently, it is usually possible to decide for a set of studies how their *P* values ought to turn out, as far as their relative rank is concerned. This is the alternative against which we would like a test to be powerful.

THE TEST

One might consider ignoring Fisher's test of *P* values completely and simply calculating the correlation between the observed ranking of the *P* values and the expected ranking under the alternative, rejecting the null hypothesis if that correlation is significantly greater than zero. Unfortunately, this is not intuitively reasonable because one might get a correlation that was significantly greater than zero but still quite small. In such a case one would want to look at the *P* values before deciding to reject the hypothesis of no harmful effect.

What is needed is some way to look both at whether the *P* values are small and whether they tend to be in the right order. We propose a test that accepts the hypothesis of no harmful effect when the rank correlation is not positive, and rejects it when the rank correlation is positive, provided that Fisher's test of *P* values is then significant. However, the level at which we use Fisher's test depends on how positive the rank correlation is. If it is small we use Fisher's test at a low level of significance, that is, we ask for a lot of evidence against the null hypothesis before we reject it. If the correlation is large we use Fisher's test at a high level of significance, that is, we insist only on weak evidence before we reject the null hypothesis. The different levels of significance for each value of the rank correlation coefficient are so adjusted that the overall level of significance of the test is equal to whatever level we have decided upon (usually 5%).

The test is performed as follows:

1. From what is known about the sample sizes and the exposure levels, decide what the expected ranking of the P values would be if there were in fact a harmful exposure effect.

2. Calculate the Spearman rank correlation coefficient between the expected ranking and the ranking that is actually observed. Accept the null hypothesis if the correlation is not positive.

3. If the rank correlation is positive, obtain the following three values from the distribution of the Spearman rank correlation coefficient (tables of the distribution are obtainable from Kendall (2) for up to 13 studies and from elementary textbooks such as Dixon and Massey (3) for up to 10 studies):

   B = probability that the rank correlation is greater than 0 and less than the observed value.

   C = probability that the rank correlation equals the observed value.

   D = probability that the rank correlation is positive.

4. Reject the null hypothesis if Fisher's test of P values is significant at the level αA, where α is the desired level of significance of the overall test and

$$A = (2B + C)/D^2.$$

The test is derived in the Appendix.

EXAMPLES

We illustrate the test using hypothetical examples for two reasons. First, such examples can best demonstrate how the test performs under clear-cut conditions where Fisher's test of P values alone is obviously inappropriate. Second, the problem of assessing the overall significance of multiple studies in real life involves a detailed prior assessment of the quality and suitability of each such study, and it would be misleading to imply that a statistical test alone enables us to pass judgement on them.

Table 1 shows five cohort studies. The expected deaths are the numbers of deaths from a particular cancer, say after 15 years latency,

Table 1

Five hypothetical cohort studies in which Fisher's original test overstates the significance

Expected deaths	Exposure	Observed *P* value	Observed rank	Expected rank
100	High	0.135	3	1
75	High	0.122	2	2
75	Medium	0.143	5	3
75	Low	0.111	1	4
50	Low	0.141	4	5

Calculated value of Fisher's test = 20.44; therefore, the unadjusted Fisher's test is judged to be significant at the 5% level of significance by this test. $P = 0.026$.
Rank correlation = 0.10
Probability that rank correlation is > 0 and $< 0.10 = B = 0$.
Probability that rank correlation equals $0.10 = C = 0.083$.
Probability that rank correlation is positive = $D = 0.475$.
$A = 0.367$.
$\alpha A = (0.05)(0.367) = 0.018$.
The critical value of chi-square with 10 degrees of freedom at the (adjusted) 1.8% level of significance is between 20.48 and 23.21.

and the observed *P* values are obtained from statistical tests of the standard mortality ratios (SMRs). The actual values of the SMRs are not relevant here. The expected ranking is the obvious one, ranging from the smallest to the largest, and requires no assumption about the dose–response relationship other than that higher exposures lead to higher SMRs, all other things being equal.

The correlation between the observed and expected rankings is small, which leads to an adjusted level of significance of 1.8%. The calculated value of Fisher's test of *P* values is not significant at this level, so the overall test is not significant at the 5% level. Note that the critical value for the 5% level of Fisher's test is 18.31, so that if we had applied that test without adjustment we would have gotten a statistically significant result, ignoring the fact that the *P* values did not have the rank order that they should have had.

Table 2 shows the other side of the coin. The studies are the same, but the resulting observed *P* values are different. The correlation

Table 2

Five hypothetical cohort studies in which Fisher's original test understates the significance

Expected deaths	Exposure	Observed *P* value	Observed rank	Expected rank
100	High	0.111	1	1
75	High	0.135	3	2
75	Medium	0.122	2	3
75	Low	0.237	4	4
50	Low	0.300	5	5

Calculated value of Fisher's test = 17.73; therefore, the unadjusted Fisher's test is not judged to be significant at the 5% level of significance by this test. $P = 0.062$.
Rank correlation = 0.90.
Probability that rank correlation is > 0 and $< 0.90 = B = 0.433$.
Probability that rank correlation equals $0.90 = C = 0.034$.
Probability that rank correlation is positive = $D = 0.475$.
$A = 3.98$.
$\alpha A = (0.05)(3.98) = 0.20$.
The critical value of chi-square with 10 degrees of freedom at the (adjusted) 20% level of significance is 13.44.

between the observed and expected rankings is large, so that the adjusted level of significance is 20%. The calculated value of Fisher's test of *P* values is significant at this level, so that the overall test is significant at the 5% level. If we had applied Fisher's test of *P* values without adjustment the result would not have been significant at the 5% level because Fisher's test ignores the fact that the *P* values came out in pretty much the expected order.

DISCUSSION

Different epidemiologic studies usually differ in their circumstances of exposure as well as in their sample sizes and their designs. Even among studies with the same design, for example historical cohort studies, there are usually differences in exposure which make epidemiologists reluctant to pool them in any naive way such as by totalling the

observed and expected deaths. Fisher's test of *P* values, although it is convenient to use, does not address the problem of how to deal either with varying sample sizes or varying circumstances of exposure.

Instead of viewing these differences as problems, they can be viewed as additional information because they tell us what we should expect if there is in fact an exposure effect. The modified Fisher's test of *P* values presented here makes use of differences in exposure and sample size from study to study to produce a test that is sensitive, as Fisher's original test of *P* values is not, to those alternatives that we would expect to occur if there were an exposure effect in studies with different exposures and samples sizes.

Any test against ordered alternatives invites comparison with the one proposed by Jonckheere (5). However, the present test is designed for use in situations where the individual observations in each study are not available, so that Jonckheere's test is not usable. On the other hand, the *P* values required for the present test do not appear to be available in the typical situation described by Jonckheere.

In any case, the present test manages to combine the probabilities yielded by the rank correlation and the chi-square statistic without resorting to a new distribution, so that the test can be used with the existing tables for these two statistics.

REFERENCES

1. Fisher RA. Statistical methods for research workers. 4th ed. Edinburgh: Oliver and Boyd, 1932.
2. Kendall MG. Rank correlation methods. 4th ed. London: Griffin, 1970.
3. Dixon WJ, Massey FJ. Introduction to statistical analysis. 4th ed. New York: McGraw-Hill, 1983.
4. Bartholomew DJ. A test of homogeneity for ordered alternatives. Biometrika 1959; 46: 36–48.
5. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. Biometrika 1954; 41: 133–145.

APPENDIX

The modified Fisher's test of *P* values proceeds by calculating the correlation between the observed and expected rankings of *P* values, and then using Fisher's test at an adjusted significance level α that depends on the value of the correlation coefficient that is calculated.

The adjusted level is zero when the correlation is zero or less, and, for reasons explained in the text, gets higher as the correlation gets higher. At the same time, the adjusted significance levels have to be such that the overall level of significance of the test has to equal whatever significance level has been decided upon.

More precisely, if there are n independent studies the correlation between the observed and expected rankings of their P values can take on values $r_1, r_2, \ldots, r_m$ ranging from -1 to +1. For each possible value r_i we use Fisher's test at an adjusted level of significance α_i. The overall level of significance of the test is therefore

$$\Sigma\, \alpha_i P\,(r = r_i) = \alpha, \tag{1}$$

where $P\,(r = r_i)$ is the probability that the rank correlation equals r_i when there is no exposure effect, either favorable or unfavorable.

We do not reject at all for $r_i \leq 0$, that is, $\alpha_i = 0 \leq$ for $r_i \leq 0$, and we want α_i to increase as r_i increases. Also, we want eq. 1 to hold for whatever α we choose. All this occurs if we choose α_i as the appropriate function of the probability that the corresponding value r_i occurs.

For any value $r_i > 0$, the probability that the rank correlation is positive but less than r_i under the hypothesis of no exposure effect is

$$B_i = \sum_{}^{i-1}{}' P\,(r = r_j), \tag{2}$$

where the prime denotes summation over the positive values of r_j. The probability that the correlation equals r_i is, under the hypothesis,

$$C_i = P\,(r = r_i). \tag{3}$$

Then

$$\Sigma'\, P\,(r = r_i)\,(2B_i + C_i) = 2\sum_{j<i}{}' P\,(r = r_i)\,P\,(r = r_j) + \Sigma'\, P^2\,(r = r_i) \tag{4}$$

$$= [\Sigma'\, P\,(r = r_i)]^2 = D^2,$$

where D is the probability that the rank correlation is positive under the hypothesis.

Consequently, if

$$\alpha_i = \alpha\,(2B_i + C_i)/\,D^2 \qquad (5)$$

when $r_i > 0$ and zero otherwise, it fulfills our requirements.

If there is a favorable exposure effect the true rank correlation between the observed and expected rankings will be negative. This means that the calculated r is more likely to be negative than positive, so that the actual probabilities for positive values of r will be less than the ones in eq. 1. Therefore the probability of concluding that there is an unfavorable effect when in fact there is a favorable effect is less than α, so that the test is indeed one-sided.

An apparent problem arises because when the rank correlation equals a particular value r_i, Fisher's statistic has a conditional distribution given r_i. However, Fisher's statistic is the sum of independent chi-square variables, and Bartholomew (4) has shown that such a random variable is independent of the ranking of the component chi-square variables. Therefore, no matter what the value r_i, the distribution of Fisher's statistic is still chi-square with degrees of freedom equal to twice the number of studies. Therefore for a particular observed value of r_i, if it is positive, we can carry out the test by calculating α_i, finding the critical value of Fisher's statistic for that α_i from a chi-square table, and rejecting the hypothesis if the calculated value of Fisher's test exceeds that critical value.