Psychiatry Research, **10**, 207-216
Elsevier

Inferential Statistical Methods for Strengthening the Interpretation of Laboratory Test Results

Neil Kalter, Michael Feinberg, and Bernard J. Carroll

Received April 22, 1983; revised version received July 14, 1983; accepted October 1, 1983.

Abstract. Laboratory test results for the diagnosis of psychiatric illness usually are reported descriptively despite the ready availability of appropriate inferential statistics. A test's sensitivity, specificity, and diagnostic confidence are conditional probabilities. Confidence intervals may be calculated for these probabilities in any given study. Statistical tests for comparing the results of several studies use techniques for planned and posterior comparisons applied to contingency tables. These established statistical methods aid in the interpretation of laboratory test findings.

Key Words. Statistics, laboratory tests, depression, dexamethasone suppression test.

In the past few years, authors from different research centers have published work on the use of biological tests for the diagnosis of psychiatric illness, particularly endogenous depression (melancholia). The results of these investigations are often related to the ability of the diagnostic test to identify patients with the given illness (*sensitivity*) and the ability of the test to exclude patients who do not have this diagnosis (*specificity*). Given the prevalence of the disease in the population sampled, the predictive value (*diagnostic confidence*) of the procedure can be given (Galen and Gambino, 1975). Despite the ready availability of inferential statistical methods for just such data, research reports have remained at the descriptive level. The purpose of this article is to call attention to statistical techniques that can enhance our understanding of these diagnostic test results.

Casting Laboratory Results as Conditional Probabilities

The sensitivity, specificity, and diagnostic confidence of a test are, in effect, conditional probabilities which may be displayed in the familiar 2 × 2 contingency table. In Table 1, *a*, *b*, *c*, and *d* refer to observed frequencies of joint events, and N the total number of cases. Thus, *a* represents the number of cases both diagnosed as I (the index disorder) and for whom a positive test result is obtained. Similarly, cell *d* represents the number of patients diagnosed as not having the index disorder ($\bar{I}$) and for whom the test is negative. Sensitivity is defined as $a/(a + b)$, specificity as $d/(c + d)$, and

Neil Kalter, Ph.D., is Associate Professor, Departments of Psychology and Psychiatry, and Michael Feinberg, M.D., Ph.D., is Associate Professor, Department of Psychiatry, The University of Michigan, Ann Arbor, MI. Bernard J. Carroll, M.D., Ph.D., is Professor and Chairman, Department of Psychiatry, Duke University, Durham, NC. (Reprint requests to Dr. N. Kalter, P7202 Children's Psychiatric Hospital, University of Michigan Medical School, Ann Arbor, MI 48109, USA.)

0165-1781/83/$03.00

diagnostic confidence as $a/(a+c)$. All are conditional probabilities whereas prevalence of the disease, which is $(a+b)/N$, is not.

Table 1. Sensitivity, specificity, and diagnostic confidence cast in contingency table form

		Test results		
		+	—	
Patient group	I	a	b	(a + b)
	$\bar{I}$	c	d	(c + d)
		(a ± c)	(b + d)	N

Sensitivity = a/(a + b)
Specificity = d/(c + d)
Diagnostic confidence = a/(a + c)

The report that a diagnostic test has a sensitivity of, say, 60% thus represents the conditional probability a/(a + b), i.e., the rate of occurrence of positive test results in the index disorder group of patients. This descriptive statistic is helpful, but may be either a highly variable or, conversely, a stable estimate of the true value of a test's sensitivity in some defined population of patients. And if one wishes to compare sensitivity figures across studies, more than descriptive reports are needed.

Confidence Intervals for Conditional Probabilities

It is readily apparent that the conditional probabilities referred to as sensitivity, specificity, and diagnostic confidence are descriptive, *sample* statistics analogous to a sample mean. As is well known, a sample mean, $\bar{x}$, may be used to describe some N observed cases and also to draw inferences about the value of the population mean, μ. In a similar manner, a sample conditional probability describes some aspect of the N cases observed and can serve to estimate the value of the corresponding population conditional probability. For convenience, let P_c represent any sample conditional probability, e.g., sensitivity, and π_c the same conditional probability in the population from which the sample was drawn. As is the case for any summary statistic, P_c will show variation in value across repeated samples from the same population because P_c contains an error term. The question that arises is how good an estimate of π_c is any particular value of P_c? Put in another way, are the P_c figures across replications of a study likely to be very similar or markedly divergent? How much confidence can a researcher or practitioner have in a particular report of the sensitivity, specificity, and/or diagnostic confidence of a particular diagnostic test for some well-defined population?

The familiar method of calculating a confidence interval for a sample statistic rather than relying solely on a point estimate addresses these questions (Galen and Gambino, 1975). In studies of endogenous depression, diagnosis often is a dichotomous variable (has the index disorder or does not, I versus $\bar{I}$) as is the laboratory test result (positive

versus negative). In such instances, the usual equation for calculating confidence intervals for a binomial proportion may be used with N relatively large.

$$P_c \pm z_{\alpha/2} [P_c (1-P_c)/N]^{1/2} \quad (1)$$

For example, Carroll et al. (1981) report a sensitivity of 43%, a specificity of 96%, and diagnostic confidence of 94%, using the dexamethasone suppression test (DST) for endogenous depression. These data, cast as frequencies in a contingency table, are shown in Table 2.

Table 2. Data from Carroll et al., cast in contingency table form

	Dexamethasone suppression test result		
Diagnosis	+	—	**N**
Endogenously depressed	92	123	215
Not endogenously depressed	6	147	153
	98	270	368

Applying equation (1) to sensitivity (92/215 = 0.43) permits us to calculate the 95% confidence interval for the true sensitivity value of the population of interest.

$$0.43 \pm 1.96 [(0.43)(0.57)/215]^{1/2} \quad (2a)$$

$$0.43 \pm 1.96 (0.00114)^{1/2} \quad (2b)$$

$$0.43 \pm 0.066 \quad (2c)$$

Thus, the confidence interval ranges from 0.364 to 0.496. (In this equation, "N" refers to the row marginal, the denominator of the conditional probability.) The large sample size yields a relatively stable estimate of DST sensitivity for the population. We would not expect replications of this study to show very different results. One can have "confidence" in the 43% figure reported. But suppose the sensitivity figure of 43% had been based on considerably fewer cases as is often observed. If the sample had 23 cases of endogenous depression, then 10 cases would have had a positive test result (10/23 = 0.43). The confidence interval in this example would be 23% to 63% and one is less assured that the 43% reported is a good estimate of the population sensitivity. Yet sensitivity is reported as 43% in both instances.

Clearly, as in calculating a confidence interval for a mean, the length of the interval depends in part on the size of the sample. Other things being equal, sample size determines how good an estimate P_c is of π_c. Note, however, that if one is working with other than a dichotomous variable, e.g., a test result is divided into positive, negative, and indeterminate outcomes, or more than two patient groups are compared, equation (1) will not suffice. In such instances, Goodman's (1965) method for calculating simultaneous confidence intervals for multinomial proportions is suggested.

$$\frac{(A + 2f_{ij}) \pm \left\{A[A + 4 f_{ij} (n-f_{ij})/n]\right\}^{1/2}}{2(A + n)} \quad (3)$$

Here, f_{ij} is the frequency of cases in the cell of interest in a contingency table, n is the number of cases in the relevant row or column marginal, and A is the value of chi-square with one degree of freedom cutting off the upper 100 $(1-\alpha/k)\%$ of the chi-square distribution. (Note that when k = 2, the binomial case, A is set equal to the value of chi-square cutting off the upper $100[1-\alpha]\%$ of the chi-square distribution; α is not divided by k in the special case of the binomial.) The value of α is determined by the confidence interval desired, i.e., if the interval is to be 95%, then α equals 0.05, while k represents the number of categories into which a row or column of interest is divided.

It should be noted that equation (3) or (1) may be used with equivalent results only when k = 2. For example, recalculating the 95% interval for the 43% sensitivity figure of Carroll et al. (see Table 2), $f_{ij} = 92$, n = 215, and A = 3.84, the chi-square value with one degree of freedom for the 95th percentile, i.e., 100(1-0.05) in the chi-square distribution. Since k = 2, α is not divided by k. Applying equation (3) yields the following figures.

(4a) $$\frac{[3.84 + 2(92)] \pm \{3.84\,[3.84 + 4(92)\,(215-92)/215]\}^{1/2}}{2(3.84 + 215)}$$

(4b) $$\frac{187.84 \pm \{3.84\,[3.84 + 45264/215]\}^{1/2}}{437.68}$$

(4c) $$\frac{187.84 \pm [3.84\,(3.84 + 210.53)]^{1/2}}{437.68}$$

(4d) $$\frac{187.84 \pm (823.1817)^{1/2}}{437.68}$$

(4e) $$\frac{187.84 \pm 28.6911}{437.68}$$

This yields a confidence interval from 0.364 to 0.495. When dichotomous variables are used and N is large, equation (1) or (3) is appropriate. Since (1) is easier to apply, it is preferred.

However, when variables with more than two categories are used ($k > 2$), equation (1) is not appropriate and equation (3) is recommended. An example might be when DST result is divided into positive, negative, and indeterminate outcome (e.g., Feinberg and Carroll, 1982). A constructed data set representing this possibility is shown in Table 3. Here observed sensitivity is 40% (40/100), $f_{ij} = 40$, n = 100, k = 3, and $100(1-\alpha/k)\% = 98.3\%$ with $\alpha = 0.05$. The value of A, the chi-square value cutting off the upper 98.3% of the chi-square distribution with one degree of freedom, is most easily found by taking the square of normal z score at $(1-\alpha/2k)$. So z^2 at $(1-0.05/6) = (2.394)^2 = 5.73$. The confidence interval for the observed 0.40 specificity value then can be found using equation (3).

(5a) $$\frac{[5.73 + 2(40)] \pm (5.73\,[5.73 + 4(40)\,(100-40)/100])^{1/2}}{2(5.73 + 100)}$$

(5b) $$\frac{85.73 \pm [5.73\,(5.73 + 9600/100)]^{1/2}}{211.46}$$

(5c) $$\frac{85.73 \pm [5.73\,(5.73 + 96)]^{1/2}}{211.46}$$

(5d) $$\frac{85.73 \pm (582.9129)^{1/2}}{211.46}$$

(5e) $$\frac{85.73 \pm 24.1436)}{211.46}$$

Here the confidence limits are 0.291 and 0.523. This interval appropriately is longer than the one we would have obtained using equation (1) because it takes into account the multinomial rather than binomial situation represented by $k > 2$.

Table 3. Constructed data with k > 2

		DST result +	indeterminate	—	N
Criterion	I	40	12	48	100
Group	$\bar{I}$	6	8	86	100
		46	20	134	200

Comparing Conditional Probabilities Across Studies

Though this article focuses on statistical methods, it is important to note that different results across studies may be due to some immediately apparent or subtle methodological variations. For example, study A may focus on inpatients, while study B has used outpatients. Studies may use different assay methods or collect laboratory samples at different times of the day. More subtle differences can also contribute to divergent findings. One study may have a higher base rate, or prevalence, of the index disorder than others and prevalence affects diagnostic confidence figures. The statistical methods discussed here do not stand in the place of rigorous examination of the methodological aspects of studies to be compared. Rather they supplement them as is usually the case when one uses statistics prudently.

Suppose that one wishes to compare the sensitivity figures obtained across several studies. For each study the data may be represented in the contingency table format displayed in Table 1. (Note that when conditional probabilities are cast in contingency tables, frequencies rather than probabilities appear.) When one has several such tables and wishes to compare their sensitivity figures, the relevant row may be taken from each table and these several rows constitute a new contingency table. An omnibus

chi-square test may then be calculated for this new table, and, if significant, may be followed by specific posterior comparisons.

To illustrate this method, data bearing on the DST for endogenous depression reported by Carroll et al. (1981), Nuller and Ostroumova (1980), and Schlesser et al. (1980) are presented in contingency table form in section A of Table 4. Sensitivity reported by Carroll et al. is 43% (92/215); by Nuller and Ostroumova, 69% (36/52); and by Schlesser et al., 45% (66/146). Are these differences significant?

Table 4. Comparison of Carroll et al., Nuller and Ostroumova, and Schlesser et al., for sensitivity

A.

		Carroll et al. DST result			Nuller and Ostroumova DST result			Schlesser et al. DST result		
		+	—	N	+	—	N	+	—	N
Criterion	I	92	123	215	36	16	52	66	80	146
Group	$\bar{I}$	6	147	153	8	77	85	0	151	151
		98	270	368	44	93	137	66	231	297

B.

		DST result		
		+	—	N
Carroll et al.	I	92	123	215
Nuller and Ostroumova	I	36	16	52
Schlesser et al.	I	66	80	146
	I	194	219	413

C.

		DST result		
		+	—	N
Carroll et al. plus Nuller and Ostroumova	I	128	139	267
Nuller and Ostroumova	I	66	80	146
		144	219	413

Though methods for comparing proportions are commonly available, we find it most helpful to represent such a test in the contingency table format shown in Table 4, part B. Here row I (patients diagnosed as having the index disorder, endogenous depression here) from each of the three studies are combined in a single table. This permits an omnibus test followed by posterior tests to be determined by visually inspecting the data (as is often the case in deciding upon posterior comparisons following the omnibus *F*-test in analysis of variance) or planned, orthogonal compari-

sons determined on a priori basis. A clear description of orthogonal partitioning of contingency tables and the relevant chi-square calculations is given by Bresnahan and Shapiro (1966). Applying a test to compare a collection of conditional probabilities is equivalent to an omnibus test for the table shown in Table 4, part B. Displaying the table suggests where the posterior comparisons most fruitfully should be made. If, instead, orthogonal comparisons are to be conducted, the table permits visualizing whether such comparisons are nonredundant (orthogonal).

Returning to the illustration at hand, the omnibus chi-square for the data in Table 4, part B is 12.03, $df = 2$, $p < 0.001$. As when working in an analysis of variance design, visual inspection is used to indicate which specific comparisons may be significant. The Nuller and Ostroumova data look different from figures reported in the other two studies. A posterior comparison following Smith's (1966) technique, which is analogous to a Scheffé contrast, may be done by combining the data reported by Carroll et al. and those reported by Schlesser et al., and then comparing them to Nuller and Ostroumova's data. This comparison is shown in Table 4, part C. The resulting chi-square, calculated in the usual way, is 11.82. The degrees of freedom from the original 3 × 2 table in part B are used, yielding $p < 0.01$. The additive properties of chi-square and of degrees of freedom for orthogonal contrasts reveal that the Nuller and Ostroumova-Carroll et al. comparison must be nonsignificant. The usual omnibus chi-square for Table 4, part B equals 12.03-11.82 for Table 4, part C = 0.21 which is not significant.

One may visually inspect the data displayed in Table 4, part B and decide that the sensitivity figure reported by Nuller and Ostroumova (69%) looks different from both Carroll et al. (43%) and Schlesser et al. (45%). This process leads to two nonorthogonal, posterior tests; Nuller and Ostroumova versus both Carroll et al. and Schlesser et al. separately. To avoid spurious statistical significance, it is suggested that the methods for calculating a test for partial association in a contingency table (Bresnahan and Shapiro, 1966) and the degrees of freedom from the original 3 × 2 table to assess these resulting chi-square values (Smith, 1966) both be used. Note that since these comparisons involve analyzing subsets of the overall table shown in Table 4, part B, the Bresnahan and Shapiro equations are appropriate. Tests on tables derived by *collapsing* rows and/or columns of some original table (as in the example given above in which the Carroll et al. and Schlesser et al. rows are combined and tested against the Nuller and Ostroumova row) permit calculating chi-square in the usual way, but using the degrees of freedom from the original table to assess significance. When one or more rows/columns of an initial table are *deleted* to form a comparison, e.g., Nuller and Ostroumova versus Carroll et al. disregarding Schlesser et al., the following equation applies.

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{m} n^2_{ij}/e_{ij} - \sum_{i=1}^{l} o^2_{i.}/e_{i.} - \sum_{j=1}^{m} o^2_{.j}/e.j + O^2/E \tag{6}$$

Here a subtable has been constructed from some original table and contains l rows and m columns; n_{ij} = observed frequencies in cells of the new, partitioned table; $o_{i.}$ = an observed row marginal of the new table; $o_{.j}$ = an observed column marginal in the new

table; O = total N in the new table; e_{ij} = expected cell frequencies in the new table calculated using the marginals and N of the *original* table; $e_{i.}$ = an expected row marginal in the new table obtained from summing the e_{ij} values for cells appearing in the new, reduced table; $e_{.j}$ = an expected column marginal in the new table derived similarly to $e_{i.}$; and E = the sum of all e_{ij} values in the new reduced table. Using this equation to compare Nuller and Ostroumova versus Carroll et al. for sensitivity, we obtain the following results.

(7a)
$$\chi^2 = \frac{36^2}{(52)(194)/413} + \frac{16^2}{(52)(219)/413} + \frac{92^2}{(215)(194)/413} + \frac{123^2}{(215)(219)/413} - \left[\frac{215^2}{(215)(194)/413 + (215)(219)/413} = \frac{52^2}{(52)(194)/413 + (52)(219)/413}\right] - \left[\frac{128^2}{(215)(194)/413 + (52)(194)/413} + \frac{139^2}{(215)(219)/413 + (52)(219)/413}\right] + \frac{267^2}{(215)(194)/413 + (215)(219)/413 + (52)(194)/413 + (52)(219)/413}$$

(7b)
$$\chi^2 = (53.0579 + 9.2842 + 83.808 + 132.7021) - (215 + 52) - (130.6342 + 136.4659) + 267$$

(7c)
$$\chi^2 = 278.8522 - 267 - 267.1001 + 267$$

(7d)
$$\chi^2 = 11.75,\ df = 2,\ p < 0.01$$

(Note that many of the calculations in the various denominators in (7a) are redundant so that the equation is not so laborious as it first appears.) Applying the same method to the Schlesser et al. versus Nuller and Ostroumova comparisons yields chi-square = 8.88, $df = 2$ (degrees of freedom from the original table shown in Table 4, part B), $p < 0.01$. These two comparisons are nonorthogonal (partially redundant) and thus the resulting chi-square values may not be summed or subtracted.

As with any sample statistic, observing a difference between figures reported for two or more samples is insufficient to conclude that there is a "real" difference between the populations. For example, Holsboer et al. (1980) report a sesnsitivity of 24% (14 of 59 endogenously depressed patients showing a positive DST result) while Brown et al. (1979) found DST sensitivity to be 40% (8 of 20 endogenously depressed patients had a positive test result). The findings of these two studies can be compared by constructing a 2 × 2 table in which only the "I" rows from each are used (see Table 5). The usual omnibus chi-square test may be applied to the data in Table 5, part C (as was done when comparing three studies simultaneously in Table 4, part B). The resulting chi-square is 1.97, $df = 1$, NS. Although the rate reported descriptively by Brown et al. is nearly twice as high as that found by Holsboer et al., this difference is within expectable sampling fluctuation.

Conclusions

Reports of laboratory tests to detect various psychiatric disturbances can be cast in the

form of contingency tables and appropriate conditional probabilities can be derived.

Table 5. Comparison of Brown et al. and Holsboer et al.

A. Brown et al.

		DST result		
		+	—	N
Criterion	I	8	12	20
Group	$\bar{I}$	0	29	29
		8	41	49

B. Holsboer et al.

		DST result		
		+	—	N
Criterion	I	14	45	59
Group	$\bar{I}$	6	37	43
		20	82	102

C. Sensitivity comparison of Brown et al. and Holsboer et al.

		DST result		
		+	—	N
Brown et al.	I	8	12	20
Holsboer et al.	I	14	45	59
		22	57	79

Sensitivity, specificity, and diagnostic confidence are terms for the various conditional probabilities of interest. The typical report of such data usually stops short of available inferential statistical procedures. Yet two circumstances compellingly suggest the wisdom of going beyond descriptive sample statistics. In one situation, the clinician-researcher wishes to know how much confidence to have in a particular value for sensitivity, specificity, or diagnostic confidence. This may be accomplished by computing confidence intervals around the obtained sample conditional probability of interest. The second circumstance that calls for the use of inferential statistical techniques involves the comparison of conditional probabilities obtained using a particular laboratory test across two or more studies. These tests permit researchers to decide whether or not such "differences" are explained most parsimoniously as expectable sampling fluctuations in the value of a given conditional probability. Such statistical tests consist of constructing a comparison table by "peeling off" the relevant

row or column from each of the original contingency tables, and then analyzing the resulting table for cross-study differences using the usual chi-square tests as well as posterior tests and tests on partitions of the comparison table. Together, these uses of inferential statistical methods strengthen and make more precise conclusions based on information gleaned from diagnostic laboratory tests.

References

Bresnahan, J.L., and Shapiro, M.M. A general equation and technique for the exact partitioning of chi-square contingency tables. *Psychological Bulletin,* **66,** 252 (1966).

Brown, W.A., Johnston, R., and Mayfield, D. The 24-hour dexamethasone suppression test in a clinical setting: Relationship to diagnosis, symptoms and response to treatment. *American Journal of Psychiatry,* **136,** 543 (1979).

Carroll, B.J., Feinberg, M., Greden, J.F., Tarika, J., Albala, A.A., Haskett, R.F., James, N.McI., Kronfol, Z., Lohr, N., Steiner, M., deVigne, J.P., and Young, E. A specific laboratory test for the diagnosis of melancholia: Standardization, validation and clinical utility. *Archives of General Psychiatry,* **38,** 15 (1981).

Feinberg, M., and Carroll, B.J. Separation of subtypes of depression using discriminant analysis: I. Separation of unipolar endogenous depression from non-endogenous depression. *British Journal of Psychiatry,* **140,** 384 (1982).

Galen, R.S., and Gambino, S.R. *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses.* Wiley, New York (1975).

Goodman, L.A. On simultaneous confidence intervals for multinomial proportions. *Technometrics,* **7,** 247 (1965).

Holsboer, F., Klein, H., Bender, W., and Benkert, O. Hypothalamic-pituitary-adrenal activity in a group of 100 heterogenic depressed patients; diagnostic validity and biochemical aspects of the cortisol response to dexamethasone suppression. (Abstract #297) *Progress in Neuro-Psychopharmacology,* suppl., 180 (1980).

Nuller, J.L., and Ostroumova, M.N. Resistance to inhibiting effect of dexamethasone in patients with endogenous depression. *Acta Psychiatrica Scandinavica,* **61,** 169 (1980).

Schlesser, M.A., Winokur, G., and Sherman, B.M. Hypothalamic-pituitary-adrenal axis activity in depressive illness. *Archives of General Psychiatry,* **37,** 737 (1980).

Smith, J.E.K. Posterior comparisons in contingency tables. *Michigan Mathematical Psychology Program Monographs,* **4,** 1 (1966).