AN APPLICATION TO A LABORATORY PROBLEM OF DISCRIMINANT FUNCTION ANALYSIS INVOLVING MORE THAN TWO GROUPS

BY

R. F. JENNISON, J. B. PENFOLD, and J. A. FRASER ROBERTS

From the Research Department, Royal Eastern Counties Institution, Colchester, and the Pathology Laboratory, Essex County Hospital, Colchester

We have recently attempted to assess the value of a modification of the Wassermann reaction, based upon the suggestion of Richardson (1940). The results are being described in a pathological journal (Jennison, Penfold, and Roberts, 1949). The present paper deals with certain points which emerged during the statistical analysis, and which have a rather more general interest in regard to the application of statistical methods to biological data.

MATERIAL AND METHODS

The technique for the standard Wassermann test was that described by Panton and Marrack (1945). Three tubes were made up as follows:

Tube	Serum c.cm.	Saline c.cm.	Complement (3 M.H.D. in 0.5 c.cm.)	Antigen c.cm.	Sensi- tized cells c.cm.
1	0.1	0.4	0.5	0.5	1.0
2	0.1	0.2	0.7	0.5	1.0
3 (control)	0.2	0.8	0.5	0	1.0
(4)	0.2	0.3	0.5	0.5 diluted 1 in 7	1.0)

Absence of haemolysis in tubes 1 and 2 indicates a positive result; complete haemolysis in both tubes a negative result; partial haemolysis in either or both a doubtful result. Richardson's suggested modification is very briefly this: excess of antigen as in the usual Wassermann reaction does not provide the optimum proportions of antigen and antibody for the fixation of complement; suitable dilution of the antigen gives a proportion nearer the optimum in syphilis, while the non-specific antigens are diluted, usually beyond the point at which they will give their least reaction. In these circumstances the fixation increases (less haemolysis) with syphilitic and decreases (more haemolysis) with non-specific sera,

thereby increasing respectively the sensitivity and specificity of the test. Accordingly in the modified technique a fourth tube was added as shown above. Theory would suggest that when the amount of haemolysis in tube 4 was less than or equal to that in tube 1 the result is likely to be positive; when greater than in tube 1 probably negative.

In order to assess the value of the addition of tube 4, 1,000 sera were chosen for the modified test out of some 9,000 routine Wassermann and Kahn reactions. These were deliberately selected as difficult sera:

1. Those showing incomplete fixation with 3 M.H.D. of complement.
2. Those showing a discrepancy between Wassermann and Kahn reactions.
3. Those from treated cases which had not been consistently negative previously with the modified test.
4. Those from cases giving positive routine Wassermann or Kahn reactions, but with no clinical evidence of syphilis.
5. Those from antenatal clinics giving a positive or doubtful presumptive Kahn reaction.

The 1,000 sera were classified into four clinical groups:

- Group I. Untreated syphilis.
- Group II. Treated syphilis.
- Group III. Probable or possible syphilis, i.e. some history or physical signs of the disease.
- Group IV. No history or signs of syphilis.

The four groups should thus represent decreasing proportions of true sero-positive syphilis and the criterion against which the technique is assessed is its success in differentiating between them.

Where haemolysis in any tube was partial, its amount was measured to the nearest 10 per cent.

RESULTS

The results for the 1,000 sera by the standard technique, not using tube 4, are shown in Table I Including now the additional information given by

the readings in tube 4, of the 646 sera with complete haemolysis in tubes 1 and 2, 546 also gave complete haemolysis in tube 4, while 100 showed partial haemolysis or no haemolysis. The distribution of the 646 sera between the groups is shown in Table II.

TABLE I
RESULTS OF THE STANDARD WASSERMANN TECHNIQUE

	NUMBERS					
	Group					
	I	II	III	IV	Total	
	Positive	44	95	20	3	162
Doubtful	13	137	23	19	192	
Negative	11	299	97	239	646	
Total	68	531	140	261	1,000	
	PERCENTAGES					
	Group					
	I	II	III	IV		
	Positive	65	18	14	1	
	Doubtful	19	26	17	7	
Negative	16	56	69	92		
	100	100	100	100		

TABLE II
RESULTS GIVEN BY TUBE 4 WITH 646 SERA NEGATIVE BY TUBES 1 AND 2

Group	Number	Also negative in Tube 4	Partial or no haemolysis in Tube 4	Per cent. with partial or no haemolysis in Tube 4
I	11	5	6	54.5
II	299	216	83	27.8
III	97	88	9	9.3
IV	239	237	2	0.8
Totals	646	546	100	

This result is a striking one. It is clear that fixation, complete or partial, in tube 4, when there has been complete haemolysis in tubes 1 and 2, does strongly suggest a positive serum; tube 4 has thus undoubtedly improved the sensitivity of the test.

As shown in Table I, 192 sera gave partial haemolysis in tube 1, or 2, or both. Their distribution in relation to tube 4 is shown in Table III, distinguishing those in which haemolysis in tube 4

was less than or equal to that in tube 1 from those in which it was greater. This is also a striking

TABLE III
RESULTS GIVEN BY TUBE 4 WITH 192 SERA DOUBTFUL BY TUBES 1 AND 2

Group	Numbers			Percentages	
	Haemolysis 4 ≤ 1	Haemolysis 4 > 1	Total	Haemolysis 4 ≤ 1	Haemolysis 4 > 1
I	13	0	13	100	0
II	111	26	137	81	19
III	14	9	23	61	39
IV	3	16	19	16	84
Total	141	51	192		

result. Haemolysis in tube 4 which is less than or equal to that in tube 1 suggests a positive serum; haemolysis in tube 4 which is greater than in tube 1 a negative serum; tube 4 has thus undoubtedly improved the specificity of the test.

Only 5 sera previously positive, i.e. with no haemolysis in tubes 1 or 2, showed partial haemolysis in tube 4. Three of these sera belonged to group II and two to group III.

It will be noted that the application of the simple rule of taking the difference between the readings in tubes 1 and 4 applies equally to the sera of Table II and of Table III, though it does make the 5 previous full positives "probably negative." There might be, however, an element of artificiality in applying the rule to the 105 previous full positives or full negatives; for with the full positives any haemolysis, however slight, must indicate "probably negative," while with the previous full negatives any fixation, however slight, must indicate "probably positive." Nevertheless, it is clear that a straightforward and very simple deduction from theory does lead to a considerable measure of useful discrimination, judged, as always, by the proportions in the four clinical groups, with their decreasing proportion of true sero-positive syphilis.

FURTHER STATISTICAL ANALYSIS

It seemed profitable to examine the data further, and without preconception, in order to see whether the actual readings in the three tubes could be used in such a way as to secure improved discrimination between the clinical groups. The data are now the 297 (100+192+5) sera in which there was partial haemolysis in one or more tubes, for it seems better not to complicate the analysis with the remaining

FIG. 1.—Tube 4 only—per cent. haemolysis. Variance ratio 12·10. Each rectangle represents a single individual.

TABLE IV
ANALYSIS OF VARIANCE: MEAN SQUARES AND CROSS PRODUCTS

Variation	Degrees of Freedom	1 ²	2 ²	4 ²	1·2	1·4	2·4	X ²	X ₁ ²	D ²
Between groups ..	3	5·147	6·165	138·029	-0·537	-6·778	28·076	387·739	380·239	156·732
Within groups ..	293	14·686	8·464	11·410	8·168	5·706	3·655	22·948	22·671	14·684
Variance ratio ..		<1	<1	12·10				16·90	16·77	10·67

X²=discriminant function; X₁²=rounded off function; D²=difference, tubes 1 and 4.

703 which are either fully positive or fully negative in all tubes. The basic data are given in the Appendix Tables. Looking first at the haemolysis readings in the three tubes separately, we find:

		Mean haemolysis (per cent.)		
		Tube 1	Tube 2	Tube 4
Group 1		55.3	78.4	27.9
Group 2		63.7	81.6	37.1
Group 3		62.6	85.6	48.8
Group 4		59.0	89.5	80.5

The percentage readings in tubes 1 and 2 are useless for the further classification of doubtful sera. In tube 1 the means are not even in the right order; in tube 2 the differences are too small to be of any value. In terms of the analysis of variance shown in Table IV, with both tubes the mean square between groups is less than the mean square within groups. The result with tube 4 is very different, the variance ratio of Table IV being no less than 12.10. Actually this is a better result than that given by the difference method of the previous section, which yields the somewhat lower variance ratio of 10.67. Doubtful sera would be rather more efficiently classified by using the figure for percentage haemolysis in tube 4 instead of the difference between tubes 1 and 4.

Although tubes 1 and 2 taken alone provide no useful discrimination, it remains possible that some appropriate combination of all three readings might give better discrimination than tube 4 alone. In the analysis there are four groups. The theory of discriminant function analysis with more than two groups was elucidated by Fisher (1938, 1940), though not very much use seems to have been made up to the present of a method which must have applications in a wide variety of biological problems. The subject is fully discussed, with numerical examples, in a forthcoming book by Prof. K. Mather (1949), to whom we are greatly indebted for showing us the typescript of the section appropriate to our problem.

It is clear that what is required is a single function, for the differences between the four clinical groups are quantitative; we need, therefore, that combination of the three variables which maximizes the difference between groups relative to the variation within groups. The computation involves little more than the evaluation of a determinant of order corresponding to the number of variables. It would, of course, be possible in the present problem

to reduce the groups to two by, say, combining groups I and II and omitting group III. But this means losing information, and could not be done at all in many problems to which the method is applicable, and in which the number of groups is indefinitely large: for example, if it were desired to obtain a function which maximized the variation between sibships relative to that within sibships.

FIG. 2.—(Tube 1)–(Tube 4). Variance ratio 10.67. Each rectangle represents a single individual.

Let A_{11} , A_{12} , etc., be the total sums of squares and cross products, and a_{11} , a_{12} , etc., the sums of squares and cross products between groups; then the coefficients for the readings in tubes 1, 2, and 4 are found from the equations:

$$\begin{aligned} b_1(A_{11} - \phi a_{11}) + b_2(A_{12} - \phi a_{12}) + b_4(A_{14} - \phi a_{14}) &= 0 \\ b_1(A_{21} - \phi a_{21}) + b_2(A_{22} - \phi a_{22}) + b_4(A_{24} - \phi a_{24}) &= 0 \\ b_1(A_{41} - \phi a_{41}) + b_2(A_{42} - \phi a_{42}) + b_4(A_{44} - \phi a_{44}) &= 0 \end{aligned}$$

FIG. 3.—Discriminant function. $y = 105 + (\text{Tube 1}) - \frac{1}{2}(\text{Tube 2}) - \frac{1}{2}(\text{Tube 4})$. Variance ratio 16.77. Each rectangle represents a single individual.

ϕ is estimated by evaluating the determinant

$$\begin{vmatrix} A_{11} - \phi a_{11} & A_{12} - \phi a_{12} & A_{14} - \phi a_{14} \\ A_{21} - \phi a_{21} & A_{22} - \phi a_{22} & A_{24} - \phi a_{24} \\ A_{41} - \phi a_{41} & A_{42} - \phi a_{42} & A_{44} - \phi a_{44} \end{vmatrix} = 0$$

For the present material (using the percentage haemolysis readings in units of grouping ranging from 0 to 10) this is:

$$\begin{vmatrix} 4318 \cdot 45118 - \phi 15 \cdot 43964 & 2391 \cdot 60606 - \phi 1 \cdot 61057 & 1651 \cdot 45791 - \phi 20 \cdot 33504 \\ 2391 \cdot 60606 - \phi 1 \cdot 61057 & 2498 \cdot 54545 - \phi 18 \cdot 49606 & 1155 \cdot 21212 - \phi 84 \cdot 22895 \\ 1651 \cdot 45791 - \phi 20 \cdot 33504 & 1155 \cdot 21212 - \phi 84 \cdot 22895 & 3757 \cdot 36027 - \phi 414 \cdot 08803 \end{vmatrix} = 0$$

which yields:

$$155,980,507 - 23,766,775 \cdot 9\phi + 113,045 \cdot 085\phi^2 - 55 \cdot 1013107\phi^3 = 0$$

The lowest root is the one needed. This is:

$$\phi = 6 \cdot 7809485$$

Substituting in the above equations:

$$\begin{aligned} 4213 \cdot 75578b_1 + 2402 \cdot 52725b_2 + 1789 \cdot 34877b_4 &= 0 \\ 2402 \cdot 52725b_1 + 2373 \cdot 12462b_2 + 584 \cdot 05995b_4 &= 0 \\ 1789 \cdot 34877b_1 + 584 \cdot 05995b_2 + 949 \cdot 44883b_4 &= 0 \end{aligned}$$

There are actually two equations relating the three coefficients, so only their relative magnitudes can

be determined. This, however, is all that is required. Putting $b_1 = 1$, we find:

$$\begin{aligned} b_1 &= 1 \\ b_2 &= -0 \cdot 64642653 \\ b_4 &= -1 \cdot 48696473 \end{aligned}$$

Functions of this type do not appear to be sensitive to moderate changes in the magnitudes of the coefficients, hence b_2 and b_4 can be rounded off to $-\frac{1}{2}$ and $-1\frac{1}{2}$ respectively; this reduces the variance ratio by less than 1 per cent. (In the calculations, however, it is necessary to carry a relatively large number of figures.) As is explained below, the best practical use of the function is for distinguishing "probably positive" from "probably negative" sera, instead of attempting any further sub-division. Scrutiny of the figures indicates that a suitable constant is 105; so the discriminant function is:

$$y = 105 + (\% \text{ haemolysis in tube 1}) - \frac{1}{2}(\% \text{ haemolysis in tube 2}) - 1\frac{1}{2}(\% \text{ haemolysis in tube 4})$$

A positive answer or zero is taken to indicate a

probably positive result; a negative answer a probably negative result.

Table IV gives the analysis of variance. The mean squares and cross-products of columns three to eight are calculated in the usual way. The mean squares for the discriminant function are calculated by using the coefficients determined above. Thus the mean square between groups is given by:

$$(b_1^2 \times 5 \cdot 147) + (b_2^2 \times 6 \cdot 165) + (b_3^2 \times 138 \cdot 029) + (2b_1b_2 \times -0 \cdot 537) + (2b_1b_3 \times -6 \cdot 778) + (2b_2b_3 \times 28 \cdot 076)$$

This for the squares and products of coefficients taken to five places of decimals gives the figure of 387.739 shown in the column headed X^2 and for the rounded off coefficients the figure of 380.239 shown in the column headed X_1^2 . The variance ratio of 16.77 in this column checks, as it must do, with the variance ratio calculated directly from the individual values used in preparing Fig. 3. Similarly, the last column is calculated by putting $b_1=1$ and $b_4=-1$. For the mean square between groups:

$$(1 \times 5 \cdot 147) + (1 \times 138 \cdot 029) + 2(-1 \times -6 \cdot 778) = 156 \cdot 732$$

Again the variance ratio of 10.67 checks with the variance ratio calculated directly from the individual values of Fig. 2.

We have, then, the rather remarkable result that although tubes 1 and 2 are by themselves useless for the further classification of doubtful sera, yet in combination with tube 4 they yield a function having a substantial advantage over that tube used alone. Furthermore, owing to the possibility of rounding off the coefficients with only trivial loss of information, it is a very simple function which can be applied in a moment in the laboratory. Table V shows the result of using the function in this way for the 297 doubtful sera.

TABLE V

DISCRIMINATION OF 297 DOUBTFUL SERA ATTAINED BY USE OF FUNCTION

Group	Probably positive $y \geq 0$	Probably negative $y < 0$	Per cent. probably positive
I	19	0	100
II	212	11	95
III	27	7	79
IV	5	16	24

The three figures illustrate graphically the discrimination between the clinical groups secured by

the three methods (1) the reading in tube 4, (2) the simple difference method—tubes 1 and 4, (3) the discriminant function.

It will be noted that statistical analysis confirms theoretical expectation, in that the most important feature of the function turns out to be the difference in the readings in tubes 1 and 4. The function, just like the difference method, makes all previous full negatives with anything less than complete haemolysis in tube 4 probably positive. The principal advantage lies in a resorting of some of the sera doubtful by the standard technique. A previous full positive remains probably positive unless the haemolysis in tube 4 is 80 per cent. or more. This seems a sensible result; the highest haemolysis in the five such sera of the present series was 40 per cent. An advantage of the function which is not perhaps fully reflected by the variance ratios is the greater range of values; as haemolysis is measured in steps of 10 per cent., the readings in tube 4 and the difference method lead to a limited and somewhat inflexible set of figures. This is well illustrated in the figures; group II—the most important group—is well spread out instead of being crowded closely against the boundary line.

The distribution of the percentage haemolysis figures, as shown in the appendix tables, is markedly non-normal; in fact it tends to be J- or U-shaped. Yet, as so often happens, the application of straight-forward linear methods has given a useful result. In view of this no attempt has been made to find a more complex function. There is no simple way of doing this; any appreciable improvement is very problematical; and in any event such a function would lead to a formula that would be too complicated for rapid laboratory use. The non-normal distribution of the variables has one consequence, however. It would be unsafe to rely too closely on the actual magnitude of the answer; a large figure, positive or negative, does not necessarily indicate a more positive or more negative result than one somewhat nearer zero. In practice this is no disadvantage. The recommended international reporting, using the standard technique, is to classify sera into three groups only: positive, negative, and doubtful. There could be no question of setting up a considerable number of categories; two intermediate ones are enough. At the Essex County Hospital Pathology Laboratory we now simply report the intermediate sera, according to the figure given by the function, as "doubtful (positive by the modified technique)" or "doubtful (negative by the modified technique)."

APPENDIX TABLES

HAEMOLYSIS READINGS IN THE THREE TUBES FOR THE 297 SERA WITH PARTIAL HAEMOLYSIS IN AT LEAST ONE TUBE
(Units of grouping: 0, 1, 2, 10 represent percentage haemolysis 0, 10, 20 100.)

GROUP I

Tube 2

	0	1	2	3	4	5	6	7	8	9	10	
Tube 1	0	1			2		1				1	5
1												
2					1					1		2
3	1											1
4												
5												
6										2		2
7												
8										1		1
9										2		2
10										6		6
	1	1			3		1			13		19

GROUP I

Tube 4

	0	1	2	3	4	5	6	7	8	9	10	
Tube 1	0	5										5
1												
2			2									2
3	1											1
4												
5												
6	2											2
7												
8	1											1
9							1		1			2
10	1					1	2			2		6
	10		2			1	3		1	2		19

GROUP I

Tube 2

	0	1	2	3	4	5	6	7	8	9	10	
Tube 4	0	1	1		2		1				5	10
1												
2					1					1		2
3												
4												
5										1		1
6										3		3
7												
8										1		1
9										2		2
10												
	1	1			3		1			13		19

GROUP II													
Tube 2													
	0	1	2	3	4	5	6	7	8	9	10		
Tube 1	0	3	1	8	3	6	2	3	3	3	1	2	35
	1		2	3	1		1			1	1		9
	2	1		1	1	4		2		1		5	15
	3			1				1			1		3
	4				1	1	2	1		1	2	5	13
	5								2			2	4
	6			1				2		1		6	10
	7									3	1	6	10
	8								1	1	3	26	31
	9							1			1	6	8
	10				1						1	83	85
	4	3	14	7	11	5	10	6	11	11	141	223	

GROUP II

		Tube 4											Tube 1
		0	1	2	3	4	5	6	7	8	9	10	
0	25		3	1	4	1			1				35
1	4	2		1		1	1						9
2	9	1	2	1	1							1	15
3	2							1					3
4	6		4		2					1			13
5			1					1	1		1		4
6	4		1	2		1				1		1	10
7	1	1			5		1				1	1	10
8	4	2	8	3	3	2	3			3	1	2	31
9	2		2	1				1		1		1	8
10	9	3	6	1	12	6	8	8	18	13	1		85
		66	9	27	10	27	11	16	10	24	16	7	223

GROUP III

		Tube 2											Tube 1
		0	1	2	3	4	5	6	7	8	9	10	
0	2			1		1	1						5
1													
2									2	1	1		4
3													
4						1					1		2
5													
6											4		4
7											1		1
8						1					4		5
9											4		4
10											9		9
		2		1		1	3		2	1	24		34

GROUP II

		Tube 2											Tube 4
		0	1	2	3	4	5	6	7	8	9	10	
0	1	2	11	5	8	4	6	3	3	2	21		66
1			1							1	1	6	9
2	1				3		2			1	2	18	27
3	1	1						1		1		6	10
4	1		1			1	1			1	1	21	27
5				1							1	9	11
6			1							1	1	13	16
7									1	1	1	7	10
8									1		1	22	24
9										1	1	13	16
10				1							1	5	7
		4	3	14	7	11	5	10	6	11	11	141	223

GROUP III

		Tube 4											Tube 1
		0	1	2	3	4	5	6	7	8	9	10	
0			1		1	1					1	1	5
1													
2	1		1									2	4
3													
4	2												2
5													
6				1	1		2						4
7					1								1
8	2									1		2	5
9						1					1	2	4
10	2		3		2				1	1			9
		7	5	1	5	2	2	1	2	2	2	7	34

GROUP III

Tube 2

	0	1	2	3	4	5	6	7	8	9	10	
0							1		1		5	7
1												
2	1									1	3	5
3											1	1
4	1										4	5
5						1					1	2
6											2	2
7											1	1
8											2	2
9				1							1	2
10							2		1		4	7
	2		1			1	3		2	1	24	34

GROUP IV

Tube 2

	0	1	2	3	4	5	6	7	8	9	10	
0										1		1
1												
2				1		1					1	3
3									1			1
4											2	2
5						1					1	2
6									1			1
7											4	4
8											2	2
9											3	3
10											2	2
				1		2			2	1	15	21

GROUP IV

Tube 4

	0	1	2	3	4	5	6	7	8	9	10	
0						1						1
1												
2	1									1	1	3
3											1	1
4						1					1	2
5						1					1	2
6											1	1
7											4	4
8											2	2
9											3	3
10	1									1		2
	2					2	1		1		14	21

GROUP IV

Tube 2

	0	1	2	3	4	5	6	7	8	9	10	
0					1						1	2
1												
2												
3												
4						1					1	2
5										1		1
6												
7											1	1
8												
9						1						1
10									2		12	14
					1	2			2	1	15	21

REMAINING 703 SERA

	0 (all tubes)	10 (all tubes)
Group I	44	5
Group II	92	216
Group III	18	88
Group IV	3	237

SUMMARY

An attempt has been made to assess the value of a modification of the standard Wassermann reaction by applying it to 1,000 difficult sera divided into four clinical groups representing decreasing proportions of true sero-positive syphilis. A function which maximizes the variation between groups relative to that within groups was found to give a substantial improvement in discrimination as compared with other methods of using the readings. It is suggested that more use might be made of discriminant function analysis with multiple groups:

a technique which must have many applications in a wide variety of biological problems.

REFERENCES

- Fisher, R. A. (1938). *Ann. Eugenics*, 8, 376.
——— (1940). *Ibid.*, 10, 422.
Jennison, R. F., Penfold, J. B., and Roberts, J. A. F. (1949). "The Value of Diluted Antigen in the Wassermann Reaction." In the press.
Mather, K. (1949). "Biometrical Genetics: The Study of Continuous Variation." London: Methuen. In the press.
Panton, P. N., and Marrack, J. R. (1945). "Clinical Pathology." Fifth Edition. London: Churchill.
Richardson, G. M. (1940). *Brit. J. vener. Dis.*, 16, 166.