J. Clin. Chem. Clin. Biochem.
Vol. 24, 1986, pp. 601–609

A Distribution-Free, Multivariate Discriminating Method

Initial experience with the discrimination of patient groups by 2 or more clinical chemical parameters

By *Bernhard Keller* and *Herbert Keller*

Institut für Klinische Chemie und Haematologie des Kantons St. Gallen

(Received October 21, 1985/April 6, 1986)

Summary: We communicate a distribution-free quasi graphic procedure for obtaining a linear discriminating function. The method is based on the following considerations:

Suppose a group A is to be separated from a group B using two parameters X and Y. The centre of each group is defined as the median (or mean or mode) of its points. Of all straight lines passing through any point of A and any point of B those are retained which intersect the segment joining the centres of A and B. For each of these the number of wrongly allocated points is calculated (i. e. the points which do not lie on the same side of the straight line as their group centres). In this way one obtains straight lines with maximal separating power (for the two groups given). Finally, each optimal line is rotated in such a way that its defining points are also correctly allocated.

If more than two parameters are available a stepwise procedure can be used: the distance from the separating straight line obtained from the first two parameters is introduced as a new parameter, which is then combined with the third parameter to yield a new discriminating function which depends on all three parameters. Iterating this step one can combine any number of parameters.

The method was implemented on a personal computer.

It was first applied to a textbook model (chances of survival for M. haemolyticus neonatorum estimated by concentrations of haemoglobin and bilirubin in cord blood). As a second model we used a tetravariate study (primary hyperparathyreoidism versus other hypercalcaemic diseases, estimated by the blood concentrations of calcium, phosphate and chloride, and by haematocrit).

In addition to the learning collectives, 2 test collectives were available: a group of primary hyperparathyreoidism patients from another clinic and a group of patients with secondary hyperparathyreoidism from a third clinic. In this orientating study the new method shows good discriminating capacity.

Eine verteilungsfreie, multivariate Diskriminanzmethode

Zusammenfassung: Es wird ein verteilungsfreies, quasi graphisches Verfahren zur Ermittlung einer linearen Diskriminanzfunktion beschrieben, das auf folgenden Überlegungen basiert:

Angenommen ein Kollektiv A sei von einem Kollektiv B zu trennen, wobei Parameter X and Parameter Y als Variate dienen. Nun wird im Kollektiv A und B der Schwerpunkt (oder der Median oder das Dichtemittel) festgelegt und die beiden Schwerpunkte durch eine Gerade verbunden. Der nun eingesetzte Algorithmus lautet: von jedem Punkt des Kollektivs A wird zu jedem Punkt des Kollektivs B eine Verbindungs-Gerade gezogen. Jene Geraden, welche die Schwerpunkt-verbindende Strecke schneiden, bewirken eine Trennung der

beiden Kollektive. Es wird nun für jede einzelne dieser Geraden untersucht, wie groß die Anzahl der falsch zugeordneten Punkte ist, wobei jeder Punkt als falsch zugeordnet betrachtet wird, der nicht auf der gleichen Seite der Geraden wie der Schwerpunkt seines Kollektivs liegt. In dieser Weise werden die Geraden mit der maximalen Trennwirksamkeit ermittelt, anschließend wird jede optimale Gerade so gedreht, daß auch die definierenden Punkte korrekt zugeordnet werden.

Wenn mehr als zwei Parameter zur Verfügung stehen, so muß schrittweise vorgegangen werden: die Abstände von der Trenngeraden bezüglich der ersten beiden Kenngrößen wird als neuer Parameter betrachtet und ihm ein dritter Parameter gegenübergestellt. In dieser Weise kann beliebig oft fortgefahren werden. Zur Durchführung des Rechenverfahrens wurde ein Programm für einen Personal-Computer entwickelt.

Das neue Verfahren wurde zunächst an einem Lehrbuch-Modell (Überlebenschancen bei Morbus haemolyticus neonatorum, geschätzt aus der Hämoglobin- und Bilirubin-Konzentration im Nabelschnur-Blut) erprobt.

Ferner konnte eine tetravariate klinische Studie überprüft werden (primärer Hyperparathyreoidismus versus andere hypercalcaemische Erkrankungen, abgeschätzt aus den Kenngrößen Hämatokrit, Chlorid, Calcium und Phosphat). Außer den beiden Lernkollektiven stand ein Kollektiv von primären Hyperparathyreoidismus-Patienten aus einer anderen Klinik, sowie ein Kollektiv von Patienten mit sekundärem Hyperparathyreoidismus aus einer dritten Klinik zur Verfügung. Bei diesen orientierenden Untersuchungen erwies sich das Verfahren als sehr leistungsfähig.

Introduction

If two or more distinct groups are to be discriminated by two or more features, methods of discriminant analysis are used (1–9). The "classical" linear discriminant analysis (10) presupposes a multivariate normal distribution. Frequently this type of distribution is not given (or its presence cannot be verified, e. g. because of lack of data). This may be the reason why frequently unsatisfactory results are obtained. Therefore different non-parametric methods of discriminant analysis have been developed (5, 11, 12) by many authors.

Procedure

Our new, quasi "graphic" procedure is based on the following considerations:

If two collectives are to be separated by two characteristics, a dissecting straight line (= discriminant function) is found, which separates all points so that none or a minimal number of points is wrongly classified (allocated). It is possible that several straight lines have the same power of separation (we will not enter into the possibility that a curve-linear function provides the optimal dissection).

We demonstrate our procedure by a model (tab. 1): A scatter diagram (fig. 1) of parameter X and Y shows the values of group A and B. (In fig. 1 the points are numbered for easier reference). For each group (cluster) the mean values (or the medians or the modes) are calculated (S_A, S_B respectively).

If one or both variates have a discriminating power with respect to the groups then the means (or medians or modes) cannot be identical. The greater the discriminating power, the greater the distance between S_A and S_B will be.

Now the following algorithm is applied: from each point of group A a straight line to each point of collective B is drawn. All lines which do not intersect the connecting line (S_AS_B) are ignored. For all lines that intersect the connecting line, the number of incorrectly allocated points is determined, applying the following rule: a point is considered to be incorrectly allocated if it does not lie on the same side of the straight line as its group mean (or mode or median). Figure 2 shows the six possible straight lines from A_1 to B_{1-6}; (A_1B_1), (A_1B_5) and (A_1B_6) do not intersect (S_AS_B). The discriminating effect of the three other lines is given in table 2.

Subsequently for point A_2 all dissecting lines are searched and it is checked whether the discriminating power is better than for the preceding lines. Continuing in this way the discriminating power of all possible lines is investigated. Finally the straight line (or lines) is retained, which provides the smallest number of wrong allocations. In our example this is true for two straight lines (a_1b_3) and (a_6b_1), where just one point is misclassified: B_2 (tab. 3).

The optimal straight line is now rotated in such a way that the defining points are correctly classified, too.

Tab. 1. Model collectives A and B with variables X and Y.

	X	Y
A_1	1.5	1.5
A_2	0.5	2.5
A_3	1.0	3.5
A_4	2.0	4.0
A_5	3.0	3.0
A_6	2.75	2.27

	X	Y
B_1	2.0	1.5
B_2	2.5	2.75
B_3	4.0	2.5
B_4	4.5	1.5
B_5	3.5	0.5
B_6	2.5	0.5

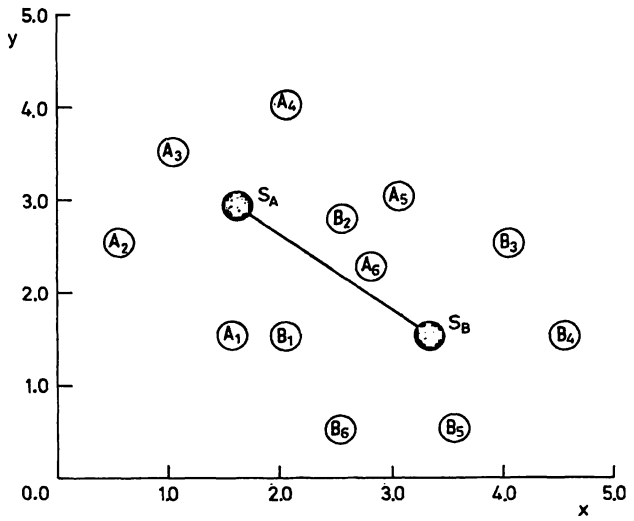

Fig. 1. Model of collectives A and B to be discriminated by parameters X and Y; graphic representation of tab. 1.
S_A = centre of collective A
S_B = centre of collective B
(S_A S_B) = straight line joining the centres

Figure 3 shows the two optimal lines which give only one incorrect classification. The discrimination by the classical linear discriminant analysis is worse, with two incorrect allocations (B_2 and A_6). The distances to the new discriminating line can be used as a new parameter (XY) combining the original parameters X and Y. This new parameter might be the X axis of a new scatter diagram where a third parameter forms the Y axis; one can now investigate whether some other discriminating line gives a better discrimination.

This new procedure produces lines which are approximately identical with those from the classical linear discriminant analysis, provided that a sufficient number of multivariate normally distributed points are

Tab. 2. Discriminating effect of different dissection straight lines.

Dissection straight-lines	False points
a_1 b_2:	A_5; A_6
a_1 b_3:	B_2
a_1 b_4:	B_2; B_3

Tab. 3. Demonstration of the discriminating steps shown by the display of the calculator. Column A and B correspond to the joining lines. The column "False" shows the false allocations of the corresponding dissecting straight line.

A	B	False
1	2	2
1	3	1
6	1	1

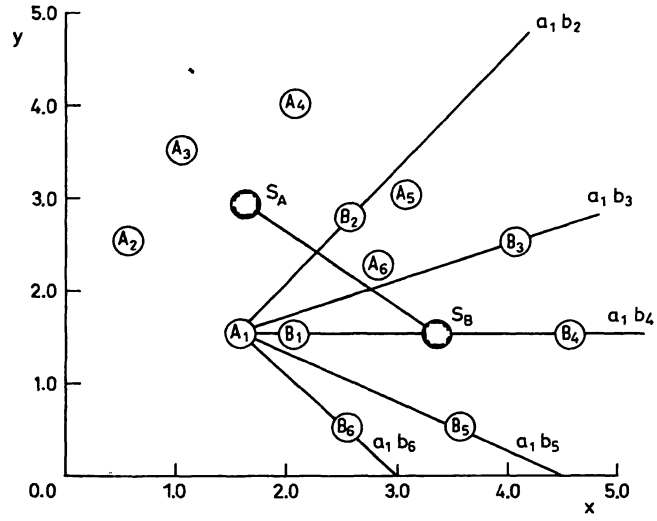

Fig. 2. Straight lines from A_1 to B_{1-6}; (S_A S_B) is dissected by (a_1b_2), (a_1b_3) and (a_1b_4).

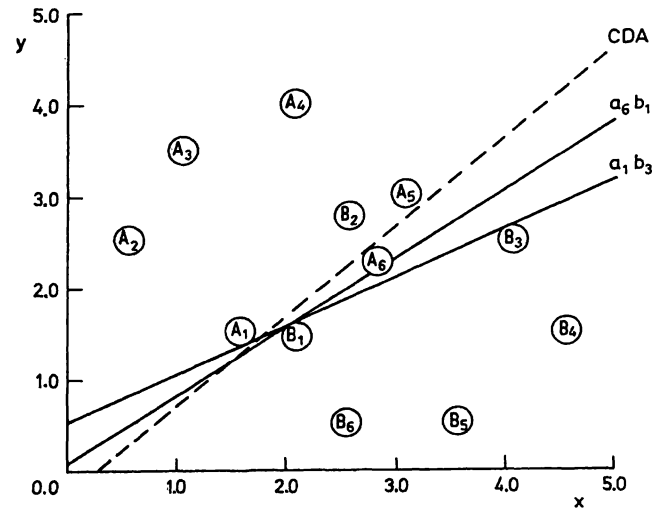

Fig. 3. The optimal dissecting straight lines (a_1b_3) and (a_6b_1) correctly discriminate all pairs except B_2. The dissecting straight line of the "classical" discriminant analysis (CDA) yields two false allocations: B_2 and A_6.
$Z = 0.52604 \cdot X - 1 \cdot Y + 0.5534$

available. In any case, the straight line, obtained in this way, is the best possible for the given "learning collective". If more than one optimal discriminating line is found, each one must be investigated to determine whether the addition of a new parameter leads to a better discrimination. It is possible that important differences occur, although the discriminating lines showed identical power of separation in the first step.

Application

For the application in practice we have developed a program ("OPTIGER") in BASIC for the HP 87XM (see Addendum)[1]).

For small collectives — e. g. 12 points in our model — the calculation is performed in a few seconds. If the number of points is considerably larger (> 100), then the procedure takes a few hours.

On the order "determination of the optimal discriminating line" the computer screen displays the individual steps of the procedure. As an example, the discriminating steps of our model are shown in table 3. Also the classical discriminant analysis can be performed on request by the same program OPTIGER. Less time is necessary for this calculation, e. g. only a few minutes for > 100 points.

During the construction of the optimal discriminating line, the predominance of false positive over false negative allocations or vice versa, is not taken into account. Consequently the diagnostic sensitivity or specificity predominates. Using the OPTIGER program a shifting of the discriminating line is possible in order to improve diagnostic sensitivity or specificity. Nevertheless in any case a shifting causes a deterioration of the discriminating power.

The efficiency of OPTIGER has been tested, using an example from a well known textbook (13), and on a clinical study from the literature. Other examples have been published separately (14, 15).

Examples of application

Textbook-model

Discriminant analysis for the assessment of survival in cases of Icterus haemolyticus neonatorum, (2 parameters available)

In this example the concentrations of haemoglobin and bilirubin in the cord blood of new born babies are determined. It is required to predict the chance of survival by means of these two parameters. Out of 79 infants affected by the haemolytic disease 63 survived and 16 died. The histogram (fig. 4) shows that with both parameters the groups overlap widely. It is not possible to assess the chance of survival by the haemoglobin or the bilirubin level alone.

Another picture shows the scatter diagram (fig. 5): obviously the survivors are concentrated in the lower right part, and those that died in the upper left.

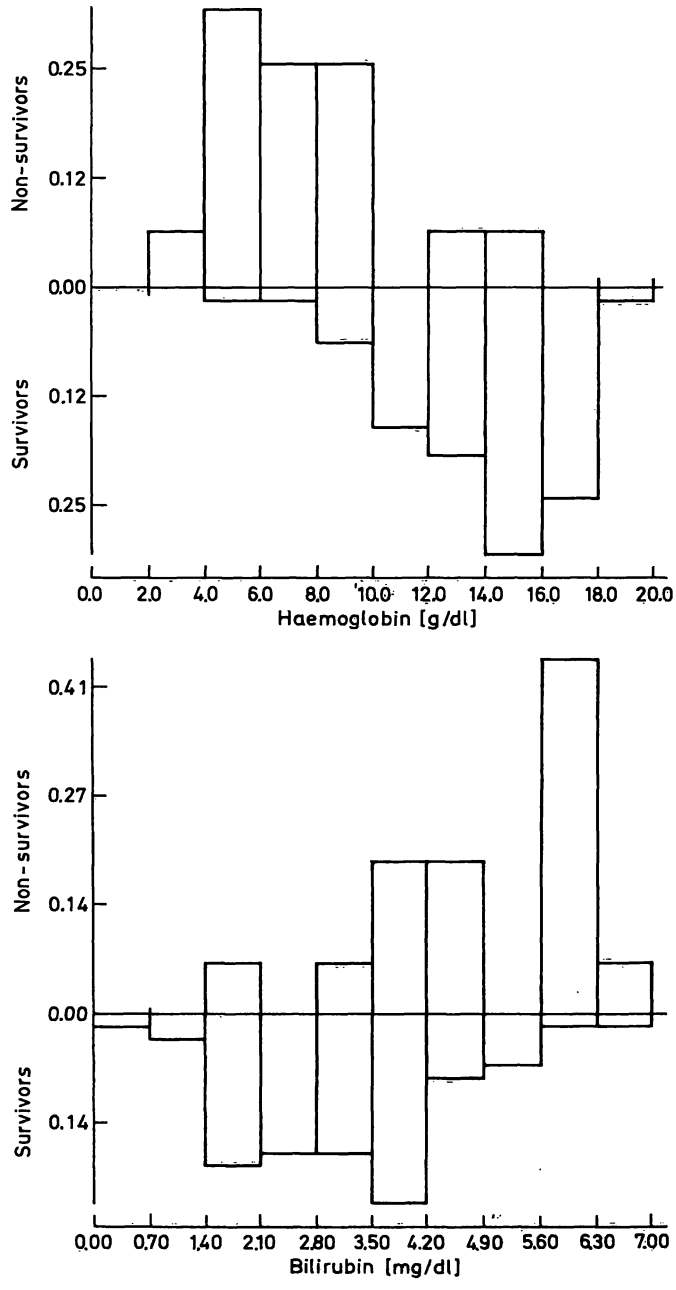

Fig. 4. Histogram of haemoglobin and bilirubin concentrations in cord blood of newborns, suffering of M. haemolyticus neonatorum. Non-survivors (N = 16) top, survivors (N = 63) bottom. Both collectives are calculated as 100%.

[1]) A listing can be received from the authors on request.

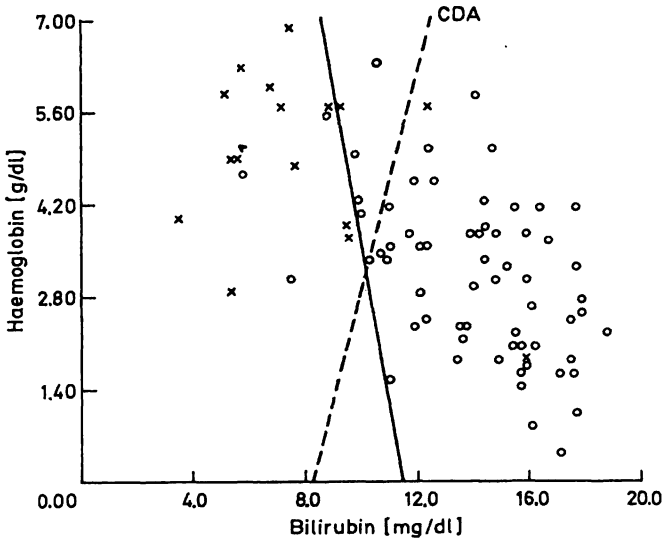

Fig. 5. Scatter diagram of figure 4. With the straight line constructed by classical discriminant analysis (CDA) 11 patients are falsely allocated; with the OPTIGER-line 5 patients are falsely allocated.
o = survivors; * = non-survivors; OPTIGER 1 = −2.5027 · haemoglobin − 1 · bilirubin + 28.703

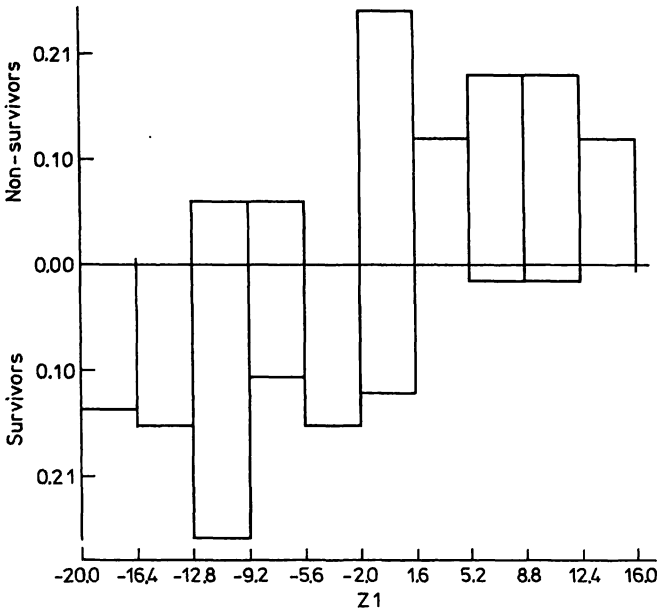

Fig. 6. Histogram in analogy to fig. 4, the function
Z = 2.5 Hb − Bilirubin + 28.7
is used as parameter.

The calculation of the classical discriminant analysis results in a discriminating function, whereby 54 of 63 survivors are correctly classified (and 9 are misclassified), and 14 of the 16 that died are correctly classified (and 2 are misclassified), i. e. 68 out of 79 (86%) are correctly classified.

The OPTIGER program finds two straight lines with correct classifications of 60/63 survivors and 14/16 non-survivors. There are only 5 misclassifications, 94% being correctly allocated by OPTIGER. It is noteworthy to state that both discriminating lines are nearly vertical on the X axis, which means that haemoglobin has a much higher discriminatory power than bilirubin.

The histogram of the distribution with the new parameter Z 1 of both collectives is shown in figure 6. By the new discriminant function,

$$Z_1 = 2.5\ \mathrm{Hb} - \mathrm{Bilirubin} + 28.7,$$

the chances of survival improve with increasing negativity of the resulting value, and vice versa.

Tetravariate clinical study

In order to discriminate between the primary hyperparathyreoidism and other hypercalcaemic disorders, several authors have recommended discriminant analysis (16—20).

Recently a study by *Lafferty* (Cleveland) (21) was published on this subject. For the discrimination of hypercalcaemic patients he used haematocrit, chloride, calcium and phosphate, and for some of the patients he also used parathormone. We received the complete data base from the author[2]). It contains 100 patients suffering from primary hyperparathyreoidism, verified by biopsy; 31 patients with bone metastases of malign diseases, 20 patients with pseudohyperparathyreoidism, 4 patients with vitamin-D intoxication, 3 patients with thyreotoxicosis and 6 patients with hypercalcaemia from other diseases. The histograms of the 4 parameters are shown in figure 7 where the patients with primary hyperparathyreoidism are at the top and all other 64 non-hyperparathyreoidism patients at the bottom. The calcium concentration in serum is lower in the hyperparathyreoidism group compared with the other patients, but there is a broad overlaping zone.

The same is true for phosphate in serum, where the hyperparathyreoidism patients also show lower values.

Within the distribution of the chloride concentrations the hyperparathyreoidism group shows a peak, significantly higher than in the non-hyperparathyreoidism group.

[2]) We thankfully acknowledge the generosity of Dr. *Lafferty*.

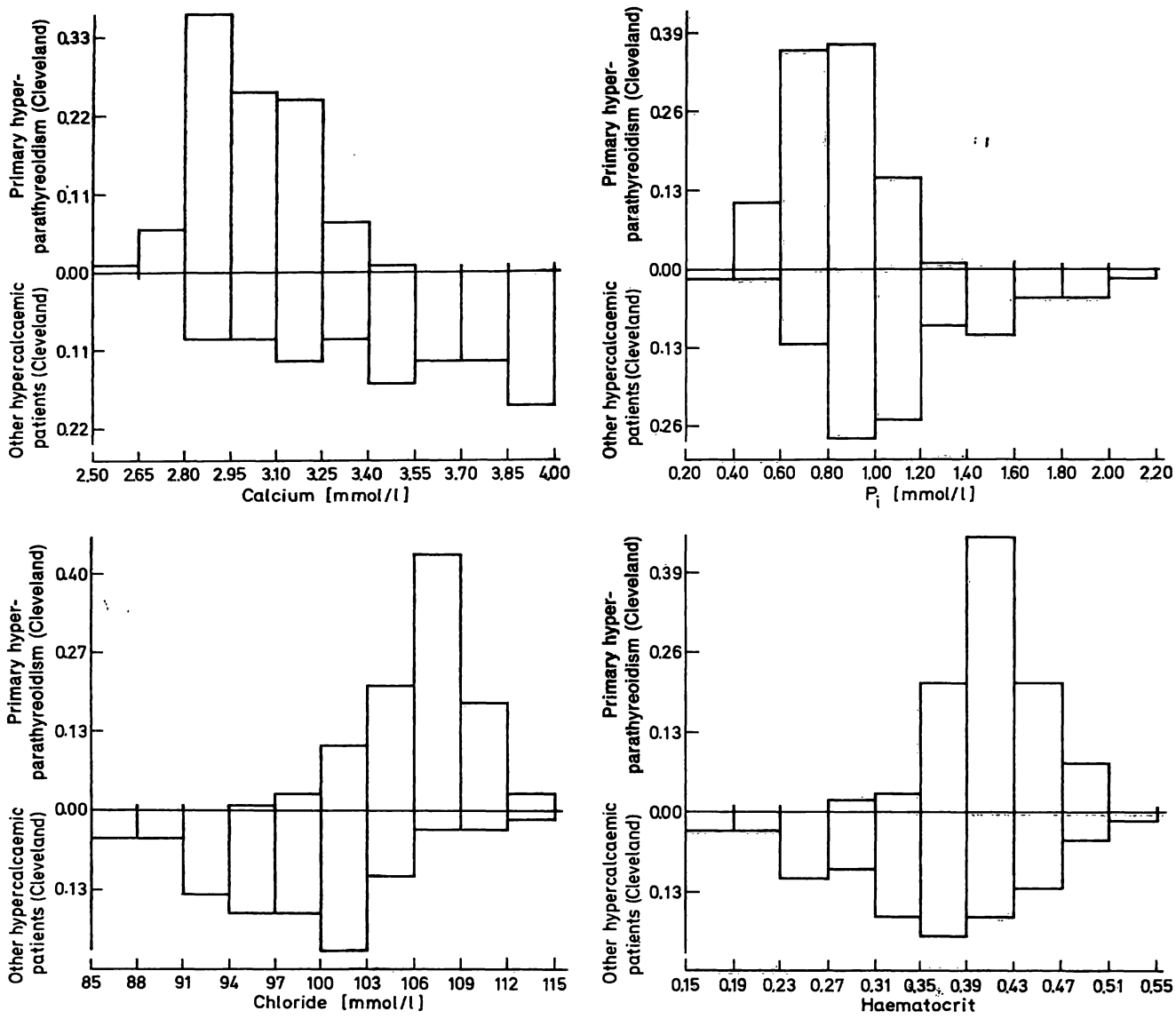

Fig. 7. Histograms of patients with primary hyperparathyreoidism from the Cleveland clinic (top), and other hypercalcaemic patients from the Cleveland clinic (bottom): both collectives are calculated as 100%.

Also the mean haematocrit value is higher in hyperparathyreoidism patients, again with complete overlaping.

Parathormone was not included in our calculations because its concentration was not determined for the majority of patients. Furthermore, parathormone is an analytically uncertain parameter with insufficient standardization and insufficient comparability from one laboratory to another.

The univariate validity of each parameter does not suffice to solve the diagnostic problem. We therefore tried a bivariate combination. With 4 parameters, 6 bivariate combinations are possible. Figure 8 shows the scatter diagram of calcium versus phosphate with the optimal discriminating line, Z_1, calculated by the computer. The discriminating line allocates 94/100 hyperparathyreoidism patients correctly, and 55/64 non-hyperparathyreoidism patients; i. e. 15 patients

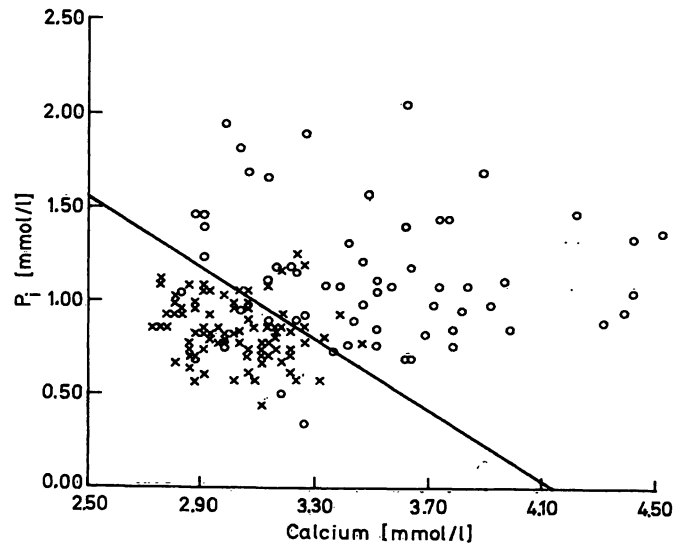

Fig. 8. Scatter diagram of primary hyperparathyreoidism collective from the Cleveland clinic (*) and other hypercalcaemic patients from the Cleveland clinic (o)
Ca vs P_i
$Z1 = -0.95697 \cdot Ca - 1 \cdot P_i + 3.9501$

were wrongly classified. The discriminating power of the classical discriminant analysis is lower, 18 patients being incorrectly allocated. A new scatter diagram was constructed with the parameter Z_1 on the X-axis and chloride on the Y-axis (fig. 9). With OPTIGER one can find 6 optimal discriminating lines of identical discriminating power: 95/100 hyperparathyreoidism patients and 58/64 non-hyperparathyreoidism patients are correctly classified: 11 patients are misclassified. The new discriminant function, Z_2, is obviously better than Z_1.

With Z_2 on the X-axis and haematocrit on the Y-axis, the new scatter diagram gives even better discriminating by Z_3 (fig. 10): now 98/100 hyperparathyreoidism patients and 62/64 non-hyperparathyreoidism patients are correctly classified, the number of misclassifications being only 4. The distribution of the "learning-group" under Z_3 is shown in figure 11.

The discriminating power of Z_3 was controlled by two other independent patient groups: 37 samples from patients with verified primary hyperparathyreoidism from the clinics of the Zürich University, and 37 samples from patients of our own hospital, suffering of secondary hyperparathyreoidism. The discrimination of the Zürich hyperparathyreoidism patients from the non-hyperparathyreoidism group is nearly as perfect as that for the learning group (fig. 12). 35/37 Zürich hyperparathyreoidism patients are correctly classified. As expected, the distribution of the patients with secondary hyperparathyreoidism is worse (fig. 13). Only 29/37 St. Gall secondary hyperparathyreoidism patients are correctly classified.

With the function,

Z_3 = 19.4 calcium (mmol/l) + 20.3 phosphate (mmol/l) − 15.5 chloride (mmol/l) − haematocrit 100 + 115,

a primary hyperparathyreoidism becomes more probable with increasing negativity of the results, and vice versa.

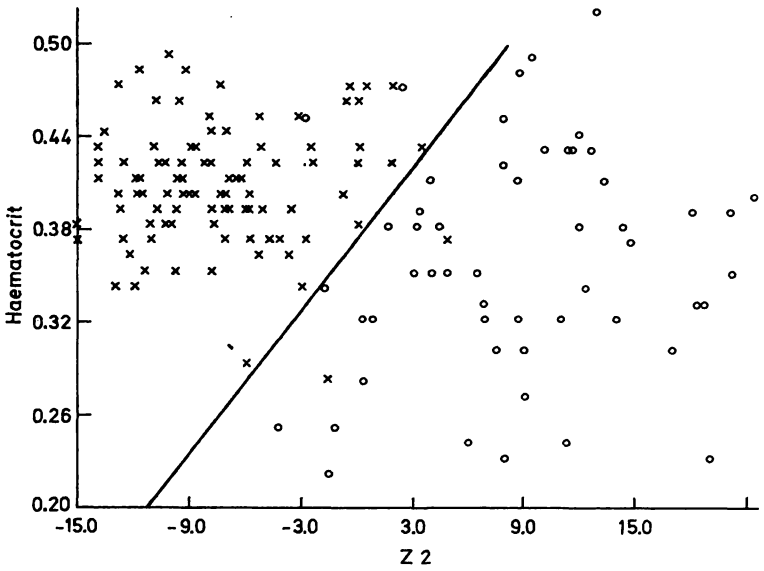

Fig. 10. Scatter diagram of primary hyperparathyreoidism collective from the Cleveland clinic (*) and other hypercalcaemic patients from the Cleveland clinic (o).
Z2 vs haematocrit
Z3 = + 19.39438 · Ca + 20.26731 · P_i −1.545 · Cl − 100 · haematocrit + 115.05043

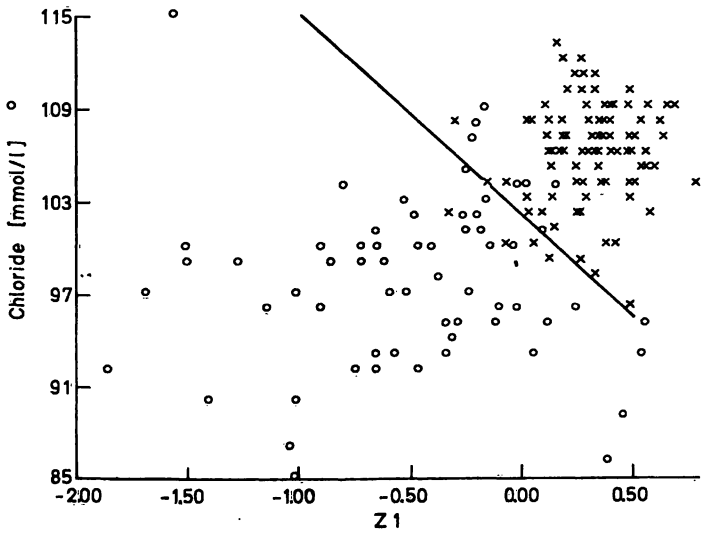

Fig. 9. Scatter diagram of primary hyperparathyreoidism collective from the Cleveland clinic (*) and other hypercalcaemic patients from the Cleveland clinic (o)
Z1 vs Cl^-
Z2 = + 12.55505 · Ca + 13.12 · P_i − 1 · Cl + 50.246

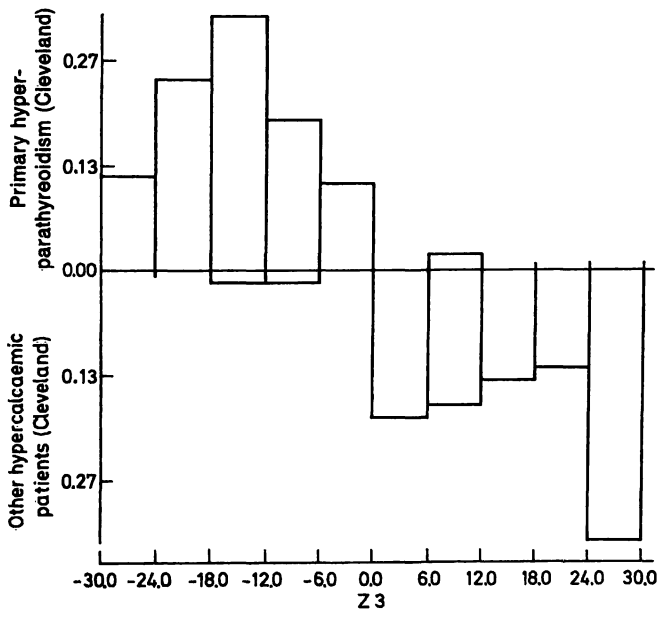

Fig. 11. Histogram of primary hyperparathyreoidism collective from the Cleveland clinic (top), and other hypercalcaemic patients from the Cleveland clinic (bottom), with Z3 as parameter.

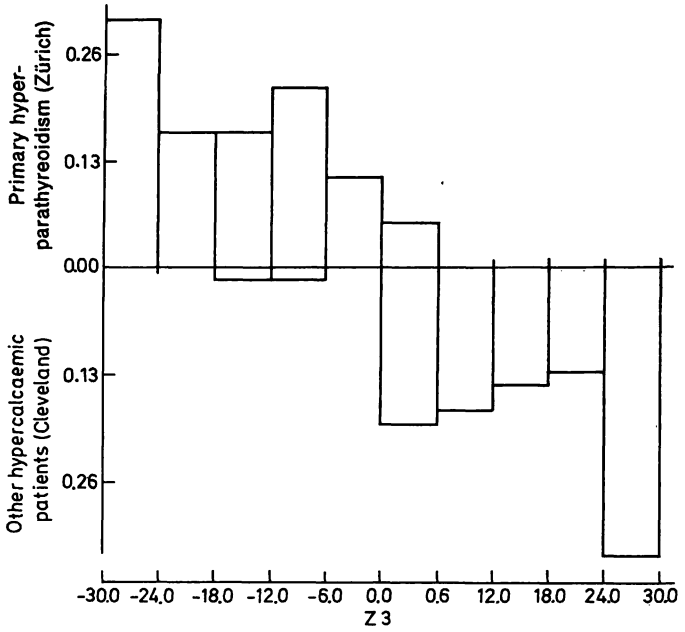

Fig. 12. Histogram of a patient collective with primary hyperparathyreoidism from the Zürich clinics (top) and other hypercalcaemic patients from the Cleveland clinic (bottom).

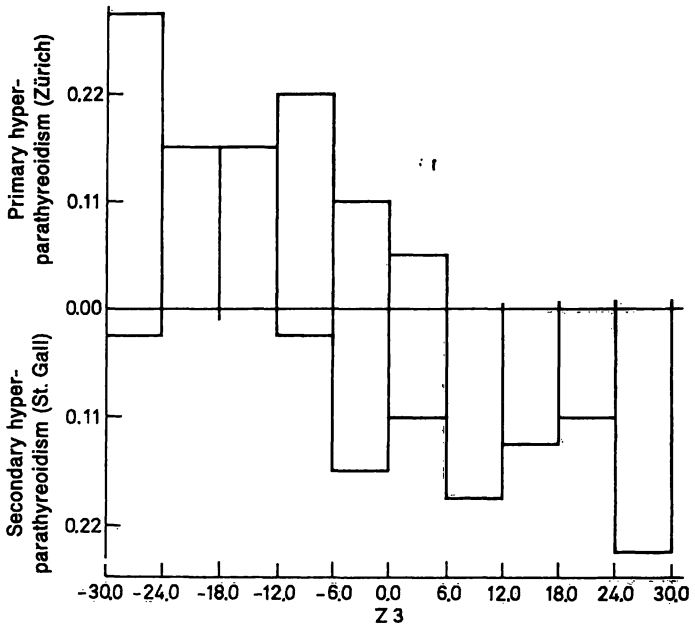

Fig. 13. Histogram of primary hyperparathyreoidism collective from the Zürich clinics (top) and a patient collective with secondary hyperparathyreoidism from the St. Gall clinic (bottom).

It should be pointed out that the function Z_3 is valid only for hypercalcaemic patients. If normocalcaemic patients are to be separated from hyperparathyreoidism patients (an illogical problem!) it would be necessary to construct a new discriminating function using another learning group.

Addendum

I) Denotations

Collective	number of points	points	index	centre
A	n	$A_1 \ldots A_n$	i	S_A
B	m	$B_1 \ldots B_n$	j	S_B

II) Algorithms

A point is false with respect to a straight line, if it is not on the side of the centre of its collective

F_g = number of false points with respect to the straight line g

F_{min} = minimum of F_g

1 calculate S_A, S_B
2 let $F_{min} = m + n$
3 i: = 1
4 j: = 1
5 determine the straight line $g = (A_i, B_j)$
6 if g does not dissect ($S_A S_B$), go to 9
7 determine F_g
8 if $F_g \leqslant F_{min}$, store g and let F_{min}: = F_g
9 j: = j + 1
10 if $j \leqslant m$, go to 5
11 i: = i + 1
12 if $i \leqslant n$, go to 4
13 rotate the optimal straight line $g = (A_i, B_j)$ by an adequate angle about the center of A_i, B_j
14 list the optimal straight lines

References

1. Lachenbruch, P. A. (1975) Discriminant analysis. Hafner Press, New York.
2. Solberg, H. E. (1975) Discriminant analysis in clinical chemistry. Scand. J. Lab. Clin. Invest. *35*, 705–712.
3. Anderson, T. W. (1958) An introduction to multivariate statistical analysis. John Wiley New York.
4. Ellis, G. & Goldberg, D. M. (1979) Discriminant function analysis applied to laboratory tests in patients with hepatobiliary disease. Comp. Biomed. Res. *12*, 483–501.
5. Flury, B. & Riedwyl, H. (1983) Angewandte multivariate Statistik. G. Fischer Verlag, Stuttgart.
6. Durbridge, T. C. (1984) Applying descriptive discriminant analysis as a visual aid for physicians interpreting biochemical test results. Clin. Biochem. *17*, 321–326.
7. Michaelis, J. (1972) Zur Anwendung der Diskriminanzanalyse für die medizinische Diagnostik. Habil. Schrift, Mainz.

8. Hermans, J., van Zomeren, B., Raatgever, J. W., Sterk, P. J. & Habbema, J. B. F. (1981) Use of posterior probabilities to evaluate methods of discriminant analysis. Meth. Inform. Med. *20*, 207–212.
9. Goldschmidt, H. M. J. & Leyten, J. F. (1985) Medicometrics – a new, promising discipline. In: Computing in Clinical Laboratories (Trendelenburg, Chr., ed.) Proc. 5th Internat. Conference, Stuttgart (FRG) June 1985.
10. Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems, Annals of Eugenics VII, part II, pp. 179–188.
11. Birkenfeld, W. (1978) One step towards a nonparametric discriminant analysis. In: COMPSTADT 1978. Proceedings in computational statistics. 3rd symposium held in Leiden 1978. Physica Verlag, Wien, pp. 162–169.
12. Hansert, E., Federkiel, H. & Stamm, D. (1984) A new procedure for discriminating between two patient populations using multivariate decision limits: Application in the detection, and exclusion of alcoholism based on clinical laboratory findings. J. Clin. Chem. Clin. Biochem. *22*, 791–810.
13. Armitage, P. (1971) Statistical methods in medical research. Blackwell Scientific Publ., Oxford (reprint 1980).
14. Riesen, W. F., Mordasini, R. & Keller, H. (1983) Chemial relevance of apolipoproteins. Proc. Intern. Sympos.: New aspects on lipoprotein metabolism, Heidelberg 1983.
15. Riesen, W. F., Mordasini, R. & Keller, H. (1985) Zur klinischen Bedeutung der Apolipoproteine. Lab. Med. *9*, 339–344.
16. Fraser, P., Healthy, M., Rose, N. & Watson, L. (1971) Discriminant functions in differential diagnosis of hypercalcemia. Lancet *II*, 1314–1319.
17. Fraser, P., Healthy, M., Rose, N. & Watson, L. (1976) Further experience with discriminant functions in differential diagnosis of hypercalcemia. Postgrad. Med. J. *52*, 254–257.
18. Transbøl, J., Jorgensen, F. S., Hornum, J. & Keiding, N. (1977) Hypercalcaemia discrimination index: A multivariate analysis of parathyroid function in 107 hypercalcaemic patients. Acta Endocrinol. *86*, 768–783.
19. Transbøl, J. (1977) On the diagnosis of so-called normocalcaemic hyperparathyreoidism. Acta. Med. Scand. *202*, 481–487.
20. LoCascio, V. H., Vallaperta, P., Adami, S., Cominacini, L., Galvanini, G., Bianchi, I., Ferrari, J. & Scuro, L. A. (1978) Discriminant analysis in differential diagnosis of hypercalcaemia. Clin. Endocrinol. *8*, 349–356.
21. Lafferty, F. W. (1981) Primary hyperparathyreoidism. Arch. Intern. Med. *141*, 1761–1766.

Professor Dr. Dr. H. Keller
Institut für Klinische Chemie
und Hämatologie des Kantons
CH-9000 St. Gallen