Reprint from

RECENT ADVANCES IN DOPING ANALYSIS (4)

W. Schänzer
H. Geyer
A. Gotzmann
U. Mareck-Engelke
(Editors)

Sport und Buch Strauß, Köln, 1997

U. FLENKER, E. NOLTEERNSTING, W. SCHÄNZER, M. DONIKE:
Detection and Classification of Different Male Steroid Profiles by Means of Multivariate Statistics
In: W. Schänzer, H. Geyer, A. Gotzmann, U. Mareck-Engelke (eds.) Recent advances in doping analysis (4). Sport und Buch Strauß, Köln, (1997) 210-203

U. Flenker, E. Nolteernsting, W. Schänzer, M. Donike

Detection and Classification of Different Male Steroid Profiles by Means of Multivariate Statistics

Institute of Biochemistry, German Sports University Cologne

Abstract

The results of this study indicate that male steroid profiles can be subdivided into two phenotypes which appear to be quite discrete. The abundance of each type is different in different parts of Eurasia. The detection and description of the specific metabolic patterns is possible after use of cluster analysis. Classification into both groups *a posteriori* is done by linear discriminant analysis (*LDA*).

1 Introduction

When doing steroid profiling, a considerable number of male persons who show remarkable low concentrations of urinary testosterone can be observed. Figure 1 shows the distribution which resulted from all male routine samples in Cologne in 1994. Two maxima clearly can be recognized. This observation indicates that males could be subdivided in two subpopulations showing different steroid metabolism.

Unfortunately the two parts of the distribution show a large region of overlap. Therefore the introduction of at least one additional variable is necessary to seperate between the two groups.

It is often a promising approach to add dimensions when a system under study cannot be described sufficiently using only one or two parameters. Looking at more than one parameter at the same time may give far better insight into a paticular situation. For example groups can appear to be discrete in space or even hyperspace, although they may overlap at lower dimensions.

Of course this method requires that additional variables exist, which are related to the system in a significant manner. A typical example for the fulfilment of these conditions are metabolic pathways, where several substances show a variety of complex interactions.

Talking in terms of statistics, multivariate methods are more suitable than univariate ones

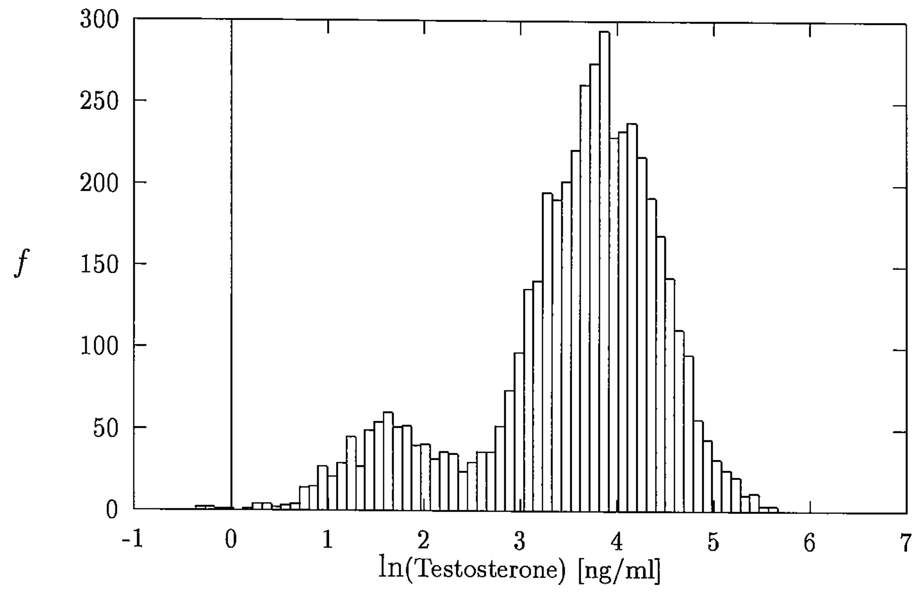

Figure 1: Distribution of natural logarithms of urinary testosterone concentrations from 4749 male athletes. The concentrations were corrected to urine density $1.02 g/cm^3$ using (1). f: Absolute frequency.

in those cases.

2 Material and Methods

The steroid profiles of all male routine samples collected in 1994 in the Cologne Laboratory were used. Samples with pH-values lower than 5 and greater than 7.6 were excluded from the study. Profiles containing clear faults in the evaluation of the GC/MS-data were excluded as well. Some profiles showed negative peak areas for instance or the selection of peaks with wrong retention times. 4749 profiles remained to be evaluated. Additionally the data of 135 male steroid profiles obtained from dope controls in the 1994 Asian Games (Hiroshima) were studied.

The variables included in the multivariate analysis were Testosterone (Test), 5α-Androstane-$3\alpha, 17\beta$-diole (5α-Diol) and 5β-Androstane-$3\alpha, 17\beta$-diole (5β-Diol).

The concentrations of all steroids were corrected for density using the Levine-Fahy equation (1), which could be shown to eliminate significant correlation between steroid concentrations and urine density [7]. C_c is the corrected density, C_m the concentration measured, ϱ_s the density used for standardization ($1.02 g/cm^3$) and ϱ is the density measured. After this proceedure all concentrations were transformed logarithmically.

(1)
$$C_c = \frac{\varrho_s - 0.998}{\varrho - 0.998} \cdot C_m$$

A randomized sample of $n = 500$ was drawn out of the whole entity to be used for clustering. This group thus was meant to constitute a representative *training set* to allow the cognition of patterns.

When performing cluster analysis standardized variables were used (Z-transformation), which avoids distorsions due to different scales of the concentrations [1,9]. To eliminate outliers a single linkage analysis was applied in a preceeding step. Arbitrarily those 15 objects that were agglomerated last were excluded from further analysis. Cluster analysis itself was performed using the centroid method. A good description of this and different other important methods is given by Vogt and Nagel [9]. To estimate the optimal number of groups in the population the elbow criterion was used. The whole run of clustering generally follows the recommendations of Backhaus *et.al.* [1].

The groups found to be valid were reclassified by linear discriminant analysis (*LDA*). *LDA* calculates a linear, $n-1$-dimensional (hyper-)plane having the property to discriminate between groups chosen in advance [4, 8]. Therfore it is a tool of *supervised pattern recognition* [6]. The resulting discriminant function then was applied on the Cologne population and in a second step on the Hiroshima population.

As the nationalities of the asian athletes were known, these samples were grouped by regional criteria. Three areas were defined: Arabia, Central Asia and East Asia. Homogenity of those regions concerning the frequencies of the groups classified by *LDA* was tested by means of χ^2-statistics (Test of Snedecor and Brandt, described in [2]).

3 Results

Figure 2 shows the change of heterogenity during the last 10 steps of the cluster analysis, where the figure has to be read from right to left. The upper graph indicates the distance of the agglomerated clusters in the respective step. The lower graph describes the increase of spanned distances from step to step. The greatest increase can be observed during the agglomeration of the last two clusters into one. A comparable increase occurs at the step when the number of clusters is reduced from five to four. A visual impression of the sample's structure after division into two groups can be obtained from figure 3. The properties of both clusters are shown in table 1. The scores of all variables are slightly larger than zero in cluster 1. In cluster 2 the mean of testosterone concentrations lies more than two standard deviations lower than in cluster 1. 5β-Diol concentrations are decreased to a similar extend in cluster 2. The latter group also shows a decrease of 5α-Diol concentrations, but it turns out lower than that of the other substances. Table 2 shows the capability of the discriminant function to classify the members of both assumed clusters. About 98% of the sample are grouped correctly. The classification is asymmetric:

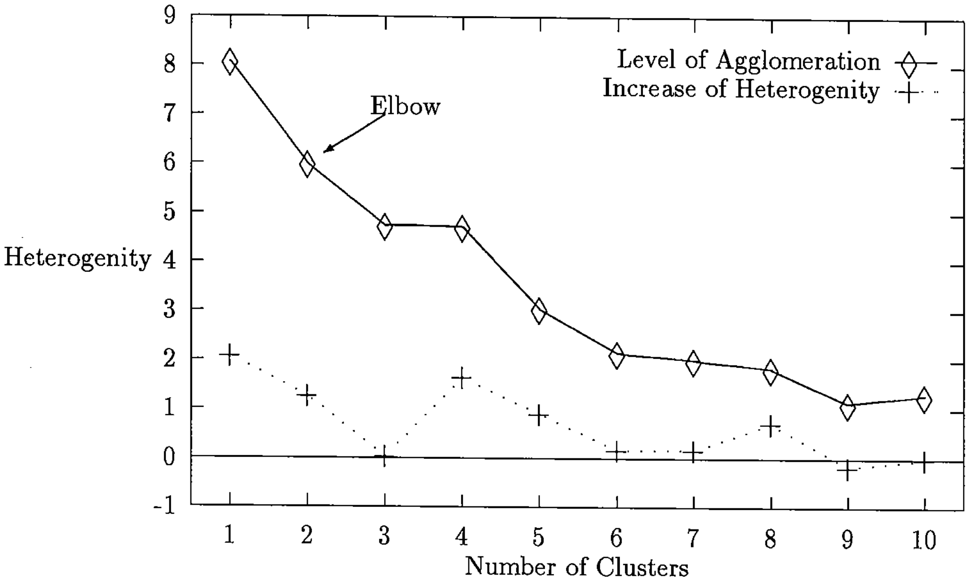

Figure 2: Change of heterogenity during agglomerative cluster anlysis. Heterogenity is expressed as squared eucledian distance of the centroids.

Table 1: Means and standard deviations of all variables (Z-scores) after cluster analysis

	Cluster 1			Cluster 2		
	Test	5α-Diol	5β-Diol	Test	5α-Diol	5β-Diol
$\bar{x}$	0.32456	0.10289	0.20689	-1.89833	-0.51666	-1.43773
s	0.58598	0.90502	0.76854	0.52165	0.71317	0.55212

A greater share of cluster 2 is classified wrong than of cluster 1. The coefficients of the function are contained in table 3. The critical value of the discriminant function L_{crit} was determined as -1.91 where the criterion was a minimal amount of misclassifications in the training set. Application of the discriminant function to Cologne routine samples led to assignment of 4141 males to group 1 (87.2%) and of 608 ones to group 2 (12.8%). Table 4 shows the observed frequencies of the two groups in different parts of Asia. The $\hat{\chi}^2$-test of Snedecor and Brandt yielded a probability of ≤ 0.0001.

4 Discussion

The hypothesis that male steroid profiles can be subdivided in at least two subpopulations was brought forth already by Rauth [7]. She performed a k-means cluster analysis and used the concentrations of testosterone and epitestosterone and the ratio of both substances as variables. One drawback of the k-means algorithm is that the number of groups has to be determined in advance. Thus it is a tool of confirmatory rather than of explorative data analysis.

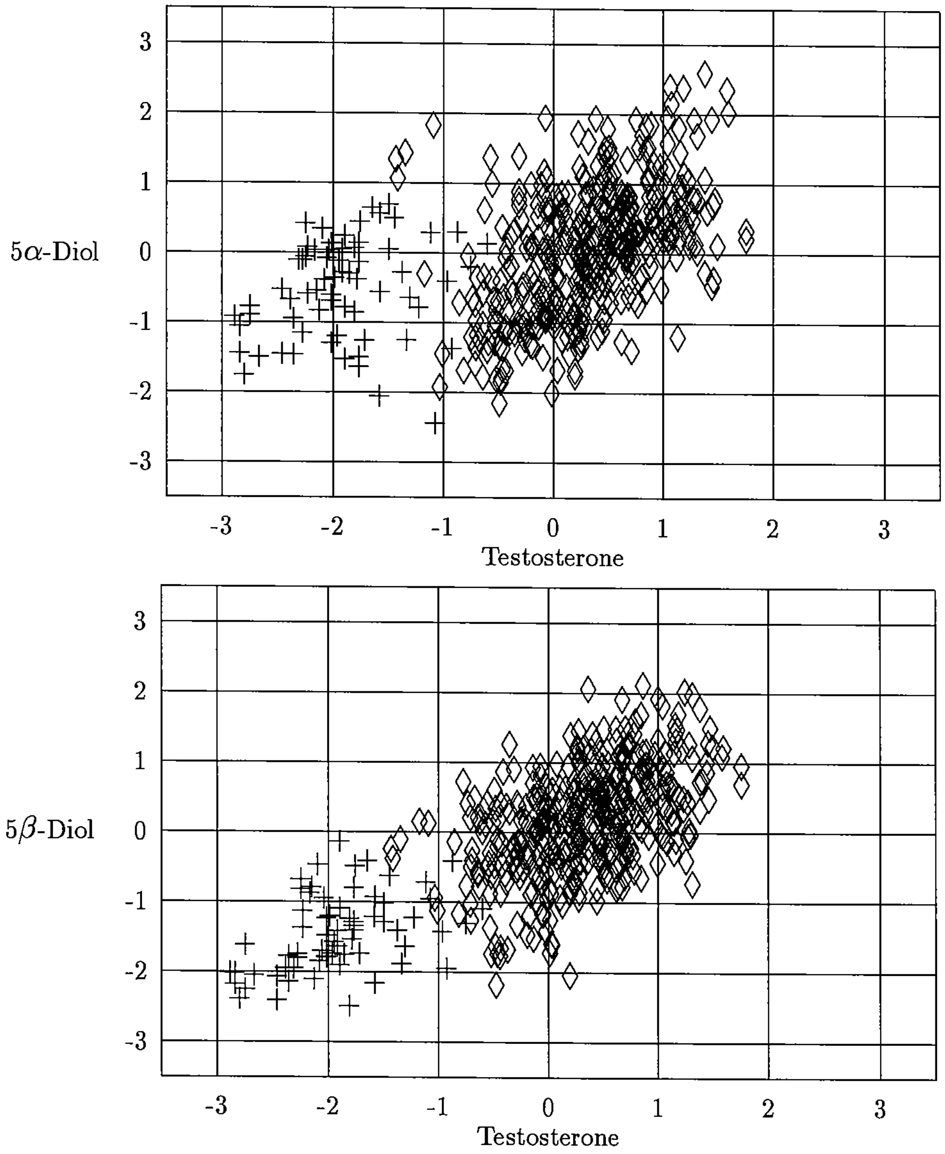

Figure 3: Relation of Testosterone to 5α-Diol and of Testosterone to 5β-Diol after classification by cluster analysis. Legend: $\Diamond$ Cluster 1, + Cluster 2.

Table 2: Ability of calculated discriminant function to reclassify the clusters.

Group	n	Prediction			
		Correct		Wrong	
1	418	414	99.04%	4	0.96%
2	67	62	92.53%	5	7.47%
Total	485	476	98.14%	9	1.86%

Table 3: Coefficients of calculated discriminant function.

Variable	Coefficient
Test	1.76986
5α-Diol	-0.77855
5β-Diol	0.45340
Constant	-5.29887

Rauth [7] used four groups. One of these proved to contain profiles with low testosterone concentrations and decreased and approximate equal concentrations of isomeric Androstane-$3\alpha, 17\beta$-diols. This group is likely to be identical with group 2 of this investigation.

The variables taken into account in this study were chosen following to the frequent observation of low testosterone concentrations. Steroids which stand close to testosterone in its metabolism appeared to be promising parameters therefore. 5α-dihydrotestosterone seemed to be appropiate as well but was not taken, as the determination of its concentration was not reliable at low values.

Deichsel and Trampisch [3] suggest to use Ward's method in cluster analysis by default. It was used succesfully in clinical chemistry by Folkerts *et.al.* [5]. As it is said, that Ward's clustering algorithm tends to produce groups of equal size, the centroid algorithm was chosen, which is not afflicted with this drawback and shows good properties otherwise [1]. Apart from the selection of appropriate algorithms and variables the choice of an optimal number of groups is a critical point in cluster analysis [5, 9]. A simple method is the elbow criterion suggested by Backhaus *et.al* [1]. In the present study this criterion appears to have led to reasonable results, as the number of two groups corresponds well to theoretical assumptions.

The discriminant function calculated to seperate between the two groups shows remarkable good power of classification. The fact that about 98% of the grouped population are reclassified correctly supports the assumption that male steroid profiles consist of two discrete phenotypes concerning the metabolism of testosterone.

The question remains whether the discriminant function succeeds in the decomposition of the bimodal distribution of testosterone concentrations. Figure 4 shows the respective

Table 4: Frequencies of detected phenotypes in different parts of Asia

	Total	Arabia	Central Asia	East Asia
Group 1	50	13	10	27
Group 2	85	8	5	72
Σ	135	21	15	99

histogram after classification of the total population by *LDA*. The result is quite evident: The distribution of testosterone is likely to be regarded as resembling two subgroups. They can be seperated in three dimensional space by using testosterone and two of its metabolites simultaneously.

The two groups found out are likely to react differently to exogenous testosterone for example. *LDA* seems to be a good tool to determine the belonging of an male athlete to one of the phenotypes with high probability. Thus after classification of the relevant persons it would be possible to investigate the metabolic patterns of both groups by experimental means.

The existance of a low urinary concentration of testosterone is often said to depend on asian origin of the male. The results of the present investigation rather suggest a genetic influence. The frequency of group 2 increases in regions of Far Eastern. Nevertheless more than one quarter of the East Asians under study show a 'normal' profile.

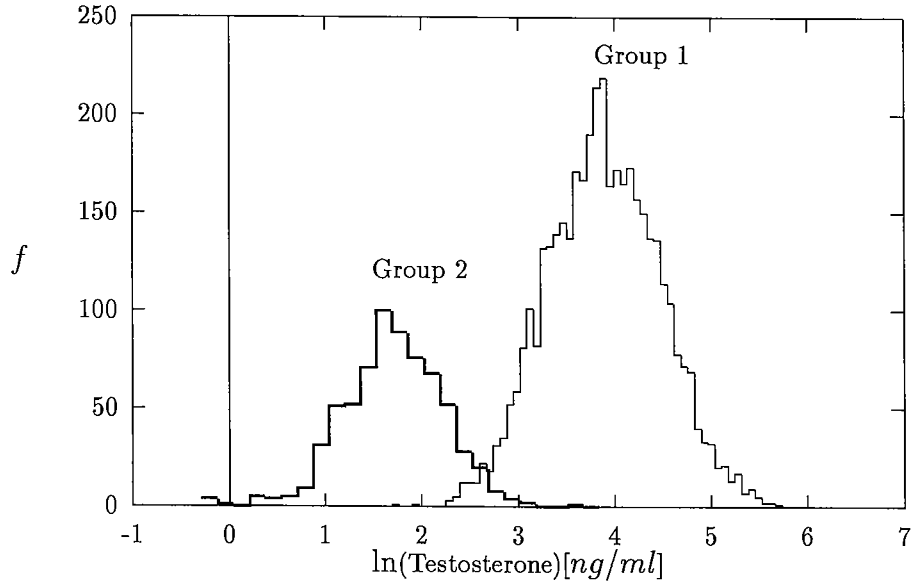

Figure 4: Distribution of urinary logarithmic testosterone concent both classified subgroups f: Absolute frequency.

References

[1] K. Backhaus, B. Erichson, W. Plinke, and R. Weber. *Multivariate Analysemethoden.* Springer-Verlag, fifth edition, 1990.

[2] Jürgen Bortz, Gustav A. Lienert, and Klaus Boehnke. *Verteilungsfreie Methoden in der Biostatistik.* Springer-Verlag, first edition, 1990.

[3] Guntram Deichsel and Hans Joachim Trampisch. *Clusteranalyse und Diskriminanzanalyse.* Biometrie. Gustav Fischer Verlag, 1985.

[4] Bernhard Flury and Hans Riedwyl. *Angewandte multivariate Statistik.* Gustav Fischer Verlag, 1983.

[5] Ulrike Folkerts, Dorothea Nagel, and Wolfgang Vogt. The use of cluster analysis in clinical chemical diagnosis of liver diseases. *Journal of Clinical Chemistry and Clinical Biochemistry*, 28(6):399–406, 1990.

[6] M. P. Merde and D. L. Massart. Supervised pattern recognition: The ideal method? *Analytica Chimica Acta*, 191:1–16, 1986.

[7] Susanne Rauth. *Referenzbereiche von urinären Steroidkonzentrationen und Steroidquotienten.* PhD thesis, Deutsche Sporthochschule Köln, 1994.

[8] Helge E. Solberg. Discriminant analysis. *Critical Reviews in Clinical Laboratory Sciences*, pages 209–242, November 1978.

[9] Wolfgang Vogt and Dorothea Nagel. Cluster analysis in diagnosis. *Clinical Chemistry*, 38(2):182–198, 1992.