Drug Information Journal, Vol. 31, pp. 433–447, 1997
Printed in the USA.

0092-8615/97

MULTIVARIATE MANY-TO-ONE PROCEDURES WITH APPLICATIONS TO PRECLINICAL TRIALS

SIEGFRIED KROPF, PHD

Research Assistant, Institute of Biometrics and Medical Informatics, Otto von Guericke University Magdeburg, Magdeburg, Germany

LUDWIG A. HOTHORN, PHD

Director, Institute of Bioinformatics, University of Hannover, Hannover, Germany

JÜRGEN LÄUTER, PHD

Director, Institute of Biometrics and Medical Informatics, Otto von Guericke University Magdeburg, Magdeburg, Germany

Comparisons of several treatments with a control represent a standard situation in preclinical trials. Usually, they are considered with a single variable, resulting in multiple test procedures such as the Dunnett test (1). Here, the multivariate many-to-one problem is considered, where several variables are observed on each individual of the control and treatment groups.

Classical MANOVA tests and their derivatives for the many-to-one problem require large sample sizes in order to be powerful if the dimension is high. In this paper, a new class of stabilized multivariate tests proposed by Läuter (2) and Läuter, Glimm, and Kropf (3) is extended to this special design. The new tests are based on linear scores which are derived in a certain way from the original variables. They utilize factorial relations among the variables.

It is shown here that the procedures keep the multiple level. In simulation experiments several versions of multivariate tests are compared with each other. Standard approaches are included as well as different score versions and a comparison of Dunnett-like procedures with Bonferroni-type procedures. Generally, an improved power of the new tests compared to standard procedures is demonstrated.

Key Words: Multivariate tests; Stabilized scores; Many-to-one procedures; Dunnett test; Principal component test

INTRODUCTION

MULTIPLE ENDPOINTS CAN occur in many different situations. A variable may be observed under different conditions or in the course of time, or there may also be different

---

Presented at the DIA Workshop "Statistical Methodology in Non-Clinical and Toxicological Studies," March 25–27, 1996, Bruges, Belgium.

Reprint address: Dr. S. Kropf, Institut für Biometrie und Medizinische Informatik, Otto-von-Guericke-Universität Magdeburg, Leipziger Straße 44, D-39120 Magdeburg, Germany.

variables related to a common object. If a testing problem with multiple endpoints is treated by more-fold univariate tests, then problems arise, for example, the type I error of the whole procedure can exceed the nominal level by far. A correction for the multiplicity is possible, for example, by the Bonferroni method, but then the correlation structure between the variables is still ignored and information is lost. Traditional multivariate methods, however, are derived under very general conditions. Thus, these tests have to handle a lot of parameters, and they require large samples when the dimension of the vectors of variables is high.

Various authors, among them O'Brien (4), have proposed tests that utilize special parameter structures and combine advantages of univariate and multivariate tests. They did it, however, mainly in a heuristic approach which yields only approximate distributions of the test statistic under the null hypothesis. In two recent papers, Läuter (2) and Läuter, Glimm, and Kropf (3) proposed a new class of so called 'stabilized' parametric tests. In a similar manner as in the tests of O'Brien, linear scores are computed from the variables, and these scores are handled in standard tests. Special rules for the derivation of the weight vectors for the scores ensure that the final tests with the scores exactly keep the type I error despite the preprocessing. Within the framework of these rules, there is a variety of special realizations, allowing for an adjustment to a broad field of practical demands, such as the utilization of symmetry assumptions among the parameters or of factorial parameter structures, the choice of one-sided or two-sided problems, the inclusion of a selection of relevant variables, and so on. For the one-way layout which is the basis for the present paper, the results of the two papers can be summarized as follows.

Let

$$x_j^{(k)} \sim N_p(\mu^{(k)}, \Sigma) \quad (k = 1, \ldots, K; j = 1, \ldots, n^{(k)}; n^{(k)} \geq 1; n = n^{(1)} + \ldots + n^{(K)} \geq K + 1)$$

be K samples of independent p-dimensional observations of size $n^{(1)}, n^{(2)}, \ldots, n^{(K)}$, let $X = (x_1^{(1)}, \ldots, x_{n^{(K)}}^{(K)})$ be the combined $(p \times n)$-sample matrix,

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n^{(k)}} x_j^{(k)} = \frac{1}{n} X 1_n \quad \text{with} \quad 1'_n = (1, \ldots, 1)$$

the total sample mean vector, and $\bar{X} = \bar{x} 1'_n = \frac{1}{n} X 1_n 1'_n$ the corresponding mean value matrix.

From the vectors $x_j^{(k)}$ of observations, the score values

$$z_j^{(k)} = d' x_j^{(k)} \quad (k = 1, \ldots, K; j = 1, \ldots, n^{(k)})$$

are computed with weight vectors d which are some unique function of the matrix

$$G_0 = (X - \bar{X})(X - \bar{X})' = X\left(I_n - \frac{1}{n} 1_n 1'_n\right)X' \quad (I_n \text{ is the } n \times n \text{ identity matrix}).$$

Then an exact level α test of the null hypothesis $H_0: \mu_1 = \ldots = \mu_k$ can be carried out in the usual ANOVA with the scores $z_j^{(k)}$ as input (F test with $K - 1$ and $n - K$ degrees of freedom).

The mathematical basis for the proof is the theory of spherical distributions (5). Due to the sample-based weights, the usual normality and independence assumptions regarding the sample elements are no longer fulfilled for the scores. Special choices for the scores are, for example:

- SS scores with weights $d_{SS} = [\text{Diag}(G_0)]^{-1/2}r$. SS scores are favorable when the mean differences of all variables have equal directions and approximately equal magnitudes (in units of the corresponding standard deviations), and
- PC scores with the weight vector d_{PC}, that is, the eigenvector belonging to the largest eigenvalue of the problem $G_0\, d_{PC} = \text{Diag}(G_0)\lambda d_{PC}$, $d'_{PC}G_0 d_{PC} = 1$. These scores do not suppose such a symmetric behavior in the p variables as in the SS method. They are advantageous if there is one latent variable behind the observed variables which is responsible for the differences among the groups. For one-sided problems the absolute values of the coefficients of weight vectors are used. As a generalization, q-dimensional scores can be derived with a weight matrix consisting of the first $q > 1$ eigenvectors of the above eigenvalue problem.

The power of the tests with the different approaches is demonstrated in the following fictional example. Suppose there are two independent samples of high-dimensional normal vectors ($p \geq 22$), each of sample size 12. All p variables have a variance of one, pairwise correlation coefficients of 0.3 (compound symmetry), and mean differences between the two populations of one. Figure 1 shows the power of the 'Bonferronized' univariate t tests (probability of at least one significance in the p tests), the power of the classical Hotelling's T^2 test (6,7), and the power of the PC test (SS test gives very similar results) when the number of variables enclosed rises from only one up to 22 (beyond this value the T^2 test would not be applicable). The nominal level of both tests is 0.05. The power of the T^2 test is derived from the noncentral F distribution; for the other tests simulations with 10,000 replications each have been done. Figure 1 demonstrates the advantages of the stabilized tests. Even in small samples, they can efficiently utilize the information from a large number of variables. In these weak correlated special data the T^2 test has no advantages with respect to the univariate t test procedure with a Bonferroni correction.

The aim of the present paper is the transfer of these ideas to the many-to-one procedures. Such procedures are usually considered when a control group has to be compared to several

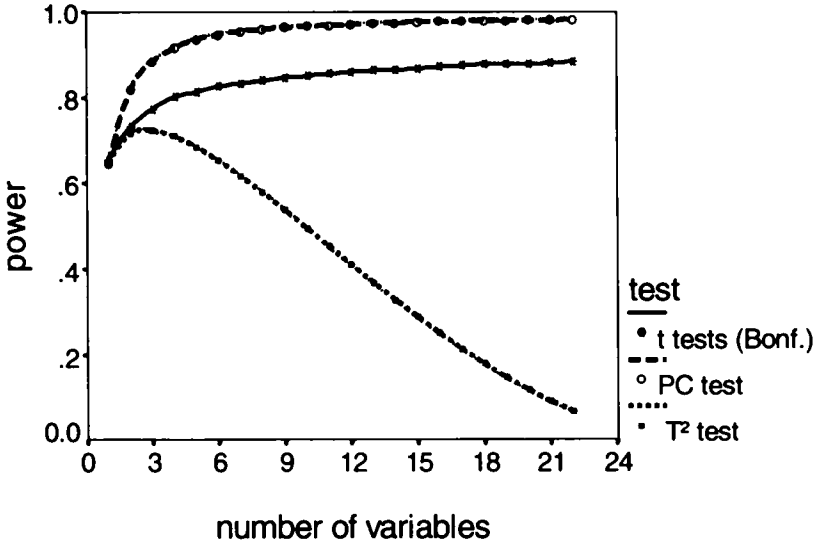

FIGURE 1. Power of Hotelling's T^2 test, PC test, and Bonferroni procedure with t tests for an increasing number of variables in the comparison of two samples of 12 each (for parameters cf. text).

treatment groups. The transfer is not a trivial step because the procedures have to keep the multiple error level, that is, even if some treatments differ from the control, then the error of the first kind has to be controlled for the remaining comparisons. In the computation of the scores, however, it is not known which treatments are equal to the control, so that the sphericity assumptions have to be restricted.

In the next section multivariate many-to-one procedures are described. They are followed by simulation studies for the power of the procedures and by an example with data from experiments with Wistar rats.

MANY-TO-ONE PROCEDURES

The following situation is considered in this paper: A control group is to be compared to K treatment groups. The sample elements are assumed to be independent p-dimensional normal vectors with common covariance matrix in all groups $x_j^{(k)} \sim N_p(\mu^{(k)}, \Sigma)$, $(k = 0, \ldots, K; j = 1, \ldots, n^{(k)})$. The hypotheses to be tested are H_k: $\mu^{(0)} = \mu^{(k)}$ $(k = 1, \ldots, K)$. As the K comparisons treat a common phenomenon, the type I error is to be kept for the whole procedure regardless of the actual classification into true and false hypotheses. Regarding the alternative several cases can be considered:

1. Unrestricted alternative: no assumptions about the order of mean values for the treatment groups, but one-sided comparisons with respect to the control group are possible,
2. Total order restriction: assumption of a strictly monotone trend in the expected mean values of all variables, starting with the control group and ending with treatment K; often used in connection with increasing doses of a drug, and
3. Partial ordering of expected mean values; not considered here (cf. [8]).

Procedures for Univariate Data

For univariate data a variety of parametric multiple test procedures has been proposed for the many-to-one problem. They are based on the multivariate t distribution (Dunnett test [1]) or on an approximate multinormal distribution as in the paper of James (9), taking the correlation structure into account, or on α-adjustment methods applied on the pairwise comparisons. Furthermore, the proposals can be classified into single-step procedures and stepwise procedures (step-down: Holm [10], Marcus, Peritz, and Gabriel [11], Dunnett and Tamhane [12]; step-up: Hochberg [13], Hommel [14], Dunnett and Tamhane [15]). For restricted alternatives, special methods are available, among them the proposal of Williams (16), stepwise tests of a priori ordered hypotheses (Hothorn and Lehmacher [17], Maurer, Hothorn, and Lehmacher [18]) or contrast tests (Fligner and Wolfe [19], see also the summary paper of Tamhane, Hochberg, and Dunnett [20]).

In this paper, the focus is on three basic methods:

1. Dunnett test: The sample means $\bar{x}^{(k)} = \frac{1}{n^{(k)}} \sum_{j=1}^{n^{(k)}} x_j^{(k)}$ $(k = 0, \ldots, K)$ and the pooled sample variance $s^2 = \frac{1}{n - K - 1} \sum_{k=0}^{K} (n^{(k)} - 1)s^{(k)2}$ with $\nu = n - K - 1$ degrees of freedom are calculated, where $n = \sum_{k=0}^{K} n^{(k)}$ and $s^{(k)2} = \frac{1}{n^{(k)} - 1} \sum_{j=1}^{n^{(k)}} (x_j^{(k)} - \bar{x}^{(k)})^2$ $(k = 0, \ldots, K)$. Then the test statistics $t^{(k)} = \frac{\bar{x}^{(k)} - \bar{x}^{(0)}}{s} \sqrt{\frac{n^{(0)} n^{(k)}}{n^{(0)} + n^{(k)}}}$ $(k = 1, \ldots, K)$ are computed for the one-sided test

(greater mean values for the treatment groups under the alternative hypotheses). For a two-sided test (test against arbitrary mean differences), the absolute values of the mean differences are taken in the above formula for $t^{(k)}$. A null hypothesis H_k ($k = 1, \ldots, K$) is rejected, if the value of the corresponding test statistics is greater than or equal to the critical value $t_0 = t_{K,\nu,\rho,1-\alpha}$. This critical value depends on the number of groups (K), the degrees of freedom in s^2 (ν), the multiple level (α), and the correlation coefficients (ρ) between the K test statistics, which are functions of the sample sizes and are equal to 0.5 if all samples have equal size. It is derived from the multivariate t distribution and can be taken from tables (eg, in Dunnett [1]) or determined with special software (eg, SAS function PROBMC [21]). An enhancement of this procedure is given by a closed testing version of the Dunnett test. Dunnett and Tamhane (12) describe the following step-down procedure: The test statistics $t^{(k)}$ ($k = 1, \ldots, K$) of the usual Dunnett test are ordered into $t_{(1)} \leq \ldots \leq t_{(K)}$. Starting with $l = K$, a null hypothesis $H_{(l)}$ (belonging to $t_{(l)}$) is rejected, if $t_{(l)} \geq t_{l,\nu,\rho,1-\alpha}$. As long as the hypotheses are rejected, the procedure continues with $l - 1$. When at a certain step $H_{(l)}$ cannot be rejected, that is, $t_{(l)} < t_{l,\nu,\rho,1-\alpha}$, then the procedure stops, and all hypotheses $H_{(1)}, \ldots, H_{(l)}$ are not rejected. Thus, the first critical value is identical to that of Dunnett's original proposal. Due to the monotony of the multivariate t quantils with respect to the dimension (first subscript), every significant outcome of the Dunnett test will be significant in the closed testing procedure, too. It is possible, however, to find more significant results in the subsequent steps. Hence, the closed testing procedure is more powerful,

2. Bonferroni/Holm adapted pairwise t tests: Pairwise t tests are carried out between the control group and each of the treatment groups. The P values of the K tests are ordered by increasing magnitude and then compared to the critical levels $\frac{\alpha}{K}, \frac{\alpha}{K-1}, \frac{\alpha}{K-2}, \ldots$ Starting with the smallest P value, the corresponding hypotheses are rejected as long as the P values are less than or equal to the corresponding critical level. If a P value does not fall below the critical level, then the procedure stops. The power of the procedure can be increased, if the variance is estimated from all samples instead of only those two samples which are compared at that moment (resulting in $n - K - 1$ degrees of freedom, 'multiple t test'). Then there is a strong connection to the Dunnett closure test: Both procedures are based on the same test statistics, and the critical values, derived from the univariate t distribution at level α/K_{red} and from the K_{red}-dimensional t distribution at the level α, respectively, are close together, especially for a small number of groups (K_{red} denotes the reduced number of groups after the foregoing steps of the procedure), and
3. Pairwise comparisons with a priori ordered alternatives: This procedure is often used for restricted alternatives. In this case, the pairwise procedure starts with a t test between the control group and treatment K and continues with the group $K - 1$, $K - 2, \ldots$ as long as all tests give significant results at the unadjusted level α. The procedure stops with the first nonsignificant pairwise test. Other a priori orderings of the pairwise tests are possible by nonstatistical arguments. Again, the power can be increased when all comparisons are done with the pooled variance estimate from all $K + 1$ groups.

T^2 Based Tests for Multivariate Data

All of the above procedures can be extended for multivariate data by replacing the t test constructions by the corresponding versions of the T^2 test (5,6). Only two-sided tests are available, however:

1. Higazi and Dayton (22,23) gave a multivariate extension of the Dunnett test. The test

statistics $T^2_{(k)} = \frac{n^{(0)}n^{(k)}}{n^{(0)}+n^{(k)}}(\bar{x}^{(k)} - \bar{x}^{(0)})'S^{-1}(\bar{x}^{(k)} - \bar{x}^{(0)})$ $(k = 1, \ldots, K)$ with mean vectors

$$\bar{x}^{(l)} = \frac{1}{n^{(l)}} \sum_{j=1}^{n^{(l)}} x_j^{(l)} \quad (l = 0, \ldots, K)$$

and the pooled covariance matrix

$$S = \frac{1}{n - K - 1} \sum_{k=0}^{K} \sum_{j=1}^{n^{(k)}} (x_j^{(k)} - \bar{x}^{(k)})^2$$

with $\nu = n - K - 1$ degrees of freedom are compared to critical values given in their paper for $p \leq 5$ and for the assumption of equal sample sizes at least in the K treatment groups. The corresponding closed testing procedure can be used analogously because the closed testing procedure works independently of the special constructions of the tests,

2. Pairwise comparisons with Bonferroni/Holm adjustment are based on the same statistics with an additional constant factor, so that the F distribution can be applied: $F_{(k)} = \frac{f^{(k)}}{(f^{(k)}+p-1)p}T^2_{(k)}$ with p and $f^{(k)}$ degrees of freedom, where $f^{(k)} = n^{(0)} + n^{(k)} - p - 1$, if the variance is estimated only from those two groups which are compared at that moment, and $f^{(k)} = n - K - p$, if the variance is computed from all groups $(k = 1, \ldots, K)$. The decision based on the P values of this pairwise test is made in the same way as in the univariate case, and
3. The same statistics and P values can be used in pairwise comparisons with a priori ordered hypotheses (without α adjustment).

Multivariate Tests Based on Linear Scores

These tests are stabilized alternatives to the T^2 based tests. The basic methodology of the score method is simple: Compute a weight vector d according to the rules given in the introduction, and use it to transform all observation vectors into scores $z_j^{(k)} = d'x_j^{(k)}$ $(k = 0, \ldots, K; j = 1, \ldots, n_k)$. These scores are treated with the standard tests for univariate data (see above) though they do not meet the standard assumptions of these procedures (independence, normality). In the same way, a $p \times q$ weight matrix can be evaluated, and q scores can be computed per sample element (one per column of the weight matrix). Then the scores have to be treated with the multivariate tests from the last subsection, but now with a reduced dimension $q < p$. This would be advantageous if more than one important latent factor is suspected behind the observed data. But this aspect is not considered here in more detail.

For pairwise t tests with the covariance estimate based only on the two enclosed groups, it seems natural to also determine the scores on the basis of these two groups only, and to recompute them for each comparison. The validity of this procedure is given by the results reviewed in the introduction.

In order to carry out tests with a variance estimate from all $K + 1$ groups, such as in the Dunnett test or in multiple t tests, the matrix G_0 from the introduction is calculated from all $K + 1$ groups, too. Hence, the scores need to be computed only once. The theoretical justification is not straightforward here, because for a comparison of one treatment group against the control group, the null hypothesis of equal expectations in these two groups does not include the assumption that all other treatment groups have the same expectation,

too. The corresponding theorems, however, can be generalized for this many-to-one situation. Thus, for the univariate tests considered here, the null distribution of the test statistics applied to the scores is the same as if they were applied to original independent univariate normal data, even though the global null hypothesis of equality among the expectations of all groups is not necessarily true. In this context, the Dunnett procedure with its K single comparisons is interpreted as a test of the intersection of all true hypotheses. The corresponding test statistic is $F(z) = \max t^{(k)}(z)$, where the maximum is to be taken over all groups k for which the null hypothesis is true. The local level of this test is the multiple level of the Dunnett procedure.

SIMULATION EXPERIMENTS

Simulation experiments have been done under the null hypothesis as well as under various alternative hypotheses for the comparison of a control group to three treatment groups. The nominal level of all tests was $\alpha = 0.05$. Under the null hypothesis the type I error is guaranteed by the theory given above if all assumptions are met. That is why only the results of two little series are given here, demonstrating the importance of these assumptions. The rejection rates of the null hypothesis are shown for the Dunnett-like tests (rejection of the global hypothesis) with SS scores and PC scores, both being special cases for weight vectors derived uniquely from G_0, and with O'Brien's ordinary least squares (OLS) scores. The weights for the latter are computed analogously to those of SS scores, but with the pooled variances instead of the total variances. Thus, they are not within the restriction of the new class of tests. Figure 2a shows the results for 10 uncorrelated variables and groupwise sample sizes increasing from 5–30, Figure 2b gives corresponding results for a fixed sample size of five per group and the number of uncorrelated variables increasing from 2–10. Both figures include the confidence intervals for the rejection rates. All results are based on 100,000 replications. Whereas the tests with SS scores and with PC scores are always close to the nominal level, the tests with the OLS scores are anticonservative to a moderate degree in small and high-dimensional samples.

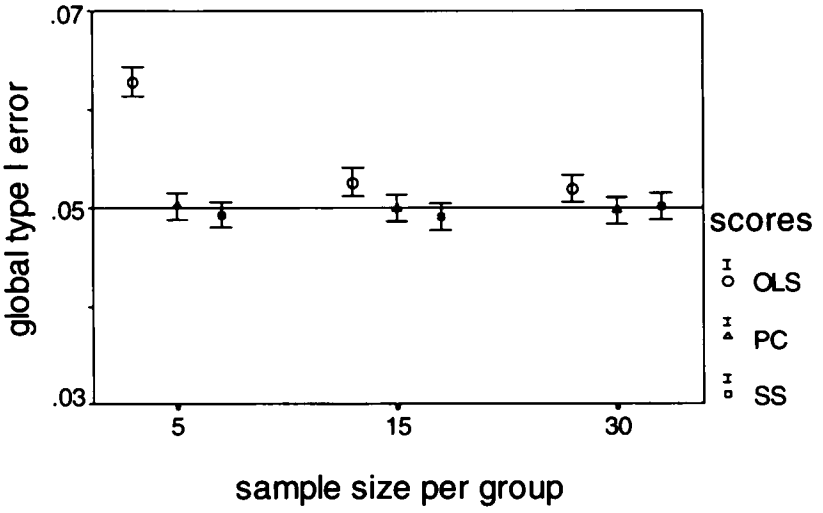

FIGURE 2A. Global type I error with confidence intervals from simulation experiments with one control group and three treatment groups of varying sample sizes in Dunnett tests based on scores. There are 10 uncorrelated variables used in the tests. One hundred thousand replications have been done for each parameter structure.

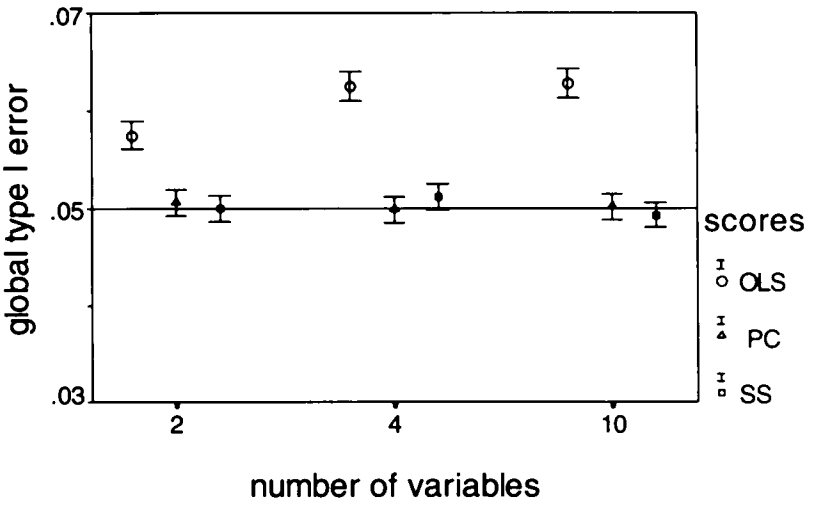

FIGURE 2B. Global type I error with confidence intervals from simulation experiments with one control group and three treatment groups with a sample size of five per group in Dunnett tests based on scores. The number of uncorrelated variables varies from 2–10. One hundred thousand replications have been done for each parameter structure.

The parameter structures under the alternative hypothesis are chosen in such a manner that a one-sided pairwise t test between the control group and treatment 3 based on the 'optimal' choice of weights for known parameters $d = \Sigma^{-1}(\mu^{(3)} - \mu^{(0)})$ would always have a power of 0.95. In structures labeled 's' (symmetric) all variables have equal mean differences and equal pairwise correlation coefficients of 0.66. In the asymmetric structures ('a') the same is true for half of the variables, whereas the other variables are uncorrelated and have no mean difference with respect to the control group. The other two treatment groups have reduced mean differences (factor 1/3 in Treatment 1 and 2/3 in Treatment 2) for all variables in the structures characterized by '0/1/2/3' and have the same parameters as Treatment 3 in the structures characterized by '0/3/3/3'.

Tables 1a and 1b show the rejection rates in 100,000 replications for the comparisons with Treatment 1 and Treatment 3 (local power). In Table 1a the sample size per group is fixed to 15 and the number of variables varies between 2, 4, and 10; in Table 1b the number of variables is fixed to four and sample size per group varies between 5, 15, and 30.

The following procedures have been included:

- Higazi/Dayton test and Dunnett-like test with SS scores and PC scores, respectively. The Higazi/Dayton test is denoted here as the T^2 version of the Dunnett test. It has not been performed for 10 variables because of missing critical values,
- The same tests using the closed testing principle with regard to the different treatments as described above for the Dunnett test,
- Pairwise tests with Bonferroni/Holm α adjustment. The variance is computed only from the two groups involved. In the score tests, the weights are computed only from the two groups involved, and
- Pairwise tests as in the foregoing, but now without α adjustment. Instead of that the ordering of the pairwise hypotheses is fixed in advance (the control group is first compared to Treatment 3, then to Treatment 2, and finally to Treatment 1; if a comparison does not yield significant differences, then the procedure stops).

TABLE 1A
Results of Simulation Experiments with a Varying Number of Variables

Treatment Spacing	p	s/a	Dunnett			Dunnett Closure			Bonferroni/Holm			A Priori Ordering		
			T^2	SS	PC	T^2	SS	PC	T^2	SS	PC	T^2	SS	PC
	2	s	.602	.813	.813	.612	.824	.824	.556	.790	.790	.722	.901	.901
	2	a	.603	.476	.476	.613	.492	.492	.555	.421	.421	.725	.614	.614
Treatment 3	4	s	.448	.811	.811	.459	.822	.822	.386	.787	.787	.568	.900	.900
0/1/2/3/	4	a	.450	.582	.786	.460	.597	.798	.386	.536	.740	.568	.718	.869
	10	s		.811	.811		.822	.822	.171	.787	.787	.322	.899	.899
	10	a		.698	.807		.711	.818	.172	.661	.777	.322	.816	.893
	2	s	.061	.142	.142	.102	.231	.231	.091	.219	.219	.079	.215	.215
	2	a	.061	.082	.082	.102	.125	.125	.091	.117	.117	.079	.103	.103
Treatment 1	4	s	43	.140	.140	.069	.227	.227	.058	.215	.215	.042	.210	.210
0/1/2/3	4	a	42	96	.133	.068	.151	.216	.058	.142	.197	.043	.132	.190
	10	s		.139	.139		.228	.228	.033	.216	.216	.014	.211	.211
	10	a		.115	.138		.183	.226	.032	.174	.212	.014	.166	.206
	2	s	.602	.813	.813	.673	.874	.874	.625	.855	.855	.722	.901	.901
	2	a	.603	.465	.465	.675	.535	.535	.626	.479	.479	.725	.614	.614
Treatment 3	4	s	.448	.811	.811	.516	.872	.872	.443	.853	.853	.568	.900	.900
0/3/3/3	4	a	.450	.575	.788	.515	.651	.853	.444	.606	.811	.568	.718	.869
	10	s		.811	.811		.873	.873	.194	.853	.853	.322	.899	.899
	10	a		.693	.807		.767	.869	.194	.736	.844	.322	.816	.893
	2	s	.599	.814	.814	.672	.874	.874	.625	.854	.854	.492	.782	.782
	2	a	.599	.463	.463	.672	.534	.534	.625	.476	.476	.494	.361	.361
Treatment 1	4	s	.450	.810	.810	.516	.872	.872	.447	.852	.852	.301	.781	.781
0/3/3/3	4	a	.451	.572	.784	.518	.650	.850	.446	.605	.809	.301	.492	.722
	10	s		.812	.812		.872	.872	.193	.852	.852	.085	.781	.781
	10	a		.694	.807		.769	.868	.193	.736	.843	.084	.636	.769

Local power of different tests for the many-to-one problem with a control group and three treatment groups ($\alpha = 0.05$). All samples have a size of 15. Results are given for Treatments 1 and 3. The rejection rates are estimated from simulation experiments with 100,000 replications per parameter structure. s = symmetric structures, a = asymmetric structures (for more detailled explanation and for the group patterns '0/1/2/3' and '0/3/3/3' see the text)

The results reflect very well the intentions and basic assumptions of the different procedures. The advantages of the score methods compared to the T^2 based methods are obvious for the structures considered. That is remarkable insofar as the relationship of sample size and number of variables is not an extreme one. The differences decrease with increasing sample size and increase with an increasing number of variables.

The SS scores and the PC scores yield similar results in the 'symmetric' structures. In the asymmetric structures the PC scores clearly have better results. Even in the asymmetric situations, however, the power of the SS versions is still greater than that of the T^2 based methods. It should be noted here that all the parameter constellations of Tables 1a and 1b are compatible with the idea of one latent variable (one-factor model, cf. Läuter [24]) and thus, support the PC scores.

Comparing the results of the Dunnett-like procedure with those of its closure, one can state a distinct improvement by the closure. The amount of improvement is small only for Treatment 3 when the other two treatments have smaller mean differences than the control group and, hence, Treatment 3 is compared as the first group to the control group in most cases. In all other situations the gain in power is considerable.

The power in the pairwise tests with a Bonferroni/Holm adjustment is less than the power

TABLE 1B
Results of Simulation Experiments with Varying Sample Sizes

Treatment Spacing	n	s/a	Dunnett			Dunnett Closure			Bonferroni/Holm			A Priori Ordering		
			T^2	SS	PC	T^2	SS	PC	T^2	SS	PC	T^2	SS	PC
	5	s	.367	.828	.828	.377	.837	.837	.172	.749	.749	.356	.899	.898
	5	a	.367	.550	.762	.377	.565	.774	.173	.414	.620	.356	.639	.811
Treatment 3	15	s	.448	.811	.811	.459	.822	.822	.386	.787	.787	.568	.900	.900
0/1/2/3	15	a	.450	.582	.786	.460	.597	.798	.386	.536	.740	.568	.718	.869
	30	s	.469	.808	.808	.479	.818	.818	.438	.792	.792	.608	.898	.898
	30	a	.468	.591	.795	.478	.607	.806	.436	.563	.768	.607	.734	.882
	5	s	.041	.146	.146	.064	.241	.241	.037	.212	.212	.018	.211	.211
	5	a	.040	.092	.129	.064	.145	.212	.036	.127	.169	.017	.118	.162
Treatment 1	15	s	.043	.140	.140	.069	.227	.227	.058	.215	.215	.042	.210	.210
0/1/2/3	15	a	.042	.096	.133	.068	.151	.216	.058	.142	.197	.043	.132	.190
	30	s	.044	.139	.139	.070	.225	.224	.064	.217	.217	.049	.211	.211
	30	a	.044	.096	.136	.070	.153	.219	.064	.145	.207	.049	.134	.201
	5	s	.367	.828	.828	.432	.888	.888	.198	.830	.829	.356	.899	.898
	5	a	.367	.525	.760	.430	.603	.829	.200	.479	.705	.356	.639	.811
Treatment 3	15	s	.448	.811	.811	.516	.872	.872	.443	.853	.853	.568	.900	.900
0/3/3/3	15	a	.450	.575	.786	.515	.651	.851	.444	.606	.811	.568	.718	.869
	30	s	.469	.808	.808	.534	.869	.869	.497	.855	.855	.608	.898	.898
	30	a	.468	.587	.794	.533	.663	.857	.496	.633	.834	.607	.734	.882
	5	s	.369	.828	.828	.433	.888	.888	.199	.830	.830	.105	.778	.778
	5	a	.368	.527	.759	.430	.605	.830	.198	.481	.705	.104	.393	.619
Treatment 1	15	s	.450	.810	.810	.516	.872	.872	.447	.852	.852	.301	.781	.781
0/3/3/3	15	a	.451	.572	.784	.518	.650	.850	.446	.605	.809	.301	.492	.722
	30	s	.467	.807	.807	.533	.869	.869	.495	.855	.855	.345	.781	.781
	30	a	.468	.585	.792	.534	.662	.857	.496	.633	.833	.345	.515	.750

Results of simulation experiments as in Table 1a, but now the number of variables is fixed to four and the sample size per groups varies.

in the Dunnett-like tests, but the differences are not very large. Of course, the power of the procedures with a priori ordering is always greater than that of the other procedures with respect to the comparison of the control group to Treatment 3, which is compared first. This difference can be considerable. For Treatment 1, which is compared last, this advantage disappears in the constellations with equal group spacing ('0/1/2/3'), and it is reversed in the cases where all treatments have equal effects ('0/3/3/3'). In the latter case, the procedures with a priori ordering are the only ones that have different results for Treatments 1 and 3 (up to random variations in the other procedures).

EXAMPLE

As an example, a chronic toxicological study in Wistar rats is considered. A control group is compared with a low dose group and a high dose group of some substance. From each animal the liver mass and the concentration of ASAT, ALAT, and AP in the serum is recorded. For the purpose of demonstration only six animals per group are used (Table 2).

To characterize the example, groupwise means and standard deviations and pooled correlation coefficients are calculated:

Variable	Groupwise Means and Standard Deviations Control	Low Dose	High Dose	Pooled Correlation Coefficients l.mass	ASAT	ALAT	AP
liver mass	12.05 ± 2.24	12.48 ± 1.81	13.57 ± 1.60	1.000	0.157	0.268	0.002
ASAT	1.96 ± 0.18	2.27 ± 0.30	3.09 ± 0.45	0.157	1.000	0.545	–0.104
ALAT	0.39 ± 0.20	0.76 ± 0.21	0.94 ± 0.28	0.268	0.545	1.000	0.161
AP	0.26 ± 0.04	0.31 ± 0.04	0.28 ± 0.07	0.002	–0.104	0.161	1.000

Thus, in the example the sample sizes are rather small, but the number of variables is small, too. The correlation between the variables is small or moderate. The means show a monotone trend with increasing values with increasing dose of the substance. This trend (when expressed in units of the standard deviation) and the pairwise correlations are not equal for the four variables, however.

The T^2 based methods use the inverse of the pooled covariance matrix

$$S = \begin{pmatrix} 3.60644 & 0.09803 & 0.11752 & 0.00024 \\ 0.09803 & 0.10746 & 0.04124 & -0.00178 \\ 0.11752 & 0.04124 & 0.05336 & 0.00194 \\ 0.00024 & -0.00178 & 0.00194 & 0.00272 \end{pmatrix}.$$

The Higazi/Dayton statistics are found to be $T^2_{(1)} = 9.021$ and $T^2_{(2)} = 38.057$. From the tables in Higazi and Dayton (22), the critical values for $p = 4$ variables and $\nu = 18 - 2 - 1 = 15$ degrees of freedom and equal sample sizes in all groups are 20.2 for $\alpha = 0.05$ or 32.4 for $\alpha = 0.01$, respectively. Thus, only the high dose group differs significantly from the control group ($P < 0.01$). Going on into a stepwise closure procedure, the critical values can now be revised for only one remaining nonsignificant comparison. As this results in 16.37 for

TABLE 2
Data for the Experiment with Wistar Rats. A Control Group is to be Compared with Two Dose Groups

Group	Liver mass	ASAT	ALAT	AP
Control	11.6	1.82	0.71	0.29
	12.6	2.30	0.46	0.24
	10.5	1.99	0.45	0.23
	16.2	1.90	0.31	0.22
	11.5	1.94	0.18	0.29
	9.9	1.81	0.20	0.30
Low dose	11.3	2.27	0.47	0.28
	14.2	2.73	0.89	0.27
	11.2	1.79	0.55	0.25
	14.5	2.31	0.77	0.34
	13.5	2.32	0.99	.036
	10.2	2.19	0.87	0.33
High dose	12.7	2.76	0.56	0.21
	14.0	2.60	0.81	0.26
	14.6	3.74	1.40	0.21
	12.8	3.00	1.05	0.32
	11.4	3.52	0.86	0.31
	15.9	2.91	0.98	0.39

$\alpha = 0.05$ and 27.67 for $\alpha = 0.01$ no further significances can be found, and the procedure stops.

If pairwise comparisons are carried out with the variance computed from only the two corresponding groups, then $F_{(1)} = 2.236$ and $F_{(2)} = 6.965$. With degrees of freedom four and seven this corresponds to P values of 0.166 and 0.014. If the order of hypotheses to be tested had been given in advance (what is useful in this example with increasing doses) then these values are the basis for the decision. Otherwise a Bonferroni/Holm correction would double the second P value according to the approach of adjusted P values of Wright (25). In any case, the results would state no significance for the low dose group and only a significance at level 0.05 for the high dose group, thus having weaker results than the Higazi/Dayton test. Using the covariance estimate from all groups, however, corrects this disadvantage. Then $F_{(1)} = 1.804$ and $F_{(2)} = 7.611$ with degrees of freedom four and 12, which yields P values 0.1929 and 0.0027 (or with Bonferroni/Holm correction 0.1929 and 0.0054). Thus, all T^2 based procedures find the effect of the high dose treatment at least at level 0.05, but not the effect of the low dose treatment.

When the comparisons are based on scores derived from all three groups, then first the matrix:

$$G_0 = (X - \bar{X})(X - \bar{X})' = \begin{pmatrix} 61.420 & 6.937 & 4.183 & 0.060 \\ 6.937 & 5.693 & 2.414 & 0.013 \\ 4.183 & 2.414 & 1.770 & 0.077 \\ 0.060 & 0.013 & 0.077 & 0.046 \end{pmatrix}$$

is computed. The inverse square root of the diagonal elements gives the weights for the SS scores

$$d_{SS}' = (0.128\ 0.419\ 0.752\ 4.642),$$

whereas the weight vector for the PC method is determined from $G_0 d_{PC} = \mathrm{Diag}(G_0)\lambda d_{PC}$, $d_{PC}' G_0 d_{PC} = 1$ as

$$d_{PC}' = (0.057\ 0.252\ 0.477\ 0.869).$$

The score values for both methods are given in Table 3.

TABLE 3
Score Values for the Three Samples Derived by the SS Method and by the PC Method on the Basis of all Samples, Respectively

SS Scores			PC Scores		
Control Group	Low Dose	High Dose	Control Group	Low Dose	High Dose
4.12	4.05	4.17	1.71	1.68	1.87
4.03	4.88	4.69	1.73	2.16	2.07
3.58	3.75	5.46	1.52	1.57	2.63
4.12	4.98	5.17	1.74	2.07	2.27
3.76	5.11	5.02	1.48	2.14	2.22
3.56	4.41	5.80	1.38	1.84	2.45

With the Dunnett test applied to these score values the following results can be obtained (equal sample sizes, d.f. = 15):

- **For SS scores**:
  low dose: $t_{(1)} = 2.390$ P value one-sided 0.0274 and two-sided 0.0549,
  high dose: $t_{(2)} = 4.265$ P value one-sided 0.0006 and two-sided 0.0013, and
- **For PC scores**:
  low dose: $t_{(1)} = 2.394$ P value one-sided 0.0273 and two-sided 0.0545,
  high dose: $t_{(2)} = 4.950$ P value one-sided 0.0002 and two-sided 0.0003.

As in both cases the high dose treatment is significant at the level 0.01 or higher, the results for the low dose group can be checked in a second step of the closure procedure with the one-dimensional t distribution, yielding P values of 0.0152 (SS) and 0.0151. Thus, the 0.01 level cannot be ensured either.

When the same test values are evaluated by the univariate t test with the same degree of freedom, then the one-sided P values are:

- SS: low dose 0.0152, high dose 0.0003, and
- PC: low dose 0.0151, high dose 0.0001.

Two-sided P values are obtained by doubling these values. If the a priori ordering of the hypotheses is not assumed, then the 'more significant' values, that is, those from the high dose, also have to be doubled (Bonferroni/Holm correction).

When the pairwise comparisons with t tests are based on scores that are derived from the two involved samples only, then in the t tests with 10 degrees of freedom the following unadjusted one-sided P values are found:

- SS: low dose 0.0081, high dose 0.0006, and
- PC: low dose 0.0047, high dose 0.0002.

The adjustment for two-sided tests and the Bonferroni/Holm correction is done in the same way as above.

Though carried out with fewer degrees of freedom, the results for the low dose group are better ($P < 0.01$) than those with scores based on all three samples. This could be a hint that the 'global' scores are dominated by the more distinct differences between the control group and the high dose group, whereas the 'pairwise' scores are better adapted to each individual comparison.

Summarizing, this example demonstrates:

- The advantage of score-based tests with respect to conventional multivariate tests in small samples,
- A further advantage by the possibility of carrying out one-sided tests,
- The higher flexibility of PC scores compared to SS scores,
- The advantage of the use of a priori ordered hypotheses (as far as possible),
- The similarity of the results of the Dunnett procedure and those of the multiple univariate t tests with Bonferroni correction, and
- The advantage of the closure procedure for the 'less significant' treatment.

DISCUSSION

The simulation experiments and the example also demonstrate the advantages of the use of stabilized linear scores. The exactness of these tests with regard to the type I error is a

consequence of the theory of spherical distributions and is guaranteed under the usual ANOVA assumptions. The rules for the derivation of the weight vectors are general enough to allow for a variety of adoptions to practical problems.

These rules utilize special restrictions in the data structures such as information on underlying factorial structures or on a 'similar behavior' of the variables. In particular, they enable one-sided tests to be conducted. The power of the resulting tests is dependent on the validity of the special assumptions. But in small samples and/or with high-dimensional data, stabilized procedures have advantages compared to T^2 based methods, even if the assumed models are given only in rough approximation. Thus, the stabilized tests are of special interest for practical problems with restricted sample sizes, as, for example, in toxicology.

Earlier proposals for stabilized tests such as those from O'Brien (4) or from Tang, Geller, and Pocock (26) have the same intentions, but are not exact tests. They should be replaced by the new versions.

As exact level α tests, the score tests can be used in a closed test procedure in order to evaluate the influence of single variables or of subgroups of variables (as proposed by Kropf [27] and Lehmacher, Wassmer, and Reitmeir [28]). In the example in the last section the difference between the control group and Treatment 3 can be found not only with all four variables, but also with ASAT or ALAT alone and with all combinations of these variables with others at a multiple 0.05 level when the one-sided Dunnett like test with PC scores is used.

PC scores are more flexible than SS scores. The score versions SS and PC are only examples, however, other proposals can be given (3). This includes the possibility of selection procedures for important variables (without loosing exactness!), otherwise summary score tests would not be effective when only a small part of the variables had the expected effects. In the end, however, summary statistics need some kind of common behavior of several variables to be effective. If only a single variable or very few variables is/are effective, then univariate tests with Bonferroni adjustment or the control of the multiple error (with regard to the p variables) by permutation or bootstrap techniques (29,30) might be more powerful.

The extension of the Spherical theory to multiple test problems given here can also be used with other tests for univariate data to make them the basis for score tests. In that way, for example, contrast tests (eg [20]) can be used with scores.

In pairwise tests, the inclusion of all samples into the estimation of variance enhances the degrees of freedom for the tests. On the other hand, with regard to robustness considerations it can be better to avoid the information from other samples, for example, in order to restrict the influence of samples with enlarged variances.

When heavy deviations from the normal distribution are known or suspected, then a rank transformation can have positive effects on the type I error and on the power of the score tests (cf. Bregenzer and Lehmacher [31]). The test with SS scores applied to ranks corresponds to the nonparametric proposal of O'Brien (4).

The calculation of scores from the data is easy with standard tools in the statistical packages, especially for SS scores and PC scores in the one-way layout. Macros for the packages SAS or SPSS can be obtained from the authors.

The results of the Bonferroni (Bonferroni/Holm) procedure are not much worse than those of the Dunnett (Dunnett closure) test. The use of a priori ordering can be very effective, but only when the order, given in advance, meets the real situation.

Acknowledgments—The authors thank the reviewer for his constructive comments on the manuscript.

REFERENCES

1. Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *J Am Stat Assoc.* 1955;50:1096–1121.
2. Läuter J. Exact *t* and *F* Tests for Analysing Studies with Multiple Endpoints. *Biometrics.* 1996;52(3):964–970.
3. Läuter J, Glimm E, Kropf S. New Multivariate Tests for Data with an Inherent Structure. *Biometrical J.* 1996; 38:5–22.
4. O'Brien PC. Procedures for Comparing Samples with Multiple Endpoints. *Biometrics.* 1984;40:1079–1087.
5. Fang KT, Zhang YT. *Generalized Multivariate Analysis.* Beijing and Springer-Verlag Berlin Heidelberg Science Press; 1990.
6. Hotelling H. A Generalised *T* Test and Measure of Multivariate Dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* University of California, Los Angeles and Berkeley; 1951:23–41.
7. Anderson TW. *An Introduction to Multivariate Statistical Analysis.* 2nd edition. New York: Wiley; 1984.
8. Shaffer JP. Modified Sequentially Rejective Multiple Test Procedures. *J Am Stat Assoc.* 1986;81:826–831.
9. James S. Approximate Multinormal Probabilities Applied to Correlated Multiple Endpoints in Clinical Trials. *Stat Med.* 1991;10:1123–1135.
10. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6:65–70.
11. Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika.* 1976;63:655–660.
12. Dunnett CW, Tamhane AC. Step-Down Multiple Tests for Comparing Treatments with a Control in Unbalanced One-Way Layout. *Stat Med.* 1991; 10:939–947.
13. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika.* 1988;75:800–802.
14. Hommel G. A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. *Biometrika.* 1988;75:383–386.
15. Dunnett CW, Tamhane AC. A Step-Up Multiple Test Procedure. *J Am Stat Assoc.* 1992;87:162–170.
16. Williams DA. A Test for Difference Between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control. *Biometrics.* 1971;27:103–117.
17. Hothorn L, Lehmacher W. A Simple Testing Procedure 'Control versus *k* Treatments' for One-Sided Ordered Alternatives, with Application in Toxicology. *Biometrical J.* 1991;33:179–189.
18. Maurer W, Hothorn LA, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: Vollmar J, ed. *Biometrie in der chemisch-pharmazeutischen Industrie.* Vol. 6. Stuttgart: Fischer Verlag; 1995:3–18.
19. Fligner MA, Wolfe DA. Distribution-Free Tests for Comparing Several Treatments with a Control. *Stat Neerl.* 1982;36:119–127.
20. Tamhane AC, Hochberg Y, Dunnett CW. Multiple test procedures for dose finding. *Biometrics.* 1996;52:21–37.
21. SAS Institute Inc. SAS/STAT Software: Changes and Enhancements through Release 6.11, SAS Institute Inc., Cary, NC; 1996.
22. Higazi SMF, Dayton CM. Comparing Several Experimental Groups with a Control in the Multivariate Case. *Comm Stat.* 1984;13:227–241.
23. Higazi SMF, Dayton CM. Tables for the Multivariate Extension of the Dunnett Test when the Control Group and Balanced Experimental Groups have Different Sample Sizes. *Comm Stat.* 1988;17:85–101.
24. Läuter J. *Stabile multivariate Verfahren: Diskriminanzanalyse—Regressionsanalyse—Faktoranalyse.* Berlin: Akademie Verlag; 1992.
25. Wright SP. Adjusted *P*-Values for Simultaneous Inference. *Biometrics.* 1992;48:1005–1013.
26. Tang DI, Geller, NL, Pocock SJ. On the Design and Analysis of Randomized Clinical Trials with Multiple Endpoints. *Biometrics.* 1993;49:23–30.
27. Kropf S. Application of Multiple Test Procedures to the Combination of Multivariate and Univariate Tests with Varying Variable Sets. *Biometrical J.* 1988;30:460–470.
28. Lehmacher W, Wassmer G, Reitmeir P. Procedures for Two-Sample Comparisons with Multiple Endpoints Controlling the Experimentwise Error Rate. *Biometrics.* 1991;47:511–521.
29. Blair RC, Troendle JF, Beck RW. Control of Familywise Errors in Multiple Endpoint Assessments via Stepwise Permutation Tests. *Stat Med.* 1996;15:1107–1121.
30. Westfall PH, Young SS. *Resampling Based Multiple Testing.* New York: Wiley; 1993.
31. Bregenzer T, Lehmacher W. Multivariate Summenstatistiken—Ergebnisse einer Simulationsstudie. In: Trampisch, HJ et al., eds. *Medizinische Forschung, Arztliches Handeln.* 40. Jahrestagung der GMDS, 1995, MMV Medizin Verlag, München, 1995:156–161.