Biometrical Journal **44** (2002) 1, 101–116

Many-to-one Comparisons in Stratified Designs[1]

EGBERT BIESHEUVEL

Biometrics Department
Organon International
The Netherlands

LUDWIG A. HOTHORN

LG Bioinformatik
University of Hannover
Germany

Summary

CHEUNG and HOLLAND (1992) extended Dunnett's procedure for comparing all active treatments with a control simultaneously within each of r groups while maintaining the Type I error rate at some designated level α allowing different sample sizes for each of the group-treatment categories. This paper shows that exact percentage points can be easily calculated with current available statistical software (SAS). This procedure is compared to resampling techniques and a Bonferroni corrected Dunnett-within-group procedure by means of a simulation study.

Key words: Comparisons with a control; Stratified Design; Resampling Techniques.

1. Introduction

DUNNETT's (1955) procedure is a widely used multiple comparisons procedure for simultaneously comparing, by interval estimation or hypothesis testing, all active treatments with a control for a one-way design when sampling from a distribution where the normality assumption is reasonable. Here we will discuss the extension to the multi-group situation. This situation occurs in biological and medical research, one can think of different lots in an experiment, or males and females or young and old patients in a clinical trial. As explained by TUKEY (1977), when a large data set undergoes extensive data splitting without careful control of the overall error rate, 'false significance' can easily result. For instance, in multiple hypotheses testing, the familywise error rate (FWE) can be substantial even though each individual hypothesis is tested with a small α level. A multiple com-

[1] Presented at the 2nd International Meeting on Multiple Comparison Procedures, Berlin, June 2000

 0323-3847/02/0102-0101 $ 17.50+.50/0

parisons procedure is said to control the FWE in the weak sense if the FWE is controlled under the complete null configuration, and is said to control the FWE in the strong sense if it controls the FWE under any of the configurations. See HOCHBERG and TAMHANE (1987).

It should also be kept in mind that other distributional characteristics such as discreteness, skewness (i.e. non-normality) and equality of variances is as important as the choice of the multiple comparison procedure itself.

When one wishes to undertake the treatment versus control tests in a multi-group situation, it is worth considering whether the investigator should control the familywise error rate:

(a) for active treatments versus control averaged over all groups via the original Dunnett procedure;
(b) for active treatments versus control separately in each group via the original Dunnett procedure (Sometimes called 'Dunnett-within-group' procedure); or
(c) globally across all groups as presented in the present paper.

CHEUNG and HOLLAND (1991) extended Dunnett's procedure for comparing all active treatments with a control simultaneously within each of r groups while maintaining the Type I error rate at some designated level α for the situation of a common sample size. In 1992 they described this procedure allowing different sample sizes for each of the group-treatment categories. However, they tabulated exact percentage points for the equicorrelated case only but showed that linear interpolation on these values yielded satisfactory approximations to the exact percentage points for most configurations of unequal sample sizes.

This paper shows that with current available statistical software (SAS), the exact percentage points can be calculated easily. In addition the determination of power calculations is addressed.

Resampling techniques as available alternative methods to perform many-to-one comparisons in the stratified design are also discussed. A clinical example is given and finally the described tests are compared by means of a simulation study.

2. Hypothesis and test statistic

Let X_{ijk} denote the k^{th} observation on treatment j in group i. Assume there are r groups and c active treatments. Let $j = 0$ denote the control treatment and the other c active treatments are labelled by $j = 1$ to c, respectively. Without loss of generalisation the number of treatments contained in each group are assumed to be equal, although the formulas will also apply in case this situation does not hold, i.e. c varies across i.

Assume that the sample values $\{X_{ijk}\}$ are independently normally distributed with mean μ_{ij} and common variance σ^2, i.e. $X_{ijk} \sim N(\mu_{ij}, \sigma^2)$.

Let s^2 be the usual pooled variance estimator of σ^2 based on ν degrees of freedom, which is independent of the sample means $\bar{X}_{ij}$.

The aim is to test the null hypothesis of no effect between any of the c active treatments versus control within each of the r groups

$$H_0 : \mu_{ij} = \mu_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

against the one-sided alternative hypothesis that an active treatment exists which is superior to control within any of the r groups

$$H_1 : \exists j : \mu_{ij} > \mu_{i0} \qquad (i = 1, \ldots, r, \;\; j = 1, \ldots, c)$$

or in case of the two-sided alternative hypothesis that an active treatment exists which is different from control within any of the r groups

$$H_1 : \exists j : \mu_{ij} \neq \mu_{i0} \qquad (i = 1, \ldots, r, \;\; j = 1, \ldots, c).$$

Similar to the test statistic as proposed by DUNNETT (1955) in his original procedure, CHEUNG and HOLLAND (1991, 1992) proposed the following pivotal statistics:

$$D_{ij} = \frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \qquad (i = 1, \ldots, r, \;\; j = 1, \ldots, c)$$

for the one-sided alternative hypothesis

and

$$D_{2ij} = \frac{|\bar{X}_{ij} - \bar{X}_{i0}|}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \qquad (i = 1, \ldots, r, \;\; j = 1, \ldots, c)$$

for the two-sided alternative hypothesis.

Notice that for $1 \leq j_1, j_2 \leq c$, the correlation between

$$D_{ij_1} \text{ and } D_{ij_2} \text{ is } \varrho_{i(j_1, j_2)} = b_{ij_1} b_{ij_2} \quad \text{where} \quad b_{ij} = \sqrt{\frac{n_{ij}}{n_{i0} + n_{ij}}}\,.$$

So the correlation matrix $\boldsymbol{R}$ is given by:

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R_1} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots \\ \boldsymbol{0} & \ldots & \boldsymbol{R_r} \end{pmatrix} \quad \text{with} \quad \boldsymbol{R}_i = \begin{pmatrix} 1 & \varrho_{i(1,2)} & \ldots & \varrho_{i(1,c)} \\ \varrho_{i(2,1)} & 1 & \ldots & \ldots \\ \ldots & \ldots & 1 & \varrho_{i(c-1,c)} \\ \varrho_{i(c,1)} & \ldots & \varrho_{i(c,c-1)} & 1 \end{pmatrix}$$

Thus given i, the product correlation structure holds.

To test H_0 use the statistic

$D = \max_{1 \leq i \leq r;\, 1 \leq j \leq c} \{D_{ij}\}$ in case of the one-sided alternative hypothesis and

$D = \max_{1 \leq i \leq r;\, 1 \leq j \leq c} \{D_{2ij}\}$ in case of the two-sided alternative hypothesis.

Similar to the original Dunnett procedure, it can easily be shown that the joint distribution of the D_{ij}'s is a rc multivariate t-distribution with ν degrees of freedom and correlation matrix $\boldsymbol{R}$.

The remaining of this paper is restricted to the one-sided test situation only. The formulas for the one-sided test situation are simpler although the extension to the two-sided test situation is rather straightforward.

3. Evaluation of the upper percentage points

The estimation and testing problem under examination requires upper percentage points $d_\alpha = d(\alpha, r, c, \nu, \{b_{ij}\})$ from the probability distribution of $D = \max_{1 \le i \le r;\, 1 \le j \le c} \{D_{ij}\}$ such that $P(D \le d_\alpha) = 1 - \alpha$ under the null hypothesis.

To find the upper percentage points it is sufficient to calculate the probability $P(D \le t)$ for arbitrary t because several methods exist to calculate a quantile given the probabilities $P(D \le t)$. Two popular methods are the %-rejection method as proposed by EDWARDS and BERRY (1987) and the class of root finding methods, like the secant method and the bisection method. BRETZ (1999) showed that the bisection method yields good results and is simple to implement.

Therefore, we focus on the problem how to calculate these probabilities $P(D \le t)$.

Below, three methods of calculating $P(D \le t)$ are illustrated.

3.1 *Multivariate t-distribution*

Make use of the probabilities of a multivariate t-distribution:

$$\begin{aligned} P(D \le t) &= P\Big(\max_{1 \le i \le r;\, 1 \le j \le c} \{D_{ij}\} \le t\Big) \\ &= P(\text{all } D_{ij} \le t) = P(D_{11} \le t \ \wedge \ \ldots \ \wedge \ D_{rc} \le t)\,. \end{aligned}$$

Under H_0, $(D_{11}, \ldots, D_{rc})$ follow a central rc-variate t-distribution with ν degrees of freedom and correlation matrix $\boldsymbol{R}$ characterised by $\{b_{ij}\}$. In this case t is the equiquantile of this t-distribution.

These points can for example be calculated if one uses the randomised lattice rule method as described by GENZ and BRETZ (1999).

3.2 *Multivariate normal distribution*

Making use of the relationship between the multivariate t-distribution and the multivariate normal distribution:

Let T_{rc} be multivariate t distributed with ν degrees of freedom and correlation matrix $\boldsymbol{R}$ characterised by $\{b_{ij}\}$.

DUNNETT (1955) showed that the distribution function of a rc-variate t-distribution with ν degrees of freedom and correlation matrix $\boldsymbol{R}$ can be transformed into a

single integral over a *rc*-dimensional standardised normal distribution with the same matrix $\boldsymbol{R}$ as covariance matrix, i.e.

$$P(T_{rc} \le t) = \int_{-\infty}^{t} \dots \int_{-\infty}^{t} h_{rc}(x; \boldsymbol{R}; \nu)\, \mathrm{d}x_{11} \dots \mathrm{d}x_{rc}$$
$$= \int_{-\infty}^{\infty} \Phi_{rc}(\sqrt{x}\, t; 0; \boldsymbol{R})\, \mathrm{d}G_{\nu}(x)$$

where $h_{rc}(x; \boldsymbol{R}; \nu)$ is the density function of the *rc*-variate *t*-distribution, $\Phi_{rc}(x; 0; \boldsymbol{R})$ is the *rc*-dimensional standardised normal cumulative density function and $G_{\nu}(.)$ is the cumulative density function of a χ^2_{ν}/ν random variable.

This leads to $P(D \le t) = P(T_{rc} \le t) = \int_{-\infty}^{\infty} \Phi_{rc}(\sqrt{x}\, t; 0; \boldsymbol{R})\, \mathrm{d}G_{\nu}(x)$.

Thus the problem of finding the equiquantile of a multivariate *t*-distribution has been reduced to the calculation of the cumulative density function of a multivariate standardised normal distribution. Several solutions are available to solve this problem.

3.3 *Univariate normal distribution*

Make use of the fact that the correlation matrix $\boldsymbol{R}$ of the D_{ij}'s satisfies the product structure correlation within each of the *r* groups (see Section 2).

The calculation of the probability of a multivariate normal distribution with a correlation matrix which satisfies the product structure does not involve the integration of the full dimensional multivariate normal distribution but can be simplified to the integration of univariate standard normal distributions only. See also BECHHOFER and TAMHANE (1974) for the easier evaluation of a multivariate normal integral having a block covariance structure.

Define $u = s^2/\sigma^2$, then:

$$P(D \le t) = P\left(\max_{1 \le i \le r;\, 1 \le j \le c} \{D_{ij}\} \le t\right)$$
$$= \int_0^{\infty} P\left(\max_{1 \le i \le r;\, 1 \le j \le c} \left\{ \frac{\bar{X}_{ij} - \bar{X}_{i0}}{\sigma \sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right\} \le t\sqrt{u}; \sqrt{\frac{s^2}{\sigma^2}} = \sqrt{u} \right) g(u)\, \mathrm{d}u$$
$$= \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j=1}^{c} \Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1 - b_{ij}^2}} \right) \varphi(y)\, \mathrm{d}y \right] g(u)\, \mathrm{d}u$$

where $g(u)$ is the density function of a χ^2_{ν}/ν distributed variable, i.e. $g(u) = \dfrac{\nu^{\nu/2}\, \mathrm{e}^{-u\nu/2}\, u^{\nu/2-1}}{\Gamma(\nu/2) 2^{\nu/2}}$, and $\Phi(.)$ and $\varphi(.)$ are the cumulative and probability density function of the univariate standard normal distribution respectively.

Notice that the inner integrand

$$\text{prob}_i = \int_{-\infty}^{\infty} \prod_{j=1}^{c} \Phi\left(\frac{b_{ij}y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}}\right) \varphi(y)\,\mathrm{d}y$$

is the probability provided by the original Dunnett procedure applying infinite degrees of freedom.

To avoid programming problems CHEUNG and HOLLAND (1992) proposed linear interpolation on the exact percentage points for the equicorrelated case to derive approximated percentage points for the case of unequal sample sizes. However, with the current available statistical software like the SAS system, the exact percentage points can also be easily calculated for this situation.

The probability provided by the original Dunnett procedure applying infinite degrees of freedom can be calculated with the function PROBMC within SAS and the outer integral can be calculated using the subroutine QUAD available within PROC IML, which performs numerical integration in one dimension. (The SAS code can be found on the homepage www.bioinf.uni-hannover.de.)

4. Implementation for testing and estimation

In practical problems the testing of an individual hypothesis whether a particular active treatment is superior to the control in one of the strata is often more relevant than the testing of the global hypothesis as described in Section 2.

Consider the finite family of rc sub-hypotheses

$$H_{0ij} : \mu_{ij} = \mu_{i0}\,,$$

against

$$H_{1ij} : \mu_{ij} > \mu_{i0}\,.$$

Clearly, $H_0 = \bigcap_{ij} H_{0ij}$ and $H_1 = \bigcup_{ij} H_{1ij}$.

Notice that the test procedure to test the global null hypothesis H_0 can be seen as an Union-Intersection test procedure as described by ROY (1953). In fact, all individual hypotheses H_{0ij} with corresponding $D_{ij} > d_\alpha$ can be rejected if the global hypothesis H_0 can be rejected because of $D > d_\alpha$ while strongly controlling the FWE rate at level α. This can be shown by considering the testing procedure as a simultaneous test procedure in the sense of GABRIEL (1969).

Corresponding upper one-sided $100(1-\alpha)\%$ simultaneous confidence intervals for $\mu_{ij} = \mu_{i0}$ are given by

$$(\bar{X}_{ij} - \bar{X}_{i0} - d_\alpha s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}, \infty) \quad \text{for all} \quad i = 1, \ldots, r \quad \text{and} \quad j = 1, \ldots, c.$$

The simultaneous coverage probability of the rc intervals is easily seen to be $1-\alpha$ (See also CHEUNG and HOLLAND (1992).)

5. Power

The power to test the global null hypothesis H_0 against the one-sided alternative hypothesis H_1 using the statistic $D = \max_{1\le i\le r;\,1\le j\le c}\{D_{ij}\}$ can be derived as follows (see also Genz and Bretz (1999)):

$$\begin{aligned}
\text{Power} &= P(D \ge d_\alpha \mid H_1) = P\Big(\max_{1\le i\le r;\,1\le j\le c}\{D_{ij}\} \ge d_\alpha \mid H_1\Big)\\
&= 1 - P\Big(\max_{1\le i\le r;\,1\le j\le c}\{D_{ij}\} < d_\alpha \mid H_1\Big)\\
&= 1 - P\left(\frac{\bar{X}_{11}-\bar{X}_{10}}{s\sqrt{n_{11}^{-1}+n_{10}^{-1}}} < d_\alpha \wedge \ldots \wedge \frac{\bar{X}_{rc}-\bar{X}_{r0}}{s\sqrt{n_{rc}^{-1}+n_{r0}^{-1}}} < d_\alpha \mid H_1\right)\\
&= 1 - P\left(\frac{\dfrac{(\bar{X}_{11}-\mu_{11})-(\bar{X}_{10}-\mu_{10})}{\sigma\sqrt{n_{11}^{-1}+n_{10}^{-1}}}+\delta_{10}}{s/\sigma}\right.\\
&\qquad \left. < d_\alpha \wedge \ldots \wedge \frac{\dfrac{(\bar{X}_{rc}-\mu_{rc})-(\bar{X}_{r0}-\mu_{r0})}{\sigma\sqrt{n_{rc}^{-1}+n_{r0}^{-1}}}+\delta_{rc}}{s/\sigma} < d_\alpha\right).
\end{aligned}$$

The probability term described above has a rc-variate noncentral t-distribution with correlation matrix $\boldsymbol{R}$, ν degrees of freedom and non-centrality vector

$$\delta = (\delta_{ij})_{1\le i\le r;\,1\le j\le c} = \left(\frac{\mu_{ij}-\mu_{i0}}{\sigma\sqrt{n_{ij}^{-1}+n_{i0}^{-1}}}\right)_{1\le i\le r;\,1\le j\le c}.$$

This definition of power, often called global power, is in line with the definition given by HAYTER and LIU (1992). They define the power as the probability of rejecting the global hypothesis H_0, if $\max_{1\le i\le r;\,1\le j\le c}(\mu_{ij}-\mu_{i0}) \ge \Delta$ for pre-assigned Δ. So H_0 is rejected if Dunnett's test rejects at least one of the sub-hypotheses $H_{0ij}: \mu_{ij} \ne \mu_{i0}$, no matter which one. That is, a rejected H_{0ij} need not belong to a treatment with $(\mu_{ij}-\mu_{i0}) \ge \Delta$. However, the hypothesis belonging to the largest difference from the control will have the greatest chance of being rejected.

Another way to define power is to look at the finite family of rc sub-hypotheses H_{0ij} and H_{1ij}.

Then different types of power can be defined. The most common used definitions are the so called any-pair power, the probability of detecting at least one true difference among all pairs, the all-pairs power, the probability of detecting all true differences among all pairs, and the so called per-pair power, referring to one particular pair.

Let S be the subset of $\{ij\}$ such that the null hypotheses H_{0ij} are not true. Then the all-pairs power can be written as

$$\text{Power}_{\text{all-pairs}} = P(D_{ij} \geq d_\alpha \, \forall \, ij \in S) = P\left(\frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \geq d_\alpha \, \forall \, ij \in S \right)$$

$$= P\left(\frac{\dfrac{(\bar{X}_{ij} - \mu_{ij}) - (\bar{X}_{i0} - \mu_{i0})}{\sigma\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} + \dfrac{\mu_{ij} - \mu_{i0}}{\sigma\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}}}{s/\sigma} \geq d_\alpha \, \forall ij \in S \right).$$

with the probability of a k-variate noncentral t-distribution with ν degrees of freedom and non-centrality vector

$$\delta = (\delta_{ij})_{ij \in S} \left(\frac{\mu_{ij} - \mu_{i0}}{\sigma\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right)_{ij \in S}.$$

Here k is the dimension of S, i.e. the number of false null hypotheses H_{0ij}.

Making use of the relationship between the multivariate t-distribution and the normal distribution results in the expression:

$$\text{Power}_{\text{all}-\text{pairs}} = P(D_{ij} \geq d_\alpha \, \forall \, ij \in S)$$

$$= \int_0^\infty \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j:ij \in S} \Phi\left(\frac{-b_{ij} y - d_\alpha \sqrt{u} + \delta_{ij}}{\sqrt{1 - b_{ij}^2}} \right) \varphi(y) \, \mathrm{d}y \right] g(u) \, \mathrm{d}u \, .$$

Analogously the any-pair power can be written as

$$\text{Power}_{\text{any}-\text{pair}} = P(D_{ij} \geq d_\alpha \, \exists \, ij \in S) = 1 - P(D_{ij} < d_\alpha \, \forall \, ij \in S)$$

$$= 1 - P\left(\frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} < d_\alpha \forall \, ij \in S \right)$$

$$= 1 - P\left(\frac{\dfrac{(\bar{X}_{ij} - \mu_{ij}) - (\bar{X}_{i0} - \mu_{i0})}{\sigma\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} + \delta_{ij}}{s/\sigma} < d_\alpha \, \forall \, ij \in S \right).$$

Analogously to the calculations of the upper percentage points as described in Section 3, the power probabilities can be calculated, although it is more complicated.

Sample size determination is important in the planning of the experiments and studies. It demands that a minimum difference Δ (>0) between μ_{ij} and μ_{i0} is preassigned which is worth detecting. Assume that there are k hypotheses with $\mu_{ij} - \mu_{i0} \geq$. In most cases k is completely unknown, i.e. $1 \leq k \leq rc$. However, it is easy and useful to consider the more general case, where a priori knowledge provides that $g \leq k \leq h$ for some integers g and h with $1 \leq g \leq h \leq rc$. Thus, the most common situation where no a priori knowledge is available is regarded as the special case $g = 1$ and $h = rc$. A least favourable configuration (LFC) can be determined that provides a minimum probability to reject all/any of these k hypotheses, if $g \leq k \leq h$. Based on the LFC, the sample size can be calculated that guarantees a minimum pre-assigned all-pairs/any-pair power. See e.g. HORN and VOLLANDT (1998) for more details.

6. Resampling techniques

This section describes three resampling techniques that can be used to calculate p-values in the stratified many-to-one comparison situation. All three methods can be easily calculated within SAS. (The SAS code of this section can be found on the homepage www.bioinf.uni-hannover.de.) See WESTFALL and YOUNG (1993) for an extensive overview of resampling-based multiple testing issues.

6.1 *Stochastic approximation*

EDWARDS and BERRY (1987) proposed the first method we describe. This method approximates a quantile by parametric simulation.

The idea is to substitute a random variable D_α obtained by computer simulation instead of the upper-α percentile point d_α itself, in much the same way as s is substituted for σ.

The upper-α percentile points are exact if one considers the simulation to be an integral part of the experiment. They are approximate, with the degree of approximation under control, if one prefers to consider the obtained simulated critical point as a fixed constant, i.e. conditional on the results of the simulation.

The simulation run size m can be set so that the tail area for the simulated quantile q, say $\hat{q}$, is within γ of $1 - \alpha$ with $100(1 - \varepsilon)\%$ confidence. In equation form: $P(|F(\hat{q}) - (1 - \alpha)| \leq \gamma) = 1 - \varepsilon$, where F is the cumulative distribution function of the maximum.

This method can be applied within SAS by making use of the SIMULATE adjustment option within the procedure PROC MIXED.

6.2 *Bootstrap*

The bootstrap method creates pseudo-data sets by sampling observations with replacement from the original set of observations. An entire data set is thus created, and p-values for all tests are computed on this pseudo-data set. A counter is set up for each individual test to record whether the minimum p-value from the pseudo-data set is less than or equal to the actual p-value for each base test. This process is to be repeated a large number of times, and the proportions of occurrences of the minimum pseudo-p-values being less than or equal to each of the actual p-values is the adjusted p-value reported.

The pooling of the treatment groups is not likely to recreate the shape of the null hypothesis distribution, since the pooled data are likely to be multimodal. Therefore the variables are mean-centred before resampling. The bootstrap method controls the FWE in the strong sense approximately, being more precise with larger sample sizes. These details as well as others are well described by WESTFALL and YOUNG (1993).

This method can be applied within SAS by making use of the BOOTSTRAP option within the procedure PROC MULTTEST.

6.3 *Permutation*

The permutation-style adjusted p-values are computed in identical fashion as the bootstrap adjusted p-values, with the exception that the resampling is to be performed without replacement instead of with replacement. In the spirit of re-randomisation analyses, the variables are not centred prior to resampling, in contrast to the bootstrap method.

This method can also be applied within SAS by making use of the PERMUTATION option within the procedure PROC MULTTEST. However, note that PROC MULTTEST uses the global permutation distribution in the calculation of all the p-values. Therefore, the p-value of the global hypothesis is correct, but the p-values of the individual sub-hypotheses can be invalid. This problem has been illustrated by WESTFALL and WOLFINGER (2000).

In addition, it should be emphasized that the permutation approach is conditional with respect to the data and so it gives rise to conditional inferences, whereas bootstrap is not a strictly conditional procedure, in fact it is asymptotically unconditional.

7. Example

Irritable Bowel Syndrome (IBS) represents one of the most common functional gastro-intestinal disorders. This disease occurs in both genders, although more frequent in females. Last year the Food and Drug Administration (FDA) has ap-

proved the first compound for treatment of female IBS patients only. In males no clear efficacy could be demonstrated. As a result of this, there was high interest to perform subgroup analyses by gender in running clinical trials conducted for a similar compound. These trials were designed to show efficacy of several active treatment arms versus placebo across both genders. The data described here are part of the results of a randomised double-blind placebo controlled 12-weeks dose ranging clinical trial in IBS. An important efficacy endpoint in the treatment of IBS is the improvement in abdominal pain. The intensity of the abdominal pain was assessed on a daily basis by the patients on a 5 points scale (ranging from none (0) to incapacitating (4)). The endpoint of interest is the baseline adjusted average daily abdominal pain during the last week of treatment. (See Table 1).

The assumption of normally distributed data with a common variance is not unusual in the planned analysis of such data. Tables 2 and 3 show the analyses results.

Table 1
Reduction in baseline adjusted abdominal pain scores in last week on drug

Treatment	Subgroup					
	Males			Females		
	N	Mean	SD	N	Mean	SD
Placebo	21	0.206	0.639	59	0.221	0.724
Dose 1	24	0.662	0.761	59	0.430	0.876
Dose 2	26	0.512	0.706	56	0.515	0.710
Dose 3	27	0.482	0.866	52	0.619	0.714
Dose 4	20	0.530	0.856	59	0.578	0.803

Table 2
Analysis results of the stratified many-to-one comparisons for the IBS example (σ^2= 0.5865; Df = 359; d_α= 2.443)

Group	Contrast	Estimate	Std error	Pr(t)*	P-value**	One-sided 95% CI***
M	Dose1-Plac	0.455	0.229	0.024	0.141	$(-0.104, \infty)$
	Dose2-Plac	0.306	0.225	0.087	0.404	$(-0.243, \infty)$
	Dose3-Plac	0.276	0.223	0.108	0.471	$(-0.269, \infty)$
	Dose4-Plac	0.324	0.239	0.088	0.407	$(-0.261, \infty)$
F	Dose1-Plac	0.209	0.150	0.083	0.388	$(-0.158, \infty)$
	Dose2-Plac	0.293	0.154	0.029	0.166	$(-0.083, \infty)$
	Dose3-Plac	0.398	0.157	0.006	0.040	$(0.013, \infty)$
	Dose4-Plac	0.356	0.151	0.009	0.061	$(-0.012, \infty)$

* one-sided unadjusted p-value
** one-sided Dunnett adjusted p-values across groups
*** simultaneous one-sided confidence intervals $(\bar{X}_{ij} - \bar{X}_{i0} - d_{1\alpha} s \sqrt{n_{ij}^{-1} + n_{i0}^{-1}}, \infty)$

Table 3
P-values for different methods for the IBS example

Group	Contrast	Dunnett	Edwards	Bootstrap	Permutation	Bonferroni-Dunnett*
M	Dose1-Plac	0.141	0.139	0.136	0.139	0.138
	Dose2-Plac	0.404	0.403	0.399	0.402	0.438
	Dose3-Plac	0.471	0.470	0.467	0.468	0.525
	Dose4-Plac	0.407	0.407	0.402	0.405	0.442
F	Dose1-Plac	0.388	0.386	0.382	0.386	0.449
	Dose2-Plac	0.166	0.165	0.161	0.164	0.182
	Dose3-Plac	0.040	0.039	0.038	0.040	0.044
	Dose4-Plac	0.061	0.061	0.058	0.061	0.067

* original Dunnett procedure within both genders at a significance level of ½ α

8. Simulations

A simulation study was conducted to compare the behaviour of the four procedures described above as well as the Bonferroni corrected Dunnett-within-group procedure. This procedure consists of applying the original Dunnett procedure within each of the groups at a significance level of α/r to control the FWE in the strong sense. Some of the many configurations that were investigated are reported here.

The number of observations for each of the control treatments is chosen to be equal, say n_0, and the number of observations for each of the active treatments within each of the groups is also taken to be equal, say n_a. The following pairs of combinations of (n_0, n_a) were considered: (2, 5), (5, 5), (5, 10), (10, 5) and (10, 10). The number of groups is taken to be two ($r = 2$) and there are three active treatment arms within each of the groups ($c = 3$). The random error terms are taken as independently identically distributed random variables from the standard normal distribution or from the lognormal distribution, which are being generated as the exponential of a standard normal random variable. The performance of these methods are compared under the null hypothesis to check whether the FWE is correctly kept at an alpha level of $\alpha = 0.05$ using 10.000 replications for each of the settings. The results are shown in Table 4.

All five methods approximate the alpha level in the situation of normally distributed data quit well. However, in case of lognormal distributed data, the stratified Dunnett procedure, the Bonferroni corrected Dunnett-within-group method as well as the stochastic approximation technique as proposed by EDWARDS and BERRY (1987) become much too liberal. This is not a surprise because it is well known that Dunnett's procedure doesn't control the FWE in case of non-normal distributed data. The other two resampling techniques (bootstrap and permutation) behave much better in this situation.

Table 4
Empirical Type I errors ($\alpha = 0.05$) for $r = 2$ groups and $c = 3$ active treatment arms within each group based on 10000 replications

		Method				
n_0	n_a	Dunnett	Edwards	Bootstrap	Permutation	Bonferoni-Dunnett
Normal data						
2	5	0.0507	0.0508	0.0487	0.0510	0.0512
5	5	0.0534	0.0525	0.0493	0.0527	0.0483
5	10	0.0519	0.0529	0.0495	0.0516	0.0565
10	5	0.0565	0.0563	0.0529	0.0564	0.0508
10	10	0.0497	0.0494	0.0487	0.0502	0.0512
Lognormal data						
2	5	0.0170	0.0173	0.0220	0.0337	0.0121
5	5	0.0730	0.0737	0.0630	0.0593	0.0442
5	10	0.0283	0.0280	0.0403	0.0403	0.0835
10	5	0.1087	0.1097	0.0637	0.0567	0.0155
10	10	0.0627	0.0630	0.0560	0.0540	0.0468

In addition, the methods are compared under the alternative hypothesis using the any-pair power. Results of the all-pairs power can be found in BIESHEUVEL (2002). The setting of the one-sided alternative hypothesis represents a linear shift of 0.5 for each of the active treatment means; i.e. the means of the three active treatments have a positive shift of 0.5, 1.0 and 1.5, respectively compared to placebo. The simulations are conducted using 3000 replications for each of the settings.

Table 5 shows the results for the any-pair power.

In the situation of normally distributed data, the stratified Dunnett and the resampling techniques have very similar power results, although the bootstrap technique seems to have the lowest power of these four techniques. Even in the case of only two groups, the Bonferroni corrected Dunnett-within-group procedure shows the lowest power of these five methods for all settings as shown above. This is line with our expectation and one can image that the loss of power in comparison to the other methods will increase if the number of groups (r) increases.

Before one compares the power of these five methods directly in the situation of lognormal distributed data, one should keep in mind that the stratified Dunnett, the stochastic approximation and the Bonferroni corrected Dunnett-within-group methods are not maintaining the FWE. On the other hand, Table 5 doesn't show that the bootstrap and permutation techniques have a much lower power in these settings.

These simulation results indicate that the stratified Dunnett procedure maintains the FWE in the situation of normal distributed data, as do the other proposed

Table 5
Any-pair power a linear shift and for $r = 2$ groups, $c = 3$ active treatment arms within each group based on 3000 replications

		Method				
n_0	n_a	Dunnett	Edwards	Bootstrap	Permutation	Bonferoni-Dunnett
Normal data						
2	5	0.5327	0.5317	0.5117	0.5353	0.5017
5	5	0.7687	0.7707	0.7477	0.7717	0.7343
5	10	0.8833	0.8820	0.8793	0.8853	0.8640
10	5	0.8843	0.8840	0.8713	0.8867	0.8800
10	10	0.9760	0.9760	0.9743	0.9757	0.9730
Lognormal data						
2	5	0.2307	0.2280	0.2970	0.3190	0.2670
5	5	0.4280	0.4277	0.3943	0.3927	0.4707
5	10	0.4370	0.4367	0.5143	0.5070	0.5700
10	5	0.5127	0.5107	0.4017	0.4050	0.4930
10	10	0.5923	0.5913	0.5557	0.5603	0.6487

methods. The power of all five methods, except the Bonferroni style adjusted method, are similar for the situation of normal distributed data. The simulation study also indicates that the FWE of the stratified Dunnett procedure, the Bonferroni corrected Dunnett-within-group method and the stochastic approximation technique are inflated in case of lognormal distributed data. Both the bootstrap and permutation resampling techniques seem to behave better without substantial loss in power.

9. Discussion

The comparison of all active treatments with a control simultaneously within each of r groups while maintaining the probability of making any Type I error at some designated level α in the situation of unequal sample sizes can be performed with current available statistical software. Adjusted p-values, exact percentage points and simultaneous confidence intervals can be easily derived in case of normal distributed data. Power formulas are also derived for this situation.

Other techniques, like resampling techniques, are also available within current statistical software to perform many-to-one comparisons in a stratified design. In particular, the bootstrap method seems to be worthwhile to calculate p-values in case of non-normal distributed data. Although, PROC MULTTEST doesn't provide you directly with simultaneous confidence intervals. See for example WESTFALL and YOUNG (1993) how to construct simultaneous confidence intervals in this situation.

This paper has focused on single-step multiple testing procedures. However, if one is only interested in hypothesis testing and not in determination of confidence intervals, CHEUNG and HOLLAND (1994) proposed a step-down multiple testing procedure for the multi-group situation which has more power than the single-step procedure as described in this paper. The procedure PROC MULTTEST provides also an option of step-down bootstrap and permutation resampling.

Acknowledgements

The authors would like to thank the two reviewers for their helpful suggestions.

References

BECHHOFER, R. E. and TAMHANE, A. C., 1974: An iterated integral representation for a multivariate normal integral having block covariance structure. *Biometrika* **61**, 615–619.

BIESHEUVEL, E. H. E., 2002: Many-to-one comparisons in stratified designs. Ph.D. thesis, University of Hannover. (in preparation)

BRETZ, F., 1999: Powerful modifications of Williams' test on trend. Ph.D. thesis, University of Hannover. (see also homepage www.bioinf.uni-hannover.de)

CHEUNG, S. H. and HOLLAND, B., 1991: Extension of Dunnett's multiple comparison procedure to the case of several groups. *Biometrics* **47**, 21–32.

CHEUNG, S. H. and HOLLAND, B., 1992: Extension of Dunnett's multiple comparison procedure with differing sample sizes in the case of several groups. *Computational Statistics & Data Analysis* **14**, 165–182.

CHEUNG, S. H. and HOLLAND, B., 1994: A step-down procedure for multiple tests of treatment versus control in each of several groups. *Statistics in Medicine* **13**, 2261–2267.

DUNNETT, C. W., 1955: A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.

EDWARDS, D. and BERRY, J. J., 1987: The efficiency of simulation-based multiple comparisons. *Biometrics* **43**, 913–928.

GABRIEL, K. R., 1969: Simultaneous test procedures – some theory of multiple comparisons. *Annals of Mathematical Statistics* **40**, 224–250.

GENZ, A. and BRETZ, F., 1999: Numerical computation of multivariate t-probabilities with application to power calculations of multiple contrasts. *Journal of Statistical Computation and Simulation* **63**, 361–378.

HAYTER, A. J. and LIU, W., 1992: A method of power assessment for tests comparing several treatments with a control. *Communications in Statistics – Theory Methods* **21**, 1871–1889 .

HORN, M. and VOLLANDT, R., 1998: Sample sizes for comparisons of k treatments with a control based on different definitions of the power. *Biometrical Journal* **40**, 5, 589–612.

ROY, S. N., 1953: On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* **24**, 220–238.

SAS/STAT® Software, 1996: Changes and Enhancements through Release 6.11. SAS Institute Inc., Cary, NC, USA.

TUKEY, J. W., 1977: Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198**, 679–684.

WESTFALL, P. H. and YOUNG, S. S., 1993: Resampling based multiple testing: examples and methods for p-value adjustment. John Wiley & Sons, New York.

WESTFALL, P. H. and WOLFINGER, R. D., 2000: Closed multiple testing procedures and

PROC MULTTEST, Observations: The Technical Journal for SAS Software Users, SAS Institute Inc., June 2000, issue 23, www.sas.com/service/library/periodicals/obs/obswww23.

EGBERT BIESHEUVEL
Biometrics Department
Organon International
Grickenweg 29
PP 500
NL-5340 Amoss
The Netherlands

Received, September 2000
Revised, April 2001
Revised, August 2001
Accepted, October 2001