Journal of Biopharmaceutical Statistics, 24: 378–397, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2013.860156

POWER AND SAMPLE SIZE DETERMINATION IN CLINICAL TRIALS WITH MULTIPLE PRIMARY CONTINUOUS CORRELATED ENDPOINTS

Pierre Lafaye de Micheaux[1], Benoit Liquet[2,3,4], Sébastien Marque[5], and Jérémie Riou[2,3,5]

[1]*Department of Mathematics and Statistics, Université de Montréal, Quebec, Canada*
[2]*University of Bordeaux, ISPED, INSERM, Bordeaux, France*
[3]*INSERM, ISPED, Bordeaux, France*
[4]*The University of Queensland, Brisbane, Australia*
[5]*Danone Research, Palaiseau Cedex, France*

The use of two or more primary correlated endpoints is becoming increasingly common. A mandatory approach when analyzing data from such clinical trials is to control the family-wise error rate (FWER). In this context, we provide formulas for computation of sample size and for data analysis. Two approaches are discussed: an individual method based on a union–intersection procedure and a global procedure, based on a multivariate model that can take into account adjustment variables. These methods are illustrated with simulation studies and applications. An R package known as `rPowerSampleSize` is also available.

Key Words: Correlated endpoints; Family-Wise error rate; Multiple continuous endpoints; Multivariate normal distribution; Power; Sample size determination.

1. INTRODUCTION

The use of multiple endpoints to characterize product efficacy and safety measures is an increasingly common feature in recent clinical trials. Efficacy is often defined not by a unique endpoint but by a combination of several parameters. Regulatory agencies commonly require more than one endpoint to measure different aspects of product efficacy in confirmatory clinical trials. However, the use of multiple endpoints is a source of debate (Sankoh, 1997), and a lot of statistical literature on the subject has been published. In general, national health authorities recommend, on the basis of the biostatistics guideline developed by the International Conference on Harmonization (ICH E9 Expert Working Group, 1999), the selection of one primary endpoint to provide strong scientific evidence of the efficacy of a test treatment. However, this strategy has clear limitations, notably when it leads to arbitrary classification of different endpoints. While it reduces the dimension of the

Received April 18, 2012; Accepted October 22, 2012

Address correspondence to Pierre Lafaye de Micheaux, Département de Mathématiques et de Statistique, 2920 Chemin de la Tour, Montréal, Québec H3T 1J4, Canada; E-mail: lafaye@dms.umontreal.ca

problem, if the classification is not sufficiently robust, real effects may be ignored or left undetected. Consequently, many clinical trials incorporate multiple primary endpoints to demonstrate the efficacy of the product.

Consideration of multiple endpoints nonetheless brings with it several challenges to the design and analysis of trial data (Cook and Farewell, 1996; O'Brien, 1984; Pocock et al., 1987). Several authors have discussed power calculations in clinical trials when two or more primary endpoints are given as continuous variables (Chuang-Stein et al., 2007; Dunnett and Tamhane, 1992; Senn and Bretz, 2007). On the one hand, one strategy following the same philosophy as that of the guidelines is to reduce as far as possible the number of endpoints (Neuhäuser, 2006). However, this strategy may result in a loss of information concerning endpoints and does not address the scientific problem of the selection of parameters. On the other hand, an alternative strategy is to consider all primary endpoints. We have focused on the situation of treating all endpoints equally, thus referred to as co-primary endpoints. However, in clinical practice, it can also be interesting to consider a different weighting for each endpoint (Bretz et al., 2009, 2011; Burman et al., 2009). Depending on the scientific question raised, statisticians may be interested in "or" comparisons (detecting at least one significant primary endpoint) or in multiple must-win comparisons (detecting at least r among m comparisons); see Julious and McIntyre (2012). Several authors developed multiple testing procedures in the context of a "win" on all co-primary endpoints; see for example, Berger (1982) and Sozu et al. (2006, 2010, 2011). We have limited this report to the detection of at least one primary endpoint for the "two treatments" case. In this context, the most common strategy is to use either single-step (Simes, 1986; Sidak, 1967) or stepwise (Holm, 1979; Hochberg, 1988; Hommel, 1988) procedures. For single-step methods, the rejection or nonrejection of a single hypothesis does not account for the outcome of any other hypotheses. A well-known example of single-step procedures is the Bonferroni test. In contrast, for stepwise methods, the rejection or nonrejection of a particular hypothesis may take into account the outcome of other hypotheses. Stepwise methods are more powerful than single-step procedures. The equally well-known Holm procedure is a stepwise extension of the Bonferroni test using a closure principle. Both types of procedures are conservative (lead to wrongly "accepting" the null hypothesis) and might lead to biased test decisions, as information about correlations of the endpoints is not exploited. This implies a strong control of the type I error probability and consequently, a decrease in the power of each test. An extensive work has been done by Sankoh (1997) in order to characterize the advantages and limits of adjusted methods. Gatekeeping procedures (Dmitrienko et al., 2003), which consist of scheduling the hypotheses and analyzing the data with multiple families of null hypotheses, suffer from similar problems and need an order of priority among the endpoints. Another alternative is to use the union–intersection test procedure (Roy, 1953). This method can control the family wise error rate (FWER), and correlations among endpoints can be taken into account. Finally, global methods that take correlations into account, such as the T^2 test of Hotelling (1953), can be used "where the endpoints are alternative measures of the same fundamental quantity" (Sankoh, 1997). One limitation of this procedure is that it gives a global and nondirectional result. This problem has been pointed out by Sankoh et al. (1999). Furthermore, in the context of data missing completely at random (MCAR), Yoon et al. (2011) recently showed that this is a less powerful method.

The aim of this article is to provide sample size calculation methods, as well as corrections, for type I error probabilities based either on a global method with a multivariate linear model or on an individual method involving a union–intersection procedure. The approach of the global method is to generalize the T^2 test of Hotelling to deal with adjustment variables. Finally, we compare power and FWER control of both methods with common methods for different scenarios of correlation and adjustment. In section 2, we present the statistical methods related to simultaneous testing, as well as power and sample size calculations. In section 3, we present the results of a simulation study, and an application in two nutritional clinical studies. Lastly, the results are discussed and a conclusion including limitations and perspectives is provided in section 4.

2. METHODS

Two different approaches are presented in this section. First, we present an individual testing procedure with an exact control of the FWER. In this context, the power and the sample size determination are defined under different assumptions. Second, we propose a global procedure based on a multivariate model involving adjustment variables.

2.1. Overview

We consider the context where a vector $\mathbf{X} = (X_1, \dots, X_m)^\mathsf{T}$ of m quantitative variables (endpoints) is measured in a group of $2n$ subjects taken at random in two subpopulations: a control group (C) and a test group (T). Let $\mathbf{X}_1^j, \dots, \mathbf{X}_n^j$ be n independent and identically distributed (conditional on group j) random vectors, with expectation $\boldsymbol{\mu}^j$ and some covariance matrix Σ, where $j = C$ stands for the control group and $j = T$ stands for the test product. The kth component $X_{i,k}^j$ of vector $\mathbf{X}_i^j$ denotes the ith observation ($1 \leq i \leq n$), on the kth endpoint ($1 \leq k \leq m$) for product j. Let $\boldsymbol{\delta} = (\delta_1, \dots, \delta_m)^\mathsf{T} = \boldsymbol{\mu}^T - \boldsymbol{\mu}^C$, with $\delta_k = \mu_k^T - \mu_k^C$ be the vector of true mean differences between the test and control products respectively, where $^\mathsf{T}$ denotes vector or matrix transposition. The test product will be considered to be different from the control product on the kth endpoint, if $\delta_k \neq 0$. The clinical aim is to be able to detect a mean difference between the test and the control product for at least one endpoint among m. This can be stated under a statistical hypothesis formalism as:

$$\mathscr{H}^0\colon \boldsymbol{\delta} = \boldsymbol{0}_m \quad \text{versus} \quad \mathscr{H}^1\colon \boldsymbol{\delta} \neq \boldsymbol{0}_m, \tag{1}$$

where $\mathbf{0}_m = (0, \dots, 0)^\mathsf{T}$ is the null vector of length m. In Section 2.3, we use a global test of $\mathscr{H}^0$ to address this problem. Another avenue is to consider a so-called individual testing procedure based on the m following single hypotheses:

$$\mathscr{H}_k^0\colon \delta_k = 0 \quad \text{versus} \quad \mathscr{H}_k^1\colon \delta_k \neq 0, \tag{2}$$

noting that

$$\mathscr{H}^0 = \bigcap_{k=1}^{m} \mathscr{H}_k^0 \quad \text{and} \quad \mathscr{H}^1 = \bigcup_{k=1}^{m} \mathscr{H}_k^1. \tag{3}$$

This latter approach, based on the family hypothesis $\{\mathscr{H}_1^0, \ldots, \mathscr{H}_m^0\}$, is considered first.

2.2. Individual Testing Procedure

In the context of individual testing procedures, we need to define all the test statistics used. When the variances $\sigma_k^2 = \mathbb{V}\mathrm{ar}(X_{1,k}^j)$, $1 \leq k \leq m$, are known, the standardized test statistic that will be used to test (2) is:

$$Z_k^{(n)} = \frac{\overline{X}_k^T - \overline{X}_k^C}{\sqrt{\frac{2}{n}}\sigma_k}, \tag{4}$$

where $\overline{X}_k^j = \frac{1}{n}\sum_{i=1}^{n} X_{i,k}^j$ is the sample mean for group j.

When the σ_k^2's are unknown, they will be estimated by the pooled variance

$$\hat{\sigma}_k^2 = \frac{1}{2n-2}\sum_{i=1}^{n}\left[\left(X_{i,k}^C - \overline{X}_k^C\right)^2 + \left(X_{i,k}^T - \overline{X}_k^T\right)^2\right]$$

and the "studentized" test will be used instead

$$T_k^{(n)} = \frac{\overline{X}_k^T - \overline{X}_k^C}{\sqrt{\frac{2}{n}}\hat{\sigma}_k}. \tag{5}$$

In the sequel, $Z_k^{(n)}$ will be replaced with $T_k^{(n)}$ when the σ_k's are unknown.

2.2.1. A Direct Approach to Control the FWER. We reject the individual null hypothesis $\mathscr{H}_k^0$ if $|Z_k^{(n)}|$ is larger than a suitable multiplicity adjusted critical point c_α. Since $\mathscr{H}^1 = \cup_{k=1}^{m}\mathscr{H}_k^1$, it seems natural to decide $\mathscr{H}^1$ if at least one member of the family $\{\mathscr{H}_1^0, \ldots, \mathscr{H}_m^0\}$ is rejected using an individual procedure. The type-I error probability of the global procedure is then exactly equal to the FWER of the multiple procedure, defined as:

$$\begin{aligned} FWER &= \mathsf{P}(\text{reject at least one } \mathscr{H}_k^0,\ 1 \leq k \leq m \,|\, \mathscr{H}^0 \text{ is true}) \\ &= 1 - \mathsf{P}\left\{\left(|Z_1^{(n)}| \leq c_\alpha\right) \cap \cdots \cap \left(|Z_m^{(n)}| \leq c_\alpha\right) \,|\, \mathscr{H}^0 \text{ is true}\right\}. \end{aligned} \tag{6}$$

The adjusted critical value c_α is chosen to satisfy $FWER = \alpha$, for a fixed significance level α. Obviously, the joint distribution of $\mathbf{Z}_n = (Z_1^{(n)}, \ldots, Z_m^{(n)})^\mathsf{T}$, or of $\mathbf{T}_n = (T_1^{(n)}, \ldots, T_m^{(n)})^\mathsf{T}$ when the σ_k's are estimated, has to be known or at least approximated to some degree (see Section 2.2.3). Note that this procedure allows us to explicitly specify the value of the FWER, which is better than controlling its value using an upper limit, as is usually the case.

2.2.2. Power and Sample Size Determination.

An important task in the design phase of clinical trials, is to determine the sample size n that guarantees a prespecified power, noted hereafter as $1-\beta$. In single testing situations, power is defined as the probability of rejecting the null hypothesis under investigation, when it is false. For multiple testing and multiple comparisons, Westfall et al. (1999) propose other definitions of power. The clinical interest here is to detect at least one significant endpoint among m with a given power, so we use the the so-called minimal power (referred to as disjunctive power by Senn and Bretz (2007)), which is given by

$$\begin{aligned} 1-\beta &= \mathsf{P}\left(\text{reject at least one } \mathscr{H}_k^0,\ 1 \le k \le m \,|\, \mathscr{H}^1 \text{ is true}\right) \\ &= 1-\mathsf{P}\left\{\left(|Z_1^{(n)}| \le c_\alpha\right) \cap \cdots \cap \left(|Z_m^{(n)}| \le c_\alpha\right) \,|\, \mathscr{H}^1 \text{ is true}\right\}. \end{aligned} \tag{7}$$

In this article, we want to determine the common adjusted critical value c_α, as well as the sample size n, in order to control the FWER at a fixed significance level α and to guarantee a prespecified minimal power $1-\beta$. We use an iterative procedure based on equations (6) and (7) with two unknown parameters (c_α and n). Clearly, the joint distribution of the test statistics used in equations (6) and (7) has to be known under $\mathscr{H}^0$, as well as under $\mathscr{H}^1$. In the latter case, this distribution will depend on the value of the vector of mean differences between the test and control products (reported hereafter as $\boldsymbol{\delta}^* \neq 0$), and also on Σ or an estimate of it. This is investigated thereafter.

Remark. Equation (6) can also be used alone for determining c_α in order to control the FWER when the aim is to analyze a data set.

2.2.3. Distribution of $\mathbf{Z}_n$ and $\mathbf{T}_n$.

Normality assumption and known covariance matrix. We assume that the random vectors $\mathbf{X}_1^j, \ldots, \mathbf{X}_n^j$ follow a $\mathcal{N}_m(\boldsymbol{\mu}^j, \Sigma)$ distribution with Σ known. In this context, it is easy to show that

$$\mathbf{Z}_n \overset{\mathscr{H}^0}{\sim} \mathcal{N}_m\left(\boldsymbol{0}_m, R\right) \quad \text{and} \quad \mathbf{Z}_n \overset{\mathscr{H}^1}{\sim} \mathcal{N}_m\left(\sqrt{\frac{n}{2}} P\boldsymbol{\delta}^*, R\right),$$

where $\boldsymbol{\delta}^* \neq \boldsymbol{0}_m$ is the value of $\boldsymbol{\delta}$ under $\mathscr{H}^1$ and where $R = P\Sigma P$ is the $m \times m$ correlation matrix associated with Σ, with P the diagonal matrix whose kth element is $1/\sigma_k$.

Remark. Senn and Bretz (2007) proposed an alternative method based on a common latent variable in the case where you have a single unvarying pairwise correlation and if the components of $P\boldsymbol{\delta}^*$ (noncentrality parameters) are all the same.

Normality assumption and unknown covariance matrix. In this context, allowing $Y_k = \frac{\overline{X}_k^T - \overline{X}_k^C}{\sqrt{\frac{2}{n}}\sigma_k}$ and $U_k = v\frac{\hat{\sigma}_k^2}{\sigma_k^2}$, we use the vector

$$\mathbf{T}_n = \left(T_1^{(n)}, \ldots, T_m^{(n)}\right)^{\mathsf{T}} = \left(\frac{Y_1}{\sqrt{U_1/v}}, \ldots, \frac{Y_m}{\sqrt{U_m/v}}\right)^{\mathsf{T}},$$

where, under the global null hypothesis $\mathscr{H}^0$, the vector $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathsf{T}}$ follows an m-dimensional normal distribution with correlation matrix R and where $U_1, \ldots, U_m$ are dependent χ^2 random variables with $v = 2n - 2$ degrees of freedom. The distribution of the vector $\mathbf{T}_n$ is a type II multivariate Student distribution with v degrees of freedom, generalization of a bivariate t-distribution considered by Siddiqui (1967), representing the situation of two endpoints. It has not been possible, as far as we know, to obtain an expression of the density or distribution function of this law in a closed form. Hasler and Hothorn (2011) propose, without justification, to approximate this distribution by an m-variate type I t-distribution with $v = 2n - 2$ degrees of freedom and with correlation matrix $\widehat{R}$, an estimate of R. Using the same approximation as these authors, the distribution of the vector $\mathbf{T}_n$ under the alternative hypothesis is approximated by an m-variate type I t-distribution with the noncentrality parameter $\sqrt{\frac{n}{2}}P\boldsymbol{\delta}^*$ and with $v = 2n - 2$ degrees of freedom.

Asymptotic context. In order to be more general, we can consider that the covariance matrices differ between the control and test group respectively, namely, that we have $\mathbb{V}\text{ar}\left(\mathbf{X}_1^j\right) = \Sigma^j$, $j = C, T$. Then, the usual individual test statistic $T_k^{(n)}$ is defined by

$$T_k^{(n)} = \frac{\overline{X}_k^T - \overline{X}_k^C}{\sqrt{\frac{\hat{\sigma}_{k,C}^2}{n} + \frac{\hat{\sigma}_{k,T}^2}{n}}}, \tag{8}$$

where $\hat{\sigma}_{k,j}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_{i,k}^j - \overline{X}_k^j)^2$ for $j = C, T$. The multivariate central limit theorem enables us to state

$$\sqrt{n}\left[(\overline{\mathbf{X}}^T - \overline{\mathbf{X}}^C) - (\boldsymbol{\mu}^T - \boldsymbol{\mu}^C)\right] \xrightarrow{L} \mathcal{N}_m(\mathbf{0}_m, \Sigma),$$

where $\overline{\mathbf{X}}^j = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^j$ and where $\Sigma = \Sigma^C + \Sigma^T$ since the two groups are independent. We thus have

$$R^{-1/2}\left[\sqrt{n}V(\overline{\mathbf{X}}^T - \overline{\mathbf{X}}^C) - \sqrt{n}V\boldsymbol{\delta}^*\right] \xrightarrow{L} \mathcal{N}_m(\mathbf{0}_m, I_m), \tag{9}$$

where here $R = V\Sigma V^{\mathsf{T}}$ with $V = \operatorname{diag}\left(1/\sqrt{\sigma_{k,C}^2 + \sigma_{k,T}^2}\right)$. In this context, under very general conditions (Cox and Hinkley, 1994, pp. 258–266), we can estimate Σ^j by :

$$\widehat{\Sigma}^j = \frac{1}{n-1}\sum_{i=1}^n \left(\mathbf{X}_i^j - \overline{\mathbf{X}}^j\right)\left(\mathbf{X}_i^j - \overline{\mathbf{X}}^j\right)^{\mathsf{T}}.$$

Then $\widehat{R} = \widehat{V}\ \widehat{\Sigma}\ \widehat{V}$ is a consistent estimator of R, the correlation matrix of $\mathbf{T}_n = \sqrt{n}\widehat{V}(\overline{\mathbf{X}}^T - \overline{\mathbf{X}}^C)$, where $\widehat{V} = \mathrm{diag}\left(1/\sqrt{\hat{\sigma}^2_{k,C} + \hat{\sigma}^2_{k,T}}\right)$ and $\widehat{\Sigma} = \widehat{\Sigma}^C + \widehat{\Sigma}^T$. Now, using Slutsky's theorem, we obtain

$$\widehat{R}^{-1/2}\mathbf{T}_n \xrightarrow{L} \mathcal{N}_m(\boldsymbol{0}_m, I_m), \text{ under } \mathcal{H}^0,$$

and

$$\widehat{R}^{-1/2}\left(\mathbf{T}_n - \sqrt{n}\widehat{V}\boldsymbol{\delta}^*\right) \xrightarrow{L} \mathcal{N}_m(\boldsymbol{0}_m, I_m), \text{ under } \mathcal{H}^1\text{: } \boldsymbol{\delta} = \boldsymbol{\delta}^* \neq \boldsymbol{0}_m.$$

2.2.4. Practical Implementation.

Computation of the adjusted critical value c_α and determination of the sample size n are done in R, an open-source statistical software (R Development Core Team, 2011). We used the `pmvnorm( )` and `pmvt( )` functions from the `mvtnorm` package (Genz and Bretz, 2009; Genz et al., 2012) for the computation of the multivariate normal and of the multivariate type I t-distribution probabilities. The sample size computation involves an effect size parameter. We recall that the effect size for the kth endpoint is defined as $\Delta_k = \frac{\mu_k^T - \mu_k^C}{\sigma_k^*}$, where σ_k^* is the population standard deviation of variable X_k. Note that σ_k^* can be expressed in terms of the standard deviations of the variables X_k^C and X_k^T. In our framework of normality assumption with known or unknown covariance matrix, the standard deviation σ_k^* equals σ_k and the vector of effect size for the m endpoints corresponds to the term $P\boldsymbol{\delta}^*$. In the asymptotic context, as we consider a standard deviation in the control group that is different from the test group ($\sigma_{k,T} \neq \sigma_{k,C}$), the vector of effect size corresponds to $\sqrt{2}V\boldsymbol{\delta}^*$. In this latter definition, we consider that $\sigma_k^* = \sqrt{\frac{\sigma^2_{k,C}+\sigma^2_{k,T}}{2}}$.

Two iterative procedures have been defined according to the assumptions made.

Normality assumption and known covariance matrix. Briefly, the procedure based on the `pmvnorm( )` function consists of performing the following steps:

(i) Specifying the effect size $P\boldsymbol{\delta}^*$ for all endpoints, the correlation matrix R, the significance level α and the desired power $1 - \beta$.
(ii) Determining c_α as a solution of $FWER = \alpha$ in equation (6).
(iii) For a starting value n_0 of sample size, computing the minimal power $1 - \beta$ using equation (7) with n_0 and c_α from step (ii).
(iv) Going back to step (iii) with an incremented or decremented sample size n_0 until the desired power.

Normality assumption and unknown covariance matrix. The procedure (based on the `pmvt( )` function) is slightly different because the distribution of $\mathbf{T}_n$ under the null hypothesis depends on the sample size:

(i) Specifying the effect size for all endpoints defined by $P\boldsymbol{\delta}^*$, the correlation matrix $\widehat{R}$ (which can be given by a pilot study), the significance level α, and the desired power $1 - \beta$.

(ii) For a starting value n_0 of sample size, determining c_α as a solution of $FWER = \alpha$ in equation (6).

(iii) Computing the minimal power $1 - \beta$ using equation (7) with n_0 and c_α from step (ii).

(iv) Going back to step (ii) with an incremented or decremented sample size n_0 until the desired power.

Asymptotic context. The principle of the procedure is the same as the procedure for the *normality assumption and known covariance matrix* case. Only the specification of the effect size for all endpoints defined by $\sqrt{2}V\boldsymbol{\delta}^*$ and of the correlation matrix $\widehat{R}$ change and can be defined by a pilot study. This definition of the effect size parameter permits a homogeneous notation with both previous cases.

2.3. Global Procedure

2.3.1. Model. We propose the following multivariate linear regression model to represent the data generation process:

$$\mathbf{Y} = \Gamma B + \mathbf{E}, \tag{10}$$

where $\mathbf{Y}^\mathsf{T} = [\mathbf{X}_1^C; \dots; \mathbf{X}_n^C; \mathbf{X}_1^T; \dots; \mathbf{X}_n^T]$ is a $m \times 2n$ matrix, $\Gamma = [\mathbf{1}_{2n}; \mathbf{g}; \mathbf{A}]$ is the $2n \times (p+2)$ design matrix, with $\mathbf{1}_{2n} = (1, \dots, 1)^\mathsf{T}$, $\mathbf{g} = (\mathbf{1}_n^\mathsf{T}, \boldsymbol{0}_n^\mathsf{T})^\mathsf{T}$, being an indicator variable of each group (1 for the control group and 0 for the treatment group), $\mathbf{A}$ is a $2n \times p$ matrix whose lth column $\boldsymbol{a}_l = (a_{l1}^C, \dots, a_{ln}^C, a_{l1}^T, \dots, a_{ln}^T)^\mathsf{T}$ contains the measurements of the lth adjustment variable ($1 \leq l \leq p$) on the $2n$ subjects, B is a $(p+2) \times m$ matrix of unknown coefficients associated with the design matrix, and $\mathbf{E}$ is a $2n \times m$ random matrix of errors such that vec($\mathbf{E}$) follows a $\mathcal{N}_{2n\times m}(\boldsymbol{0}_m; I_{2n} \otimes \Sigma)$ distribution, where vec($\cdot$) denotes the column-stacking operator and $\otimes$ denotes the Kronecker symbol. We let $\boldsymbol{\delta}$ be the second row of the matrix B, which represents the adjusted group effect for the m endpoints. It is worthwhile mentioning that $\boldsymbol{\delta} = \mathbb{E}[\mathbf{X}_i^C - \mathbf{X}_i^T | \mathbf{a}_i]$. In the sequel, we note $\bar{a}_l^j = (1/n)\sum_{i=1}^n a_{li}^j$ and $\overline{a_l a_{l'}}^j = (1/n)\sum_{i=1}^n a_{li}^j a_{l'i}^j$, $j = C, T$.

To overcome the multiple testing problem, we propose using a global test of the hypothesis

$$\mathcal{H}^0\colon \boldsymbol{\delta} = \boldsymbol{0}_m \quad \text{versus} \quad \mathcal{H}^1\colon \boldsymbol{\delta} \neq \boldsymbol{0}_m.$$

In the framework of model (10), this can be restated as

$$\mathcal{H}^0\colon \boldsymbol{C}\boldsymbol{B} = \boldsymbol{0}_m^\mathsf{T} \quad \text{versus} \quad \mathcal{H}^1\colon \boldsymbol{C}\boldsymbol{B} \neq \boldsymbol{0}_m^\mathsf{T},$$

using the contrast row vector $\boldsymbol{C} = (0, 1, \boldsymbol{0}_p^\mathsf{T})$ of size $1 \times (p+2)$.

Under the multinormality assumption of the disturbances, the least square estimator of matrix B is given by:

$$\widehat{\mathbf{B}} = \left(\Gamma^\mathsf{T}\Gamma\right)^{-1}\Gamma^\mathsf{T}\mathbf{Y} \sim \mathcal{N}_{(p+2)\times m}\left(B, \left(\Gamma^\mathsf{T}\Gamma\right)^{-1} \otimes \Sigma\right).$$

We can thus state

$$\boldsymbol{C}\widehat{\mathbf{B}} = \boldsymbol{C}\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y} \sim \mathcal{N}_m\left(\boldsymbol{C}B, W\right),$$

where

$$W = (\boldsymbol{C} \otimes I_m)\left[\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1} \otimes \Sigma\right]\left(\boldsymbol{C}^{\mathsf{T}} \otimes I_m\right) : m \times m. \tag{11}$$

This leads to

$$W^{-1/2}\left[\boldsymbol{C}\widehat{\mathbf{B}} - \boldsymbol{C}B\right] = W^{-1/2}\left[\boldsymbol{C}\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y} - \boldsymbol{C}B\right] \sim \mathcal{N}_m(\boldsymbol{0}_m, I_m).$$

2.3.2. Statistical Procedure and Distribution.

Known covariance matrix Σ. In this context, the test statistic considered is:

$$Z_n^2 = \left(\boldsymbol{C}\widehat{\mathbf{B}}\right) W^{-1}\left(\boldsymbol{C}\widehat{\mathbf{B}}\right)^{\mathsf{T}}.$$

After some relatively easy but tedious computations, we were able to show, using a formula for matrix inversion in block form, that W from equation (11) can be written as

$$W = (\Gamma^{\mathsf{T}}\Gamma)^{-1}_{2,2}\Sigma = \frac{1}{n}\left(2 + \boldsymbol{v}^{\mathsf{T}}M^{-1}\boldsymbol{v}\right)\Sigma,$$

where $\boldsymbol{v}$ is a $p \times 1$ vector whose lth component is $v_l = \bar{a}_l^C - \bar{a}_l^T$, and where M is a $p \times p$ matrix with general term $M_{l,l'} = \left(\overline{a_l a_{l'}}^C - \bar{a}_l^C \bar{a}_{l'}^C\right) + \left(\overline{a_l a_{l'}}^T - \bar{a}_l^T \bar{a}_{l'}^T\right)$.

It can be shown (Bilodeau and Brenner, 1999) that, under the null hypothesis, Z_n^2 follows a χ^2 distribution with m degrees of freedom, reported as χ_m^2. We then reject the null hypothesis $\mathcal{H}^0$ if the observed value of the test statistic Z_n^2 is greater than $q_{1-\alpha}^m$, the quantile of order $1-\alpha$ of the χ_m^2. Under the alternative hypothesis $\mathcal{H}^1 : \boldsymbol{C}B = \boldsymbol{\delta}^{*\mathsf{T}}$ (with $\boldsymbol{\delta}^* \neq \boldsymbol{0}_m$), the test statistic Z_n^2 follows a noncentral χ^2 distribution with m degrees of freedom and decentrality parameter

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}}W^{-1}\boldsymbol{\delta}^*. \tag{12}$$

Unknown covariance matrix Σ. In this context, the test statistic considered is

$$T_n^2 = \left(\boldsymbol{C}\widehat{\mathbf{B}}\right)\widehat{W}_n^{-1}\left(\boldsymbol{C}\widehat{\mathbf{B}}\right)^{\mathsf{T}},$$

where $\widehat{W}_n = (\boldsymbol{C} \otimes I_m)[(\Gamma^{\mathsf{T}}\Gamma)^{-1} \otimes \widehat{\Sigma}](\boldsymbol{C}^{\mathsf{T}} \otimes I_m) = \frac{1}{n}(2 + \boldsymbol{v}^{\mathsf{T}}M^{-1}\boldsymbol{v})\widehat{\Sigma}$, with $\widehat{\Sigma}$ an unbiased estimator of Σ (see Appendix 2 for an explicit definition). In this case, under the null hypothesis, T_n^2 converges in distribution to a χ^2 distribution with m degrees of freedom. We also prove that, under the alternative hypothesis $\mathcal{H}^1 : \boldsymbol{C}B = \boldsymbol{\delta}^{*\mathsf{T}} \neq \boldsymbol{0}_m^{\mathsf{T}}$, the test statistic T_n^2 converges in distribution to a noncentral χ^2 distribution with m degrees of freedom and decentrality parameter

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}}\widehat{W}_n^{-1}\boldsymbol{\delta}^*. \tag{13}$$

We note that, without adjustment variables in model (10), the test statistic T_n^2 reduces to the classical Hotelling's test statistic (see Appendix 2 for a proof of this result).

2.3.3. Power and Sample Size Determination.

In the context of this global test, the power function for the statistic Z_n^2 (or T_n^2) is

$$1 - \beta = \mathsf{P}\left(Z_n^2 > q_{1-\alpha}^m | \mathscr{H}^1\right) = 1 - F_{\chi_m^2(\lambda_n)}(q_{1-\alpha}^m), \tag{14}$$

where $F_{\chi_m^2(\lambda_n)}(\cdot)$ is the cumulative distribution function of the noncentral χ_m^2 with decentrality parameter λ_n. The sample size required to achieve the desired power $1 - \beta$ is given as the smallest integer satisfying equation (14), using the decentrality parameter λ_n as given in equation (12) (or (13)).

2.3.4. Practical Implementation.

To achieve the sample size computation, the user specifies the vector $\boldsymbol{\delta}^*$ of mean differences between the test and the control products, the covariance matrix Σ between the outcomes, the desired significance level α, and the desired power $1 - \beta$. The R program we developed enables computation of the decentrality parameter λ_n using equation (12), and the sample size using equation (14) (or (13)).

In the presence of a single adjustment variable, the decentrality parameter λ_n from equation (13) reduces to

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}} \left\{ \frac{1}{n} \left(2 + \frac{(\bar{a}_1^C - \bar{a}_1^T)^2}{\overline{a_1^2}^C - (\bar{a}_1^C)^2 + \overline{a_1^2}^T - (\bar{a}_1^T)^2} \right) \widehat{\Sigma} \right\}^{-1} \boldsymbol{\delta}^*,$$

where $\bar{a}_1^j = (1/n)\sum_{i=1}^n a_{1i}^j$, $\overline{a_1^2}^j = (1/n)\sum_{i=1}^n a_{1i}^{2\,j}$ and where a_{1i}^j is the value of the adjustment variable for the i-th subject for treatment j ($j = C$: control; $j = T$: treatment). From a practical point of view, the user has to specify this parameter, which can be evaluated after a pilot study has been conducted. Note that for a binary adjustment variable, for example, gender (1 = women and 0 = men), the user only has to specify the frequency of women in each group (since in this case $\overline{a_1^2}^C = \bar{a}_1^C$). Moreover, if the number of women in each group is the same, the computation of the decentrality parameter λ_n reduces to the case without an adjustment variable.

3. RESULTS

3.1. Simulations

A simulation study was performed to evaluate the performance of the two proposed approaches. We first studied how the individual testing procedure was able to control the FWER. We investigated three different assumptions: normality and known covariance matrix (termed "NormKnown"), normality and unknown covariance matrix (termed "NormUnKnown"), and the asymptotic context (termed "Asympt"). We also compared the results obtained with the Bonferroni method and with the “naive method,” which consists of choosing the most significant test without any correction of the FWER. We then investigated the power of our proposed approaches and compared them to standard approaches such as Holm (1979), Hochberg (1988), and Bonferroni and Hotelling (1953); methods that are implemented using the `multtest` R package (Pollard et al., 2005).

All data for these simulations came from the model defined in equation (10) with one adjustment variable ($p = 1$), which follows a Bernoulli distribution with a probability of success $\pi = 0.6$. Each simulation was carried out on 200 subjects in each group. To simplify the interpretation and shorten the simulation study, we considered a compound symmetric covariance matrix Σ with $\text{diag}(\Sigma) = \mathbf{1}_m$. Note that we thus have a constant pairwise correlation ρ for all pairs of outcomes. Moreover, as multiple endpoints are often correlated in the same direction, we only investigated positive correlations ($\rho > 0$). We used 5,000 replications for all simulations. In the sequel, we define $B^{\mathsf{T}} = [\boldsymbol{b}_0, \boldsymbol{\delta}, \boldsymbol{b}_1]$ where the vector $\boldsymbol{b}_0$ $(m \times 1)$ represents the intercept of the model and where $\boldsymbol{b}_1$ represents the coefficient vector associated with the adjustment variable.

3.1.1. FWER. We first investigated, for different numbers of endpoints ($m \in \{1, \ldots, 15\}$), the control of the FWER for the individual testing procedure (under the three assumptions), for the Bonferroni method and for the naive approach. Under the null hypothesis ($\boldsymbol{\delta} = \boldsymbol{0}_m$), the FWER was estimated by the proportion of the Monte Carlo experiments that lead to a rejection of the null hypothesis ($p_{value} < 0.05$). In this simulation, we considered a model without an adjustment variable ($\boldsymbol{b}_1 = \boldsymbol{0}_m$), the intercept vector was fixed at $\boldsymbol{b}_0 = \mathbf{1}_m$, and finally ρ was set to 0.6. Figure 1 shows the evolution of the FWER for different numbers of endpoints.

The *naive* method, without correction for the multiple testing problem, increased with the number of endpoints. The error rate calculated by the Bonferroni method decreased with the number of endpoints. This correction was therefore too conservative whereas the individual testing procedure gave a type I error rate close to the nominal 0.05 value, under the three assumptions.

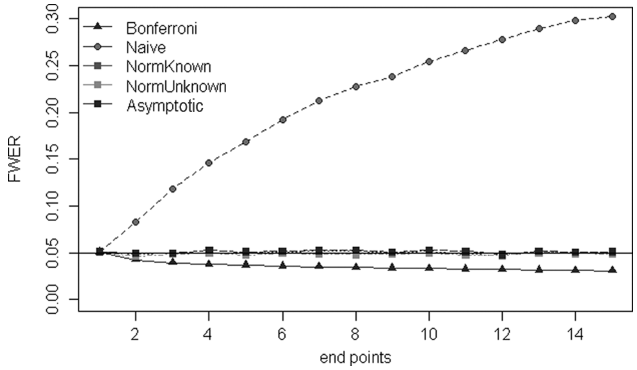

Figure 1 FWER as a function of the number of endpoints for the naive, Bonferroni, and individual testing procedure.

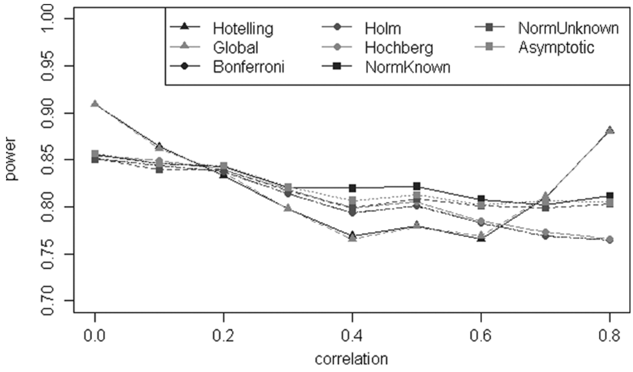

Figure 2 Minimal power study of different methods as a function of the correlation coefficient ρ from data that came from a model with no adjustment variable effect.

3.1.2. Power. We investigated the performance of the proposed approaches with varying correlations among the outcomes. We considered $m = 3$ correlated endpoints. First, we investigated the case where the adjustment variable has no effect ($\boldsymbol{b}_1 = \boldsymbol{0}_m$) on the outcomes. We then used $\boldsymbol{b}_1 = (1.0, 0.8, 0.6)^{\mathsf{T}}$ in order to determine the effect of an adjustment variable. In the two simulations cases, we fixed the treatment effect at $\boldsymbol{\delta} = (0.31, 0.13, 0.19)^{\mathsf{T}}$ and the intercept at $\boldsymbol{b}_0 = (2, 3, 1)^{\mathsf{T}}$.

No effect of the adjustment variable. The results of the minimal power for the different methods are presented in Fig. 2. We can see that for low correlation ($\rho < 0.2$) and high correlation ($\rho > 0.7$), the global method seems to be more powerful than the individual testing procedure, which was more powerful than the other methods. For medium correlation, the individual testing method was the most powerful and the global procedures were less so. Finally, for high correlation ($\rho > 0.6$), as expected, the methods that do not take into account the correlation between the endpoints were the least powerful.

Effect of the adjustment variable. The results of the minimal power for the different methods are presented in Fig. 3. In this simulation, we can see that the global procedure and the individual testing procedure with known covariance matrix assumption were more powerful than the others. The global method was more powerful for low and high correlation, whereas the individual testing procedure for known covariance matrix assumption gave better results for medium correlation. In this situation, the global method outperformed Hotelling's test. Taking into account an adjustment variable improves the estimation of the covariance matrix, which results in an increase of the power with the global method. The other methods are less powerful in this situation since they do not take into account the adjustment variable. However, the individual testing procedure was still more powerful than methods that do not take into account the correlation between endpoints.

3.2. Sample Size Computation

We present some results about sample size calculation in the context of three endpoints with the following parameters: the vector $\boldsymbol{\delta}^* = (0.2, 0.3, 0.4)^{\mathsf{T}}$ of mean

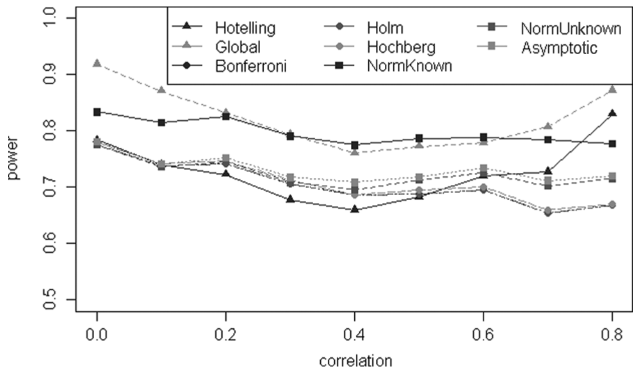

Figure 3 Minimal power study of different methods as a function of the correlation coefficient ρ for data generated from a model involving an adjustment variable.

differences between the test and the control products, and a compound symmetric covariance matrix Σ between the outcomes, with $\text{diag}(\Sigma) = (1.1^2, 1.2^2, 2.3^2)^\mathsf{T}$. The desired significance level was chosen at $\alpha = 0.05$. For the global method with a binary adjustment variable, the frequency of this variable in each group is fixed for samples of any size to $\bar{a}^C = 0.4$ and $\bar{a}^T = 0.6$. The results in Table 1 are coherent with the previous power simulation study. The sample size of each group required to reach a desired minimal power increases with the correlation coefficient for the Bonferroni procedure. We can also observe that for low ($\rho < 0.2$) and high ($\rho > 0.6$) correlations, the procedure based on the model requires fewer subjects than the other methods. For medium correlations, it is the individual testing procedure which requires the smallest sample size. We note that the global method with an adjustment variable (MA) requires more subjects than a model without an adjustment variable (M). While this may appear strange in regards to the previous

Table 1 Sample size n of each group required to achieve the desired level of minimal power: for Bonferroni (B), for our individual testing procedure for known (K) or unknown (U) covariance matrix, and in the asymptotic context (A), for our global method based on a multivariate model without an adjustment variable (M), and with a binary adjustment variable (MA), for various correlations ρ and with FWER= 0.05.

	Power (0.80)					Power (0.90)				
ρ	B	K/A	U	M	MA	B	K/A	U	M	MA
0	221	219	222	174	181	287	285	288	226	235
0.1	233	231	233	207	215	304	303	305	268	280
0.2	246	243	245	238	248	322	319	321	309	322
0.3	258	255	256	268	279	340	336	337	349	363
0.4	272	265	267	296	308	358	350	352	385	401
0.5	285	276	277	320	334	376	365	366	416	434
0.6	299	285	286	339	354	393	376	378	441	459
0.7	312	292	293	349	363	409	386	387	453	472
0.8	325	295	297	338	352	423	390	391	440	458
0.9	333	291	292	278	289	431	383	385	361	376

power simulation study, in fact, the sample size calculation assumes that Σ is known. Therefore, adding an adjustment variable to improve the estimation of Σ is not useful. However, in practice Σ is unknown (even if it has been estimated in a previous study with a very large sample size). We recommend that the sample size be determined using an adjustment variable if further data analysis is to be performed.

3.3. Application in Clinical Studies in Nutrition

The purpose of this section is to present the results obtained using our methods, in terms of sample size determination and the statistical data analysis. The two following applications deal with clinical studies performed in nutrition. Both studies were double-blind randomized controlled trials (DB-RCT) performed according to Good Clinical Practices (ICH-GCP).

3.3.1. Example 1: Sample Size Calculation. The first application is the sample size determination of a new DB-RCT with the objective of demonstrating the efficacy of the consumption of a dairy product on seric antibody titers for three strains of Influenza virus. For kth strain, the individual null hypothesis is $\mathscr{H}_k^0 : \delta_k = \mu_k^T - \mu_k^C = 0$. The product will be considered as effective if at least one out of the three strains is statistically significant. According to the usual standard, the type II error probability is fixed at 20% in order to obtain a power of 80% and the family-wise error rate must be controlled at 5%.

Two pilot studies were planned to define the product effects and variability. Both were DB-RCT multicentric studies conducted in France among elderly volunteers during the two vaccination seasons 2005 and 2006. Details are reported in Boge et al. (2009). Based on the results, the product's effects defined by means and correlations were calculated. The mean differences between both groups is $\hat{\boldsymbol{\delta}} = (0.35, 0.28, 0.46)^{\mathrm{T}}$ and the covariance matrix was $\widehat{\Sigma} = \begin{pmatrix} 5.58 & 2.00 & 1.24 \\ 2.00 & 4.29 & 1.59 \\ 1.24 & 1.59 & 4.09 \end{pmatrix}$. Following experts consensus, the mean differences obtained could be considered as clinically relevant. Based on these assumptions, we compared the most powerful methods, namely, the global and the individual procedure for known covariance matrix. Table 2 shows that the sample size may be reduced significantly depending on the method used. Indeed, with an individual procedure for known covariance matrix and an adjusted type I error probability at 0.0178, the sample size falls to 336 subjects required to see a significant difference for at least one outcome, versus 359 for the global method.

Table 2 Sample size computation with global method and individual procedure

Method	Type I error	Sample size (n)
Global	0.05	359
Indiv	0.0178	336

Note. Global: Global method based on multivariate model. Indiv: Individual procedure for known covariance matrix.

Table 3 Adjusted p_{value} estimation on 11 immunological markers for various multiple testing procedures, in an efficacy study of fermented dairy product

Endpoint	Naive	Bonferroni	Holm	Hochberg	Asympt
1	0.13	1.00	0.74	0.60	0.62
2	0.15	1.00	0.74	0.60	0.67
3	0.45	1.00	0.90	0.67	0.98
4	0.67	1.00	0.90	0.67	1.00
5	0.10	1.00	0.70	0.60	0.52
6	0.02*	0.21	0.19	0.18	0.14
7	0.00*	0.01*	0.01*	0.01*	0.01*
8	0.30	1.00	0.90	0.67	0.91
9	0.12	1.00	0.74	0.60	0.59
10	0.07	0.76	0.55	0.55	0.40
11	0.02*	0.22	0.19	0.18	0.14

*: Significant association.

3.3.2. Example 2: Analysis of Clinical Study Data. In order to demonstrate the effect of a fermented dairy product on the immune system, a monocentric, DB-RCT, parallel study with two groups was performed in 1,000 healthy subjects. Results are reported in Guillemard et al. (2009). As an exploratory analysis, the immune function of interest was characterized by a set of 11 biomarkers. According to the exploratory concept of this analysis, the product efficacy was assumed if at least one out of the 11 markers was statistically significant. During this analysis, the covariance matrix between parameters and the means vector was estimated on the actual data.

The statistical analysis was performed using a test of comparison of means with common multiple testing procedures and with the proposed asymptotic individual testing procedure defined in Section 2. A global procedure involving a model without an adjustment variable was also used. The functions `indiv.lm.analysis( )` and `global.lm.analysis( )` from the R package `rPowerSampleSize` were used to perform the analysis.

The results for the individual method (three assumptions) in terms of "adjusted p_{value}" estimation are summarized in Table 3.

The importance of using a type I error correction can be seen in this table. Without any correction ("naive method"), we could conclude that there are three significant endpoints (markers 6, 7, and 11). But, when a correction method is used, only one significant endpoint is found (marker 7). This conducts to conclude to the efficacy of the product. The global method, which gives us only a single result, is also significant with a p_{value} less than 0.01 and confirms that we have at least one marker that is significant.

4. CONCLUDING REMARKS

In this article, we considered two approaches (individual and global) for sample size determination and for the analysis of data with multiple continuous endpoints. The global method we have developed leads to a generalization of the well-known Hotelling (1953) statistic, involving adjustment variables. The methods developed allow consideration of cases when the covariance matrix is known or estimated. When designing clinical studies, assumptions for the sample size calculation may come from different sources that are more or less biased. The best situation is to be able to gather data from a well-designed pilot study, among defined populations on relevant and well-measured endpoints. In this case, the estimator of the mean differences could be considered as nonbiased or slightly biased. An interesting approach would be to consider the lower and upper boundaries of the confidence interval of the estimator and to calculate the two sample sizes associated to them. Regarding the covariance matrix, the bias may be more sensitive but several approaches might be used in order to correct this bias, as described, for instance, by Julious and Owen (2006). In any case, it is important to consider the results from literature or previous studies not as "known values" but always as estimations with their variability.

Simulation studies showed that the method based on the global model with adjustment variables is a powerful method when the true covariance matrix is unknown. Consequently, better sample size computations are possible. When adjustment variables are not available, the individual method seems to be more powerful, except for low and high correlation cases. Furthermore, in this context, the methods developed perform favorably compared to common procedures. Therefore, the choice of the method depends mainly on the aim of the study; for example, the global method gives only a global result and not a directional one. However, the choice also depends on the value of the correlation coefficient between the endpoints. Note that work on the generalization of the individual method in the context of detecting r significant endpoints among m is ongoing. Note also that we have implemented various methods in an `R` package called `rPowerSampleSize`. In this article, we have focused our analysis on bilateral tests, however, our package also takes into account the unilateral case for the individual procedure.

ACKNOWLEDGEMENT

We thank the Danone Research Clinical Study Platform for making the data available.

FUNDING

The work of the first author was supported by a grant from the Natural Science and Engineering Research Council of Canada.

REFERENCES

Berger, R. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 4:295–300.

Bilodeau, M., Brenner, D. (1999). *Theory of Multivariate Statistics*. New York, NY: Springer.

Boge, T., Rémigy, M., Vaudaine, S., Tanguy, J., Bourdet-Sicard, R., Van der Werf, S. (2009). A probiotic fermented dairy drink improves antibody response to influenza vaccination in the elderly in two randomised controlled trials. *Vaccine* 27:5677–5684.

Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*. 28:586–604.

Bretz, F., Maurer, W., Hommel, G. (2011). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* 30:1489–1501.

Burman, C., Sonesson, C., Guilbaud, O. (2009). A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in Medicine* 28:739–761.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., Offen, W. (2007). Challenge of multiple co-primary endpoints: A new approach. *Statistics in Medicine* 26:1181–1192.

Cook, R., Farewell, V. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 159:93–110.

Cox, D., Hinkley, D. (1994). *Theoretical Statistics*. Boca Raton, FL: Chapman & Hall.

Dmitrienko, A., Offen, W., Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 22:2387–2400.

Dunnett, C., Tamhane, A. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* 87:162–170.

Genz, A., Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Heidelberg, Germany: Springer-Verlag. Available at: http://CRAN.R-project.org/package=mvtnorm

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. (2012). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9992.

Guillemard, E., Tondu, F., Lacoin, F., Schrezenmeir, J. (2009). Consumption of a fermented dairy product containing the probiotic lactobacillus casei dn-114001 reduces the duration of respiratory infections in the elderly in a randomised controlled trial. *British Journal of Nutrition* 103:58–68.

Hasler, M., Hothorn, L. A. (2011). A Dunnett-type procedure for multiple endpoints. *International Journal of Biostatistics* 7:74–81.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* 75:383–386.

Hotelling, H. (1953). A generalised *t* test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 23–41.

Julious, S., McIntyre, N. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. *Pharmaceutical Statistics* 11:177–185.

Julious, S., Owen, R. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics* 5:29–37.

Neuhäuser, M. (2006). How to deal with multiple endpoints in clinical trials. *Fundamental & Clinical Pharmacology* 20:515–523.

O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* 40:1079–1087.

Pocock, S., Geller, N., Tsiatis, A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* 43:487–498.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., Dudoit, S. (2005). multtest: Resampling-based multiple hypothesis testing. R package version 2.6.0. Available at: http://www.bioconductor.org/packages/release/bioc/html/multtest.html

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* 24:2387–2400.

Sankoh, A., Huque, M., Dubey, S. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine* 16:2529–2542.

Sankoh, A., Huque, M., Russel, H., D'Agostino, R. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal* 33:119–140.

Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 6:161–170.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62:626–633.

Siddiqui, M. M. (1967). A bivariate *t*-distribution. *Annals of Mathematical Statistics* 38:162–166.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–754.

Sozu, T., Kanou, T., Hamada, C., Yoshimura, I. (2006). Power and sample size calculations in clinical trials with multiple primary variables. *Japanese Journal of Biometrics* 27:83–96.

Sozu, T., Sugimoto, T., Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine* 29:2169–79.

Sozu, T., Sugimoto, T., Hamasaki, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics* 21:650–668.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R., Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests*. Cary, NC: SAS Press.

Williams, J. D., Woodall, W. H., Birch, J. B., Sullivan, J. H. (2004). Distributional Properties of the Multivariate T2 Statistic Based on the Successive Differences Covariance Matrix Estimator. Technical report. Department of Statistics, Virginia Polytechnic Institute and State University.

Yoon, F. B., Fitzmaurice, G. M., Lipsitz, S. R., Horton, N. J., Laird, N. M., Normand, S. L. T. (2011). Alternative methods for testing treatment effects on the basis of multiple outcomes: Simulation and case study. *Statistics in Medicine* 30:1917–1932.

APPENDIX 1: `rPowerSampleSize` PACKAGE

The `rPowerSampleSize` package was developed in R, an open source statistical software available at `http://www.r-project.org`. It contains, for the moment, five functions: three for the sample size computation (one for the individual procedure, one for the global method, and one for the Bonferroni procedure) and two for the analysis of real data in order to solve the multiple testing problem (one for the individual procedure, one for the global method). We present an illustration of the main `rPowerSampleSize` functions next.

Briefly, concerning the sample size determination, the user needs to specify in the `indiv.1m.ssc( )` function the alternative hypothesis (bilateral or unilateral), the effect size and the correlation between the endpoints. However, for the global method, the user also needs to specify in the `global.1m.ssc( )` function, the difference of means between the two groups (δ^*), and the vector of the standard deviations of each endpoint, instead of the effect size. With an adjustment variable, the user also needs to enter the mean of the adjustment variable for each group. The functions for sample size computation provide the adjusted significance level and the required sample size. The `bonferroni.1m.ssc( )` function is not displayed in the following.

```
 # Sample size computation for the individual method:
 > indiv.1m.ssc(method = "Known",ES=c(0.1,0.2,0.3),
 cor=diag(1,3))

 Sample size: 183
 Adjusted significance level: 0.0170

 # Sample size computation for the global method:
 > global.1m.ssc(method = "Adj.Model",mean.diff=c(0.1,0.2,
0.3), sd=c(1,1,1),cor=diag(1,3),v=-0.2,M=0.46)

 Sample size: 163
```

Concerning the analysis of the data, the user needs to specify in the `indiv.1m.analysis( )` function, the alternative hypothesis (bilateral or unilateral), and the assumption used: asymptotic test or normality assumption. The function provides the adjusted p_{value} associated to each endpoint. For the analysis based on the multivariate model, the user specifies the adjustment covariable in the `global.1m.analysis( )` function that returns the p_{value} of the global test.

```
 > data(data.sim)
 > n <- nrow(data)/2
 > XC <- data[1:n,1:3]
 > XT <- data[(n+1):(2*n),1:3]

 # Data analysis for the individual method:
 > indiv.1m.analysis(method = "UnKnown",XC=XC,XT=XT,n=n)

 Endpoint 1 2 3
 Adjusted p-value 0.4164 0.1419 0.0076

 # Data analysis for the global method:
 > global.1m.analysis(XC=XC,XT=XT,A=data[,5],n=n)

 p-value: 0.0023
```

APPENDIX 2: STATISTIC T_n^2 REDUCES TO HOTELLING'S TEST STATISTIC

Without adjustment covariates in the multivariate model, the matrix $\widehat{W}_n = (\boldsymbol{C} \otimes I_m)[(\Gamma^{\mathsf{T}}\Gamma)^{-1} \otimes \widehat{\Sigma}](\boldsymbol{C}^{\mathsf{T}} \otimes I_m)$ reduces to $\widehat{W}_n = \frac{2}{n}\widehat{\Sigma}$. Then the statistic T_n^2 is defined as

$$T_n^2 = \frac{n}{2}(\boldsymbol{C}\widehat{\mathbf{B}})\widehat{\Sigma}^{-1}(\boldsymbol{C}\widehat{\mathbf{B}})^{\mathsf{T}},$$

where

$$\widehat{\mathbf{B}} = (\Gamma^{\mathsf{T}}\Gamma)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y}$$

and

$$\widehat{\Sigma} = \frac{1}{2n-(2+p)}(\mathbf{Y}-\Gamma\widehat{\mathbf{B}})^{\mathsf{T}}(\mathbf{Y}-\Gamma\widehat{\mathbf{B}}),$$

with p the number of adjustment variables. In the context of no adjustment variable, we obtain

$$\widehat{\mathbf{B}} = \left[\overline{\mathbf{X}}^{T}, \overline{\mathbf{X}}^{C} - \overline{\mathbf{X}}^{T}\right]^{\mathsf{T}}$$

and thus

$$\widehat{\Sigma} = \frac{1}{2n-2}\sum_{i=1}^{n}\left[\left(\mathbf{X}_i^C - \overline{\mathbf{X}}^C\right)\left(\mathbf{X}_i^C - \overline{\mathbf{X}}^C\right)^{\mathsf{T}} + \left(\mathbf{X}_i^T - \overline{\mathbf{X}}^T\right)\left(\mathbf{X}_i^T - \overline{\mathbf{X}}^T\right)^{\mathsf{T}}\right].$$

Finally, the statistic T_n^2 can be written as

$$T_n^2 = \frac{n}{2}(\overline{\mathbf{X}}^C - \overline{\mathbf{X}}^T)^{\mathsf{T}}\widehat{\Sigma}^{-1}(\overline{\mathbf{X}}^C - \overline{\mathbf{X}}^T).$$

This statistic is also known as Hotelling's two-sample T-squared statistic T^2 when the two groups have the same sample size ($n_T = n_C$) and follows Hotelling's T-squared distribution:

$$T^2 = \frac{n_T n_C}{n_T + n_C}\left(\overline{\mathbf{X}}^C - \overline{\mathbf{X}}^T\right)^{\mathsf{T}}\widehat{\Sigma}^{-1}\left(\overline{\mathbf{X}}^C - \overline{\mathbf{X}}^T\right) \sim T^2(m, n_T + n_C - 2).$$

Note that Williams et al. (2004) showed that

$$T^2 \xrightarrow{L} \chi_m^2.$$

Thus, using a quantile based on the chi-squared distribution or based on Hotelling's T-squared distribution will logically give the same results.