ITALIAN JOURNAL OF PUBLIC HEALTH

Power estimation for multiple co-primary endpoints: a comparison among conservative solutions

Antonio Lucadamo[(1)], Nadia Accoto[(2)], Daniele De Martini[(3)]

ABSTRACT

The problem of estimating the power of the multivariate Intersection Union test (IUT) is studied. Four classical parametric solutions and a bootstrap non-parametric one, providing statistical lower bounds (i.e. one directional confidence intervals) for the power, are considered. The performances of these techniques in several bivariate IUT settings are compared through a simulation study. All solutions are biased, since their actual coverage probabilities are higher than the nominal one. The bootstrap solution shows the smallest bias, and the variability of its estimates is the lowest. Moreover, the bias of the bootstrap solution reduces faster than those of the other techniques when the pilot sample size, or the correlation, or the rate between the two non-centrality parameters increases. Also, the non-parametric bootstrap solution can be improved by calibration, with a considerable bias reduction.

Key words: Sample size estimation; Conservative approach; Bootstrap solution

(1) SEGIS – Università del Sannio
(2) Department of Decision Sciences – Università L. Bocconi – Milan, Italy
(3) DIMEQUANT Department – Università Milano Bicocca – Milan, Italy

CORRESPONDING AUTHOR: Antonio Lucadamo, SEGIS – Università del Sannio, via delle Puglie 82, 82100 Benevento. e-mail: alucadam@unisannio.it. Tel: 0824305763

DOI: ***10.2427/8671***

INTRODUCTION

In the last decade multiple endpoint statistical problems have received increasing attention. On many occasions, indeed, clinical research aims to demonstrate the efficacy of a new drug on more than one endpoint. Related techniques for statistical analysis (1, 2) and for sample size computation (3, 4) have then been developed.

For certain disorders, indeed quite a few, a new treatment is required by regulatory agencies to demonstrate efficacy on multiple co-primary endpoints, all significant at the one-sided 2.5% level. The adequate statistical test for treating this latter problem is the so-called (5, 6) Intersection Union Test (IUT). Berger (7) has proposed the use of IUT for acceptance sampling problems. Recently, many authors provided interesting contributes on IUT and its applications to biomedical statistics. Among others, Chuang-Stein et al. (8) proposed an approach based on the notion of controlling the maximum false positive rate over the restricted null space in order to use a higher significance level to test individual endpoints; Offen et al. (9), who formed a team of experts from the Pharmaceutical Research and Manufacturers of America, provided medical and statistical solutions for multiple co-primary endpoints.

As regards power and sample size computation for the IUT, a conservative approach based on a mathematical sharp lower bound for the power function can be found in Eaton and Muirhead (10); these authors obtained the lower bound induced under no correlation and showed other interesting results. Song (11) provided sample size formulas for IU testing of rate differences in non-inferiority trials. Considering a mathematical lower bound for the power, Yeo and Qu (12) adopted the plug-in pointwise sample size estimation for the IUT , but they did not take the variability of pilot data into account.

In biomedical statistics further experiments are usually planned on the basis of the results of previous studies available in literature. In particular, phase III clinical trials are planned referring to phase I and II results. This kind of sample size computation technique falls under Sample Size Estimation methodology (SSE), which is often adopted in many different applied research contexts (13-15). In these situations, if one forgets to take the variability of pilot data into account, wrong experimental planning may occur. Conservative approaches to sample size estimation (CSSE) have, therefore, been proposed (16-18).

It is worth noting that the core of CSSE is the estimation of the unknown true power of the test, i.e. the power function computed in correspondence to the unknown true value(s) of the parameter(s). One sided confidence intervals, i.e. statistical lower bounds, for the true power are then needed.

In this paper the problem of estimating the true power (or, simply, the power) of the IUT, and consequently its sample size, on the basis of a set of pilot data is considered, accounting for the variability of these latter. The aim is, therefore, to provide tools to compute statistical lower bounds for the power of the IUT. It is anticipated that technical difficulties will be due to the multidimensional nature of the parameter, which is the argument of the power function, and to the bias of IUT.

In this paper, the theoretical framework of IUT, together with its power, are recalled, some different approaches for estimating the power of IUT are introduced. A comparison of the performances of the different techniques is shown in two different sections of the paper, where the best estimation technique is refined through calibration. In the final part of the paper, a numerical example of CSSE for IUT is presented and conclusions are reported. Computational details follow in Appendix A.

THEORETICAL FRAMEWORK OF IU TEST AND POWER

Let $X = (X_1, \ldots, X_\ell)$ be the observations of the ℓ endpoints for an individual receiving the new drug and $Y = (Y_1, \ldots, Y_\ell)$ be those for an individual who received the control drug. Furthermore, assume that $X \sim \mathrm{N}_\ell(\mu_x, \Sigma_x)$ and that $Y \sim \mathrm{N}_\ell(\mu_y, \Sigma_y)$, where $\Sigma_\bullet$ are the covariance matrices.

The correlation coefficients are the off diagonal elements of the matrices $\Sigma_\bullet$, i.e. $\Sigma_{\bullet,ij} = \rho_{\bullet,ij}$ with $i \neq j$. Without loss of generality we can assume that the diagonal elements of $\Sigma_\bullet$ are all equal to 1. Being $\delta = (\delta_1, \ldots, \delta_\ell) = \mu_x - \mu_y$ the vector of the effect sizes, the statistical hypotheses for the non-inferiority multiple test are:

$$\begin{cases} \mathrm{H}_0 : \delta_j \leq 0 & \textit{for at least one } j \\ \mathrm{H}_1 : \delta_j > 0 & \textit{for all } j \end{cases} \qquad [1]$$

In practice, the multivariate null hypothesis is rejected if, and only if, all univariate null hypotheses are rejected.

Now, consider drawing a sample of m individuals from each group, that is X_i and Y_i, $i = 1, \ldots, m$ are i.i.d. with common distribution function $\mathrm{N}_\ell(\mu_x, \Sigma_x)$ and $\mathrm{N}_\ell(\mu_y, \Sigma_y)$ respectively. Then, compute the vector of the test statistics

$$T_m = \sqrt{\frac{m}{2}}\left(\bar{X}_m - \bar{Y}_m\right) = \left(\mathrm{T}_{1,m}, \ldots, \mathrm{T}_{lm}\right)$$

where $\bar{X}_m = \sum_{i=1,m} X_i / m$, and $\bar{Y}_m$ is analogous. Moreover, $\alpha \in (0,1)$ is the type I error probability, $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ and Φ is the cumulative distribution function of the standard Normal law.

In accordance with [1], the so-called Intersection-Union Test (IUT) introduced by Berger (7) is the following:

$$\Psi_m(\mathrm{T}_m) = \begin{cases} 0 & \textit{if } T_{j,m} \leq z_{1-\alpha} \textit{ for at least one } j \\ 1 & \textit{if } T_{j,m} > z_{1-\alpha} \textit{ for all } j = 1, \ldots, \ell \end{cases} \qquad [2]$$

Berger (7) showed that if all the ℓ univariate tests are α-level tests, then the global test too is α-level. In fact, under the null hypothesis the Sup of the power is α and is achieved when $\delta_i = 0$ and δ_j tends to +∞ with $j = 1, \ldots, \ell$ $j \neq i$.

Nevertheless, it is very important to note that the IUT is biased, because under the alternative hypothesis the power can be lower than α. Indeed, when $\delta_j = \varepsilon > 0 \ \forall j$ we fall under H_1 and if ε is "very small" then the power of each univariate test is $\approx \alpha$; consequently, when $\rho_{.,ij} = 0$, the power of the IUT turns out to be $\approx \alpha^\ell$ which is lower than α_ since $\alpha < 1$.

Through simple algebra we obtain that

$T_m = T_1,\ldots,T_\ell \sim \mathrm{N}_\ell\left(\sqrt{m/2}\delta,\Sigma_t\right)$, where $\Sigma_{T,ii} = 1$ and

$\Sigma_{T,ij} = \left(\rho_{x,ij} + \rho_{y,ij}\right)/2, \quad 1 \le i,j \le \ell$.

Then, the power of [2] is

$$E\left[\Psi_m\left(T_m\right)\right] = P\left(T_{1,m} > z_{1-\alpha},\ldots,T_{\ell,m} > z_{1-\alpha}\right) \quad [3]$$

and this can be computed as a function of $\delta,\Sigma_x,\Sigma_y,m\ and\ \alpha$.

We begin studying power estimation techniques for IUT under the simplest non-trivial situation, that is the bivariate case (i.e. $\ell = 2$) with equal dependence structure in the treatment and control groups. This implies $\rho_{x,12} = \rho_{y,12} = \rho$ so that $\Sigma_X = \Sigma_Y = \Sigma$.

Hence, the power function in [3] simplifies to:

$$\pi\left(\delta_1,\delta_2,\rho,m,\alpha\right) = P\left(T_{1,m} > z_{1-\alpha},T_{2,m} > z_{1-\alpha}\right) \quad [4]$$

and this can be computed as a function of the effect sizes $\delta_1\ and\ \delta_2$, of the correlation ρ, of m and α, becoming:

$$\pi\left(\delta_1,\delta_2,\rho,m,\alpha\right) = 1 - \Phi_{\delta_1\sqrt{m/2},1}\left(z_{1-\alpha}\right) - \Phi_{\delta_2\sqrt{m/2},1}\left(z_{1-\alpha}\right) + \Phi_{\delta\sqrt{m/2},\Sigma}\left(z_{1-\alpha},z_{1-\alpha}\right)$$

[5]

Given α, and being $1-\beta$ the power to achieve, the ideal sample size is:

$$M_I = \min\left\{m \mid \pi\left(\delta_1,\delta_2,\rho,m,\alpha\right) > 1-\beta\right\}$$

[6]

Note that ρ plays the role of a nuisance parameter in the power function and, consequently, in sample size computation. In some papers tables showing how M_I varies with different ρs are presented (8, 12), and differences are not negligible.

In practice, $\delta_1,\delta_2\ and\ \rho$ are unknown and so is M_I. If pilot samples are available, then M_I can be estimated and the conservative approach is suggested. So, let us suppose two samples of size n are drawn from the treatment and the control group, respectively, i.e. $X_i, i = 1,\ldots,n$, i.i.d., $X_i \sim N_2\left(\mu_X,\Sigma\right)$, and $Y_i, i = 1,\ldots,n$, i.i.d., $X_i \sim N_2\left(\mu_X,\Sigma\right)$, The challenge is now to estimate $\pi\left(\delta_1,\delta_2,\rho,m,\alpha\right)$ given m and α and, so, indirectly to estimate M_I given $1-\beta$

SOME DIFFERENT APPROACHES FOR ESTIMATING IUT POWER

The parametric approach and related techniques

Being ρ a nuisance parameter in this testing context, the power does not depend primarily on ρ. Consequently, the conservative estimation approach is here applied to the vector $\left(\delta_1,\delta_2\right)$ and its lower bounds are plugged-into the power function for obtaining conservative estimates of the true power. As regards the correlation coefficient, two solutions are considered: the first one consists in plugging-into the power function the pointwise estimate of ρ, say r_n, using the pooled estimator proposed by Donner and Rosner (19); the second one adopts the mathematical lower bound for the power proposed by Eaton and Muirhead (10), which considers $\rho = 0$.

Specifically, considering a confidence region D_n^γ for $\left(\delta_1,\delta_2\right)$ where γ is the amount of conservativeness, i.e. $P\left(\left(\delta_1,\delta_2\right) \in D_n^\gamma\right) = \gamma$ the lower bound of the power is given by $\min_{D_n^\gamma}\left\{\pi\left(\delta_1,\delta_2,\bullet,m,\alpha\right)\right\}$, where $\bullet$ stands for the generic solution adopted for substituting the unknown value of ρ in the power function.

Remark 1. Unusefulness of IUT inversion

The logical direct way for conservatively estimating $\left(\delta_1,\delta_2\right)$ is by inverting the IUT at a level γ (20). In practice, when the point estimate $d_n = \bar{X}_n - \bar{Y}_n = \left(d_{n,1},d_{n,2}\right)$ is observed, the confidence region is given by the points $\left(\bar{\delta}_1,\bar{\delta}_2\right)$ for which the IUT with null hypothesis

$H_0 : \left\{\delta_1 \le \bar{\delta}_1\ or\ \delta_2 \le \bar{\delta}_2\right\}$ is non significant.

This region turns out to be $D_n^\gamma = R^2 - \left\{\left(\delta_1,\delta_2\right) < \left(d_{n,1} - z_\gamma\sqrt{2/n},d_{n,2} - z_\gamma\sqrt{2/n}\right)\right\}$, i.e. the entire plane without an open square in the low-left part. Consequently, we have that $\min_{D_n^\gamma}\left\{\pi\left(\delta_1,\delta_2,\bullet,m,\alpha\right)\right\} = 0$ so that this region is not useful for conservatively estimating the power.

Two different approaches for D_n^γ are here adopted: the first one consists in the classical elliptical confidence region for $\left(\delta_1,\delta_2\right)$ (21); the second is based on two simultaneous lower bounds for $\delta_1\ and\ \delta_2$, according to Anderson (22) and Roy & Bose (23). In the following both are briefly recalled.

The elliptical confidence region

This region is based on the joint distribution of the sample difference mean vector with the sample covariance matrix of the bivariate normal distribution, given by the pooled within-groups covariance estimators, i.e. the distribution of

$\left(d_n,S_n\right)$, where $d_n = \bar{X}_n - \bar{Y}_n \sim \mathrm{N}_2\left(\delta,\frac{2}{n}\Sigma\right)$ and

$$S_n = \left[(n-1)S_{n,1} + (n-1)S_{n,2}\right]/2(n-1)$$

Being $\tau^2 = \frac{n}{2}\left(\bar{X}_n - \bar{Y}_n\right)' S_n^{-1}\left(\bar{X}_n - \bar{Y}_n\right)'$ in consequence of the Hotelling and Wishart distribution properties, it is obtained that $\tau^2 \sim \mathrm{T}_2^2\left(2(n-1);\varsigma^2\right)$ where $\varsigma^2 = \frac{n}{2}\delta'\Sigma^{-1}\delta$ is the non-centrality parameter. If $\delta = 0$, then $\varsigma^2 = 0$ and $\tau^2 \sim \mathrm{T}_2^2\left(2(n-1)\right)$ that is

$$(2n-3)\tau^2/4(n-1) \sim F(2,2n-3).$$

Hence, the boundary of the elliptical region for δ, centered at d_n, with $100(\gamma)$ per cent (approximated) confidence, is given by the following equation: $\frac{n}{2}\left(\delta - d_n\right)' S_n^{-1}\left(\delta - d_n\right) = 4(n-1)F_{2,2n-3}^{-1}(1-\gamma)/(2n-3)$

Note that the confidence region has, in this case, only approximated confidence level, since the diagonal elements of Σ (i.e. the variances σ_i^2) are supposed to be known and equal to 1.

We will refer to the techniques stemming from the elliptical shape of D_n^{γ} as ELLP and ELLM, when ρ is either estimated pointwise or set, according to the mathematical Minoration, equal to zero, respectively.

Remark 2. Inversion of UIT

It can be noted that the elliptical region corresponds to the inversion of the Union Intersection Test (UIT), not the IUT in study. Indeed, this confidence region is given by the points

$\left(\bar{\delta}_1,\bar{\delta}_2\right)$ for which the UIT with null hypothesis

$$H_0:(\delta_1,\delta_2) = \left(\bar{\delta}_1,\bar{\delta}_2\right) \quad \left(\textit{versus } H_1:(\delta_1,\delta_2) \neq \left(\bar{\delta}_1,\bar{\delta}_2\right)\right)$$

is non significant.

The simultaneous bounds region

The simultaneous bounds are obtained through simultaneous one-directional confidence intervals. Recall first that bi-directional simultaneous confidence intervals for the mean of a bivariate Normal distribution are

$$\left(\bar{X}_i - K_{\frac{1-\gamma}{2}}\sqrt{\frac{s_i^2}{n}};\bar{X}_i + K_{\frac{1-\gamma}{2}}\sqrt{\frac{s_i^2}{n}}\right) \qquad i = 1,2$$

where s_i^2 are estimated variance and

$$K_{\frac{1-\gamma}{2}} = \sqrt{2(n-1)F_{2,n-2}^{-1}(1-\gamma)/(n-2)}$$

Then supposing $\sigma_i^2 = 1$ to be known, the pivotal distribution simplifies to a χ_2^2, and the $100(\gamma)$ per cent conservative estimate for the effect size δ is $\left(d_{n,1} - \sqrt{c_{2,2(1-\gamma)}/n}, d_{n,2} - \sqrt{c_{2,2(1-\gamma)}/n}\right)$

where $c_{2,2(1-\gamma)}$ is such that $P\left(\chi_2^2 \le c_{2,2(1-\gamma)}\right) = 2\gamma - 1$

We will refer to these techniques as SIMP and SIMM, when ρ is either estimated pointwise or set equal to zero, respectively.

Parametric computational algorithm

As regards the calculations, $\min_{D_n^{\gamma}}\left\{\pi\left(\delta_1,\delta_2,r_n,m,\alpha\right)\right\}$ is obtained through an algorithm detecting the level curve of the power (say the iso-power curve), which is tangent to the low-left part of the (elliptical or rectangular) confidence region. Note that iso-power curves behave almost like hyperboles (see also Appendix for related computational details). In practice, the problem consists in detecting the curve tangent to the ellipse/open-rectangle centered in $\left(d_{n,1},d_{n,2}\right)$ among the family of "hyperboles". Since we use the elliptical confidence region for computing just a lower bound for $\left(\delta_1,\delta_2\right)$, note also that the real coverage probability of the ellipse is $\gamma_r = \frac{\gamma+1}{2}$ with $\gamma \in [0,1]$.

THE NON-PARAMETRIC APPROACH AND THE BOOTSTRAP TECHNIQUE

Sometimes, quite severe technical difficulties arise within parametric frameworks.

On these occasions, non-parametric methods might be useful to solve parametric problems. As has just been shown, the task of providing bounds for the power of the IUT is, actually, problematic. It is widely known that Efron's bootstrap is a highly versatile non-parametric method. Recently, a non-parametric bootstrap technique for estimating the power of statistical tests (even conservatively) has been presented by De Martini (24), where applications to sample size estimation for the Wilcoxon rank-sum test are shown. We, therefore, adopt this bootstrap technique for conservatively estimating the power of the IUT and we here provide a brief reminder.

Let us denote the bivariate empirical distribution functions of the treatment group and of the control group $F_{T,n}$ and $F_{C,n}$ respectively. Note that these functions contain information both on the shifts $\left(\delta_1,\delta_2\right)$ and on the correlation ρ. Then, when the power is viewed as a functional of the distributions (i.e., $\pi=\pi(N_2(\mu_X,\Sigma),N_2(\mu_Y,\Sigma),m,\alpha)$ the simple bootstrap plug-in estimate of the true power is $\pi(F_{T,n},F_{C,n},m,\alpha)$.

Now, in order to provide a lower bound for the true power, draw two samples of size n from $F_{T,n}$ and $F_{C,n}$ respectively, and let $F^*_{T,n}$ *and* $F^*_{C,n}$ be the empirical distribution functions so obtained. Hence, $\pi(F^*_{T,n},F^*_{C,n},m,\alpha)$ is the bootstrap estimator of the true power. Finally, denote with $\pi^\gamma(F_{T,n},F_{C,n},m,\alpha)$ the $1-\gamma$ p-tile of this latter estimator: this is the (approximated) -lower bound for the true power: $P(\pi^\gamma(F_{T,n},F_{C,n},m,\alpha)\leq\pi(N_2(\mu_X,\Sigma),N_2(\mu_Y,\Sigma),m,\alpha))\simeq\gamma$.

A theoretical justification of this bootstrap solution, which will be denoted by BO, can be found in De Martini (24).

SIMULATION STUDY

In this section we evaluate and compare the performances of the five different estimation techniques for the power of IUT introduced in the previous Section (viz. ELLP, ELLM, SIMP, SIMM and BO).

Design of the study

In order to evaluate the performances of different techniques we vary $\delta_1,k=\delta_1/\delta_2,\rho$ and also the ideal sample size M_I in such a way that the power is always 90%. We consider a small sample situation, with $M_I=60$, and a larger one, with $M_I=180$. For each M_I we consider $k=1,1.5,2$ and $\rho=0.2,0.4,0.5,0.6,0.8$. However, we do not evaluate estimation performances in all the 15 possible cases (i.e. 3 ks $\times 5_\rho$s): we just consider the 8 couples (k,ρ) that are reported in Table 1, together with the corresponding values of δ_1 for each M_I (see Appendix A for computational details).

As regards the size n of pilot samples, recent works on CSSE (17, 18) indicate that, in order to obtain sufficiently accurate power estimates, n should be of the same order of magnitude as M_I. So, for every setting here considered, we evaluate the performances of our techniques with pilot samples of size around M_I. Specifically, we set $n=2M_I/3,4M_I/3$. Hence, the total number of experimental points is 32 (2 M_I s $\times 8\delta_1 s\times 2ns$). In Table 2 the 32 Scenarios so obtained are defined in detail.

For every Scenario we simulate the behavior of our techniques by generating $B_0=5000$ samples from the bivariate normal distributions of the treatment and the control groups. We, thus, obtain 5000 conservative estimates of the power for each one of the 5 techniques, for each conservative level $\gamma=0.5,\ldots,0.99$, with step 0.01. The resulting estimates are evaluated by considering the correctness of conservative levels and the variability of the estimates.

As regards the former point, for every γ the bias is intended to be the difference between the actual coverage probability (ACP) and the nominal one (NCP, viz. γ); hence, the average bias is computed.

As regards the variability of the estimates, the weighted average of the means of the absolute standardized differences between the γ -conservative estimators π^γ and the true power π, namely I2, is adopted in accordance with De Martini (24). In practice, being $D_\gamma=(\pi^\gamma-\pi)/(1-\pi)$ if $\pi^\gamma>\pi$, and $D_\gamma=(\pi-\pi^\gamma)/\pi$ otherwise, we have that I2 is the weighted average of $E[D_\gamma]$ over the set of γ s considered, that is:

$$I2=\sum_{i=0}^{4}w_{.5+.1i}E[D_{.5+.1i}]$$

Since $\pi^{50\%}$ is merely pointwise (and not conservative) and $\gamma=90\%$ may be too severe a conservativeness, in practice the most used conservative levels are around $60-80\%$. Consequently, we used $w_{50\%}=w_{90\%}=0.125$ and $w_{60\%}=w_{70\%}=w_{80\%}=0.25$.

RESULTS

Our five techniques present a clear bias, since they all are, in most Scenarios, too conservative.

As regards parametric techniques, those based on the so-called mathematical lower bounds (viz. ELLM and SIMM) provide results similar to the respective ones obtained with the pointwise estimate of ρ (the former approach is a little more conservative than the latter). Since all parametric techniques were too conservative, estimating ρ pointwise provides less biased (i.e. better) results. Moreover, SIMP lower bounds are less biased than ELLP ones in all Scenarios, leaving SIMP the best parametric performer.

In general, the bias of the non-parametric BO is somewhat lower than that given by SIMP in all settings. The non-parametric BO should, therefore, be preferred.

The bias of all techniques decreases as n, ρ, and k increase. To show these behaviors, we focus on small sample settings (i.e. $M_I = 60$); moreover, Scenario #1 represents the basic setting, whereas Scenarios #2, #5 and #21 show the behavior of power estimation techniques as n , ρ , and k increase, respectively. In particular, for the above settings the ACPs of the different techniques against NCP are plotted in Figures 1-4. The average bias of the five techniques under these Scenarios, together with the rate of improvement, and the $I2$ index are reported in Table 3.

In Scenario #1 the biases of the different techniques are quite similar, around 15% , and the lowest is BO (13.84%). As n increases passing from $2M_I/3 = 40$ to $4M_I/3 = 80$, all techniques improve, albeit marginally (Figures 1-2); nevertheless, the improvement of BO is the highest (12.3%). As ρ passes from 0.2 to 0.8 , the bias of all techniques clearly decreases, and the average bias of BO remains somewhat lower than the others (Figures 1 and 3); once again the improvement of BO is the highest (59.3%). Finally, the highest bias reductions of all five techniques can be observed comparing Figures 1 and 4, i.e. when k passes from 1 to 2: in practice, the bias of BO disappears (average bias 0.34% , i.e. 97.5% improvement).

As far as the variability of the estimates is concerned, under Scenarios 1, 2, 5 and 21 BO presented lower $I2$ s than those of the parametric techniques in all cases but one. Moreover, the values shown by BO are very similar to those observed in estimating the power of the widely used Wilcoxon rank sum test, with the same n s and M_I (24) (Table 2 and 3).

Focusing now on BO (i.e. the best conservative power estimator among those here considered), the most interesting results from the practical point of view are those with $n = 2M_I/3$. In fact, BO performances under the eight Scenarios with $M_I = 60$ are reported in Table 4 (2nd and 6th columns). It can be noted that the highest biases are observed with $k = 1$ and small values of ρ (viz. Scenarios 1 and 3), whereas bias is very small when $k = 2$ (viz. Scenarios 21, 23 and 25). On the contrary, the variability index $I2$ shows small differences among these 8 settings.

When n passes to 80 , the average bias decreases a little, where $I2$ decreases significantly. For example, in Scenarios 4 and 14 (to be compared with Scenarios 3 and 13) the average bias is 8.62% and 2.72% , where the values of $I2$ are 0.3771 and 0.3862 , respectively.

Finally, biases and variabilities observed under large sample settings (i.e. $M_I = 180$) are similar to the corresponding ones with small samples: the bias is a little larger, where $I2$ is a little smaller. For example, Scenarios 7 and 27 (to be compared with Scenarios 1 and 21) provide average biases of 14.13% and 1.43% , and $I2$ values of 0.5081 , 0.5006 , respectively.

DISCUSSION

The clear improvement shown by the techniques increasing ρ or k is due to the fact that the IUT becomes univariate when ρ tends to 1 or when k diverges. In these cases, indeed, the test is unbiased, as are power estimation techniques (BO only approximately). When n diverges power estimators theoretically converge, and in fact the variability index $I2$ decreases; the bias also decreases, just a little.

In parametric techniques, the bias is mainly due to the different shapes of iso-power and confidence region curves. The former are quite similar to hyperboles, whereas the latter (i.e. D_γ) are either elliptical or rectangular. This implies that a certain amount of probability mass lies between D_γ and the tangent iso-power curve. Consequently, the resulting γ -conservative power estimate, i.e. the level of the iso-power curve tangent to D_γ , is more conservative than its actual nominal coverage (i.e. γ).

The cause of BO bias is mainly the biasedness of IUT. In particular, even if the two n sized samples fit $N_2(\mu_X,\Sigma)$ and $N_2(\mu_Y,\Sigma)$, respectively, well, many re-sampled n sized samples can fall close to each other, generating a couple of empirical distribution still under H_1 , but in reality close to H_0 . These re-samples inherit the bias of IUT and carry it into bootstrap estimation. For large M_I s analogous estimation problems do exist.

IMPROVING BOOTSTRAP PERFORMANCES THROUGH CALIBRATION

Calibration is usually adopted for correcting the bias of asymptotic confidence intervals. The NCP (viz. γ) of asymptotic confidence intervals is achieved when the sample size n tends to ∞ . In practice, with finite n s the ACP is different from γ . Nevertheless, there exists a correct coverage, say γ_c , which provides the confidence interval with the desired NCP of γ .

Practical calibration at first makes use of the available sample to estimate γ_c . Once the estimate $\hat{\gamma}_c$ is calculated, it is adopted to compute a confidence interval with nominal coverage $\hat{\gamma}_c$. The ACP of the confidence interval so obtained is, then, closer to γ than that of the simple confidence interval with NCP$=\gamma$.

For an introduction to calibration see Efron and Tibshirani (25). Here, we adopt calibration in the context of IUT power estimation. Moreover, since the bias of BO is lower than those of parametric techniques, we apply calibration to BO.

From the results of the simulation study in the previous Section, it is worth noting that the ACP of BO presents a parabolic shape. So, we evaluate here if a parabolic model for ACP fits the bias well. It is natural to assume that there

FIGURE 1

BIAS OF ACPS OF THE FIVE TECHNIQUES, TOGETHER WITH THAT OF BOC, UNDER SCENARIO 1

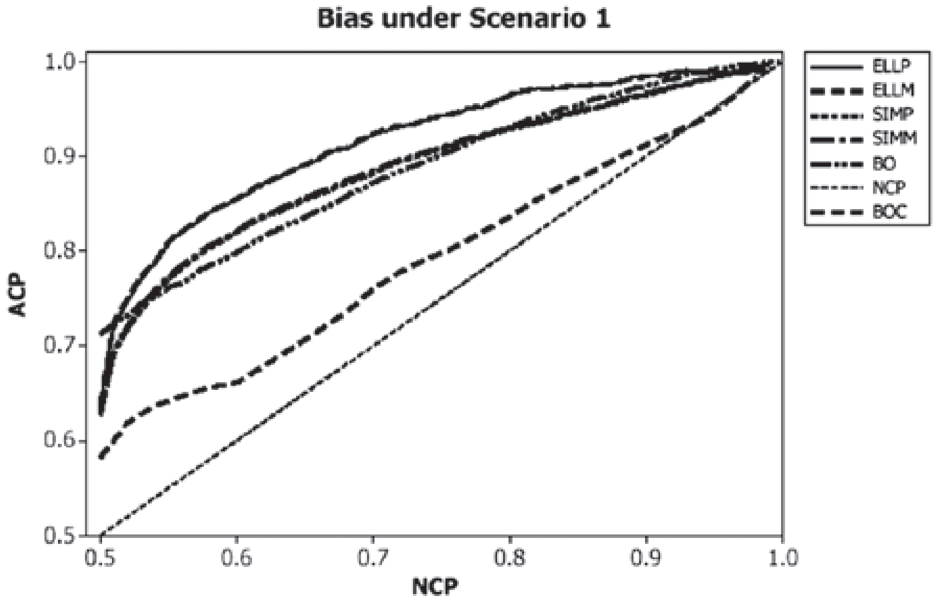
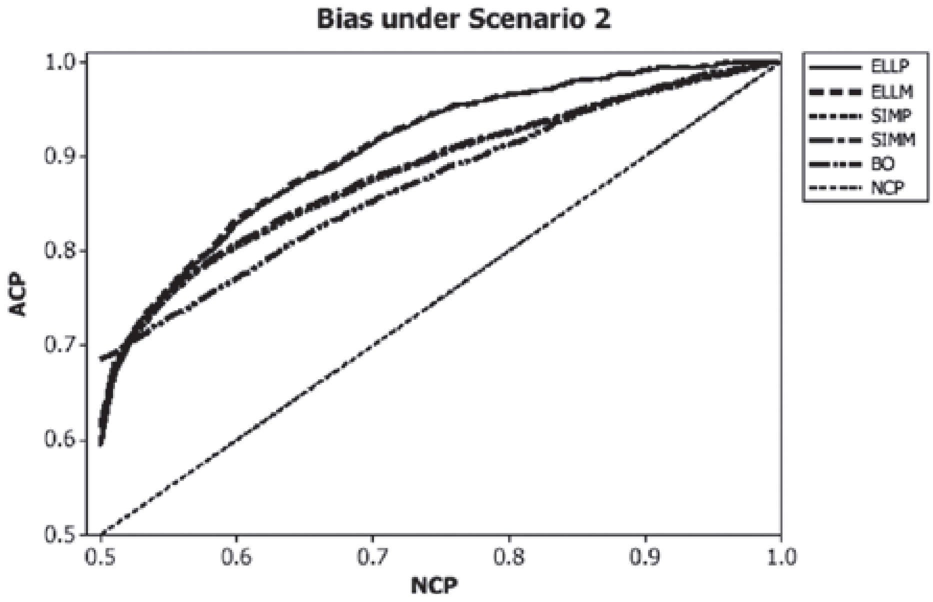


FIGURE 2

BIAS OF ACPS OF THE FIVE TECHNIQUES UNDER SCENARIO 2



is no bias with extreme coverage probabilities, i.e. setting the bias at zero when $NCP = 0$ or 1. The following model for the ACP of BO is, hence, derived:

$$ACP = aNCP^2 + (1-a)NCP \quad [7]$$

The values of a are, then, computed with the classic least squares method for all the 32 scenarios of Section 4 (a subset of a values is reported in Table 4). The model [7] fits the ACP data very well: the correlation between the observed ACP and those provided by the model is, approximately, 100% in all 32 scenarios.

FIGURE 3

BIAS OF ACPS OF THE FIVE TECHNIQUES, TOGETHER WITH THAT OF BOC, UNDER SCENARIO 5

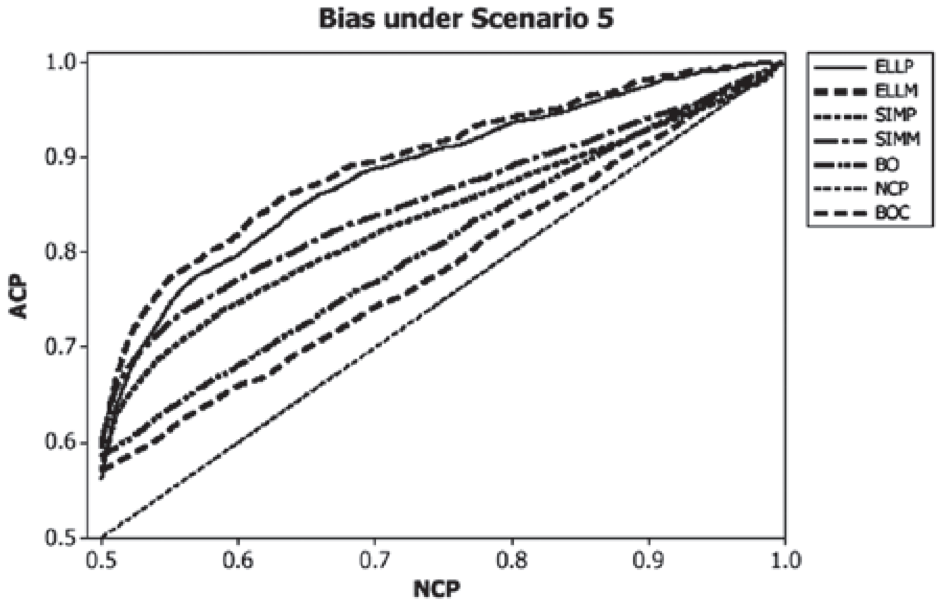


FIGURE 4

BIAS OF ACPS OF THE FIVE TECHNIQUES, TOGETHER WITH THAT OF BOC, UNDER SCENARIO 21

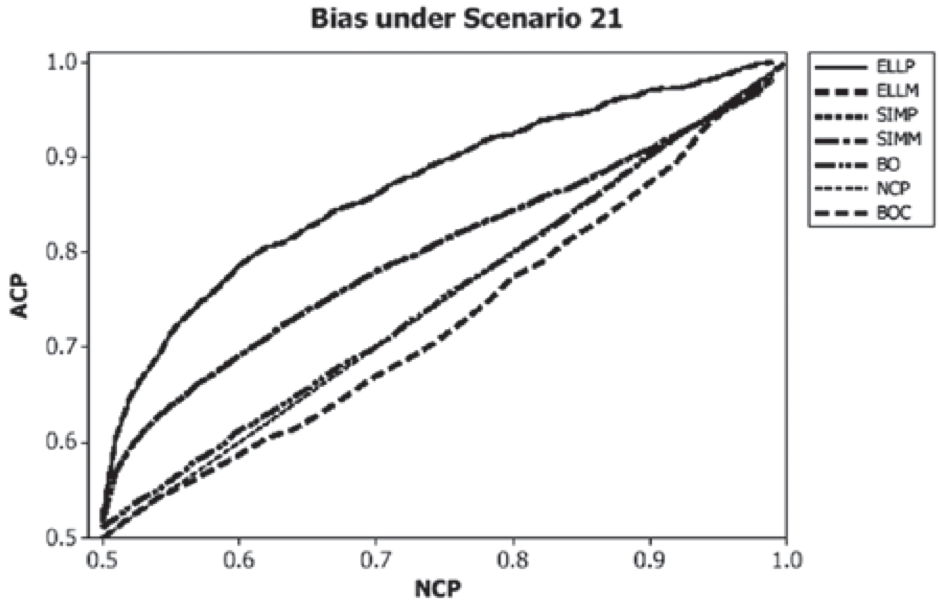


TABLE 1

DESIGN OF THE SIMULATION STUDY AND SHIFT PARAMETERS

	DISTRIBUTIONAL PARAMETERS δ_1S PROVIDING 90% POWER				
$m_1 = 60$	ρ				
	0.2	0.4	0.5	0.6	0.8
k=1	.65315		.64576		.63044
k=1.5		.88916		.88835	
k=2	1.18366		1.18363		1.18363
$m_1 = 180$	ρ				
	0.2	0.4	0.5	0.6	0.8
k=1	.37710		.37283		.36399
k=1.5		.51336		.51289	
k=2	.68337		.68337		.68339

The parabolic behavior of ACP can be exploited in our calibration: γ_c should not be computed separately for every single γ, but it can be provided in a general way by inverting the parabola in [7]. In practice, the bias is at first estimated on the basis of the pilot sample by estimating the parameter a; once $\hat{a}$ is obtained, $\hat{\gamma}_c$ is computed through the inversion of [7] at the given γ (i.e. $\hat{\gamma}_c = (\hat{a} - 1 + \sqrt{(1-\hat{a})^2 + 4\hat{a}\gamma}) / 2\hat{a}$); finally, the γ-conservative estimate of the power is computed by adopting $\hat{\gamma}_c$.

Remark 3. Bias smoothing

It is worth noting that the use of [7] for modelling the bias can be viewed as a kind of bias smoothing, which also allows considerable saving of computational time.

Simulation Study

The aim is to evaluate the performances of this BO calibrated technique (namely BOC). Although computing power estimates with calibration for a single practical case can be completed in a few minutes, to perform a simulation study becomes, computationally, quite heavy. In order to evaluate the improvement given by calibration we, therefore, consider just 8 scenarios among the 32 of the previous Section, i.e. only those with $M_I = 60$ and $n = 40$. The number of power estimates is also decreased to $B_0 = 1000$, and γ varies from 0.5 to 0.98 with a step of 0.02.

The results are most favorable: the global average of the absolute values of mean biases provided by BO (i.e. 4.86%) is reduced by calibration to 2.11%, with an improvement rate higher than 50%. The improvement rate is around 70% for the two highest average biases in particular (viz. those under Scenarios 1 and 3). Detailed results are reported in Table 4. The bias of BOC can also be observed in Figures 1, 3 and 4, where the ACP of BOC is shown.

It should be noted that calibration can sometimes invert the sign of the bias, since it tends to balance the bias itself. Moreover, when BO bias is small that of BOC can be a little higher (viz. Scenarios 21 e 23), but on these occasions calibration is not needed. These biases of BOC may be reduced by increasing B_0 and the Monte Carlo parameters of bootstrap calibration. It should also be remembered that calibration can be iterated to obtain further reductions of the bias (26). Finally, the variability $I2$ index of BOC is substantially equal to that of BO (only slightly smaller).

Hence, calibration improves BO bias significantly, but not estimation variability.

AN EXAMPLE OF CONSERVATIVE SAMPLE SIZE ESTIMATION

The problem of estimating the sample size for a clinical study on sleep disorders is studied. In the phase II trial two groups of $n = 48$ patients are recruited; these same undergo the drug and placebo treatment, respectively. Two clinical parameters concerning the quality of sleeping are recorded before and after the treatment period, and the post-pre differences represent the clinical variables of statistical interest. In practice, $\ell = 2$ and $\mathbf{X}_i$, $\mathbf{Y}_i$, $i = 1,\ldots,48$ are observed.

In order to show the efficacy of the treatment drug the statistical significance of the differences between groups of both variables should be obtained. Consequently, the IUT should be used.

The values of the standardized differences between the means are $d_{48,1} = 0.827$ and $d_{48,2} = 0.553$, and those of the correlation coefficients result $\rho_X = 0.358$ and $\rho_X = 0.396$. Since the research team considers these results to be scientifically relevant, the phase III trial is launched. The sample size can, then, be computed on the basis of preliminary data and the conservative approach is adopted.

In the light of the above results, the bootstrap calibrated technique (BOC) is used. The conservative estimated power curves are shown in Figure 5, where four conservative γ levels are considered, i.e. $\gamma = 50\%$ (viz. pointwise approach), 60%, 70% and 80%. In order to achieve a power of 90% the conservative estimates of the

TABLE 2

ALL SCENARIO SETTINGS

SCENARIO #	*n*	*m*	δ_1	ρ	*k*
1	40	60	0.65315	0.2	1
2	80	60	0.65315	0.2	1
3	40	60	0.64576	0.5	1
4	80	60	0.64576	0.5	1
5	40	60	0.63044	0.8	1
6	80	60	0.63044	0.8	1
7	120	180	0.3771	0.2	1
8	240	180	0.3771	0.2	1
9	120	180	0.37283	0.5	1
10	240	180	0.37283	0.5	1
11	120	180	0.36399	0.8	1
12	240	180	0.36399	0.8	1
13	40	60	0.88916	0.4	1.5
14	80	60	0.88916	0.4	1.5
15	40	60	0.88835	0.6	1.5
16	80	60	0.88835	0.6	1.5
17	120	180	0.51336	0.4	1.5
18	240	180	0.51366	0.4	1.5
19	120	180	0.51289	0.6	1.5
20	240	180	0.51289	0.6	1.5
21	40	60	1.18366	0.2	2
22	80	60	1.18366	0.2	2
23	40	60	1.18363	0.5	2
24	80	60	1.18363	0.5	2
25	40	60	1.18363	0.8	2
26	80	60	1.18363	0.8	2
27	120	180	0.68339	0.2	2
28	240	180	0.68339	0.2	2
29	120	180	0.68339	0.5	2
30	240	180	0.68339	0.5	2
31	120	180	0.68339	0.8	2
32	240	180	0.68339	0.8	2

sample size are: 65, 78, 98 and 129.

Moreover, the simple bootstrap conservative power estimates (BO) are computed, and are also reported in Figure 5. With the same conservative γ levels and power to be achieved, the resulting sample size estimates are: 76, 95, 117 and 161. It can be noted that these estimates are higher than BOC ones, in keeping with simulation results of previous sections.

CONCLUSIONS

The parametric techniques we considered have provided poor performances, both when ρ is estimated pointwise and when it is set, conservatively, equal to 0. In our opinion, instead of studying other estimation or bounding solutions for ρ, it would be better to focus on confidence regions of the same (or similar) shape as the iso-power IUT curves.

TABLE 3

AVERAGE BIASES OF ACP OF THE FIVE TECHNIQUES WITH RESPECT TO NCP AND I2 VALUES UNDER FOUR SCENARIOS WITH m_1 = 60

	AVERAGE BIAS			
SCENARIO	#1	#2	#5	#21
ELLP	16.88%	15.86%	13.64%	12.06%
ELLM	16.93%	16.02%	14.88%	12.07%
SIMP	14.09%	13.40%	8.68%	5.21%
SIMM	14.26%	13.66%	10.47%	5.25%
BO	13.84%	12.13%	5.63%	0.34%
	RATE OF IMPROVEMENT			
		W.R.T. *n*	W.R.T. ρ	W.R.T. *K*
ELLP		6.1%	19.2%	28.5%
ELLM		5.4%	12.1%	28.7%
SIMP		4.9%	38.4%	63.0%
SIMM		4.2%	26.6%	63.2%
BO		12.3%	59.3%	97.5%
	VALUES OF I2 INDEX			
ELLP	0.5906	0.4145	0.5756	0.5743
ELLM	0.5971	0.4224	0.6300	0.5747
SIMP	0.5484	0.3923	0.5209	0.5171
SIMM	0.5593	0.4035	0.5612	0.5178
BO	0.5176	0.3706	0.5252	0.5132

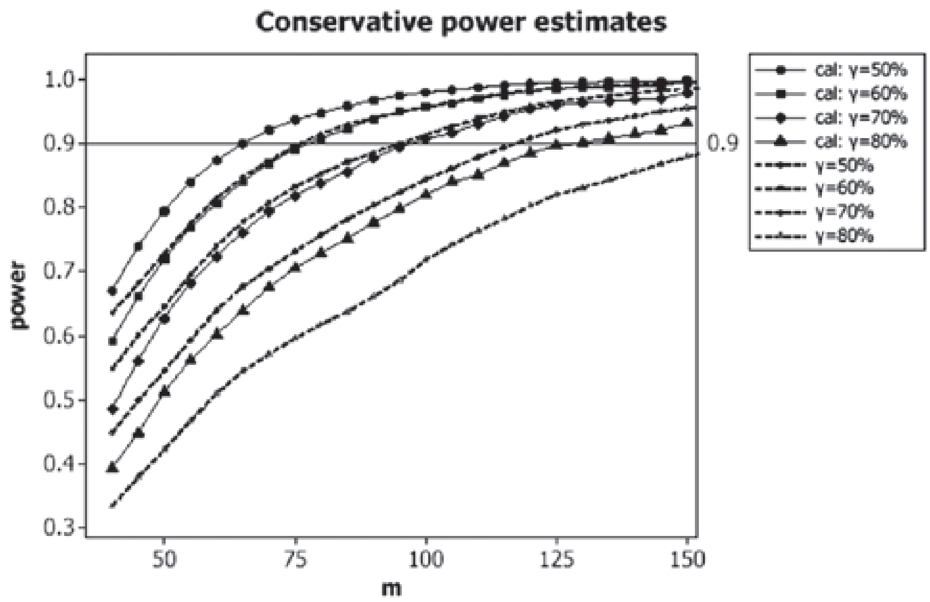
TABLE 4

AVERAGE BIASES OF ACP OF BO WITH RESPECT TO NCP AND *I2* VALUES UNDER THE EIGHT SCENARIOS WITH *n* = 40, IN COMPARISON WITH THOSE OF BOC

	AVERAGE BIASES				*I2 values*	
SCENARIO	BO	(A)	BOC	% OF IMPR.	BO	BOC
1	13.84%	(-0.8690)	4.56%	67.0%	0.5176	0.4772
3	9.79%	(-0.6163)	2.67%	72.8%	0.5207	0.5169
5	5.63%	(-0.3576)	3.69%	34.5%	0.5252	0.5231
13	4.89%	(-0.2743)	-2.53%	48.2%	0.5030	0.5034
15	2.41%	(-0.1442)	-0.27%	88.9%	0.5132	0.5182
21	0.34%	(-0.0277)	-2.10%	-512.7%	0.5132	0.5028
23	-0.69%	(0.0476)	1.04%	-51.0%	0.5114	0.5157
25	-1.24%	(0.0663)	-0.04%	96.5%	0.5178	0.5189

FIGURE 5

CONSERVATIVE POWER ESTIMATES FOR THE IU TEST OBTAINED WITH BOC AND BO, WITH γ =50%, 60%, 70% AND 80%*



*Based on n=48 phase II data of the example in "An example of conservative sample size estimation" section

Although such curves are defined by complicated equations, providing analytical solutions of this kind is an interesting challenge for the future.

Conversely, a general non-parametric solution for power estimation was available, and it can be applied to univariate or multivariate tests. This technique is based on bootstrap, it has already provided satisfactory results when applied to the Wilcoxon rank-sum test, and it can also be useful when applied to complex parametric situations. Here, this bootstrap technique has been applied to the IU test and yielded favorable performances; it presented a certain amount of coverage probability bias merely in some circumstances. Nevertheless, its performances can be improved through calibration, obtaining a considerable bias reduction. Finally, bootstrap power estimation of IUT can be applied in the same way even when different correlations within groups or deviations from normality of data distributions arise.

APPENDIX A: COMPUTATIONAL DETAILS

Computation of effect sizes and of iso-power curves

As shown in section 2 power of the test depends on $\delta_1, \delta_2, \rho, m$ and α. For every value of ρ, M_I and α there are infinite couples (δ_1, δ_2) providing a given power $1-\beta$. For this reason, we built the iso-power curves starting from the couples where $\delta_1 = \delta_2$ and used an algorithm with subsequent approximations which recalls the multivariate normal distribution. The curves so obtained turn out to behave almost like hyperboles and the maximum error with respect to the chosen value of $1-\beta$ is 0.000001. Finally, a bisection method is applied to compute the values of δ_1 and δ_2 for which the ratio $k = \delta_1 / \delta_2 = 1,1.5,2$.

Computation of the bounds for the parametric approaches

In order to build elliptical confidence regions some functions have been implemented in R package (27), but many of them do not correspond to the inversion of the UIT. For this reason, we preferred to build our elliptical confidence region ex-novo. As stated before the boundary of the elliptical region for $\delta = (\delta_1, \delta_2)$ is given by the equation:

$$\frac{n}{2}(\delta - d_n)' S^{-1}(\delta - d_n) = 4(n-1)F^{-1}_{2,2n-3}(1-\gamma)/(2n-3)$$

which can be written as:

$$F^{-1}_{2,2n-3}(1-\gamma) = \frac{n(2n-3)}{8(n-1)}[\delta - d_n]' \mathbf{S}_n^{-1}[\delta - d_n]$$

where $d_n = (d_{n,1}, d_{n,2})$. If we consider $\frac{n(2n-3)}{8(n-1)} = \lambda$ and $[\delta - d_n] = \psi$ we have:

$$F^{-1}_{2,2n-3}(1-\gamma) = \lambda \left(\psi_1 \, \psi_2\right) \begin{bmatrix} S^{-1}_{11} & S^{-1}_{12} \\ S^{-1}_{21} & S^{-1}_{22} \end{bmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}$$

where S^{-1} are the elements of the inverse of the covariance matrix. After appropriate algebraic and matricial operations, we obtained this second degree equation:

$$\lambda S^{-1}_{22}\psi_2^2 + 2\lambda S^{-1}_{12}\psi_1\psi_2 + \lambda S^{-1}_{11}\psi_1^2 - F^{-1}_{2,2n-3}(1-\gamma) = 0$$

whose solution provides the boundary of the ellipse centered in $(d_{n,1}, d_{n,2})$. The power lower bound can be found by looking for the iso-power curve that is tangent to the low-left part of the elliptical confidence region. Then, we considered the point of the ellipse that has the minimum value on the x-axis, we calculated the power at this point and considered the corresponding iso-power curve. If the curve did not intersect the ellipse at other points, then this point represented the lower bound for the power, otherwise we moved on the ellipse and we repeated the operation until we found the tangent curve.

For the simultaneous confidence intervals the computation of the lower bound of the power was obtained considering simply the power estimated through [2] at the point

$$(d_{n,1} - \sqrt{c_{2,2(1-\gamma)} / n}, d_{n,2} - \sqrt{c_{2,2(1-\gamma)} / n}) \ .$$

Bootstrap computational details

The distribution of $\pi(F^*_{T,n}, F^*_{C,n}, m, \alpha)$ was approximated with $B_1 = 500$ points generated with the Monte Carlo technique. Each point was computed on the basis of a couple of samples of size n drawn from $F_{T,n}$ and $F_{C,n}$, whose empirical distribution functions, namely $F^*_{T,n}, F^*_{C,n}$, provided $\pi(F^*_{T,n}, F^*_{C,n}, m, \alpha)$. This latter power value was computed by generating $B_2 = 500$ couples of samples of size m from $F^*_{T,n}$ and $F^*_{C,n}$ which underwent the IUT, and by considering the rate of statistically significant tests.

To implement calibration, we first computed the simple plug-in estimate of the power, i.e. $\pi(F_{T,n}, F_{C,n}, m, \alpha)$. Then, assuming this latter value to be the true power, we generated $B_c = 500$ estimates of the power for each conservative level $\gamma \in (0,1)$ and we computed the ACP. Finally, through least squares formulas, we computed the parameter a in [7] and the related corrected level γ_c for each γ-level of interest was obtained by inverting [7]. Hence, $\pi^{\gamma_c}(F_{T,n}, F_{C,n}, m, \alpha)$ was computed.

ACKNOWLEDGEMENTS: *partial support was provided by the Italian Ministry of Research (MIUR) (protocol 2007 AYHZWC Statistical methods for learning in clinical research).*

References

(1) Sankoh AJ, D'Agostino RB, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Statistics in Medicine 2003; 22: 3133-50

(2) Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics in Medicine 2003; 22: 2387-400

(3) Dilba G, Bretz F, Hothorn LA, Guiard V. Power and sample size computations in simultaneous tests for non-inferiority based on relative margins. Statistics in Medicine 2006; 25: 1131-47

(4) Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics 2007; 6: 161-70

(5) Gleser LJ. On a theory of Intersection-Union Tests, Institute of Mathematical Statistics Bullettin, 1973; 2: 233

(6) Lehmann EL. Testing multiple hypotheses. Annals of Mathematical Statistics 1952; 23: 541-52

(7) Berger RL. Multiparameter hypothesis testing and acceptance sampling. Technometrics 1982; 24: 295-300

(8) Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: A new approach. Statistics in Medicine 2007; 26: 1181-92

(9) Offen W, Chuang-Stein C; Dmitrienko A; et al. Multiple co-primary endpoints: Medical and statistical solutions - A report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. Drug Information Journal 2007; 41(1): 31-46

(10) Eaton ML, Muirhead RJ. On a multiple endpoints testing problem. JSPI 137, 2007; 11: 3416-29

(11) Song JX. Sample size for simultaneous testing of rate differences in non-inferiority trials with multiple endpoints. Computational Statistics and Data Analysis 2009; 53: 1201-7

(12) Yeo A, Qu Y. Evaluation of statistical power for multiple tests: a case study. Pharmaceutical Statistics 2009; 8: 5-11

(13) Johnston MF, Hays RD, Hui KK. Evidence-based effect size estimation: An illustration using the case of acupuncture for cancer-related fatigue BMC Complementary and Alternative Medicine 2009; 9: 1-9

(14) Eng J. Sample size estimation: how many individuals should be studied? Radiology 2003; 227: 309-3

(15) Devane D, Begley CM, Clarke M. How many do I need? Basic principles of sample size estimation. Journal of Advanced Nursing 2004; 47: 297-302

(16) Chuang-Stein C. Sample size and the probability of a successful trial. Pharmaceutical Statistics 2006; 5: 305-9

(17) Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. Pharmaceutical Statistics 2006; 5: 85-97

(18) De Martini D. Adapting by calibration the sample size of a phase III trial on the basis of phase II data. Pharmaceutical Statistics 2011; 10(2): 89-95

(19) Donner A, Rosner B. On inferences concerning a common correlation coefficient. Journal of the Royal Statistical Society 1980; Series C 29: 69-76

(20) Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 1927; 22(158): 209-12

(21) Morrison DF. Multivariate Statistical Methods. New York: Thomson/Brooks/Cole, 2005

(22) Anderson TW. An introduction to multivariate statistical analysis. New York: John Wiley & Sons, 1958

(23) Roy SN, Bose RC. Simultaneous confidence interval estimation. Annals of Mathematical Statistics 1953; 24(4): 513-36

(24) De Martini D. Conservative Sample Size Estimation in Non-parametrics. Journal of Biopharmaceutical Statistics 2011; 21(1): 24-41

(25) Efron B, Tibshirani RJ. An introduction to the Bootstrap. New York: Chapman & Hall, 1993

(26) Hall P, Martin MA. On bootstrap resampling and iteration. Biometrika 1988; 75(4): 661-671

(27) R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2005. URL http://www.R-project.org

*