STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3137–3151 (DOI: 10.1002/sim.1238)

A practical approach for comparing means of two groups without equal variance assumption

Hansheng Wang* and Shein-Chung Chow

StatPlus, Inc., 1790 Yardley-Langhorne Road, Yardley, PA 19067, U.S.A.

SUMMARY

In this paper we consider two-groups of i.i.d. normally distributed random variables ($N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$) without assuming equal variance ($\sigma_x^2 = \sigma_y^2$). We propose a simple method for constructing confidence bounds based on Howe's approximation I. Its applications in parallel clinical trial (testing $H_0 : \mu_x - \mu_y = 0$ versus $H_1 : \mu_x - \mu_y < 0$) and parallel bioequivalence (BE) trial (testing $H_0 : |\mu_x - \mu_y| \geqslant \delta$ versus $H_1 : |\mu_x - \mu_y| < \delta$) are studied. Sample size calculation formulae for both cases are derived. Their performances are evaluated by simulation. Our study shows that the proposed procedure can control type I error satisfactorily compared with Cochran–Cox's and Satterthwaite's approximations while maintaining a relatively high power. The proposed approach is not only simple for constructing the confidence limit, but also provides a simple and accurate formula for sample size calculation.

KEY WORDS: Behrens–Fisher problem; Cochran–Cox's approximation; Howe's approximation I; parallel-group design; parallel bioequivalence; Satterthwaite's approximation

1. INTRODUCTION

Testing the mean difference between two groups of independent and identically distributed (i.i.d.) random variables with unequal variance is well known as the Behrens–Fisher problem. Extensive literatures have been devoted to this topic in the past several decades. Two commonly used methods in practice are the Satterthwaite approximation and Cochran–Cox approximation. Miller [1] gave a comprehensive review of research work done in this area. Some other works include those by Benerjee [2, 3]. Statistical methods available in the literature can be roughly divided into either exact methods or approximation methods. The exact methods can control the significance level exactly using the so-called generalized tests, p-value and confidence intervals (Weerahandi [4]). On the other hand, the approximation methods attempt to control the significance level approximately within an acceptable range of the nominal level. Many exact methods can be viewed as randomized tests (Shao [5]), which have limited clinical application due to their limitation of reproducibility.

*Correspondence to: Hansheng Wang, StatPlus, Inc., 1790 Yardley-Langhorne Road, Yardley, PA 19067, U.S.A.

Received January 2001
Accepted January 2002

Among the approximation methods, Howe [6] proposed two types of approximations, which are referred to as Howe's approximation I and Howe's approximation II. Note that Howe's approximations are nothing but an improved and simplified Cornish–Fisher expansion (Cornish and Fisher [7]). As indicated in Howe [6], his method can be applied to obtain an approximate upper confidence bound for positive linear combination of chi-square random variables. This approach has been generalized to any linear combination of chi-square random variables by Graybill and Wang [8], and Ting *et al.* [9]. Hyslop *et al.* [10] obtained a 95 per cent confidence upper bound for the linearized individual bioequivalence (IBE) criteria by applying this approach to a 2×4 cross-over design. This approach is recommended by the United States Food and Drug Administration (FDA) for establishment of IBE between two drugs (FDA [11]). Chow, Shao and Wang applied the same approach to IBE study under an alternative 2×3 cross-over design and derived an approximate formula for sample size calculation.

For a parallel-group design, it is generally assumed that the data are normally distributed in raw scale or after transformation, such as log-transformation. It is often of interest to provide statistical inference on the mean difference of the two groups. In practice scientists may want to detect a clinically meaningful mean difference, or to establish therapeutic equivalence or bioequivalence between two drugs. In terms of the underlying statistical problem, they can be unified by obtaining a confidence bound for the mean difference. Under the assumption of equal variance, the problem is well solved. Where there are unequal variances, no satisfactory approach is available due to the following reasons: (i) undesirable properties, for example, randomized test; (ii) complexity; (iii) inability for controlling significance level; (iv) insufficient or little power; (v) difficulty for sample size calculation. In this paper we construct a method by applying Howe's approximation I to the linear combination of two normal random variables. We will show that the method is not only simple but also controls the significance level at the nominal level while maintaining relatively high power. In addition, a simple and accurate formula for sample size calculation is derived.

The rest of the article is organized as follows. In Section 2 we study the sensitivity of the naive method with respect to equal variances. In Section 3 we propose a method by applying the idea of Howe's approximation I to a linear combination of two normal random variables. In Section 4 we first give a short review for some existing methods, and then discuss in detail the application of the proposed confidence bound to parallel clinical trial for detecting a clinical meaningful difference. In Section 5 the proposed approach is applied to the problem for assessing bioequivalence between two drugs under a parallel-group design. Sample size calculation formulae are derived in both cases of detecting a meaningful difference and establishing bioequivalence. Simulation studies are described in Section 6 to evaluate the performance of proposed procedure.

2. SENSITIVITY ANALYSIS

For a parallel-group design, let $X_i,\ i = 1, \ldots, n$ be the i.i.d. normal observations $(\mathrm{N}(\mu_x, \sigma_x^2))$ from group 1, and $Y_i,\ i = 1, \ldots, m$ be the i.i.d. normal observations $(\mathrm{N}(\mu_y, \sigma_y^2))$ from group 2. Let $\bar{X}, \bar{Y}, s_X$ and s_Y be sample mean and sample standard deviation for X_i and Y_i, respectively. The naive methods, for example, testing hypothesis and constructing confidence bounds, are carried out by assuming that the following statistic is t-distributed with $n + m - 2$ degrees of

freedom:

$$T = \frac{\sqrt{mn(m+n-2)}(\bar{X} - \bar{Y} - \mu_X + \mu_Y)}{\sqrt{(m+n)((n-1)s_X^2 + (m-1)s_Y^2)}} \tag{1}$$

Hence, a pth naive upper confidence bound can be constructed below.

$$\bar{X} - \bar{Y} + t_{p,n+m-2}\sqrt{\frac{(m+n)((n-1)s_X^2 + (m-1)s_Y^2)}{mn(m+n-2)}} \tag{2}$$

where $t_{p,n+m-2}$ denotes the pth quantile of a standard t-random variable with $n+m-2$ degrees of freedom.

The above upper confidence bound is valid under the assumption of equal variances. However, when the variances are not equal, the test statistic given in (1) is no longer t-distributed. Quantiles of the distribution cannot be found in standard statistical tables. As an alternative, we propose using its asymptotic distribution as an approximation to its exact distribution. In order to study its asymptotic distribution, we assume

$$\lim_{n\to\infty} \frac{m}{n} = \rho$$

The naive method will compare T as given in (1) with the t-distribution with $n + m - 2$ degrees of freedom. As $n \to \infty$ and $m \to \infty$, the quantile of the t-distribution will converge to the quantile of the standard normal distribution. Therefore, if the asymptotic distribution of T is a standard normal, the naive method is asymptotically correct. Consequently, its small sample size performance is expected to be reasonably well. Hence, the naive method is insensitive to the assumption of equal variances. The asymptotic distribution of T can be obtained by

$$\begin{aligned}
T &= \frac{\sqrt{mn(m+n-2)}(\bar{X} - \bar{Y} - \mu_X + \mu_Y)}{\sqrt{(m+n)((n-1)s_X^2 + (m-1)s_Y^2)}} \\
&= \frac{\sqrt{m}(\bar{X} - \bar{Y} - \mu_X + \mu_Y)}{\sqrt{(\frac{m}{n}+1)(\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2})}} \\
&= \frac{\sqrt{m}(\bar{X} - \bar{Y} - \mu_X + \mu_Y)}{\sqrt{(\rho+1)(\frac{\sigma_X^2 + \rho\sigma_Y^2}{1+\rho})}} + o_p(1) \\
&=_d \frac{N(0, \rho\sigma_X^2 + \sigma_Y^2)}{\sqrt{\sigma_X^2 + \rho\sigma_Y^2}} + o_p(1) \\
&=_d N\left(0, \frac{\rho\sigma_X^2 + \sigma_Y^2}{\sigma_X^2 + \rho\sigma_Y^2}\right) + o_p(1)
\end{aligned}$$

As can be seen from the above, the asymptotic distribution of T is a standard normal if and only if

$$\frac{\rho\frac{\sigma_X^2}{\sigma_Y^2} + 1}{\frac{\sigma_X^2}{\sigma_Y^2} + \rho} = 1$$

The above equality holds when (i) $\sigma_X^2 = \sigma_Y^2$ or (ii) $\rho = 1$. When $\sigma_X^2 = \sigma_Y^2$, the 'naive' method is correct. On the other hand, when $m = n$, which implies $\rho = 1$, T may not exactly follow a t-distribution. However, the naive method is still asymptotically correct. As a result, we may expect that it also provides a reasonably good approximation for small sample size. Our simulation result in Section 6 confirms this observation.

3. HOWE'S APPROXIMATION I

Suppose there are two independent random variables $X(E(X) = \mu_X)$ and $Y(E(Y) = \mu_y)$. For a given confidence level α, let X'_α and Y'_α be the other two independent random variables such that

$$P(\mu_X \leqslant X'_\alpha) = P(\mu_Y \leqslant Y'_\alpha) = \alpha$$

By Howe's approximation I, an approximate αth upper confidence bound for $\mu_X + \mu_Y$ is given by

$$X + Y + \text{sign}(V)|V|^{1/2} \tag{3}$$

where

$$V = (X'_\alpha - X)|X'_\alpha - X| + (Y'_\alpha - Y)|Y'_\alpha - Y|$$

For a parallel-group design, we consider constructing an approximate upper confidence limit based on $\bar{X}$, $\bar{Y}$, s_X and s_Y. For a given α, the αth upper confidence bound for μ_X and $-\mu_Y$ are given by

$$X'_\alpha = \bar{X} + t_{\alpha,n-1}\frac{s_X}{\sqrt{n}} \quad \text{and} \quad Y'_\alpha = -\bar{Y} + t_{\alpha,m-1}\frac{s_Y}{\sqrt{m}}$$

Therefore, the V in (3) can be obtained by

$$\begin{aligned} V &= (X'_\alpha - \bar{X})|X'_\alpha - \bar{X}| + (Y'_\alpha + \bar{Y})|Y'_\alpha + \bar{Y}| \\ &= t_{\alpha,n-1}\frac{s_X}{\sqrt{n}}\left|t_{\alpha,n-1}\frac{s_X}{\sqrt{n}}\right| + t_{\alpha,m-1}\frac{s_Y}{\sqrt{m}}\left|t_{\alpha,m-1}\frac{s_Y}{\sqrt{m}}\right| \\ &= (-1)^{I\{\alpha<0.5\}}\left(t_{\alpha,n-1}^2\frac{s_X^2}{n} + t_{\alpha,m-1}^2\frac{s_Y^2}{m}\right) \end{aligned}$$

The last equation holds by the fact that $t_{\alpha,n-1} < 0$ and $t_{\alpha,m-1} < 0$ if and only if $\alpha < 0.5$. Therefore, applying equation (3), an approximate αth upper confidence bound is given by

$$U_\alpha = \bar{X} - \bar{Y} + (-1)^{I\{\alpha<0.5\}}\sqrt{t_{\alpha,n-1}^2\frac{s_X^2}{n} + t_{\alpha,m-1}^2\frac{s_Y^2}{m}} \tag{4}$$

Similar results were also obtained by McCullough *et al.* [12]. However, the methodology we used in deriving (4) is different. McCullough *et al.* considered the following type of upper

bound for $\mu_X - \mu_Y$:

$$\bar{X} - \bar{Y} + (-1)^{I\{\alpha<0.5\}}\sqrt{r_1^2\frac{s_X^2}{n} + r_2^2\frac{s_Y^2}{m}}$$

The constants r_1 and r_2 are chosen such that the upper bound is exact when $\sigma_X^2 = 0$ or $\sigma_Y^2 = 0$.

4. PARALLEL CLINICAL TRIAL

In this section we first give a brief review for two commonly used methods for testing mean difference with unequal variances for the purpose of comparison. They are, Cochran–Cox's approximation and Satterthwaite's approximation. After that we will discuss in detail how to apply Howe's approximation I to parallel clinical trials with sample size calculation formula.

4.1. Cochran–Cox's approximation

According to Cochran–Cox's approximation, at a given level of significance α, the null hypothesis of $H_0{:}\mu_X = \mu_Y$ will be rejected if

$$\frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/n + s_Y^2/m}} > \frac{t_{1-\alpha,n-1}s_X^2/n + t_{1-\alpha,m-1}s_Y^2/m}{s_X^2/n + s_Y^2/m} \tag{5}$$

4.2. Satterthwaite's approximation

According to Satterthwaite's approximation, at a given level of significance α, the null hypothesis of $H_0{:}\mu_X = \mu_Y$ will be rejected if

$$\bar{X} - \bar{Y} + t_{r,\alpha}\sqrt{s_X^2/n + s_Y^2/m} < 0 \tag{6}$$

where the approximate degrees of freedom r is given by

$$r = \frac{(s_X^2/n + s_Y^2/m)^2}{s_X^4/(n^2(n-1)) + s_Y^2/(m^2(m-1))}$$

Typically r is not an integer. As a result, it is a common practice to round r down to the nearest integer to be somewhat conservative.

4.3. Howe's approximation

For a parallel clinical trial, the hypothesis of interest could be one-sided or two-sided. For the purpose of illustration, we consider only one-sided tests as an example to show the proposed methodology for testing hypothesis and the derivation of a simple formula for sample size calculation. It should be noted that the proposed method can be easily generalized to the situation where a two-sided hypothesis is tested. Consider the following one-sided hypotheses

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y < 0$$

For a given significance level $\alpha < 0.5$, a $(1-\alpha)$th upper confidence bound (U) can be constructed according to (4)

$$U = \bar{X} - \bar{Y} + \sqrt{t^2_{1-\alpha,n-1}\frac{s_X^2}{n} + t^2_{1-\alpha,m-1}\frac{s_Y^2}{m}} \tag{7}$$

We reject the null hypothesis if and only if $U < 0$. Under the null hypothesis, the proposed test has approximately the significance level (type I error) of α.

Under the alternative hypothesis, for example, $\mu_X - \mu_Y = \Delta < 0$, the sample size required for achieving a desired power $(1-\beta)$ can be calculated based on (7). Assuming $m = kn$, for some integer $k > 0$, the upper bound in (7) becomes

$$\begin{aligned} U &= \bar{X} - \bar{Y} + \sqrt{t^2_{1-\alpha,n-1}\frac{s_X^2}{n} + t^2_{1-\alpha,kn-1}\frac{s_Y^2}{kn}} \\ &\approx \bar{X} - \bar{Y} + \frac{z_{1-\alpha}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k} \\ &\approx \Delta + (\bar{X} - \mu_X) - (\bar{Y} - \mu_Y) + \frac{z_{1-\alpha}}{\sqrt{n}}\sqrt{\sigma_X^2 + \sigma_Y^2/k} \end{aligned}$$

Applying (4) on $(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)$, we can obtain an approximate $(1-\beta)$th upper bound, which is given by

$$\sqrt{t^2_{1-\beta,n-1}\frac{s_X^2}{n} + t^2_{1-\beta,kn-1}\frac{s_Y^2}{kn}} \approx \frac{z_{1-\beta}}{\sqrt{n}}\sqrt{\sigma_X^2 + \sigma_Y^2/k}$$

Thus, we have

$$P\left(U < \Delta + \frac{z_{1-\alpha}}{\sqrt{n}}\sqrt{\sigma_X^2 + \sigma_Y^2/k} + \frac{z_{1-\beta}}{\sqrt{n}}\sqrt{\sigma_X^2 + \sigma_Y^2/k}\right) \approx 1 - \beta$$

Consequently, in the interest of having a $(1-\beta)$ power for rejecting the null hypothesis, it is necessary and sufficient to have

$$\Delta + \frac{z_{1-\alpha} + z_{1-\beta}}{\sqrt{n}}\sqrt{\sigma_X^2 + \sigma_Y^2/k} \leqslant 0$$

As a result, the sample size is given by

$$n \geqslant \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\Delta^2}(\sigma_X^2 + \sigma_Y^2/k) \tag{8}$$

5. PARALLEL BIOEQUIVALENCE TRIAL

For assessment of bioequivalence, pharmacokinetic responses such as area under the plasma or blood concentration-time curve (AUC) and peak concentration ($C_{\max}$) are assumed to be normally distributed after log-transformation. The 1992 FDA Guidance indicates that two drugs (for example, a test and a reference product) are said to be average bioequivalent (ABE) if the

mean difference after log-transformation lies in $(-0.223, 0.223)$. In practice, ABE is established by constructing an approximate 90 per cent confidence interval for the mean difference. If the confidence interval falls entirely within $(-0.223, 0.223)$, ABE is concluded. In practice, BE studies are usually carried out by cross-over design, which has the merit of having individual subjects serving as their own controls and consequently removing the intersubject variability from the comparison (Chow and Wang [13]). However, in some pharmacokinetic studies, the drug under study may have a relatively long half-life or the dosing regimens require a long time to complete. In these cases it is highly impractical for a cross-over design in view of the long washout period between does administration and the inevitable high drop-out rate in such a design. As an alternative, parallel design is preferred. A detailed discussion regarding statistical analysis and sample size calculation for parallel design in BE trials under the assumption of equal variances can be found in Chow and Wang [13]. In this section, statistical analysis and sample size calculation for parallel BE trials under the assumption of unequal variances will be explored.

In order to obtain a 90 per cent confidence interval for $\mu_X - \mu_Y$, it is equivalent to obtaining a 5 per cent upper confidence bound ($U_{0.05}$) and a 95 per cent upper confidence bound ($U_{0.95}$). Then, $(U_{0.05}, U_{0.95})$ will be a 90 per cent confidence interval for $\mu_X - \mu_Y$. By applying (4), $U_{0.05}$ and $U_{0.95}$ can be obtained by

$$\begin{aligned} U_{0.05} &= \bar{X} - \bar{Y} - \sqrt{t^2_{0.05,n-1}\frac{s_X^2}{n} + t^2_{0.05,m-1}\frac{s_Y^2}{m}} \\ U_{0.95} &= \bar{X} - \bar{Y} + \sqrt{t^2_{0.95,n-1}\frac{s_X^2}{n} + t^2_{0.95,m-1}\frac{s_Y^2}{m}} \\ &= \bar{X} - \bar{Y} + \sqrt{t^2_{0.05,n-1}\frac{s_X^2}{n} + t^2_{0.05,m-1}\frac{s_Y^2}{m}} \end{aligned} \tag{9}$$

Therefore, the ABE can be concluded if and only if $(U_{0.05}, U_{0.95}) \subset (-0.223, 0.223)$. It is equivalent to

$$\sqrt{t^2_{0.95,n-1}\frac{s_X^2}{n} + t^2_{0.95,m-1}\frac{s_Y^2}{m}} - 0.223 < \bar{X} - \bar{Y} < 0.223 - \sqrt{t^2_{0.95,n-1}\frac{s_X^2}{n} + t^2_{0.95,m-1}\frac{s_Y^2}{m}} \tag{10}$$

Assume $m = kn$ for some integer $k > 0$, it is desirable to have a sample size such that (10) holds with probability $(1 - \beta)$, that is

$$1 - \beta \approx P\left(\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k} - 0.223 < \bar{X} - \bar{Y} < 0.223 - \frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k}\right)$$

Let $\Delta = \mu_X - \mu_Y$. If $\Delta = 0$, it follows that

$$\begin{aligned} &P\left(\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k} - 0.223 < \bar{X} - \bar{Y} < 0.223 - \frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k}\right) \\ &= 1 - 2P\left(\bar{X} - \bar{Y} > 0.223 - \frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2 + s_Y^2/k}\right) \approx 1 - \beta \end{aligned}$$

It implies that

$$P\left(\bar{X}-\bar{Y}\leqslant 0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)$$
$$=P\left(\bar{X}-\bar{Y}+\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}<0.223\right)\approx 1-\beta/2$$

Similarly, by applying (4) to $\bar{X}-\bar{Y}$, it follows that

$$P\left(\bar{X}-\bar{Y}+\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\leqslant\frac{z_{1-\beta/2}+z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)\approx 1-\beta/2$$

This implies that the sample size n should satisfy

$$\frac{z_{1-\beta/2}+z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\approx\frac{z_{1-\beta/2}+z_{0.95}}{\sqrt{n}}\sqrt{\sigma_X^2+\sigma_Y^2/k}\leqslant 0.223$$

Consequently, n can be estimated by

$$n\geqslant\frac{(z_{1-\beta/2}+z_{0.95})^2(\sigma_X^2+\sigma_Y^2/k)}{0.223^2} \tag{11}$$

If $\Delta\neq 0$, without loss of generality, assume $0<\Delta<0.223$. It follows that

$$P\left(\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}-0.223<\bar{X}-\bar{Y}<0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)$$
$$=1-P\left(\bar{X}-\bar{Y}>0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)$$
$$-P\left(\bar{X}-\bar{Y}<\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}-0.223\right)\approx 1-\beta$$

Since $\Delta>0$

$$P\left(\bar{X}-\bar{Y}>0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)\gg P\left(\bar{X}-\bar{Y}<\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}-0.223\right)$$

where $\gg$ means 'much larger than'. Therefore, the power can be further approximated by (see Chow and Liu [14])

$$1-P\left(\bar{X}-\bar{Y}>0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)\approx 1-\beta$$

It implies that

$$P\left(\bar{X}-\bar{Y}\leqslant 0.223-\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)$$
$$=P\left(\Delta+(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)+\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\leqslant 0.223\right)\approx 1-\beta$$

Similarly, by applying (4) to $(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)$, it follows that

$$1-\beta \approx P\left(\Delta+(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)+\frac{z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right.$$
$$\left.\leqslant \Delta+\frac{z_{1-\beta}+z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\right)$$

This implies that the sample size n should satisfy

$$\frac{z_{1-\beta}+z_{0.95}}{\sqrt{n}}\sqrt{s_X^2+s_Y^2/k}\approx\frac{z_{1-\beta}+z_{0.95}}{\sqrt{n}}\sqrt{\sigma_X^2+\sigma_Y^2/k}\leqslant 0.223-\Delta$$

Consequently, n can be estimated by

$$n\geqslant\frac{(z_{1-\beta}+z_{0.95})^2(\sigma_X^2+\sigma_Y^2/k)}{(0.223-|\Delta|)^2} \tag{12}$$

Finally, it can be noted from the above proof that formula (12) can also be used when $\Delta<0$.

6. SIMULATION STUDY

6.1. Size and sensitivity

A simulation study was performed to evaluate the performance of (4) and the sensitivity of the equal variance assumption. The simulation was conducted using S-plus. Since the proposed upper bound (4) is invariant for μ_X and μ_Y, that is, the coverage probability does not depend on μ_X and μ_Y, in the simulation we set $\mu_X=\mu_Y\equiv 0$. For each parameter setting (σ_X, σ_Y, m, and n), 10 000 i.i.d. random samples were generated from $\mathrm{N}(\mu_X,\sigma_X^2/n)$, $\mathrm{N}(\mu_Y,\sigma_Y^2/k/n)$, $\sigma_X^2\chi^2(n-1)/(n-1)$, and $\sigma_Y^2/k\chi^2(m-1)/(m-1)$, respectively. They are used to simulate $\bar{X}, \bar{Y}$, s_X^2 and s_Y^2, respectively. For each of the 10 000 iterations, the confidence bounds were calculated based on (2) (the naive method) and also (4) (the proposed method). The coverage probability was calculated as the proportion of the cases that $\mu_X-\mu_Y$ is covered by the upper bounds. The results are summarized in Tables I to III. Table I summarizes the results when $\sigma_X^2=\sigma_Y^2$. The results show that both methods perform well, while the naive method is better than the proposed approximation, because in this case the naive method is exact. Table II summarizes the results when $n=m$. In this case, both methods perform reasonably well. This is because the naive method is asymptotically correct. Table III summarizes the results when $\sigma_X^2\neq\sigma_Y^2$ and $n\neq m$. In this case, the proposed method works well but the naive method becomes worse. The poor performance of the naive method may be explained in part by the fact that it is not asymptotically correct.

A similar simulation study was also carried out to compare the confidence interval length of the naive method and our method when $\sigma_X=\sigma_Y$. The results are summarized in Table IV. As it can be seen, the confidence interval length of both methods are comparable while the confidence interval length of the naive method is relatively shorter.

Table I. Size and sensitivity study results when $\sigma_X^2 = \sigma_Y^2$.

Parameter				$\beta = 0.05$		$\beta = 0.95$		Parameter				$\beta = 0.05$		$\beta = 0.95$	
n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N	n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N
5	5	0.25	0.25	0.031	0.050	0.965	0.951	5	5	0.75	0.75	0.032	0.047	0.966	0.950
5	10	0.25	0.25	0.039	0.047	0.958	0.953	5	10	0.75	0.75	0.038	0.052	0.959	0.952
5	15	0.25	0.25	0.036	0.048	0.961	0.951	5	15	0.75	0.75	0.042	0.050	0.957	0.954
5	20	0.25	0.25	0.040	0.049	0.956	0.951	5	20	0.75	0.75	0.044	0.048	0.955	0.945
10	10	0.25	0.25	0.040	0.051	0.962	0.946	10	10	0.75	0.75	0.040	0.051	0.959	0.950
10	15	0.25	0.25	0.041	0.049	0.962	0.947	10	15	0.75	0.75	0.042	0.054	0.958	0.951
10	20	0.25	0.25	0.048	0.049	0.955	0.951	10	20	0.75	0.75	0.045	0.050	0.957	0.944
15	15	0.25	0.25	0.044	0.050	0.957	0.946	15	15	0.75	0.75	0.046	0.052	0.953	0.950
15	20	0.25	0.25	0.044	0.049	0.954	0.952	15	20	0.75	0.75	0.046	0.046	0.956	0.950
20	20	0.25	0.25	0.047	0.053	0.954	0.949	20	20	0.75	0.75	0.043	0.050	0.952	0.951
5	5	0.50	0.50	0.033	0.052	0.968	0.948	5	5	1.00	1.00	0.033	0.052	0.970	0.950
5	10	0.50	0.50	0.037	0.050	0.962	0.951	5	10	1.00	1.00	0.042	0.049	0.964	0.948
5	15	0.50	0.50	0.043	0.046	0.962	0.949	5	15	1.00	1.00	0.042	0.049	0.958	0.949
5	20	0.50	0.50	0.043	0.048	0.959	0.951	5	20	1.00	1.00	0.044	0.049	0.958	0.952
10	10	0.50	0.50	0.041	0.053	0.957	0.949	10	10	1.00	1.00	0.042	0.050	0.959	0.952
10	15	0.50	0.50	0.041	0.051	0.956	0.948	10	15	1.00	1.00	0.045	0.048	0.956	0.947
10	20	0.50	0.50	0.044	0.051	0.956	0.948	10	20	1.00	1.00	0.046	0.051	0.956	0.948
15	15	0.50	0.50	0.046	0.052	0.957	0.948	15	15	1.00	1.00	0.043	0.053	0.956	0.948
15	20	0.50	0.50	0.046	0.050	0.955	0.951	15	20	1.00	1.00	0.048	0.052	0.955	0.946
20	20	0.50	0.50	0.046	0.048	0.954	0.948	20	20	1.00	1.00	0.046	0.052	0.952	0.947

Note that p_{H} represents the estimated coverage probability based on the proposed approximation method, while p_{N} is the estimated coverage probability based on the naive method.

Table II. Size and sensitivity study results when $n = m$.

Parameter				$\beta = 0.05$		$\beta = 0.95$		Parameter				$\beta = 0.05$		$\beta = 0.95$	
n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N	n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N
5	5	0.25	0.25	0.031	0.050	0.965	0.951	15	15	0.25	0.25	0.048	0.052	0.954	0.950
5	5	0.25	0.50	0.038	0.055	0.964	0.942	15	15	0.25	0.50	0.046	0.052	0.956	0.946
5	5	0.25	0.75	0.043	0.062	0.959	0.941	15	15	0.25	0.75	0.049	0.054	0.950	0.947
5	5	0.25	1.00	0.046	0.064	0.954	0.938	15	15	0.25	1.00	0.047	0.054	0.954	0.944
5	5	0.50	0.50	0.032	0.046	0.971	0.950	15	15	0.50	0.50	0.043	0.050	0.953	0.952
5	5	0.50	0.75	0.037	0.048	0.964	0.949	15	15	0.50	0.75	0.046	0.048	0.953	0.950
5	5	0.50	1.00	0.038	0.055	0.963	0.946	15	15	0.50	1.00	0.045	0.053	0.956	0.951
5	5	0.75	0.75	0.031	0.048	0.968	0.953	15	15	0.75	0.75	0.043	0.047	0.955	0.951
5	5	0.75	1.00	0.036	0.054	0.968	0.948	15	15	0.75	1.00	0.046	0.049	0.953	0.946
5	5	1.00	1.00	0.034	0.051	0.967	0.953	15	15	1.00	1.00	0.045	0.049	0.955	0.951
10	10	0.25	0.25	0.040	0.050	0.959	0.950	20	20	0.25	0.25	0.048	0.051	0.952	0.953
10	10	0.25	0.50	0.046	0.050	0.955	0.947	20	20	0.25	0.50	0.050	0.052	0.950	0.952
10	10	0.25	0.75	0.048	0.057	0.954	0.942	20	20	0.25	0.75	0.046	0.049	0.954	0.951
10	10	0.25	1.00	0.045	0.058	0.949	0.944	20	20	0.25	1.00	0.048	0.051	0.948	0.948
10	10	0.50	0.50	0.042	0.052	0.955	0.946	20	20	0.50	0.50	0.048	0.049	0.956	0.950
10	10	0.50	0.75	0.044	0.051	0.956	0.945	20	20	0.50	0.75	0.048	0.048	0.954	0.950
10	10	0.50	1.00	0.045	0.052	0.950	0.949	20	20	0.50	1.00	0.046	0.052	0.955	0.948
10	10	0.75	0.75	0.042	0.048	0.958	0.951	20	20	0.75	0.75	0.047	0.050	0.956	0.950
10	10	0.75	1.00	0.042	0.049	0.957	0.950	20	20	0.75	1.00	0.049	0.054	0.955	0.950
10	10	1.00	1.00	0.042	0.049	0.956	0.952	20	20	1.00	1.00	0.045	0.055	0.956	0.949

Note that p_{H} represents the estimated coverage probability based on the proposed approximation method, while p_{N} is the estimated coverage probability based on the naive method.

Table III. Size and sensitivity study results when $\sigma_X \neq \sigma_Y$ and $n \neq m$.

Parameter				$\beta = 0.05$		$\beta = 0.95$		Parameter				$\beta = 0.05$		$\beta = 0.95$	
n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N	n	m	σ_X	σ_Y	p_H	p_N	p_H	p_N
5	10	0.25	0.50	0.037	0.026	0.964	0.975	10	15	0.50	0.75	0.048	0.038	0.955	0.964
5	10	0.25	0.75	0.042	0.020	0.956	0.978	10	15	0.75	0.75	0.047	0.049	0.957	0.952
5	10	0.50	0.75	0.037	0.033	0.962	0.969	10	20	0.25	0.50	0.046	0.022	0.953	0.973
5	10	0.75	0.75	0.042	0.052	0.960	0.951	10	20	0.25	0.75	0.046	0.017	0.954	0.982
5	15	0.25	0.50	0.038	0.014	0.961	0.984	10	20	0.50	0.75	0.043	0.033	0.957	0.970
5	15	0.25	0.75	0.036	0.009	0.962	0.992	10	20	0.75	0.75	0.045	0.052	0.959	0.950
5	15	0.50	0.75	0.035	0.027	0.963	0.976	15	20	0.25	0.50	0.048	0.037	0.951	0.959
5	15	0.75	0.75	0.041	0.052	0.961	0.952	15	20	0.25	0.75	0.046	0.037	0.952	0.964
5	20	0.25	0.50	0.039	0.012	0.960	0.991	15	20	0.50	0.75	0.045	0.045	0.953	0.957
5	20	0.25	0.75	0.042	0.005	0.960	0.996	15	20	0.75	0.75	0.044	0.048	0.956	0.950
5	20	0.50	0.75	0.041	0.018	0.963	0.980	20	20	0.25	0.50	0.046	0.051	0.953	0.948
5	20	0.75	0.75	0.040	0.052	0.958	0.953	20	20	0.25	0.75	0.050	0.052	0.951	0.948
10	15	0.25	0.50	0.044	0.034	0.956	0.966	20	20	0.50	0.75	0.047	0.053	0.954	0.950
10	15	0.25	0.75	0.046	0.029	0.951	0.967	20	20	0.75	0.75	0.050	0.051	0.956	0.951

Note that p_{H} represents the estimated coverage probability based on the proposed approximation method, while p_{N} is the estimated coverage probability based on the naive method.

Table IV. Length of confidence intervals when $\sigma_X^2 = \sigma_Y^2$.

n	m	σ_X	σ_Y	H80	N80	H90	N90	n	m	σ_X	σ_Y	H80	N80	H90	N90
5	5	0.25	0.25	0.471	0.428	0.654	0.572	5	5	0.75	0.75	1.412	1.286	1.964	1.700
5	10	0.25	0.25	0.392	0.363	0.540	0.476	5	10	0.75	0.75	1.181	1.081	1.618	1.424
5	15	0.25	0.25	0.370	0.338	0.507	0.441	5	15	0.75	0.75	1.112	1.019	1.531	1.325
5	20	0.25	0.25	0.358	0.326	0.494	0.424	5	20	0.75	0.75	1.073	0.980	1.485	1.271
10	10	0.25	0.25	0.305	0.293	0.404	0.382	10	10	0.75	0.75	0.917	0.880	1.215	1.145
10	15	0.25	0.25	0.276	0.267	0.364	0.346	10	15	0.75	0.75	0.827	0.800	1.092	1.036
10	20	0.25	0.25	0.260	0.253	0.343	0.326	10	20	0.75	0.75	0.781	0.757	1.032	0.980
15	15	0.25	0.25	0.244	0.238	0.318	0.308	15	15	0.75	0.75	0.732	0.713	0.955	0.923
15	20	0.25	0.25	0.227	0.222	0.296	0.287	15	20	0.75	0.75	0.679	0.666	0.889	0.860
20	20	0.25	0.25	0.208	0.205	0.272	0.265	20	20	0.75	0.75	0.626	0.614	0.814	0.793
5	5	0.50	0.50	0.938	0.854	1.308	1.136	5	5	1.00	1.00	1.879	1.714	2.620	2.276
5	10	0.50	0.50	0.786	0.725	1.082	0.950	5	10	1.00	1.00	1.571	1.450	2.158	1.907
5	15	0.50	0.50	0.738	0.676	1.014	0.882	5	15	1.00	1.00	1.475	1.352	2.036	1.766
5	20	0.50	0.50	0.719	0.653	0.986	0.850	5	20	1.00	1.00	1.436	1.306	1.969	1.696
10	10	0.50	0.50	0.610	0.587	0.811	0.765	10	10	1.00	1.00	1.221	1.173	1.614	1.531
10	15	0.50	0.50	0.552	0.532	0.727	0.691	10	15	1.00	1.00	1.103	1.066	1.452	1.388
10	20	0.50	0.50	0.522	0.503	0.686	0.653	10	20	1.00	1.00	1.043	1.009	1.375	1.305
15	15	0.50	0.50	0.486	0.476	0.636	0.616	15	15	1.00	1.00	0.975	0.951	1.276	1.231
15	20	0.50	0.50	0.453	0.444	0.593	0.574	15	20	1.00	1.00	0.907	0.887	1.184	1.146
20	20	0.50	0.50	0.417	0.409	0.544	0.530	20	20	1.00	1.00	0.832	0.819	1.086	1.060

Note that N80 is the estimated 80 per cent confidence interval based on the naive method, while H80 is the estimated 80 per cent confidence interval based on Howe's approximation. N90 and H90 are similarly defined for the 90 per cent confidence interval.

Table V. Type I error of different tests when $\sigma_X^2 \neq \sigma_Y^2$ and $n \neq m$.

n	m	σ_X	σ_Y	p_N	p_H	p_C	p_S	n	m	σ_X	σ_Y	p_N	p_H	p_C	p_S
5	10	0.25	0.50	0.023	0.036	0.039	0.034	10	15	0.50	0.75	0.041	0.045	0.042	0.039
5	10	0.25	0.75	0.021	0.044	0.041	0.038	10	15	0.75	0.75	0.050	0.043	0.043	0.039
5	10	0.50	0.75	0.030	0.038	0.038	0.032	10	20	0.25	0.50	0.020	0.047	0.043	0.047
5	10	0.75	0.75	0.047	0.040	0.040	0.028	10	20	0.25	0.75	0.015	0.046	0.046	0.043
5	15	0.25	0.50	0.015	0.039	0.037	0.042	10	20	0.50	0.75	0.029	0.043	0.045	0.044
5	15	0.25	0.75	0.010	0.038	0.041	0.040	10	20	0.75	0.75	0.048	0.041	0.045	0.042
5	15	0.50	0.75	0.024	0.037	0.042	0.034	15	20	0.25	0.50	0.038	0.049	0.045	0.047
5	15	0.75	0.75	0.052	0.039	0.042	0.029	15	20	0.25	0.75	0.034	0.048	0.048	0.047
5	20	0.25	0.50	0.011	0.040	0.042	0.036	15	20	0.50	0.75	0.045	0.047	0.050	0.045
5	20	0.25	0.75	0.004	0.040	0.040	0.040	15	20	0.75	0.75	0.045	0.044	0.043	0.045
5	20	0.50	0.75	0.022	0.037	0.040	0.034	20	20	0.25	0.50	0.054	0.047	0.046	0.045
5	20	0.75	0.75	0.048	0.042	0.044	0.026	20	20	0.25	0.75	0.049	0.049	0.048	0.044
10	15	0.25	0.50	0.033	0.044	0.042	0.044	20	20	0.50	0.75	0.049	0.046	0.048	0.044
10	15	0.25	0.75	0.029	0.049	0.046	0.043	20	20	0.75	0.75	0.052	0.044	0.048	0.045

Note that p_N is the estimated coverage probability based on the naive method, while p_H, p_C and p_S are the estimated coverage probability based on Howe's approximation, Cochran–Cox's approximation and Satterthwaite's approximation, respectively.

Table VI. Sensitivity analysis for normal assumption.

n	m	σ_X	σ_Y	p_N	p_H	p_C	p_S	n	m	σ_X	σ_Y	p_N	p_H	p_C	p_S
5	10	0.25	0.50	0.006	0.024	0.048	0.040	10	15	0.50	0.75	0.019	0.034	0.052	0.043
5	10	0.25	0.75	0.002	0.019	0.084	0.046	10	15	0.75	0.75	0.042	0.054	0.030	0.038
5	10	0.50	0.75	0.009	0.031	0.028	0.035	10	20	0.25	0.50	0.007	0.033	0.054	0.047
5	10	0.75	0.75	0.032	0.065	0.016	0.029	10	20	0.25	0.75	0.004	0.024	0.079	0.047
5	15	0.25	0.50	0.002	0.035	0.035	0.040	10	20	0.50	0.75	0.013	0.046	0.039	0.045
5	15	0.25	0.75	0.001	0.026	0.055	0.044	10	20	0.75	0.75	0.043	0.069	0.025	0.039
5	15	0.50	0.75	0.006	0.056	0.025	0.034	15	20	0.25	0.50	0.015	0.026	0.072	0.048
5	15	0.75	0.75	0.028	0.090	0.013	0.022	15	20	0.25	0.75	0.013	0.022	0.090	0.050
5	20	0.25	0.50	0.000	0.048	0.025	0.036	15	20	0.50	0.75	0.025	0.035	0.056	0.049
5	20	0.25	0.75	0.000	0.030	0.048	0.046	15	20	0.75	0.75	0.049	0.054	0.034	0.042
5	20	0.50	0.75	0.003	0.071	0.021	0.028	20	20	0.25	0.50	0.028	0.025	0.078	0.051
5	20	0.75	0.75	0.026	0.105	0.012	0.021	20	20	0.25	0.75	0.025	0.021	0.097	0.054
10	15	0.25	0.50	0.010	0.024	0.063	0.048	20	20	0.50	0.75	0.036	0.028	0.064	0.050
10	15	0.25	0.75	0.010	0.023	0.091	0.051	20	20	0.75	0.75	0.049	0.047	0.047	0.047

Note that p_N is the estimated coverage probability based on the naive method, while p_H, p_C, and p_S are the estimated coverage probability based on Howe's approximation, Cochran–Cox's approximation and Satterthwaite's approximation, respectively.

6.2. *Parallel clinical trial*

The objective of the simulation study in this section is multi-fold. The first is to evaluate the accuracy of the derived formula for sample size calculation (8). Secondly, it is to compare our proposed method with the Cochran–Cox and Satterthwaite approximation. The last objective is to perform a sensitivity analysis when normal assumption does not hold.

Table V presents some simulation results, which compares our proposed methods to not only the naive method but also two other commonly used methods (the Cochran–Cox approximation

Table VII. Sample size study for parallel clinical trials with $k = 2$.

σ_X	σ_Y	$\Delta = -15\%$		$\Delta = -20\%$		$\Delta = -25\%$	
		n	p	n	p	n	p
0.20	0.20	16	0.762	9	0.730	6	0.713
0.20	0.25	20	0.783	11	0.761	7	0.730
0.20	0.30	23	0.773	13	0.768	8	0.725
0.20	0.35	28	0.790	16	0.789	10	0.766
0.20	0.40	33	0.792	19	0.790	12	0.760
0.25	0.20	23	0.792	13	0.770	8	0.723
0.25	0.25	26	0.790	14	0.751	9	0.736
0.25	0.30	30	0.794	17	0.784	11	0.769
0.25	0.35	34	0.790	19	0.772	12	0.765
0.25	0.40	39	0.792	22	0.774	14	0.774
0.30	0.20	30	0.787	17	0.779	11	0.756
0.30	0.25	33	0.786	19	0.780	12	0.761
0.30	0.30	37	0.788	21	0.782	13	0.754
0.30	0.35	42	0.800	23	0.783	15	0.769
0.30	0.40	47	0.806	26	0.782	17	0.783
0.35	0.20	39	0.785	22	0.775	14	0.759
0.35	0.25	42	0.786	24	0.780	15	0.763
0.35	0.30	46	0.795	26	0.791	17	0.778
0.35	0.35	50	0.793	28	0.778	18	0.776
0.35	0.40	56	0.798	31	0.780	20	0.771
0.40	0.20	49	0.781	28	0.790	18	0.780
0.40	0.25	53	0.798	30	0.788	19	0.762
0.40	0.30	56	0.790	32	0.785	20	0.775
0.40	0.35	61	0.792	34	0.785	22	0.782
0.40	0.40	66	0.798	37	0.784	24	0.782

Note that n is the estimated sample size required in X group according to (8) and p is the estimated power.

and Satterthwaite approximation) for unequal variances in terms of type I error. The results show that our method's performance is comparable to both the Cochran–Cox approximation and Satterthwaite approximation. However, our method enjoys a simple formula for sample size calculation.

Table VI reports the empirical significance level of not only our proposed method but also the naive, Cochran–Cox and Satterthwaite approximation when the normal assumption does not hold. Basically, for each parameter setting 10 000 data sets were generated. Each data set contained n random variables from centralized exponential distribution ($\exp(\lambda) - \lambda$) with a standard deviation σ_X and m random variables from the same type of distribution but with a standard deviation of σ_Y. Then for each data set, the null hypothesis of $H_0 : \mu_X = \mu_Y$ is tested by all the four methods. The empirical significance level is estimated by the proportion of the data sets which rejected the null hypothesis. The results show that all the above methods except the Satterthwaite approximation performs worse when the normal assumption is not true. However, with larger sample size the situation is better. This indicates our proposed method should be used with caution with small sample size when the normal assumption is problematic.

For a given parameter setting ($\Delta = \mu_X - \mu_Y < 0$, σ_X^2 and σ_Y^2/k), the sample size needed for achieving an 80 per cent power is calculated according to (8). A data set with the calculated

Table VIII. Sample size study for parallel BE trials with $k = 2$.

σ_X	σ_Y	$\Delta = 0\%$		$\Delta = 5\%$		$\Delta = 10\%$		$\Delta = 15\%$	
		n	p	n	p	n	p	n	p
0.20	0.20	10	0.717	12	0.720	25	0.790	70	0.801
0.20	0.25	12	0.735	15	0.769	29	0.787	83	0.798
0.20	0.30	15	0.780	18	0.782	35	0.788	99	0.797
0.20	0.35	17	0.762	21	0.779	41	0.790	117	0.793
0.20	0.40	21	0.790	25	0.779	49	0.793	139	0.791
0.25	0.20	14	0.741	17	0.758	34	0.786	96	0.800
0.25	0.25	16	0.757	19	0.756	38	0.788	109	0.794
0.25	0.30	19	0.781	22	0.762	44	0.796	125	0.800
0.25	0.35	21	0.771	26	0.778	51	0.798	144	0.801
0.25	0.40	25	0.790	29	0.773	58	0.799	165	0.797
0.30	0.20	19	0.768	23	0.771	45	0.792	128	0.794
0.30	0.25	21	0.772	25	0.765	50	0.799	141	0.799
0.30	0.30	23	0.774	28	0.767	55	0.790	157	0.797
0.30	0.35	26	0.774	31	0.782	62	0.795	175	0.795
0.30	0.40	29	0.775	35	0.776	69	0.786	197	0.802
0.35	0.20	25	0.781	29	0.768	58	0.786	165	0.802
0.35	0.25	26	0.760	32	0.771	63	0.795	178	0.794
0.35	0.30	29	0.777	35	0.779	68	0.791	194	0.794
0.35	0.35	32	0.788	38	0.779	75	0.794	213	0.803
0.35	0.40	35	0.790	42	0.781	83	0.798	235	0.798
0.40	0.20	31	0.778	37	0.773	74	0.788	209	0.800
0.40	0.25	33	0.778	40	0.775	78	0.793	222	0.799
0.40	0.30	35	0.778	42	0.772	84	0.793	238	0.801
0.40	0.35	38	0.786	46	0.778	90	0.791	257	0.810
0.40	0.40	41	0.773	50	0.786	98	0.802	278	0.796

Note that n is the estimated sample size required in X group according to (11) and (12), and p is the estimated power.

sample size are generated according to the given parameter setting. For the selected data set, the upper bound of $\mu_X - \mu_Y$ is calculated based on (7). A total of 10 000 simulation runs were done. The power is estimated by the proportion of the simulation runs with negative upper bound, which corresponds to the rejection of the null hypothesis. The results are summarized in Table VII. The results show that the sample size formula works satisfactorily.

6.3. *Parallel BE trial*

In this simulation study, the objective is to examine the accuracy of the derived formulae for sample size calculations (11) and (12). For a given parameter setting ($\Delta = |\mu_X - \mu_Y| < 0.223, \sigma_X^2$ and σ_Y^2/k), the sample size required for achieving an 80 per cent power is calculated according to (11) or (12). The confidence interval of $\mu_X - \mu_Y$ is calculated based on (9). A total of 10 000 simulation runs was considered. The power is estimated by the proportion of the simulation runs that the 90 per cent confidence interval fall entirely within $(-0.223, 0.223)$, which corresponds to the conclusion of ABE. The results are summarized in Table VIII. The results show that the sample size formula works satisfactorily.

7. SUMMARY

In this paper, we propose a simple solution for comparing means between two groups without equal variance assumption. The methodology is based on Howe's approximation I. Our method is compared with the Cochran–Cox and Satterthwaite methods. The simulation results show that under the normal assumption, our method's performance is comparable with the Cochran–Cox and Satterthwaite methods. However, if the normal assumption is not true, our method and Cochran-Cox's performances become worse while the Satterthwaite's performance is still relatively good. However, compared with the Cochran–Cox and Satterthwaite methods, our method enjoys simple yet accurate sample size calculation formulae when normal assumption is true.

REFERENCES

1. Miller RG. Jun. *Beyond ANOVA*, *Basics of Applied Statistics*, 2nd edn. Springer-Verlag: New York, 1997.
2. Benerjee SK. Approximate confidence interval for linear functions of means of k populations when the population variances are not equal. *Sankhya* 1960.
3. Benerjee SK. On confidence interval for two-means problem based on separate estimate of variances of t-table. *Sankhya* 1961.
4. Weerahandi S. *Exact Statistical Methods for Data Analysis*. Springer-Verlag: New York, 1995.
5. Shao J. *Mathematical Statistics*. Springer-Verlag: New York, 1998.
6. Howe WG. Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 1974; **69**:789–794.
7. Cornish EA, Fisher RA. Moments and cumulants in the specification of distribution. *Revue de l'Institute International Statistics* 1937; **5**:307–320.
8. Graybill F, Wang CM. Confidence intervals on nonnegative linear combination of variances. *Journal of American Statistical Association* 1980; **75**:869–873.
9. Ting N, Burdick RK, Graybill FA, Jeyaratnam S, Lu T-FC. Confidence intervals on linear combination of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation* 1990; **35**:135–143.
10. Hyslop T, Hsuan F, Holder DJ. A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* 2000; **19**:2885–2897.
11. FDA. Average, population, and individual approaches to establishing bioequivalence. U.S. Food and Drug Administration, Rockville, Maryland, 1999.
12. Gurland J, McCollough RS. Testing equality of means after a preliminary text of equality of variances. *Biometrika* 1962; **49**:403–417.
13. Chow SC, Wang H. On sample size calculation in bioequivalence trials. *Journal of Pharmacokinetics & Biopharmacodynamics* 2001; **28**:155–169.
14. Chow SC, Liu JP. *Design and Analysis of Bioavailability and Bioequivalence Studies—Revised and Expanded*. 2nd edn. Marcel Dekker, Inc.: New York, 1999.

