Anal Bioanal Chem (2006) 385: 975–981
DOI 10.1007/s00216-006-0417-2

TECHNICAL NOTE

Shixia Feng · Qiwei Liang · Robin D. Kinser ·
Kirk Newland · Rudolf Guilbaud

Testing equivalence between two laboratories or two methods using paired-sample analysis and interval hypothesis testing

Received: 30 January 2006 / Revised: 3 March 2006 / Accepted: 4 March 2006 / Published online: 7 June 2006

Abstract A modified interval hypothesis testing procedure based on paired-sample analysis is described, as well as its application in testing equivalence between two bioanalytical laboratories or two methods. This testing procedure has the advantage of reducing the risk of wrongly concluding equivalence when in fact two laboratories or two methods are not equivalent. The advantage of using paired-sample analysis is that the test is less confounded by the intersample variability than unpaired-sample analysis when incurred biological samples with a wide range of concentrations are included in the experiments. Practical aspects including experimental design, sample size calculation and power estimation are also discussed through examples.

Keywords Equivalence · Interval hypothesis testing · Paired-sample analysis · Bioanalysis · Comparability

Introduction

Bioanalytical methods of quantitatively measuring chemical compounds and their metabolites in biological samples must be validated to ensure that such methods yield reliable results [1]. To achieve necessary analytical throughput, bioanalytical methods are often transferred from one laboratory to another, and sometimes the samples from a clinical or animal study are analyzed using different methods or at different laboratories.

To ensure that comparable results can be achieved between two laboratories or methods, it is important to conduct cross-validation [1] or comparability studies [2] before a new laboratory or method is permitted to analyze the biological samples. There should be three objectives of the experimental design in such studies: to establish statistical equivalence [2] between the two laboratories or methods, to identify the sources of any differences, and to resolve these differences. When conducting such studies, it is important to maintain the traceability of the results through proper documentation and to determine the uncertainties (variance) in the measurement [3].

Historically, several techniques have been used to compare the data generated by two laboratories or methods using either paired- or unpaired-sample analysis (e.g., the general *t*-test, Youden matched pair plots [4], and regression analysis). However, each of these techniques has certain drawbacks [5, 6]. For example, Youden plots only provide visual, qualitative assessment, and regression analysis is based on the false assumption that the errors in the *x*-values are negligible and that all of the error resides in the *y*-values. For the general *t*-test, the most widely used technique, the hypotheses are:

$$\begin{aligned} &\mathrm{H}_0 : \mu_2/\mu_1 = 1 \text{ (there is no bias) versus} \\ &\qquad\mathrm{H}_1 : \mu_2/\mu_1 \neq 1 \text{ (bias exists)} \end{aligned} \tag{1}$$

If there is insufficient evidence to reject the null hypothesis, the two means are declared equal. It has been shown [6] that this test has the undesirable property of penalizing higher precision. If the sample size is small and there are relatively large variations within two sets of data in comparison to the relatively small difference between the two means, the *t*-test may be unable to detect a possible difference, and sometimes when the precision of each data set is very high, differences of little practical importance could be considered significant.

S. Feng (✉) · Q. Liang · R. D. Kinser
Philip Morris USA, Research Center,
4201 Commerce Road,
Richmond, VA 23234, USA
e-mail: shixia.feng@pmusa.com

K. Newland
MDS Pharma Services,
621 Rose Street,
Lincoln, NE 68502, USA

R. Guilbaud
MDS Pharma Services,
2350 Cohen Street,
St. Laurent (Montreal), Quebec, H4R 2N6, Canada

Hartmann et al. [7] compared the general *t*-test with the interval hypothesis test for testing the equivalence of two laboratories or two methods, and concluded that for laboratory or method validation purposes, the latter is preferable because the β-error of the general *t*-test can be controlled. The interval hypothesis test is based on Schuirmann's [6] two one-sided *t*-tests (TOST), and this has become the standard approach in bioequivalence studies. This approach requires the predetermination of an acceptance interval (lower limit θ_1 and upper limit θ_2) and then involves testing whether the measured bias is within the acceptance interval. In statistical terms, the hypotheses to be tested are:

$$\mathrm{H}_0 : \mu_2/\mu_1 \leq \theta_1 \text{ or } \mu_2/\mu_1 \geq \theta_2 \text{ versus } \mathrm{H}_1 : \theta_1 < \mu_2/\mu_1 < \theta_2 \tag{2}$$

The hypotheses can also be expressed as two one-sided tests:

$$\mathrm{H}_{01} : \mu_2/\mu_1 \leq \theta_1 \text{ versus } \mathrm{H}_{11} : \mu_2/\mu_1 > \theta_1 \tag{3}$$

$$\mathrm{H}_{02} : \mu_2/\mu_1 \geq \theta_2 \text{ versus } \mathrm{H}_{12} : \mu_2/\mu_1 < \theta_2 \tag{4}$$

Compared to the general *t*-test, the null and alternative hypotheses have been reversed. Consequently, the definitions of type I and type II errors in the interval hypothesis test are exactly switched with respect to the general *t*-test. The type I error is now the probability (α) of the erroneous acceptance of equivalence when in fact the two laboratories or methods are not equivalent. The type II error is the probability (β) of erroneous acceptance of nonequivalence when in fact the two laboratories or methods are equivalent. The acceptance of null hypotheses leads to the conclusion that the bias is not acceptable, and the rejection of null hypotheses leads to the conclusion that the bias is acceptable. This test has the advantage of limiting the risk of erroneously accepting a new laboratory or new method as being unbiased (when it is actually biased) to very low degree.

However, close examination of the algorithm and procedures described by Hartmann et al. [7] led us to believe that they are only suitable for testing two means from two independent sets of samples (unpaired sample analysis). Unpaired-sample experimental design in a bioanalytical cross-validation study may confound this statistical test because of a possible large pooled variance that is actually due to the intersample variability, especially for incurred biological samples obtained from clinical or animal studies. We believe that this problem can be overcome by applying paired-sample analysis.

In this manuscript, we present a modified-interval hypothesis testing procedure and practical experimental design (based on paired-sample analysis) for testing the equivalence between two bioanalytical laboratories or methods. The example used is the transfer of an LC–MS/MS method for *S*-phenylmercapturic acid (*S*-PMA), a benzene metabolite in human urine and a biomarker for exposure to benzene [8, 9]. Both spiked quality control (QC) and incurred human samples were included in this study.

Experimental part

The interval hypothesis test procedure

The acceptance interval (θ_1 and θ_2) should be defined before the experiments begin. As it is now widely accepted [1] that a validated bioanalytical method should have less than ±15 % bias at concentration levels other than the lower limit of quantification (LLOQ), where at most ±20 % bias is acceptable, the θ_1 and θ_2 values adopted in this study were therefore 0.85 and 1.15, respectively (note that although we typically find these limits reasonable, one should define the acceptance interval based on his or her own experience with the method). A paired-sample design was employed and the concentration ratio for each sample was calculated. Let $R = \frac{Conc_{Lab2}(measured)}{Conc_{Lab1}(measured)}$ for each sample and $\overline{R}_{Lab2/Lab1}$ be the mean ratio of each sample set, while R_0 is the the true ratio between the two laboratories (Lab2/Lab1) for the entire sample population. The hypotheses to be tested can be stated as:

$$\mathrm{H}_0 : R_0 \leq 0.85 \text{ or } R_0 \geq 1.15 \text{ versus } \mathrm{H}_1 : 0.85 < R_0 < 1.15 \tag{5}$$

or:

$$\mathrm{H}_{01} : R_0 \leq 0.85 \text{ versus } \mathrm{H}_{11} : R_0 > 0.85 \tag{6}$$

and

$$\mathrm{H}_{02} : R_0 \geq 1.15 \text{ versus } \mathrm{H}_{12} : R_0 < 1.15 \tag{7}$$

The H_0 will be rejected if

$$t_{cal(1)} = \frac{\overline{R}_{Lab2/Lab1} - 0.85}{SD_R/\sqrt{n}} > t_{\alpha,n-1} \tag{8}$$

where $t_{\alpha,n-1}$ is the 1−α quantile of the *t* distribution with *n*−1 degrees of freedom. $\mathrm{SD_R}$ represents the standard deviation of the mean ratio for the sample set, or by rearrangement, if

$$\overline{R}_{Lab2/Lab1} - t_{\alpha,n-1}\, SD_R/\sqrt{n} > 0.85 \tag{9}$$

and if

$$t_{cal(2)} = \frac{\overline{R}_{Lab2/Lab1} - 1.15}{SD_R/\sqrt{n}} < -t_{\alpha,n-1} \tag{10}$$

or by rearrangement, if

$$\overline{R}_{Lab2/Lab1} + t_{\alpha,n-1}\, SD_R \big/ \sqrt{n} < 1.15 \qquad (11)$$

Notice that $\overline{R}_{Lab2/Lab1} \pm t_{\alpha,n-1}\, SD_R \big/ \sqrt{n}$ actually represents the confidence interval of the mean ratio at $(1-2\alpha)\times 100\%$ level. Thus, H_0 will be rejected at a significance level of α if the $(1-2\alpha)\times 100$ % confidence interval occurs entirely between 0.85 and 1.15.

In this study, basic data calculations were performed with Microsoft Excel 2002. The interval hypothesis test and statistical power were computed with Statistical Analytical System (SAS) software.

The LC–MS/MS method

An LC–MS/MS method for *S*-PMA in human urine was initially developed and validated at Lab1, the originating laboratory. The full details of the method will be published separately. Briefly, the method utilized 1 mL of a human urine sample and a solid phase extraction procedure. The LC–MS/MS system included a Thermo Hypersil BioBasic AX column (Thermo Electron Corporation, Waltham, MA, USA) with a guard column and a PE Sciex API 4000 MS detector (Applied Biosystems, Foster City, CA) with an electrospray ionization interface. Negative ions were monitored in the multiple reaction monitoring (MRM) mode.

After the method had been validated and used to analyze approximately 1,800 human urine samples at Lab1, it was transferred to Lab2 (accepting laboratory) to accommodate the need for higher throughput. The method used at Lab2 was identical to that at Lab1 except that the final step of the extraction procedure was slightly modified to allow for differences in available analytical equipment.

Table 1 Validation summary for the two laboratories

Performance parameter	Lab1*	Lab2**
Lower limit of quantitation	20 pg mL^{-1}	20 pg mL^{-1}
Linear range	20–20,000 pg mL^{-1}	20–20000 pg mL^{-1}
Between-run precision (CV%)		
Low QC	9.7 (*n*=24)	7.8 (*n*=17)
Medium QC	1.8 (*n*=24)	2.6 (*n*=18)
High QC	2.8 (*n*=24)	2.6 (*n*=18)
Between-run accuracy (% Nominal)		
Low QC	100.0 (*n*=24)	103.3 (*n*=17)
Medium QC	97.7 (*n*=24)	96.0 (*n*=18)
High QC	99.3 (*n*=24)	96.1 (*n*=18)

*The nominal values for low, medium, and high QC levels for the Lab1 were 142, 2,140, and 15,100 pg mL^{-1}, respectively

**The nominal values for low, medium, and high QC levels for the receiving Lab2 were 48.6, 2,070, and 15,206 pg mL^{-1}, respectively

Cross-validation experimental design

The method transfer was conducted according to a method transfer protocol which outlined the sample types and source and the cross-validation acceptance criteria. Lab2 performed the initial between-site qualification batches to re-establish the lower and upper limits of quantification, interbatch accuracy and precision, and the linear range. The performance characteristics at both laboratories are summarized in Table 1.

To establish equivalence between the two laboratories, a paired-sample design was employed: paired analyses of both spiked quality control (QC) samples at three fixed concentrations (low, medium, and high), and human incurred samples. QC samples at each concentration level

Table 2 The analytical results from a cross-validation study

N	QC 1 (Nominal=89.0 pg mL^{-1})			QC 2 (Nominal=2,140 pg mL^{-1})			QC 3 (Nominal=15,100 pg mL^{-1})			Incurred (pg mL^{-1})		
	Cocn$_1$	Cocn$_2$	$R_{2/1}$	Cocn$_1$	Cocn$_2$	$R_{2/1}$	Cocn$_1$	Cocn$_2$	$R_{2/1}$	Cocn$_1$	Cocn$_2$	$R_{2/1}$
1	83.7	83.6	1.00	2,090	2,070	0.99	14,500	14,700	1.01	391	372	0.95
2	92.8	91.3	0.98	2,150	2,140	1.00	14,700	14,700	1.00	432	362	0.84
3	87.4	93.7	1.07	2,100	2,110	1.01	14,700	14,900	1.01	4430	4230	0.96
4	89.4	86.1	0.96	2,070	2,120	1.02	14,600	14,900	1.02	141	119	0.84
5	99.0	91.2	0.92	2,040	2,190	1.07	14,500	14,300	0.99	80.7	82.7	1.03
6	96.1	85.9	0.89	2,230	2,190	0.98	15,100	14,500	0.96	12800	11600	0.91
7	102	93.4	0.92	2,110	2,130	1.01	14,400	14,100	0.98	232	243	1.05
8	89.5	94.4	1.06	2,000	2,050	1.03	14,800	14,400	0.97	638	660	1.03
9	102	87.4	0.86	2,180	2,030	0.93	14,900	13,800	0.93	10800	9930	0.92
10	102	81.3	0.80	2,150	2,150	1.00	14,600	14,800	1.01	1120	1100	0.98
11	94.0	95.2	1.01	2,170	2,020	0.93	14,800	14,900	1.01	9750	8620	0.88
12	93.0	95.1	1.02	2,020	2,010	1.00	14,500	14,300	0.99	3710	3500	0.94
Mean	94.2	89.9	0.96	2,109	2,101	1.00	14,675	14,525	0.99	3710	3402	0.94
SD$_R$	6.14	4.83	0.083	69.60	63.60	0.039	200.57	354.52	0.028	4727.18	4272.11	0.07
CV(%)	6.51	5.37	8.64	3.30	3.03	3.90	1.37	2.44	2.79	127.40	125.59	7.41

were prepared by Lab1 and a portion of each sample was sent to Lab2. The nominal values for each QC level were 89.0, 2,140, and 15,100 pg mL^{-1}, respectively. Each laboratory measured QC samples at each concentration level in twelve replicates in paired fashion. In addition, 12 incurred human urine samples were collected at Lab1 and a portion of each sample was also sent to Lab2. The ratio of measured concentrations was then calculated for each sample pair. The interval hypothesis test was performed on each QC set as well as the incurred human sample set. Equivalence between the two laboratories was declared at the α=0.025 level if the 95 % confidence interval of the mean ratio occurred entirely between 0.85 and 1.15 for all QC levels and incurred human samples.

Results and discussion

Sample selection

The calibration curve of a bioanalytical method should represent the normal concentration distribution of the real biological samples. To establish equivalence between the two laboratories, we believe that at least three concentration levels which represent the entire range of the calibration curve should be studied: one near the lower boundary of the curve, one near the center, and one near the upper boundary of the curve. Real biological samples that represent the normal concentration distribution of the analyte should also be included. The selection of QC levels and incurred human samples in this study was based on these considerations. The paired-sample design is based on the properties of the interval hypothesis test, which will be discussed later in the section on paired versus unpaired sample analysis. The analytical results from the two laboratories are shown in Table 2.

Type I and II errors, power and sample size

Similar to bioequivalence studies [10, 11], the primary concern of bioanalytical chemists should be the control of α (equivalent to β in the general t-test) because results for clinical samples generated by nonequivalent laboratories or methods that are erroneously taken to be equivalent will cause comparability problems and confusion in data interpretation. The α level should be defined a priori as the acceptance interval (θ_1 and θ_2). In this study, we predefined in the cross-validation protocol that the 95 % confidence interval of the mean ratio must occur entirely between 0.85 and 1.15 for all QC levels and incurred human samples. Therefore, the risk of wrongly concluding that the two labs are equivalent is limited to no more than 2.5 %. The actual probability of this risk is represented by the p-value, and can be calculated by SAS software using the two one-sided tests approach (Eqs. 6 and 7). The p-value of the interval hypothesis test (Eqs. 5, 6 and 7) is the maximum p-value of the two one-sided tests. For example, as shown in Table 3, the p-values for the two one-sided tests for the incurred human sample set were 0.0008 and <0.0001, respectively. The p-value of the interval hypothesis test is therefore 0.0008, which is much lower than the predefined acceptable limit of 0.025. To reject H_0, both p_1 and p_2 should be less than the α-value. In this example, H_0 (nonequivalence) was rejected and H_1 (equivalence) was accepted at α=0.025. The risks of declaring equivalence between the two laboratories while the two laboratories are in fact not equivalent were extremely small for all sample sets tested.

For the incurred sample set, we plotted the ratio versus the mean concentration from the two laboratories (not shown here), which showed no correlation. This indicated that the two laboratories were not systematically different at various points in the concentration range.

A basic question bioanalytical chemists often ask when they design experiments is: how many samples need to be analyzed in order to make the conclusions of the experiment statistically meaningful? Sample size (number of sample pairs in our case) is one of the key factors that affects the statistical power, and should be estimated based on the desired power and estimated variability. The power ($1-\beta$) for the interval hypothesis test should be:

$$\begin{aligned}\text{Power}\left(\overline{R}_{\text{Lab2/Lab1}}\right) &= P\left\{t_{\text{cal}(1)} \geq t_{\alpha,n-1} \text{ and } t_{\text{cal}(2)} \leq -t_{\alpha,n-1}\right\} \\ &= P\left\{0.85 + t_{\alpha,n-1}\sigma_{\text{R}}\big/\sqrt{n} < 1.15 - t_{\alpha,n-1}\sigma_{\text{R}}\big/\sqrt{n}\right\}\end{aligned} \tag{12}$$

Table 3 The interval hypothesis test* results for mean ratios

95 % CI (lower, upper)	QC 1 (n=12) 0.91, 1.01	QC 2 (n=12) 0.97, 1.02	QC 3 (n=12) 0.97, 1.01	Incurred (n=12) 0.90, 0.99
t_1	4.57	13.13	18.45	4.61
t_2	−8.29	−13.73	−21.09	−10.33
$t_{\alpha,(n-1)}$	2.20	2.20	2.20	2.20
p_1	<0.0001	<0.0001	<0.0001	0.0008
p_2	<0.0001	<0.0001	<0.0001	<0.0001
p	<0.0001	<0.0001	<0.0001	0.0008
Power (%)	99 %	>99 %	>99 %	98 %

*The hypotheses tested are formulated in Eqs. 5, 6 and 7, and the rejection criteria for null hypotheses are formulated in Eqs. 8, 9, 10 and 11 where P is the probability and σ_R is the standard deviation of the entire sample population.

Using the same procedure described by Chow and Liu [12, 13], we obtained Eqs. 13 and 14, which can be used to estimate the minimum number of sample pairs needed to achieve the desired power:

$$n \geq (t_{\alpha,n-1} + t_{\beta,n-1})^2 \left(\frac{CV \times \overline{R}_{\mathrm{Lab2/Lab1}}}{1.15 - \overline{R}_{\mathrm{Lab2/Lab1}}} \right)^2 \quad \text{for } 1.00 < \overline{R}_{\mathrm{Lab2/Lab1}} < 1.15 \tag{13}$$

and

$$n \geq (t_{\alpha,n-1} + t_{\beta,n-1})^2 \left(\frac{CV \times \overline{R}_{\mathrm{Lab2/Lab1}}}{\overline{R}_{\mathrm{Lab2/Lab1}} - 0.85} \right)^2 \quad \text{for } 0.85 < \overline{R}_{\mathrm{Lab2/Lab1}} < 1.00 \tag{14}$$

where CV is the coefficient of variation of the sample set and $t_{\beta,n-1}$ is the $1-\beta$ quantile of the t distribution with $n-1$ degrees of freedom.

Figure 1 illustrates the effect of the number of sample pairs, the CV, and the ratio on the power at the α=0.025 level, based on the paired-sample design (when θ_1 and θ_2 were set at 0.85 and 1.15, respectively). The power increases as the CV decreases or the number of sample pairs increases.

The relationships shown in Fig. 1 should be used to guide experimental planning. However, because computing the power requires computer programming and also because the main concern of bioanalytical chemists in cross-validation is to control α, the post-study calculation of the actual power does not have to be part of the test. If the confidence interval is within the acceptance interval after the experiment, a simple way to verify whether the desired power has been achieved is to compare the calculated number of sample pairs for a desired power to the actual number of sample pairs. If the calculated number of sample pairs is less than the actual value, the desired power has been achieved. It is common in clinical studies to set β to be less than 0.2, which means that the power of the statistical test is greater than 80 %. Therefore, for example, for the incurred sample set in this study, the estimated number of sample pairs for 80 % power is: $n = (2.20 + 0.88)^2 \times \left[(0.07)/(0.94 - 0.85)\right]^2 \approx 6$. Since

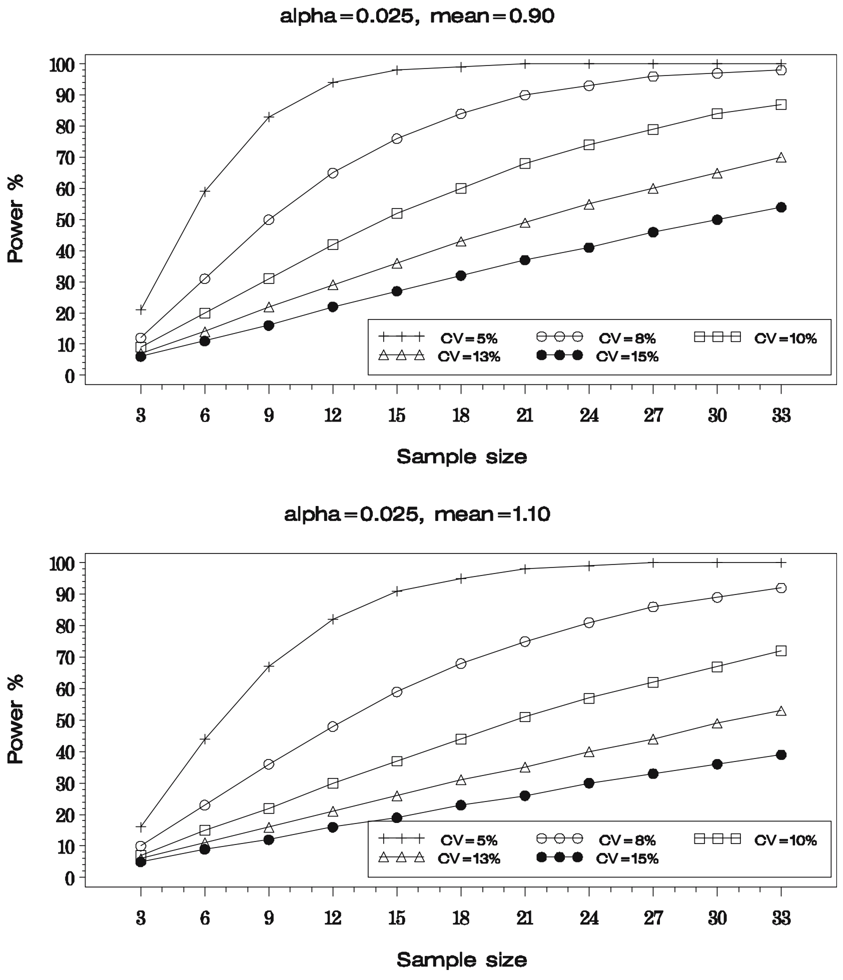

Fig. 1 The effect of the number of sample pairs, CV (%) and mean ratio on the power at the α=0.025 level. The *top figure* is for a mean ratio of 0.90, and the *bottom figure* is for a mean ratio of 1.10

the actual number was 12, the power must have exceeded 80 %. The power for this sample set is 98 %.

Each laboratory or method is associated with two types of errors: systematic and random errors. A ratio that is close to the acceptance boundary is an indication of systematic bias between the two laboratories or methods. One should first focus on finding the sources of systematic bias. For example, comparing the calibration standards, stock solutions and working solutions in both laboratories often helps resolve the problem. Sometimes in practice it is necessary to conduct pilot experiments with a limited number of samples to generate preliminary data in order to guide the full-scale comparison. The CV of the ratio is a function of random errors of the two laboratories or methods being compared. Assuming that there is no variation in sampling, the CV of the ratio can be estimated using Eq. 15 from the precision data of each laboratory or method for a given concentration level, which should have been generated during the method validation by repeated measurements of QC samples at different concentration levels.

$$CV = \frac{\sigma_R}{R_0} \approx \sqrt{\left(\frac{\sigma_{Lab1}}{Conc_{Lab1}}\right)^2 + \left(\frac{\sigma_{Lab2}}{Conc_{Lab2}}\right)^2} = \sqrt{CV_{Lab1}^2 + CV_{Lab2}^2} \tag{15}$$

Note that Eq. 15 does not apply to the incurred human samples, which may vary over a wide concentration range. Because the accuracy and precision at different concentration levels are often different for a laboratory or method, the incurred samples should have an approximately even distribution over the entire concentration range. If the incurred samples are not distributed evenly, both the ratio and CV may be skewed.

Paired versus unpaired sample analysis

The interval hypothesis test procedures used to test either the ratio of or the difference (δ) between two means from two independent random samples (unpaired sample analysis) have been described by previous authors [7, 14, 15]. When testing the ratio of two means, the rejection criteria for the null hypotheses for two independent series of data with equal variance $\left(\sigma_1^2 = \sigma_2^2\right)$ are:

$$t_{cal(1)} = \frac{(\overline{x}_2 - \theta_1 \overline{x}_1)}{\sqrt{s_p^2\left(1/n_2 + \theta_1^2/n_1\right)}} \geq t_{\alpha, n_1+n_2-2} \tag{16}$$

where s_p^2 is the pooled variance of the two means:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{17}$$

and

$$t_{cal(2)} = \frac{(\theta_2 \overline{x}_1 - \overline{x}_2)}{\sqrt{s_p^2\left(1/n_2 + \theta_1^2/n_1\right)}} \geq t_{\alpha, n_1+n_2+2} \tag{18}$$

It should be noted that the unpaired-sample design in bioanalytical cross-validation studies may confound this statistical test due to an incidence of very large pooled variance that is actually due to the intersample variability, especially for incurred biological samples. Two actually equivalent laboratories or methods may be determined to be nonequivalent unless the sample size is extremely large. Using the paired-sample design and testing the ratios allows one to remove such variability and therefore increase the sensitivity of the test. For similar reasons, it is inappropriate to test the mean differences (δ) for incurred human samples in paired-sample analysis. For example, for the incurred human sample set in this study, $\overline{\delta}$ =−308.8; SD=469.1; CV=151.9 %; 95 % CI=−606.9 (lower), −10.8 (upper); t_1=1.83 and t_2=−6.39 (compared to $t_{\alpha,(n-1)}$=2.20). Notice that the variability in the mean difference $\left(\overline{\delta}\right)$ of this sample set was very high (CV=151.9 %). This variability is mainly due to the variable sample concentrations. Therefore, the test performed on the mean difference led to the wrong conclusion that the two laboratories were not equivalent, because the test was confounded by the variability in the incurred sample itself.

Some potential issues

Compared to unpaired sample analysis, the degree of freedom in paired sample analysis is lower (n−1 versus n_1+n_2−2), which seemingly implies lower power. However, with good planning the desired power can be achieved, as discussed above.

Another issue is that the statistical testing procedure described herein assumes that the ratio has a normal distribution around the mean values, which is true in this study. The normality of the sample distributions should be tested before applying the algorithms described here, because it may affect the sensitivity of this test. For non-normal ratio data, if the sample size is sufficiently large and the variability is small, this procedure is still applicable. Log-transformation may also be considered if the data are not normally distributed.

Sometimes, especially when the sample size is small, one may encounter a situation where outliers (significantly larger or smaller ratios than the others in a data set) may occur, and the outcomes of the analysis can be influenced depending on how the outliers are treated. Potential outliers can be identified using statistical procedures such as graphical diagnostic tools (e.g., the box-and-whisker plot), Dixon's test, Grubbs' test [5], or simply by finding the values outside of three standard deviations of the mean. We recommend that the root causes be investigated before a decision is made to exclude outliers from the statistical analysis. The exclusion of an outlier may be justified if a review of the documentation clearly indicates an error in the sample processing or analysis. One should bear in mind that excluding the outliers will result in reduced sample size and will therefore affect both types of statistical errors. An alternative is to reanalyze the outlying samples in either or both laboratories using both methods. However, if the repeat analyses verify the original values, these values should not be excluded in the statistical analysis, and unfavorable conclusions may be reached in such cases. However, it is our opinion that this situation is likely due to large variability between the two laboratories or methods, which is one indication that they are not equivalent.

Sample source, stability and homogeneity are other factors that may affect the cross-validation results. The shipping and storage conditions must be proven to be effective at preserving the sample integrity. Proper documentation should be maintained to trace the samples back to their origins.

Conclusion

The interval hypothesis test and paired-sample experimental design described in this manuscript can be used to test for equivalence between two bioanalytical laboratories or methods. The premise for using this approach is that the two laboratories or methods being compared have been validated independently prior to the cross-validation. The acceptance intervals and the limits for the two types of risks associated with this statistical test should be defined prior to the experiments. Equations for estimating the number of sample pairs based on the power were also described. The number of sample pairs should be large enough to achieve the desired power. One of the advantages of this test and the experimental design based on this test is that the risk of wrongly concluding equivalence when in fact two laboratories or methods are not equivalent can be limited to a low level. In addition, as the number of sample pairs are determined based on power calculations, the risk of wrongly concluding nonequivalence when in fact the two laboratories or methods are equivalent can also be limited to a low level (<20 %). Finally, since both systematic bias and random errors from the two laboratories or methods are taken into account, this procedure should be suitable for practical use in the real world.

Acknowledgement The authors would like to thank Drs. Hans Roethig and Mohamadi Sarkar for helpful discussion and suggestions.

References

1. Shah VP, Midha KK, Findlay JW, Hill HM, Hulse JD, McGilveray IJ, McKay G, Miller KJ, Patnaik RN, Powell ML, Tonelli A, Viswanathan CT, Yacobi A (2000) Pharm Res 17:1551–1557 (also see FDA/CDER (2001) Guidance for industry bioanalytical method validation (online document). FDA/CDER, Rockville, MD, see http://www.fda.gov/cder/guidance/4252fnl.pdf, last accessed 24th April 2006)
2. Kuselman I (2006) Accred Qual Assur 10:466–470
3. EURACHEM/CITAC (2003) Traceability in chemical measurement: A guide to achieving comparable results in chemical measurement (online document). EURACHEM/CITAC, Budapest, Hungary (see http://www/eurachem.il.pt, last accessed 24th April 2006)
4. Youden WJ, Steiner EH (1975) Statistical manual of the Association of Official Analytical Chemists. AOAC, Washington, DC
5. Miller JN, Miller JC (2000) (eds) Statistics and chemometrics for analytical chemistry, 4th edn. Prentice Hall, New York
6. Schuirmann DJ (1987) J Pharmacokinet Biop 15:657–680
7. Hartmann C, Smeyersverbeke J, Penninckx W, Vanderheyden Y, Vankeerberghen P, Massart DL (1995) Anal Chem 67: 4491–4499
8. Boogaard PJ, van Sittert NJ (1995) Occup Environ Med 52:611–620
9. Boogaard PJ, van Sittert NJ (1996) Environ Health Perspect 104 (Suppl 6):1151–1157
10. Diletti E, Hauschke D, Steinijans VW (1991) Int J Clin Pharmacol Ther Toxicol 29:1–8
11. Phillips KF (1990) J Pharmacokinet Biop 18:137–144
12. Chow SC, Liu JP (eds) (1999) Design and analysis of bioavailability and bioequivalence studies, 2nd edn. Marcel Dekker, New York
13. Liu JP, Chow SC (1992) J Pharmacokinet Biop 20:101–104
14. Locke CS (1984) J Pharmacokinet Biop 12:649–655
15. Richter SJ, Richter C (2002) Qual Eng 14:375–380