Analytical Method Comparison Based upon Statistical Power Calculations

David J. Mazzo[1,3] and Margaret Connolly[2]

Received July 23, 1991; accepted November 13, 1991

Testing for the equivalence of results generated by different analytical methodology is a common practice in the pharmaceutical sciences. Methodology changes are implemented for both scientific and economic reasons during a scientific study. Thus, the need to demonstrate the appropriateness of considering data generated by distinct methods as part of a single information population arises. This paper describes a rapid and simple approach to the statistical design and interpretation of method comparison experiments. The approach presented is based upon a statistical power calculation technique, a knowledge of the variability associated with the methods to be compared and the criteria for equivalence (the limits within which differences become immeasurable or, for practical purposes, insignificant). Reference tables are included which show necessary sample sizes for comparison experiments for common combinations of these three variables.

KEY WORDS: method equivalence; statistics; power; comparison.

INTRODUCTION

It is a common occurrence in studies performed as part of the development of a new drug product or the monitoring of a marketed drug product that a change occurs in the analytical methodology. These changes usually are manifested as technique improvements (i.e., subtle but significant changes in method operating parameters, supplies, or equipment) or as technique substitution (i.e., a complete change of methodology) and are implemented for scientific and/or economic reasons. Since methodology adjustments of these types do not always occur prior to the initiation of a study, the methods used must be tested for similarity of accuracy and precision. This must be done to allow data generated in one portion of a study by the original method to be compared statistically to data generated in another portion of the same study using a distinct technique. Many factors must be considered when designing an experiment to test method equivalence including method variability, human and equipment resources, the criteria for equivalence (the limits within which differences become immeasurable or, for practical purposes, insignificant), sample availability, and time constraints.

Several approaches have been documented in the design and/or interpretation of method comparison experiments. One of the more commonly used techniques involves a linear regression of data of results by one method plotted against the corresponding results from a second method. A subsequent equivalence determination is made based upon the magnitude of the resulting correlation coefficient. While this technique is usually simple and can be experimentally quick, a number of authors have cited deficiencies and/or limitations in its application (1–5). Other approaches include principal-component analysis techniques (6,7), graphical techniques (8), and other advanced statistical techniques (9,10). Unfortunately, these approaches require relatively large sample sizes, may be time and/or resource intensive, and/or may require sophisticated statistical analysis.

An alternate approach to testing method equivalence is presented in this paper. The experimental design is based upon a statistical power calculation and has been constructed such that it will provide a relatively rapid and simple statistically valid test of method equivalence based on replicate assays of a single sample set. Once a sample size is chosen which gives a minimum adequate power to detect a difference in method mean accuracy, a sampling design is used which will generally provide additional information about the assays (e.g., relative precision, proportionality) while requiring no additional testing. The technique as described is flexible, allowing the analyst to perform the minimum amount of work necessary to determine method equivalence within the constraints of predetermined tolerable differences, method variability, and statistical probability.

EXPERIMENTAL

Sample Size Determination

Two types of errors are associated with statistical tests to determine significant difference between two sample means: Type I and Type II. Type I error results when a null hypothesis (e.g., no method difference) is rejected even though it is true. Type II error occurs when a null hypothesis is accepted even though it is false. The probability that a false null hypothesis is rejected is known as the power of the test [i.e., Power = 1 − (the probability of a Type II error)] (11).

The testing of equivalence of results generated by distinct analytical methodologies can also be described as the testing for differences of a given magnitude among the data under various experimental conditions (e.g., intrinsic method variability, sample size, resource availability, etc.). The power of a statistical test of significance, in this case, is the probability of detecting this predefined difference. In a given experimental design, additional power (versus a fixed difference) is generally gained by increasing the sample size.

Equivalence of results in the context of analytical method comparison is not equivalence in the pure mathematical sense. Rather, it is defined as the limits within which differences become immeasurable or, for practical purposes, insignificant. Since it is often unnecessary to detect a difference between methods which is smaller than the variability associated with any one of the methods, the range of differences is quickly reduced for most pharmaceutical analyses from approximately 0.5 to 5%. If we choose to apply the statistical significance test at the $\alpha = 0.05$ level (95% prob-

[1] Department of Analytical and Physical Chemistry, Rhône-Poulenc Rorer Central Research, 500 Virginia Drive, Fort Washington, Pennsylvania 19034.

[2] Department of Biostatistics, Rhône-Poulenc Rorer Central Research, 500 Virginia Drive, Fort Washington, Pennsylvania 19034.

[3] To whom correspondence should be addressed.

ability that a null hypothesis will be accepted if it is true), with a power of 0.90 or 0.95, and if the intrinsic variability of the two methods to be compared is known, then the number of samples required per method can be calculated:

$$D = (Z_\alpha + Z_\beta)[(\sigma_1^2 + \sigma_2^2)/n]^{1/2} \qquad (1)$$

Following algebraic rearrangement,

$$n = (\sigma_1^2 + \sigma_2^2)\{[(D/(Z_\alpha + Z_\beta)]^2\}^{-1} \qquad (2)$$

where

D = minimum difference to detect with given power

Z_α = critical value of the standard normal distribution (1.96 for $\alpha = 0.05$, two-sided)

Z_β = power critical value of the standard normal distribution (1.64 for $\beta = 0.05$, one-sided)

σ_1 = standard deviation of method 1

σ_2 = standard deviation of method 2

n = number of samples to be tested by each method

A slightly more exact estimate of n is obtained by an iterative calculation of Eq. (2) where $t(1 - \alpha/2, 2n - 2)$ is substituted for Z_α and $t(1 - \beta, 2n - 2)$ is substituted for Z_β. Here $t(x, \text{df})$ is the positive critical value at x probability of a t statistic with df degrees of freedom. When the required sample size is less than 20, the exact calculation generally indicates the need for one additional sample per analytical method. Otherwise, the exact method will give the same solution as that obtained using Eq. (2). Tables I and II list the exact sample sizes needed to determine a given minimum difference between method means based upon a knowledge of the standard deviation of the methods compared and predetermined values for power and statistical significance.

Table I. Required Sample Sizes (Number of Replicates by *Each* Method) When All Assay Variability Is Within-Day

CV%		Minimum difference to detect (%)			
New method	Standard method[a]	0.5	1	2	5
	Required n when power of test is 90%				
0.1	0.1	3	1	1	1
	0.2	4	3	1	1
	0.5	12	4	3	1
	1	44	12	4	2
	2	170	44	12	4
0.2	0.2	5	3	1	1
	0.5	14	5	3	1
	1	45	12	4	2
	2	171	44	12	4
0.5	0.5	23	7	3	1
	1	54	15	5	3
	2	180	46	13	4
1	1	86	23	7	3
	2	212	54	15	4
2	2	338	86	23	5
	Required n when power of test is 95%				
0.1	0.1	3	2	1	1
	0.2	4	3	1	1
	0.5	15	5	3	1
	1	54	15	5	3
	2	210	54	15	4
0.2	0.2	6	3	2	1
	0.5	17	5	3	1
	1	56	15	5	3
	2	211	54	15	4
0.5	0.5	27	8	4	2
	1	66	18	6	3
	2	222	57	15	4
1	1	105	27	8	3
	2	261	66	18	4
2	2	417	105	27	6

[a] Method which is to be replaced or augmented by new method.

Analytical Sample Preparation

After an appropriate sample size has been chosen using Table I or II, one of two cases usually is encountered. In the first case (Case 1), the methods to be compared employ identical sample preparation techniques with differences in the technique of determination of the analyte. Case 2, then, is the situation where different sample preparation techniques are employed with or without similar means of analyte determination. In order to avoid any contribution to experimental error due to differences among test samples [usually some pharmaceutical dosage form (e.g., tablet, capsule, suspension, etc.) with an intrinsic variability in the content of the analyte(s)], it is desirable to use "identical" samples when applying the methods to be compared. For Case 1, a single dosage form (tablet, capsule, etc.) is prepared for assay according to the sample preparation instructions of the methods. The resulting solution is then divided into two sets of n fractions which are to be assayed and results calculated according to each of the two methods. (Note that n is the appropriate sample size as chosen from either Table I or Table II.) For case 2, n spiked placebo samples are prepared according to the sample preparation instructions from each of the two methods. The placebo formulations for all n samples must be spiked by the addition of a *solution* (to avoid errors introduced due to lack of homogenicity when mixing two solids) containing the analyte to each of the placebos. Each of the two sets of n samples is then treated according to the sample preparation, assay, and calculation instructions of the respective methods.

It should be noted that the sample chosen for the method comparison experiment should represent, physically and chemically, the types of samples that would be encountered in the study in which the analytical methodology change is proposed. For example, for a stability study it may be appropriate to use a stability sample (i.e., aged and/or degraded sample) as the sample for the method comparison experiment. The procedures detailed herein for the preparation of samples for the method comparison experiment, and, more importantly, the statistical treatment of the method comparison data are independent of the type of sample used.

Method Comparison Data Treatment

After the choice of the appropriate number of samples and the determination of the analyte(s) according to the two

Table II. Required Samples Sizes (Number of Replicates by *Each* Method) When Assay Variability Has Between- and Within-Day Components[a]

		Required n when power of test is 90%								
New method between-day CV%		0	0	0	0	0.2	0.2	0.2	0.2	1
Standard method between-day CV%		0.2	0.2	0.2	1	0.2	0.2	0.2	1	1
Within-day CV%		Minimum difference to detect (%)								
New method	Standard method	1	2	5	5	1	2	5	5	5
0.1	0.1	*[b]	*	*	*	*	*	*	*	*
	0.2	*	*	*	*	*	*	*	*	*
	0.5	*	*	*	*	*	*	*	*	*
	1	*	*	*	*	*	*	*	*	*
	2	*	*	*	*	*	*	*	*	*
0.2	0.2	*	*	*	*	*	*	*	*	*
	0.5	8	3	1	*	26	3	1	*	*
	1	21	5	3	*	75	5	3	4	10
	2	75	14	4	5	273	15	4	6	18
0.5	0.5	11	3	1	3	40	4	2	*	*
	1	25	5	3	4	89	6	3	4	*
	2	79	14	4	6	287	16	4	6	*
1	1	38	8	3	4	138	9	3	4	*
	2	93	16	4	6	336	18	4	6	20
2	2	147	25	5	8	534	28	5	8	28

		Required n when power of test is 95%						
New method between-day CV%		0	0	0	0	0.2	0.2	0.2
Standard method between-day CV%		0.2	0.2	0.2	1	0.2	0.2	1
Within-day CV%		Minimum difference to detect (%)						
New method	Standard method	1	2	5	5	2	5	5
0.1	0.1	*	*	*	*	*	*	*
	0.2	*	*	*	*	*	*	*
	0.5	*	*	*	*	*	*	*
	1	*	*	*	*	*	*	*
	2	*	*	*	*	*	*	*
0.2	0.2	*	*	*	*	*	*	*
	0.5	11	3	1	*	4	1	*
	1	31	6	3	*	7	3	4
	2	112	17	4	7	20	4	8
0.5	0.5	16	4	2	4	4	2	*
	1	36	6	3	5	8	3	5
	2	118	18	4	7	21	4	8
1	1	57	9	3	5	11	3	5
	2	138	20	4	8	24	5	9
2	2	219	32	6	11	37	6	12

[a] Standard method = method which is to be replaced or augmented by new method.

[b] Although the sample size equation has a solution, it is not accurate unless both within-day CV% values are as large as the largest between-day CV% and at least one within-day CV% is larger than the largest between-day CV%.

methods, the data obtained are tabulated in a one-to-one correspondence fashion and the mean result (±relative standard deviation) is calculated. The two means are then compared using the appropriate (paired or unpaired) t test to determine statistical equivalence. If the result of this statistical exercise indicates equivalence, it is then acceptable to interchange the two analytical methods freely and consider the resulting data part of a single information population. If the means are shown to be not statistically equivalent and the difference is less than that decided a priori to be significant or relevant, the methods may still be freely interchanged with the data treated as if arising from a single information population. Obviously, if the statistical analyses indicates that the means are not statistically equivalent and the determined difference is greater than or equal to the difference decided a priori to be significant or relevant, the methods may not be freely interchanged since the resulting data sets will not belong to the same information population.

Statistical Simulation

Often the information gained from the method comparison experiment can be greatly increased with a small additional expenditure of resources by using analytical samples with a range of concentrations which spans the desired working range of the assay. Such data can be plotted to show whether assay variance is homogeneous over the range and whether the relative bias between methods is fixed, relative, or otherwise dependent upon concentration.

The standard sample size calculations which are used here assume that all of the analytical samples will be replicates of a single pharmaceutical formulation and the method comparison is based on a two-tailed, parametric, unpaired t test. If the analytical samples can be grouped, as, for example, when split samples taken over the range of the standard curve are assayed by two methods, then the appropriate analysis is by a paired t test. The relation between these statistical tests can be expressed in terms of σ_1^2 and σ_2^2, the intrinsic variances of methods 1 and 2, respectively, measured at a single concentration or as dispersion about a regression line (assay value vs true concentration). The paired t test, then, is

$$n^{1/2}(x_1 - x_2)/(\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2)^{1/2} \qquad (3)$$

which is compared to a table value with $n - 1$ degrees of freedom, and the unpaired t test is

$$n^{1/2}(x_1 - x_2)/(\sigma_1^2 + \sigma_2^2)^{1/2} \qquad (4)$$

which is compared to a table value with $2n - 2$ degrees of freedom.

At different values of the sample size (n) and the average correlation (r) in the paired results, either the unpaired or the paired t test will have greater power to detect an assay method difference (D). (Reference 5 discusses the correlation coefficient as a function of the intrinsic variability in the assay results and the range of sample concentrations assayed.) The analyst generally benefits by choosing a design which is appropriate for analysis by the paired t test. For this reason, it is useful to verify that the sample size chosen to have a given power against a fixed method difference based on the unpaired t test will be sufficient to give similar Type I and Type II error rates in a paired design.

A simulation study was conducted to compare unpaired t-test and paired t-test statistics. Designs were selected from combinations reported in Tables I and II. Each design was simulated 400 times with zero true mean difference (D) between assay methods and then 400 times with D equal to the minimum detectable difference. The random normal deviates were calculated by the SAS function, RANNOR (13). Using the same set of random deviations, data were generated to be identically distributed with a method 1 mean of 100 units or to be evenly grouped with method 1 means at 90, 96.7, 103.3, and 110 units. The homogeneous samples were compared by an unpaired t test and the grouped data were compared by a paired t test.

RESULTS AND DISCUSSION

The sample size calculations summarized in Tables I and II are based on two definitions of method precision. The first, within-day variance is a measure of the expected precision when a single analyst on a single day makes determinations of several analytical samples prepared from a single quantity of test material. Assay repeatability is a measure of within-day variance. The second type of variability is observed between days. Assays made on different days may differ because of unexceptional differences among analysts, apparatus, or sample preparations. Assay reproducibility is the measure of variance between days (12). Although sample comparisons will be most precise if made within a single assay, many practical comparisons, such as those among stability samples, must be made over several different assays conducted on several days. The observed differences in such comparisons must be evaluated relative to a measure of assay precision which includes repeatability and reproducibility components.

A popular method of analytical test performance comparison (the Greenbrier approach) has been proposed by Haynes *et al.* (9). This method is designed to compare both precision and relative accuracy and consistency of two test methods over a practical range of sample concentrations. The Greenbrier approach consists of a comprehensive test plan calling for 36 analytical tests, 6 for each test method carried out on 3 different days. The Greenbrier decision procedure leading to test method equivalence conclusions also requires that the analytical testing be repeated when results from the first set of three is within a defined range of uncertainty.

Although a 3-day trial is a minimal design for direct comparison of reproducibility between two assay methods, the Greenbrier test plan is often impractical or impossible to apply due to time, resource, and/or sample constraints. If the individual validation of each method included a measure of variability within and between days, then this information may be used to assess the statistical validity of a single day method comparison protocol. When the minimum difference to be detected is relatively large compared to variation between days, then a single-day trial, with some increase in sample size compared to the design for methods which have no variation between days, will have adequate power and sensitivity. (It should be noted that it is not possible to estimate test method reproducibility from such method equivalence test data since the experiment as specified is performed on a single day.)

Sample size estimates in Tables I and II are calculated from Eq. (2), with a modification for exact small sample probabilities. In the interest of generality, minimum differences and method standard deviations are expressed as a percentage of the average assay result (percentage difference and percentage coefficient of variation). Although technically, a method which has a constant coefficient of variation (CV) does not have the same properties as a method with a constant standard deviation (SD), this distinction is not critical in assays which are applied to a limited range of sample concentrations. The sample sizes (n) listed are the required number to be tested by each method. A percentage difference which is small relative to the assay CV can be detected

Table III. Simulation Results Showing Empirical Type I and Type II Error Rates When Nominal Power of Test Is 0.90 and Significance Level (Size) Is 0.05[a]

Between-day CV%		Within-day CV%		Minimum difference	Sample size	*t*-test Type I error rate		*t*-test power (1 − Type II rate)	
New method	Standard method	New method	Standard method	*D*	*n*	Unpaired	Paired	Unpaired	Paired
0.0	0.0	0.1	0.5	0.5	12	0.045	0.045	0.875	0.868
0.0	0.0	0.1	2.0	1.0	44	0.048	0.045	0.882	0.882
0.0	0.0	0.2	2.0	2.0	12	0.060	0.058	0.888	0.882
0.0	0.0	0.5	2.0	5.0	4	0.065	0.048	0.888	0.862
0.0	0.0	0.5	2.0	5.0	4	0.068	0.058	0.905	0.882
0.0	0.2	0.1	0.2	1.0	4	0.212	0.205	1.000	1.000
0.0	0.2	0.1	0.2	1.0	4	0.112	0.098	1.000	1.000
0.2	0.2	0.1	0.5	1.0	24	0.610	0.620	1.000	1.000
0.2	0.2	0.2	1.0	1.0	75[b]	0.038	0.040	1.000	1.000
0.2	0.2	0.5	2.0	2.0	16	0.048	0.042	0.952	0.955
0.2	0.2	0.5	2.0	2.0	16	0.075	0.083	0.898	0.900
0.2	1.0	2.0	2.0	5.0	8	0.030	0.028	0.995	0.985

[a] Each column represents a unique random seed used in the calls to the SAS random normal deviate generator. Standard method = method to be replaced or augmented by new method.

[b] In groups of 15 at 90, 95, 100, 105, and 110 units.

with power equal to 0.9 or 0.95 (at the 0.05 significance level), provided that the sample size is made sufficiently large. Table I shows that, when both methods have the same SD (or, equivalently, CV), a (percentage) difference which is equal to the method CV can be detected with power = 0.9 when the sample size is at least 23 by each method. A percentage difference which is twice as large as the common CV% can be detected with a sample of 7 by each method. A slight increase in sample size (23 to 27 and 7 to 8) will increase the power for these comparisons from 0.9 to 0.95.

Table II lists required sample sizes when a component of variation between days (σ) is considered as well as the variation within days (σ_w). Since *n* replicates done on 1 day give a standard error (or SD) for the sample mean which is $(\sigma^2_b + \sigma^2_w/n)^{-1/2}$, the standard error of the sample mean cannot be reduced below σ_b, even if infinitely many samples are tested in 1 day. Thus, in Table II, a detectable difference (*D*) must be at least large enough so that

$$D^2 > (\sigma_{b1}^2 + \sigma_{b2}^2)(Z_\alpha^2 + Z_\beta^2) \qquad (5)$$

Simulation experiments discussed below indicate that some variance configurations lead to inaccurate sample size calculations. The required sample sizes for some σ^2_w and power combinations in Table II are then larger than the corresponding sizes of Table I. For example, when σ^2_b and σ^2_w for both methods are 0.2 (CV%), then a difference of 1% can be detected at the 0.05 significance level and power = 0.9 with as few as 12 samples tested by each method. Generally, it appears that the detectable difference must be at least five times as large as the between-day coefficient of variation in order to have reasonable power to detect method differences in a 1-day experiment.

If factors which contribute to variation between days can be controlled, then sample size calculations might be taken from Table I, since results by both methods will share the random day effect. In general, however, it is expected that Table II will represent something closer to the true situation. If the variability between days is larger than the variability within days, and if the method comparison experiment cannot be designed to eliminate factors of between-day variation, then the experiment must be run on several days.

Results of a small Monte Carlo study to compare Type I and Type II error rates between the unpaired and the paired *t* test are reported in Table III. Although the number of replications is small, Table III shows good agreement between theoretical and empirical error rates for all of the designs which have only within-day components of variation and for most of the designs which have between- and within-day variability. These results demonstrate that sample size calculations of Table II are accurate only if both the within-day components of variability (σ_w) are at least as large as either between-day component (σ_b) and at least one σ_w is larger than both σ_b.

The results in Table III also show good agreement between the power and significance levels of the unpaired and those of the paired *t* tests. This suggests that sample sizes chosen for the simple method comparison between replicates at a single concentration are valid for an experiment designed to compare paired assay results over a range of concentrations. The paired design will give additional information about assay variability and evidence of proportional bias.

Note that the sensitive test for fixed bias is a paired *t* test. If the results of method 2 (new method) are regressed on method 1 (standard method) results and statistical tests are performed to evaluate null hypotheses that the intercept is zero and the slope is one, neither of these tests will be as sensitive as the paired *t* test. A test for intercept in a restricted regression with slope fixed at one is equivalent to the paired *t* test. Similarly, a statistical test that the slope is equal to one in a regression through the origin is a sensitive test for relative bias.

CONCLUSION

An approach to experimental design of method equivalence experiments has been presented. The approach emphasizes practical considerations (relevance of difference below a certain value) as well as statistically valid data analyses. The method is often less resource consuming than other approaches commonly employed as the comparison experiment may be run on a single day if the between-day variability is small relative to the within-day component and the minimum difference to detect. If sample preparation techniques allow, the method comparison might be done relative to within-day precision. Such a plan may not give such general evidence for method equivalence as an experiment run on several days. It is always efficient to perform a method comparison experiment as paired comparisons at concentrations spanning the working range of the assay.

REFERENCES

1. J. O. Westgard and M. R. Hunt. Use and interpretation of common statistical tests in method-comparison studies. *Clin. Chem.* **19**(1):49–57 (1973).
2. G. T. Wu, S. L. Twomey, and R. E. Thiers. Statistical evaluation of method-comparison data. *Clin. Chem.* **21**(3):315–320 (1975).
3. P. J. Cornbleet and M. C. Shea. Comparison of product moment and rank correlation coefficients in the assessment of laboratory method-comparison data. *Clin. Chem.* **24**(6):857–861 (1978).
4. M. Thompson. Regression methods in the comparison of accuracy. *Analyst* **107**:1169–1180 (1982).
5. M. J. Bookbinder and K. J. Panosian. Using the coefficient of correlation in method-comparison studies. *Clin. Chem.* **33**(7):1170–1176 (1987).
6. R. N. Carey, S. Wold, and J. D. Westgard. Principal component analysis: An alternative to "referee" methods in method comparison studies. *Anal. Chem.* **47**(11):1824–1829 (1975).
7. D. M. Holland and F. F. McElroy. Analytical method comparisons by estimates of precision and lower detection limit. *Environ. Sci. Technol.* **20**(11):1157–1161 (1986).
8. W. C. Griffiths, P. Camara, I. Diamond, and J. C. Pezzullo. A procedure for estimating bias between quantitative analytical methods. *J. Autom. Chem.* 8(3):147–150 (1986).
9. J. D. Haynes, J. Pauls, and R. Platt (chairman). Statistical Aspects of a Laboratory Study for Substantiation of the Validity of an Alternate Assay Procedure: The Greenbrier Procedure. Final Report of the Standing Committee on Statistics to the Pharmaceutical Manufacturer's Association (Quality Control Section), Washington, D.C., March 14, 1977.
10. M. J. Cardone, S. A. Willavize, and M. E. Lacy. Method validation revisited: A chemometric approach. *Pharm. Res.* 7(2):154–160 (1990).
11. J. C. Miller and J. N. Miller. *Statistics for Analytical Chemistry,* 2nd ed., Ellis Horwood, Chichester, England, 1988, pp. 75–77.
12. R. Caulcutt and R. Boddy. *Statistics for Analytical Chemists,* Chapman and Hall, New York, 1983.
13. SAS Institute Inc. *SAS Language Guide for Personal Computers, Release 6.03 Edition,* SAS Institute Inc., Cary, N.C., 1988.