Journal of Pharmacokinetics and Biopharmaceutics, Vol. 12, No. 1, 1984

PHARMACOMETRICS

A New Statistical Procedure for Testing Equivalence in Two-Group Comparative Bioavailability Trials

Walter W. Hauck[1,3] and Sharon Anderson[2]

Received July 13, 1983—Final December 16, 1983

The clinical problem of testing for equivalence in comparative bioavailability trials is restated in terms of the proper statistical hypotheses. A simple t-test procedure for these hypotheses has been devloped that is more powerful than the methods based on usual (shortest) and symmetric confidence intervals. In this note, this new procedure is explained and an example is given, including the method for sample size determination.

KEY WORDS: bioavailability; bioequivalence; hypothesis tests; sample size determination.

INTRODUCTION

The bioequivalence problem has received considerable attention recently in the statistical literature (1–13). Included in much of this work has been the recognition that the straightforward ANOVA F-test is not appropriate for the bioequivalence problem. By testing the null hypothesis of exact equality, the *F*-test is testing the wrong hypothesis. (Regardless, the *F*-test is still in use; see ref. 14 as an example.) Our purpose here is to report a new approach (1) that leads to a simple test procedure. The basis for the new procedure is a reformulation of the statistical hypotheses to correspond to the nature of the bioequivalence problem. We first describe and demonstrate the rationale and the method and then discuss the implications for design of comparative bioavailability studies. Finally, we compare the new method with others.

Preparation of this manuscript was supported in part by Grant # CA15145 from the National Cancer Institute.

[1]Northwestern University Cancer Center and Department of Community Health and Preventive Medicine, Northwestern University Medical School, Chicago, I11.

[2]American Critical Care, McGaw Park, Ill., and Department of Epidemiology and Biometry, University of Illinois, School of Public Health, Chicago, Ill.

[3]Address correspondence and reprint requests to Dr. Walter W. Hauck, Northwestern University Cancer Center, Biometry Section, Ward 3-332, 303 E. Chicago Ave., Chicago, IL 60611.

0090-466X/84/0200-0083$03.50/0

THE METHOD

Comparative bioavailability trials are conducted primarily for two reasons: either a new formulation of a standard drug is to be compared to the original formulation, as is the case in generic drug testing; or a new form of administration of a drug is to be compared to the presently marketed formulation. In either case, the objective is to demonstrate that the bioavailabilities, the rate and extent of absorption of the parent drug or metabolite in the circulatory system, of the two formulations do not differ.

In practice it is recognized that no two formulations will result in bioavailability profiles which are exactly alike. Therefore, clinically determined and meaningful limits are specified such that if the two formulations differ by less than the specified limits, the drugs are said to be bioequivalent. Let M_S and M_E represent the population means of one of the parameters of bioavailability such as the peak plasma level, time to peak level, area under the plasma level time curve, or urinary recovery, for the standard and experimental formulations, respectively. In notational terms, the two formulations are considered bioequivalent if $A_0 < M_E/M_S < B_0$ for some A_0 and B_0. Often these limits are taken to be within 20% of the standard (15), in which case, $A_0 = 0.8$ and $B_0 = 1.2$.

Data from comparative bioavailability trials are analyzed, at least in part, by an analysis of variance (ANOVA) appropriate for the design of the trial. For example, an ANOVA for a crossover study will have terms for subjects, order (group), and formulation as well as the error (residual) term. As part of the ANOVA there will typically be an F-test for formulations. This test is of the hypothesis that the two formulations are equal, i.e., $M_E = M_S$. But equality and bioequivalence are different concepts, and so the ANOVA test does not address the bioequivalence issue. Note in particular that the equivalence limits, A_0 and B_0, nowhere enter into the ANOVA.

To motivate our new test procedure, the inappropriateness of the ANOVA test for the bioequivalence problem must be addressed in statistical terms. Classical statistical hypothesis testing, which we are concerned with here, is very asymmetric. One states two hypotheses, a null hypothesis (denoted H_0) and an alternative hypothesis (H_A). A statistical test of significance then demonstrates the likelihood of the alternative hypothesis by measuring the strength of evidence *against* the null hypothesis. If the evidence is sufficiently strong, one rejects the null hypothesis in favor of the alternative. If the evidence is not sufficiently strong, one fails to reject the null hypothesis, but this is not evidence *for* the null hypothesis.

In the ANOVA, the test for formulations is specifically a test of the null hypothesis that the average bioavailabilities of the formulations are equal (H_0: $M_E = M_S$) against the alternative that they differ (H_A: $M_E \neq M_S$). If

the study is sufficiently sensitive, it is possible to detect significant differences of little clinical meaning. Alternatively, if the study is sufficiently insensitive, large differences will not be detected; the evidence will not be sufficiently strong to accept the alternative hypothesis.

When one has a hypothesis to demonstrate, the logic of hypothesis testing therefore requires that the hypothesis to be demonstrated be the *alternative* hypothesis. In the case of demonstrating equivalence, this means that the equivalence hypothesis should be the alternative and *not* the null hypothesis. The statistical hypotheses for the bioequivalence problem can then be stated as

$$H_0': \quad \frac{M_E}{M_S} \leqslant A_0 \quad \text{or} \quad \frac{M_E}{M_S} \geqslant B_0$$

and

$$H_A': \quad A_0 < \frac{M_E}{M_S} < B_0$$

We will now formally restrict ourselves to considering the case where a single, unspecified measure of bioavailability is of interest. Since, in applications such as comparative bioavailability studies, the data to be analyzed will often be logarithms of biological measures (10) such as the logarithm of the area under the plasma-time curve (AUC) or peak plasma level, we let μ_E and μ_S be the mean values in the *logarithmic* scale for subjects receiving the experimental and standard formulations. The equivalence hypotheses are then

$$\begin{aligned} &H_0: \quad \mu_E - \mu_S \leqslant A \quad \text{or} \quad \mu_E - \mu_S \geqslant B \\ \text{and} \qquad &H_A: \quad A < \mu_E - \mu_S < B \end{aligned} \tag{1}$$

where $A = \log A_0$ and $B = \log B_0$. The test statistic we will consider is

$$T = \frac{\bar{X}_E - \bar{X}_S - \frac{1}{2}(A+B)}{S(1/n_E + 1/n_S)^{1/2}}$$

where the $\bar{X}$'s are the respective sample means (in the logarithmic scale), n_E and n_S are the group sample sizes, and S is the error standard deviation calculated from the appropriate analysis of variance with degrees of freedom ν. (If $n_E = n_S = N$, then $\nu = N-2$ for a crossover design, and $\nu = 2N-2$ for a completely randomized design.) T can be seen as a measure of how far the difference in means, $\bar{X}_E - \bar{X}_S$, is from the center of the equivalence interval, $\frac{1}{2}(A+B)$. The desired test is to reject H_0 in favor of bioequivalence if the magnitude of T is sufficiently small, that is, if $|T| < C$ for some critical value C; the closer T is to zero, the more the data support the equivalence hypothesis. The distribution of T is, in general, not known, so C cannot

be found exactly. In (1) we've shown that the usual Student's t distribution can be used as a good approximation. Rather than determine the C appropriate for a given level test, we approach the problem through descriptive levels of significance, most commonly known as p values. A $100\alpha\%$ test is conducted by comparing the p value, ρ, to the chosen level of the test, α, and rejecting the null hypothesis if $\rho \leq \alpha$.

For the bioequivalence problem, the p value is the probability of obtaining a value of $|T|$ equal to or smaller than that observed. For problems such as this one, with null hypotheses composed of intervals, the significance level of the test is found by maximizing the type I error over the null hypothesis interval. The p value is defined similarly. For our procedure this means the p-value probability is calculated assuming that the true difference is at the boundary of the equivalence interval. (Also note that this is the reverse of most common statistical tests where the null hypothesis is rejected for sufficiently large values of the test statistic.) Skipping the nonilluminating derivation, the approximation to the p value is

$$\rho = F_\nu(|T| - \delta) - F_\nu(-|T| - \delta)$$

where

$$\delta = \frac{B - A}{2S(1/n_E + 1/n_S)^{1/2}} \tag{2}$$

and F_ν is the distribution function for Student's t with ν degrees of freedom. δ may be considered a standardized equivalence interval width; the smaller δ is, the more difficult it will be to conclude equivalence.

Before proceeding to an example, there are three notes. First, we have not needed to state the trial design. The S and ν are those from the appropriate ANOVA. Second, we have made the implicit assumption that the data, after the logarithmic transformation, follow a normal distribution with constant variance. The normality assumption is more reasonable when working with logarithms of AUCs than with the AUCs themselves. Third, calculation of ρ does require a computer program or good tables for the t distribution function. The calculations for the example in the next section were done using a program included in a statistics package for a pocket calculator.

EXAMPLE

To illustrate use of our method, we consider the analysis of data from a bioequivalence trial (14) that was also considered by Metzler and Huang (6). In Clayton and Leslie's Study II, erythromycin stearate taken immediately after a meal was compared to erythromycin base also taken immediately

after a meal in a balanced crossover design with 18 subjects. There was a one-week washout period separating the administration of the two formulations. The average AUCs were 4.117 and 5.231 mcg · hr · ml^{-1} for the stearate and base, respectively, so the availability of the stearate is 79% that of the base.

The ANOVA for base 10 logarithms of AUC is given in Table I. (Order is not included since the group membership of each subject was not given in ref. 14.) The F-test for formulations yields a p value of 0.08 indicating that the two formulations do *not* differ at the conventional 5% level. After the corresponding analysis in the original scale (yielding a p value of 0.18, but the normality assumption is very questionable), Clayton and Leslie concluded that the "two preparations are bioequivalent."

In the base 10 logarithm scale, the data are

$$\bar{X}_E = 0.512 \quad \text{(stearate)}$$

$$\bar{X}_S = 0.661 \quad \text{(base)}$$

$$S^2 = 0.057 \quad \text{(from the ANOVA; } \nu = 17\text{)}$$

To test H_A: $0.8 < M_E/M_S < 1.2$ by our method, we need T and δ:

$$T = \frac{\bar{X}_E - \bar{X}_S - \frac{1}{2}(A+B)}{S(2/N)^{1/2}}$$

$$= \frac{0.512 - 0.661 - \frac{1}{2}[\log_{10}(1.2) + \log_{10}(0.8)]}{\sqrt{0.057}\sqrt{2/18}}$$

$$= -1.984$$

$$\delta = \frac{\frac{1}{2}(B-A)}{S(2/N)^{1/2}}$$

$$= \frac{\frac{1}{2}[\log_{10}(1.2) - \log_{10}(0.8)]}{\sqrt{0.057}\sqrt{2/18}}$$

$$= 1.106$$

Table I. ANOVA for Log_{10} AUC

Source	DF	SS	MS	F	p value
Formulations	1	0.19748	0.19748	3.46	0.0801
Subjects	17	1.81986	0.10705	1.88	0.1022
Error	17	0.96943	0.05703		
Total	35	2.98677			

Then, using the t distribution with 17 degrees of freedom,

$$\begin{aligned}\rho &= F_{17}(|T|-\delta) - F_{17}(-|T|-\delta) \\ &= F_{17}(0.878) - F_{17}(-3.090) \\ &= 0.8039 - 0.0033 \\ &= 0.8006.\end{aligned}$$

We would thus be very *unwilling* to accept the hypothesis of equivalence (no more than 20% difference), since $\rho = 0.80$, indicating that the observed value of T is not unlikely from nonbioequivalent formulations. The great difference between the conclusion drawn from our approach (and that would also be reached by the confidence interval methods) and that of Clayton and Leslie serves to emphasize the inappropriateness of the usual F-test for the bioequivalence problem.

SAMPLE SIZE DETERMINATION

In order to design comparative bioavailability studies, one needs a method for determining an appropriate number of subjects. With the test procedure described above, a standard method of sample size determination is applicable. We assume that a prior estimate, σ, of the error standard deviation is known and that the study is designed so that each of the two formulations will be given to N subjects. In a crossover design there will thus be N subjects total, and a total of $2N$ subjects in a randomized block design. In addition we assume that the experimenters can specify a probability β, the probability of *not* accepting bioequivalence when there is in fact no difference (specifically $\mu_E - \mu_S = (A+B)/2$. Commonly $\beta = 0.1$ or 0.2. The power of the test is $1-\beta$, the probability of concluding that the two formulations are bioequivalent when there is no difference.

Given β, the value of C for concluding equivalence is found from

$$Pr[|T| < C] = 1 - \beta$$

When $\mu_E - \mu_S = (A+B)/2$, T has Student's t distribution, so $C = t_{\nu;\beta/2}$, the upper $\beta/2$ percentage part of the central t distribution on ν degrees of freedom. The sample size is then determined by adjusting δ to make the test the desired $100\alpha\%$ level for specified α. Once a value for δ, say $\tilde{\delta}$, is obtained, $\tilde{\delta}$ can be solved for $\tilde{N}$, the required number of subjects:

$$\tilde{N} = \tilde{\delta}^2(2\sigma^2) \Big/ \left(\frac{B-A}{2}\right)^2 = 8\tilde{\delta}^2\sigma^2/(B-A)^2$$

As happens in similar situations, if $\tilde{N}$ and the initial choice of ν (and hence

of C) are not compatible, the process can be repeated with a new ν determined from $\tilde{N}$ until compatible values are found.

As an example of a sample size determination for this method, we again use the data from ref. 14. We supposed that a sample size was required for a crossover design that would have a power of 0.8 for equivalence defined as

$$0.8 < \frac{M_E}{M_S} < 1.2$$

The S^2 from the ANOVA was 0.057, and we took this to be a previous estimate of σ^2 to be used in a sample size determination. For this example, $A = \log_{10}(0.8)$ and $B = \log_{10}(1.2)$.

The guesses at N and subsequent solutions for $\tilde{N}$ are shown in Table II for $\alpha = 0.05$. Two program runs (iterations) were needed, with multiple guesses in each iteration. After the first iteration, we know that 126 or 127 subjects would be required. The second run confirms this to be 127, or actually 128 for a balanced crossover design. The sample size calculations in Table II were done on the CYBER 170/730 at Northwestern University's Vogelback Computing Center. The IMSL subroutine MDTN was used for the noncentral t distribution function, MDNOR for the standard normal distribution function, and ZFALSE for numerically solving for $\tilde{\delta}$. A copy of the program (but not including the IMSL subroutines) is available on request (to W.H.).

Since comparative bioavailability studies are commonly conducted with about 20 subjects, a required sample size of 128 may be disconcerting. From a mathematical point of view, the large sample size reflects the variability in Clayton and Leslie's data. As in other statistical applications, the greater the variation, the greater the sample size required for a given α and β. Here $\tilde{N}$ is roughly proportional to σ^2. If Clayton and Leslie's variance were reduced by a factor of four (to 0.014), the required sample size would be only 34.

The penalty for conducting a study with too few subjects is low power. For example, taking $\sigma^2 = 0.057$ and using our method, Clayton and Leslie

Table II. Sample Size Determination Example

Iteration	Guess at $\tilde{N}(\nu+2)$	Solution for $\tilde{N}$
1	Infinity	126.0
	122	126.8
	62	127.7
	22	131.4
2	126	126.7
	127	126.7

had only about a 10% chance of concluding equivalence if there were no difference between the formulations. In ref. 1 we have shown that our procedure is more powerful than either of the two commonly used confidence interval procedures (symmetric and shortest) (11,13). Consequently, with these procedures one is less likely to conclude equivalence, or alternatively, they would require more subjects (for given α, β, and σ^2). Power comparisons are discussed further below.

DISCUSSION

A key step for progress in developing methodology for testing equivalence is the statement of the proper null and alternative hypotheses (Eq. 1). Once that is done, the ANOVA test of equality is seen to be inappropriate. We have proposed a statistic for the proper hypothesis that is easy to calculate; the p value for our statistic can be easily calculated using commonly available tables and/or computer programs. In addition, sample size determinations necessary for design of trials of equivalence are possible when using our approach.

Our method is an alternative to the two confidence interval approaches. In the usual (asymmetric) confidence interval method (13), bioequivalence can be concluded at the α% level if the $100(1-2\alpha)$% confidence interval lies wholly within the bioequivalence interval (A, B). [Note if the $100(1-\alpha)$% confidence interval is used, the nominal level of the test is $\alpha/2$, not α.] For the symmetric confidence interval method (11), the $100(1-\alpha)$% symmetric confidence interval must lie within the bioequivalence interval. As noted in ref. 13 and implicit in ref. 4, for every value of the parameter, the usual confidence interval method is more powerful than the symmetric interval method, that is, it has a greater chance of concluding bioequivalence. In ref. 1 we show that our method is similarly uniformly more likely to conclude bioequivalence than is the usual confidence interval method, so our method is the most powerful of these three methods.

It is important to note that the above power comparison is based on comparing three nominal α-level tests. None of the three tests is exactly α-level. The two confidence interval methods are conservative since their actual level is always less than the stated α (or equivalently that the calculated p values are too large), reaching α only as δ (Eq. 2) becomes large. As a corollary to the power comparison, the actual level of our test is always greater than that of the asymmetric intervals, which, in turn, are always greater than that of the symmetric intervals. Schuirmann (7) has shown that the usual confidence interval method can be extremely conservative, with actual levels near 0 for nominal 1% and 5% tests. In the extensive simulation study (1) that supports the validity of our method, we obtained

actual levels ranging from 3.9% to 6.2% for nominal 5% tests and from 0.5% to 1.7% for nominal 1%.

In summary, our method is more powerful than the two confidence interval methods. In part this is due to the avoidance of the sometimes extreme conservatism of the confidence intervals.

REFERENCES

1. S. Anderson and W. W. Hauck. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Comm. Stat.* **A12:**2663–2692 (1983).
2. C. W. Dunnett and M. Gent. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics* **33:** 593–602 (1977).
3. T. B. L. Kirkwood. Bioequivalence testing—a need to rethink. *Biometrics* **37:** 589–591 (1981).
4. D. Mandallaz and J. Mau. Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **37:** 213–222 (1981).
5. C. M. Metzler. Bioavailability—a problem in equivalence. *Biometrics* **30:** 309–317 (1974).
6. C. M. Metzler and D. C. Huang. Statistical methods for bioavailability and bioequivalence. *Clin. Res. Pract. Drug Res. Affairs* **1:** 109–132 (1983).
7. D. J. Schuirmann. Fixed sample tests for interval hypotheses associated with bioequivalence trials. Presented at Joint Statistical Meetings, Cincinnati, August, 1982.
8. M. R. Selwyn, A. P. Dempster, and N. R. Hall. A Bayesian approach to bioequivalence for the 2×2 changeover design. *Biometrics* **37:** 11–21 (1981).
9. W. J. Westlake. Use of confidence intervals in analysis of comparative bioavailability trials. *J. Pharm. Sci* **61:** 1340–1341 (1972).
10. W. J. Westlake. The design and analysis of comparative blood-level trials. In J. Swarbrick (ed), *Current Concepts in the Pharmaceutical Sciences, Dosage Form Design and Bioavailability*, Lea & Febiger, Philadelphia, 1973, pp. 149–179.
11. W. J. Westlake. Symmetric confidence intervals for bioequivalence trials. *Biometrics* **32:** 741–744 (1976).
12. W. J. Westlake. Design and statistical evaluation of bioequivalence studies in man. In J. Blanchard, R. W. Sawchuk, and B. B. Brodie (ed), *Principles and Perspectives in Drug Bioavailability*, Karger, Basel, 1979, pp. 192–210.
13. W. J. Westlake. Response to bioequivalence testing—a need to rethink. *Biometrics* **37:** 591–593 (1981).
14. D. Clayton and A. Leslie. The bioavailability of erythromycin stearate versus enteric-coated erythromycin base when taken immediately before and after food. *J. Int. Med. Res.* **9:** 470–477 (1981).
15. Food and Drug Administration. The bioavailability protocol guideline for ANDA and NDA submission. Division of Biopharmaceutics, Drug Monographs/Bureau of Drugs, Food and Drug Administration, 1977.