

## Non-inferiority trials: the ‘at least as good as’ criterion

Larry L. Laster<sup>1,\*;†</sup> and Mary F. Johnson<sup>2</sup>

<sup>1</sup>*University of Pennsylvania; School of Veterinary Medicine; 3900 De Lancey St.; Philadelphia; PA 19104; U.S.A.*

<sup>2</sup>*PharmaNet; Inc.; 504 Carnegie Center; Princeton; NJ 08540; U.S.A.*

### SUMMARY

To demonstrate in a clinical trial that a new or experimental therapy (et) is ‘at least as good as’ a standard therapy (st), a statistical test or confidence interval procedure must rule out clinical inferiority with a high probability. The term ‘at least as good as’ implies equivalent but not necessarily superior efficacy. As it is statistically impossible to demonstrate equivalence (that is, prove the null hypothesis of no difference), Blackwelder proposed a one-sided significance test to reject the null hypothesis that standard therapy is better than experimental therapy by a clinically acceptable amount,  $BW$ . In this paper, Blackwelder’s approach is redefined in terms of the ratio of two means ( $R_{True} = \frac{et}{st}$ ) based on a continuous variate with higher values denoting greater improvement. The ratio-based equivalents to Blackwelder’s hypotheses will be shown. The ratio parameter has the benefit of being available as a dimensionless percentage, not tied to a specified difference in means. Thus, a study can be sized to assure, with high probability, that the experimental therapy is ‘at least’ ( $R_{LB} \times 100$ ) per cent ‘as effective as’ the standard therapy, where  $R_{LB}$  is the selected lower bound on the percentage effectiveness. A practical rationale is given for defining non-inferiority as a high fraction or percentage of the standard drug’s efficacy, both in terms of statistical efficiency and medical relevance. For most typical ‘at least as good as’ applications (when  $R_{LB} \leq 1$ ), the ratio formatted test of  $H_0: R_{True} \leq R_{LB}$  is shown to be more efficient than Blackwelder’s test of  $H_0: st - et \geq BW$ , thereby requiring smaller sample sizes to detect the directionally based non-null alternatives contained in  $H_1: \frac{et}{st} \geq R_{LB}$  or, equivalently,  $st - et \leq BW$ . Further, when  $R_{True} = 1.0$ , tests of Blackwelder’s hypotheses, their ratio-based equivalents and conventional superiority can be evaluated for comparative efficiency. Testing  $H_0: R_{True} \leq R_{LB}$  with single-sided critical region of size  $\alpha$ , versus  $H_1: R_{True} \geq R_{LB}$ , is shown to be more efficient than excluding  $R_{LB}$  from the lower limit of a  $100(1-2\alpha)$  per cent two-sided symmetric confidence interval centred by  $\hat{R}$ . Relevant examples will be presented. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: non-inferiority; sample size; hypothesis testing; confidence interval; ratio estimator

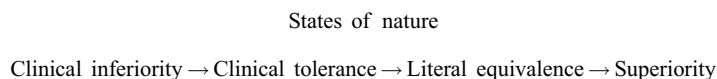
### INTRODUCTION

Clinical studies employing active or positive controls are often intended to establish that a new, experimental therapy (et) is ‘at least as good as’ the active, standard therapy (st).

\*Correspondence to: Larry L. Laster, University of Pennsylvania; School of Veterinary Medicine; 3900 De Lancey St.; Philadelphia; PA 19104; U.S.A.

†E-mail: larryl@vet.upenn.edu

The actual (true) effect of experimental therapy in contrast to standard therapy may be represented schematically as follows:



The terminology 'at least as good as' or equivalently, non-inferiority, may be interpreted as either literal equivalence or superiority. Since the statistical demonstration of literal equivalence is fruitless (that is, proving the null hypothesis of no difference), an operational definition must be considered which allows experimental therapy to be inferior to standard therapy by a clinically tolerable amount. The choice of this limit will depend on the nature of the illness treated, toxicity concerns and other risk-benefit issues. Clearly, clinicians must make this call.

In a superiority trial, the traditional hypothesis-testing framework would seek to reject the null hypothesis of equivalence with adequate power to detect some minimum, clinically meaningful difference in favour of the experimental therapy. To satisfy the 'at least as good as' criterion, and thus demonstrate clinical non-inferiority, a statistical test or confidence interval must rule out (with high probability) clinical inferiority of the experimental therapy. Such trials are designed to detect the composite alternative hypothesis 'at least as good as', now clearly defined by the union of the three possible states of nature: clinical tolerance, literal equivalence, and superiority.

By modifying the conventional null and alternative hypotheses in this way, Blackwelder [1, 2] expanded on the initial work of Makuch and Simon [3] to develop testing procedures and sample size requirements for non-inferiority trials. Using a dichotomous outcome variable as a model for presentation, Blackwelder proposed a one-sided significance test to assure, with high probability, that the difference in proportions favouring standard therapy over experimental therapy is no more than a specified clinically acceptable amount  $\delta_{BW}$  (see Blackwelder's table 1, reference [1]).

Previous work was confined to the bioequivalence framework [4–12] and approached this problem with a two-sided solution to demonstrate that the effect of a new formulation did not differ substantially in either direction from that of the original formulation. Subsequent work by Schuirmann [13], Munk [14], Berger and Hsu [15] and Brown *et al.* [16], examined different powering techniques and testing procedures for the bioequivalence approach. Metzler [17] also considered sample size projections in bioequivalence studies. Work in individual bioequivalence may be seen in Anderson and Hauck [18], Hwang and Wang [19] or more recently in Wang [20], where the hypothesis of within-patient bioequivalence is examined. Multivariate analogues in bioequivalence may be seen in Wang *et al.* [21] or Munk and Pfluger [22].

More recently, Holmgren [24] proposed a procedure using the relative risk to establish equivalence between a new treatment and an active control based on a specified percentage of the effect of the active control over that of an historical placebo. Emphasis was placed on the extent to which the benefit of the active control over placebo, as estimated from previous studies, is maintained by the new treatment.

In this paper, Blackwelder's general approach to one-sided equivalence testing will be extended to a ratio definition of the percentage effectiveness ( $\theta_{et=st}$ ) based on mean values of a *continuous response* variate. The ratio parameter has the benefit of being available as a dimensionless percentage, easily estimated and compared among studies, and not tied to a specified interval of clinical tolerance which must be selected for power and sample size

calculations. Instead, a study can be sized to assure, with high probability, that the experimental therapy is 'at least' ( $R_{LB} \times 100$ ) per cent 'as effective as' the standard therapy, where the selected lower bound on the ratio of mean effects ( $R_{LB}$ ) is dictated by clinical and/or regulatory considerations. Expressing non-inferiority as a high fraction or percentage of the standard drug's efficacy offers a useful option for planning and interpreting active control trials when the selection of an absolute value for clinical tolerance might be considered arbitrary or controversial. The ratio-based equivalents to Blackwelder's hypotheses will be shown.

This extension to Blackwelder [1], redefines the 'at least as good as' hypotheses in terms of a ratio parameter, based on a *relative difference* instead of *absolute difference*, and compares the efficiencies of the two approaches. The relative merits of other approaches (hypothesis testing versus confidence intervals, and tests of non-inferiority versus superiority) will also be discussed by contrasting sample size and efficiency formulations in the different formats. Some relevant examples will be presented.

### HYPOTHESIS TESTING

Blackwelder [1] introduced a single-sided null hypothesis for clinical inferiority to be rejected in favour of the 'at least as good as' hypothesis, here defined in terms of a continuous response variate (with higher mean values denoting greater improvement):

$$H_0: \mu_{st} - \mu_{et} \geq \Delta_{BW} \text{ versus } H_1: \mu_{st} - \mu_{et} < \Delta_{BW} \tag{1}$$

where  $\Delta_{BW}$  represents the clinical tolerance selected. A trial designed in this framework would be successful if the outcome with the test therapy was no worse than the outcome with the active control, by some clinically tolerable amount,  $\Delta_{BW}$ . To envision more clearly the nature of these hypotheses, let us display the hypothetical regions under the four states of nature depicted above:

States of nature (experimental versus standard therapy)

Clinical inferiority	→ Clinical tolerance	→ Literal equivalence	→ Superiority
$\mu_{st} - \mu_{et} \geq \Delta_{BW}$	$0 \leq \mu_{st} - \mu_{et} < \Delta_{BW}$	$\mu_{st} - \mu_{et} = 0$	$\mu_{et} - \mu_{st} > 0$
Experimental Therapy inferior by $\Delta_{BW}$ or more	Experimental Therapy inferior by less than $\Delta_{BW}$	Therapies Equivalent	Experimental Therapy superior:

Any difference less than  $\Delta_{BW}$  would have to be considered acceptable. For sizing a study and performing significance tests, the Blackwelder approach would reject the null hypothesis that the control is superior to the test therapy by  $\Delta_{BW}$  or more, in favour of the alternative that the control is better than the test therapy by less than  $\Delta_{BW}$ , this clinically acceptable amount.

### THE RATIO VIEW

Now let us consider the relationship between  $\mu_{st}$  and  $\mu_{et}$  as a ratio. A key benefit of this parameterization is that the clinical effectiveness of the experimental treatment can be viewed

as a dimensionless percentage (per cent) of the response to standard therapy as

$$R_{\text{True}} = \frac{\mu_{\text{et}}}{\mu_{\text{st}}} \tag{2}$$

Further, a firm estimate of  $\mu_{\text{st}}$  will generally be available within the ‘at least as good as’ paradigm. This estimate will be required in order to size studies with ratio-formatted hypotheses. The ratio-based equivalents to Blackwelder’s hypotheses are

$$H_0: \frac{\mu_{\text{et}}}{\mu_{\text{st}}} \leq R_{\text{LB}} \quad \text{versus} \quad H_1: \frac{\mu_{\text{et}}}{\mu_{\text{st}}} > R_{\text{LB}} \tag{3}$$

where  $R_{\text{LB}}$  is a selected *lower bound* based on a percentage of  $R_{\text{T}}$  indicating an allowance for clinical tolerance, which would usually be taken to be fairly *large* (say 80 per cent, 90 per cent etc). Again, to show ‘at least as good as’, reject  $H_0$  in favour of  $H_1$ .

The Blackwelder and ratio views of the hypotheses postulated are actually identically equivalent when  $BW$  is taken as a *small* percentage of  $\mu_{\text{st}}$  in terms of  $(1 - R_{\text{LB}})$ . To see this simply set

$$BW = (1 - R_{\text{LB}}) \mu_{\text{st}} \tag{4}$$

in  $H_0: \mu_{\text{st}} - \mu_{\text{et}} \geq BW$ , where  $\mu_{\text{st}} > 0$ , to see the result

$$\begin{aligned} \mu_{\text{st}} - (1 - R_{\text{LB}}) \mu_{\text{st}} &\geq \mu_{\text{et}} \\ \mu_{\text{st}} - \mu_{\text{st}} + R_{\text{LB}} \mu_{\text{st}} &\geq \mu_{\text{et}} \\ R_{\text{LB}} \mu_{\text{st}} &\geq \mu_{\text{et}} \end{aligned} \tag{5}$$

or

$$\frac{\mu_{\text{et}}}{\mu_{\text{st}}} \leq R_{\text{LB}}$$

that is, of course, the ratio-based  $H_0$  seen above in (3). Note, as in Blackwelder [1], to justify the ‘as least as good as’ application,  $BW$  must be positive, and thus  $R_{\text{LB}} \geq 1$  as defined here. As Blackwelder points out,  $BW$  could in theory be zero or negative, thus  $R_{\text{LB}} \geq 1$ , but these cases are atypical for non-inferiority testing.

The development, here, will continue in terms of a continuous response variate for which increases denote *improvement*. If smaller metrics were to represent *improvement*, the inequalities would simply be reversed in the hypotheses above and the ratio of means referenced to a  $R_{\text{Upper Bound}}$  instead.

### TESTING AND CONFIDENCE INTERVAL PROCEDURES

In the ratio format to demonstrate ‘at least as good as’ either reject

$$H_0: \frac{\mu_{\text{et}}}{\mu_{\text{st}}} \leq R_{\text{LB}}$$

with a single-sided significance test with critical region of size  $\alpha$ , or exclude  $R_{\text{LB}}$  with the lower limit of a  $100(1 - 2\alpha)$  per cent two-sided symmetric confidence interval for the true ratio

$$R_{\text{True}} = \frac{\mu_{\text{et}}}{\mu_{\text{st}}} \tag{2}$$

The estimator of  $R$ ,  $\hat{R} = \frac{\bar{X}_{\text{et}}}{\bar{X}_{\text{st}}}$ , is asymptotically both unbiased and normally distributed (given finite variances). The  $X_{\text{st}(i)}$ ;  $i = 1; \dots; n$  and  $X_{\text{et}(i)}$ ;  $i = 1; \dots; n$ ; are assumed to be each

independently distributed  $N(\bar{x}_{st}; \sigma^2)$  or  $N(\bar{x}_{et}; \sigma^2)$  variables with common variance  $\sigma^2 = \sigma_{et}^2 = \sigma_{st}^2$  and equal sample sizes  $n_{et} = n_{st} = n$ . Note that the sample sizes are fixed equal, simply for the purpose of sample size projection, whereas in general they need not be so. Unfortunately, the estimator  $\hat{R}$  is well known to have some difficulties associated with it. For one, the distribution theory is exceedingly complicated and not well suited to confidence interval construction or hypothesis testing (see Miller [26]).

However, a reparameterization of the ratio-based hypotheses seen in (3), due to Paulson [27]

$$H_0: \bar{x}_{et} - R_{LB} \bar{x}_{st} \leq 0 \quad \text{versus} \quad H_1: \bar{x}_{et} - R_{LB} \bar{x}_{st} > 0 \tag{6}$$

results in inferiority–non-inferiority contrasts now seen as the *difference between the mean experimental response and a high proportion or fraction ( $R_{LB}$ ) of the standard mean response*. Note, when  $R_{BW}$  is selected as a small part of  $\bar{x}_{st}$ ,  $R_{LB}$  is the complementary larger part of it. In this form the distribution theory is considerably more tractable, and allows directly for the construction of the uniformly most powerful unbiased test (UMPU, see Lehman [23])

$$(\bar{X}_{et} - R_{LB} \bar{X}_{st}) = [S^2(1 + R_{LB}^2) = n]^{1/2} \tag{7}$$

as Student's t where  $S^2$  is the two independent sample pooled estimate of  $\sigma^2$  with  $2n - 2$  degrees of freedom.

The case for lower-limit exclusionary rules (to exclude  $R_{LB}$ ) based on observed two-sided symmetric confidence intervals centred by  $\hat{R}$  (not an uncommon practice) is somewhat more complicated. The  $100(1 - 2\alpha)$  per cent Fieller-like [25] confidence interval for all values of  $R$  is based on the ratio

$$(\hat{R} - R) = [S^2(1 + R^2) = n(\bar{X}_{st})^2]^{1/2} \tag{8}$$

which should not exceed the critical constant  $t_{\alpha, 2n-2}$  in absolute value (see Miller [26]). Now, the values of  $R$  where the ratio (8) actually equals the critical constant  $t_{\alpha, 2n-2}$ , are the roots of a *complex quadratic equation*. However, for large sample sizes, the roots are approximately

$$\hat{R} \pm z [S^2(1 + \hat{R}^2) = n(\bar{X}_{st})^2]^{1/2} \tag{9}$$

where  $z$  is a single tailed normal deviate and the bracketed quantity it multiplies, the delta method estimate\* of the  $SE(\hat{R})$ . In practice, it is this approximation that is most often used.

The  $100(1 - 2\alpha)$  per cent two-sided symmetric confidence interval (equation (9)) is centred by  $\hat{R}$  and has  $\hat{R}$  contained in an estimate of its own standard error. Obviously,  $\hat{R}$  has to be observed larger than  $R_{LB}$  if its associated interval is to have any chance of excluding  $R_{LB}$ . Thus, the  $SE(\hat{R})$  will be larger using the confidence interval approach based on  $\hat{R}$ , in comparison to testing  $H_0: R_{True} \leq R_{LB}$  (or equivalently  $H_0: \bar{x}_{et} - R_{LB} \bar{x}_{st} \leq 0$ ) with single-sided critical region of size  $\alpha$ , where the  $SE(\hat{R})$  would be evaluated at  $R_{True} = R_{LB}$  ( $R_{LB} \geq 1$ ). The confidence interval based lower-limit exclusionary procedure for rejecting clinical inferiority, that is,  $H_0: R_{True} \leq R_{LB}$ ; as just defined, would in fact have the wrong size. Testing, therefore,  $H_0: \bar{x}_{et} - R_{LB} \bar{x}_{st} \leq 0$  using a single-sided Student's t-test (as seen in equation (7)) with  $2n - 2$  degrees of freedom, will be generally *more efficient*.

\*Linearization based on first-order Taylor's series expansion.

SAMPLE SIZE REQUIREMENTS IN THE RATIO FORMAT

For moderate to large samples then (for approximate normality), the solution is<sup>†</sup>

$$n_{\text{each group}} = [(CV)^2(z_{1-\alpha} - z)^2(1 + R_{LB}^2)](R_T - R_{LB})^2 \tag{10}$$

( $R_{LB} | R_T$ ) which is based on the optimal reparameterization due to Paulson [27] (see Appendix, Section A1 for formulation) where the CV is formed from  $\mu_{st}$  with  $z_{1-\alpha}$  and  $z$  the normal deviates, producing the chosen single-sided probabilities under the operationally specified hypotheses  $H_0: \mu_{st} - R_{\text{Lower Bound}} \leq 0$  and  $H_1: \mu_{st} - R_{\text{Lower Bound}} > 0$ .

*An example*

Consider planning a randomized clinical study of a new anti-hypertensive therapy known to produce fewer side-effects than a standard therapy but expected to be equally effective ( $R_{\text{True}}=1:0$ ). To accept the new therapy, clinicians want a high degree of assurance that it is at least 80 per cent as effective in lowering blood pressure as the standard agent. Reductions in seated diastolic blood pressure are expected to average 10 mmHg with standard therapy (standard deviation = 7.5 mmHg), for a  $CV=0.75$ . Using equation (10), a total of 284 randomized patients (142 per group) would provide 80 per cent power to reject the null hypothesis that the true ratio of mean blood pressure reductions is 0.80 or less, in favour of the alternative hypothesis, that the new agent provides over 80 per cent of the effect of the standard (single-tailed  $\alpha=0.05$ ). That is

$$n = \{(0.75)^2(1.645 + 0.84)^2(1 + 0.80^2)\} / \{1 - 0.80\}^2$$

$$n \approx 142 = \text{group}$$

With Blackwelder’s approach, the study would require 174 patients per group to rule out a difference in mean blood pressure reductions of 2.0 mmHg or greater (the clinical tolerance limit equivalent to ‘at least’ 80 per cent ‘as effective as’). See comparable entries in Table I.

COMPARATIVE EFFICIENCY

An efficiency ratio ( $EF = [1 + R_{LB}^2]^{-2}$ ) derived to compare sample size requirements for the ratio and Blackwelder approaches indicates that when  $R_{LB}$  is  $1:0$ , where  $R_{LB} | R_{\text{True}}$ , a typical situation in ‘at least as good as’ applications, the ratio format will result in greater efficiency (smaller sample sizes) than Blackwelder’s [1] approach, to detect the directionally based non-null alternatives contained in  $H_1: \mu_{st} > R_{LB}$  ( $R_{LB} | 1:0$ ) or, equivalently,  $\mu_{st} - \mu_{et} |_{BW} > 0$  as defined here (See Appendix, Section A2, for a demonstration of the comparative efficiencies).

A similar break-point for efficiency,  $(1 + R_{BW}^2)^{-2}$ , was reported by Makuch and Simon [3] as a function of the binomial proportion  $\mu_{st}$ , when comparing the conventional test of superiority ( $H_0: \mu_{st} - \mu_{et} \leq 0$ ) to detect the true difference  $\mu_{st} - \mu_{et}$ , with Blackwelder’s test of  $H_0: \mu_{st} - \mu_{et} \leq 0$ .

<sup>†</sup>We gratefully acknowledge the assistance of Mitchell Kotler, Colgate-Palmolive Co., in formulating this equation.

Table I. Sample size per group required to detect  $R_{True} \geq R_{LB}$  for selected values of  $R_{True}$ ,  $R_{Lower Bound}$  and CV with  $1 - \alpha = 0.8$  and one-tailed  $\beta = 0.05$

$R_{LB}$	CV									
	0.50		0.75		1.00		1.25		1.50	
$R_{True} = 0.9$										
0.50	12	(19)*	27	(43)	48	(77)	75	(121)	109	(174)
0.75	107	(137)	241	(309)	429	(549)	670	(858)	965	(1235)
0.80	253	(309)	570	(695)	1013	(1235)	1582	(1930)	2279	(2779)
$R_{True} = 0.95$										
0.50	10	(15)	21	(34)	38	(61)	60	(95)	86	(137)
0.75	60	(77)	136	(174)	241	(309)	377	(482)	543	(695)
0.80	113	(137)	253	(309)	450	(549)	703	(858)	1013	(1235)
$R_{True} = 1.0$										
0.50	8	(12)	17	(28)	31	(49)	48	(77)	69	(111)
0.75	39	(49)	87	(111)	154	(198)	241	(309)	347	(445)
0.80	63	(77)	142	(174)	253	(309)	396	(482)	570	(695)

\*Values seen in ( ) are sample size solutions for the Blackwelder equivalent hypothesis test to detect  $\mu_{st} - \mu_{et} \geq \delta_{BW}$  as defined here. Note: Owing to the distributional properties of the smaller sample size projections indicated here, they should be considered only approximate.

$\delta_{BW} \geq 0$  assuming the true difference  $(\mu_{st} - \mu_{et})$  is 0. In particular, sample size requirements for the two approaches are equal when  $\delta_{st} = (1 + \delta_{BW})\delta$ .

In the current application for ratio formatted hypotheses with continuous data, when  $R_{True} = 1.0$  (or  $\mu_{st} - \mu_{et} = 0$ ), Blackwelder's test and that of conventional superiority can be viewed as equivalent, when Blackwelder's  $\delta_{BW}$  (for clinical tolerance) is interpreted as the clinically superior difference (now  $\mu_{et} - \mu_{st}$ ) as  $[\mu_{et} - \mu_{st}]_{CSD}$ , and so in this instance, the comparative efficiency with Blackwelder's formulation applies to the standard test of superiority as well.

In order to judge the relative size of  $[\mu_{et} - \mu_{st}]_{CSD}$  that would be detectable in a superiority trial with the sample sizes required to show 'at least as good as' (with identical power but a two-sided type I error instead), the following equation can be used:

$$[\mu_{et} - \mu_{st}]_{CSD} = [\delta_{BW}(z_{1-\alpha/2} - z)] / [(EF)^{1-\beta}(z_{1-\alpha} - z)] \tag{11}$$

where  $EF = [1 + R_{LB}^2]^{-1}$ ;  $z_{1-\alpha}$ ;  $z_{1-\alpha/2}$  and  $z$  are the normal deviates chosen for the respective one-tailed and two-tailed type I errors of fixed size  $\alpha$ , and the type II error reflecting a common power.

*Example (continued)*

With  $R_{True} = 1.0$  (or  $\mu_{st} - \mu_{et} = 0$ ) the sample size per group needed to reject the null hypothesis that the true ratio of mean blood pressure reductions is 0.80 or less, was  $n \approx 142$ /group, with 80 per cent power and single-sided  $\alpha$  fixed at 0.05. In a superiority trial with the same sample sizes, the smallest clinically significant difference  $[\mu_{et} - \mu_{st}]_{CSD} = \delta_{CSD}$  (or  $R_{CSD} = \delta_{CSD} / \delta_{BW}$ ) that could be detected in favour of the new anti-hypertensive therapy is calculated using (11) as

$$[\mu_{et} - \mu_{st}]_{CSD} = [(2)(1.96 + 0.84)] / [(0.82)^{1-\beta}(1.645 + 0.84)] \approx 2.5 \text{ mmHg}$$

For the same total number of patients needed to demonstrate that the new drug is ‘at least’ 80 per cent ‘as effective as’ the standard, the study could show that the new drug is at least 25 per cent better than the standard (that is, detect  $R_{CSD} \geq 1.25$ , approximately 2.5=2:0 mmHg), with an identical type I (two-tailed) error and 80 per cent power.

The significance tests of Blackwelder and the ratio format are both UMPU tests of their respective hypotheses. The difference in efficiency of the two tests is evident based on a comparison of their *equivalent* or reparameterized contrasts when  $R_{LBj} = 1:0$  and the fact that the SE of the *equivalent* comparison is smaller for the ratio test *under the restriction imposed*. To see this, note the following: for Blackwelder (with continuous data), the sample based contrast and its variance are<sup>‡</sup>

$$\bar{X}_{st} - \bar{X}_{et} -_{BW} \tag{12}$$

with

$$\text{var}(\bar{X}_{st} - \bar{X}_{et} -_{BW}) = 2^{-2}n \tag{13}$$

For the high fractioned lower bound approach, the sample based contrast and its variance<sup>‡</sup> are

$$\bar{X}_{et} - R_{LB}\bar{X}_{st} \tag{14}$$

with

$$\text{var}(\bar{X}_{et} - R_{LB}\bar{X}_{st}) = 2^{-2}(1 + R_{LB}^2)n \tag{15}$$

Therefore, when  $R_{LBj} = 1$

$$\text{var}(\bar{X}_{st} - \bar{X}_{et} -_{BW}) \leq \text{var}(\bar{X}_{et} - R_{LB}\bar{X}_{st})$$

The resulting relative efficiency follows from the direct translation (mapping) between the two equivalent forms of contrasts and the *usual assumptions* considered for independent random variables. The identities in terms of non-centrality parameters that underpin this claim are given in the Appendix, Section A2.

A selection of comparative sample size projections for the ratio format and the corresponding Blackwelder equivalents may be seen in Table I.

Note that when smaller metrics denote improvement, thus suggesting the need for an upper bound  $R_{UB}$  ( $R_{UB} \leq 1:0$ ), Blackwelder’s approach would be more efficient. In this case, the ratio may be inverted (as  $_{st} =_{et}$ ) for testing against a lower bound to maintain the advantage of improved efficiency.

If, in this arrangement, sizing a study from the standard were still preferred, where  $CV_{st} = CV_{et}$ , an adjustment in the sample size equation for ratios would be needed, due to the change in the denominator of  $R_{True}$ . By simple substitution for  $_{et}$  in the denominator, a modified equation results as

$$n_{\text{each group}:R} = [(CV_{st})^2 R_T'^2 (z_{1-\alpha} - z)^2 (1 + R_{LB}^2)] / (R_T' - R_{LB})^2 \tag{16}$$

where  $R_T' =_{st} =_{et}$  and  $R_{LBj} = 1$ . If  $R_T' = R_T = 1$ , no adjustment to (10) is needed!

<sup>‡</sup>Expectations of functions of random variables, see, for example, Mood *et al.* [28].

We do not address the case of either  $R_{LB} \leq 1$  or  $R_{UB} \geq 1$ , for while possible arithmetically, these arrangements would usually be of little interest in non-inferiority trials.

## DISCUSSION

Unique problems are posed by active control trials intended to establish that a new, experimental treatment is 'at least as good as' the standard therapy. Unlike superiority trials, they must demonstrate that the test therapy is no worse than the active control by some clinically tolerable amount,  $\delta_{BW}$ . Clinicians have to agree on the amount of inferiority (in any metric) that they are willing to accept as medically insignificant or tolerable as a basis for non-inferiority claims. An amount of inferiority ( $\delta_{BW}$ ) must be chosen to address the statistical requirements of Blackwelder's single-sided test to establish that the experimental treatment is 'as effective as' the standard. Likewise, the 'per cent as good as' approach requires selecting a high numerical fraction of the standard mean response as a lower bound for non-inferiority ( $R_{LB}$ ). Note that this high fraction is essentially the complement to the small part or fraction of the expected response to standard therapy that Blackwelder proposed ( $\delta_{BW}$ ), as the maximum tolerable limit for a claim of non-inferiority. One could convert from one criterion to the other, using  $\delta_{BW} = (1 - R_{LB}) \mu_{st}$ , if each were referenced to the expected control response (where  $\mu_{st}$  must be positive). The quantities used for sizing studies in either formulation are interchangeable, but in those cases when clinical tolerance can be defined equally well on the ratio scale, the improved efficiency of the ratio formulation has obvious merit.

In the ratio format, selections for  $R_{True}$ ,  $R_{LB}$  and estimates of  $\mu_{st}$  and  $\sigma_{st}$  (or their CV) are required to calculate sample size requirements. Here, assumptions about *clinical effectiveness* and *tolerance* can be based on dimensionless percentages, without the need to quantify expected treatment effects in specific units of measurement. Depending on particular clinical or regulatory concerns, the rationale for sizing studies to establish that a new product is 'at least' 80 per cent or 90 per cent 'as effective as' the standard therapy may be easier to grasp and generalize across studies for a given class of drugs or a given therapeutic setting, compared to the potentially more difficult task of choosing a clinically tolerable *absolute difference* in response measurements. When good judgement and experience guide clinicians in selecting  $\delta_{BW}$ , the clinically tolerable difference for planning a non-inferiority trial, the basis for this decision will usually involve choosing some small fractional part of the expected control response.  $R_{LB}$  is determined in an identical manner. In sizing a future study, historical data for the standard therapy will play a part in either formulation.

Justification for the clinical relevance of the ratio (versus the absolute difference) to define limits for non-inferiority should be examined. It is our belief that the procedure is a valid and compelling alternative to Blackwelder's technique. The methodology presented here has proven utility in clinical applications both in terms of its simplicity and acceptance by medical and statistical personnel engaged in designing and interpreting non-inferiority trials. Clinicians often find it simpler to agree upon the *per cent as good as* or *high fractional part* of the positive control effect that the new product should achieve, than to understand and select an absolute value for clinical tolerance (Blackwelder's delta). They appear more comfortable choosing the percentage lower bound and, after minimal explanation of the hypothesis test, tend to visualize potential study outcomes for a realistic range of control responses.

Expressing non-inferiority as a high numerical fraction of the expected active control response has both practical and scientific merits. The improved efficiency of this procedure in most typical non-inferiority testing situations is a clear bonus from both statistical and clinical (patient resource) perspectives. In addition, there are certain advantages in using a percentage lower bound for testing non-inferiority *at the conclusion of the study*. If the control response is *not* predicted accurately, the amount of inferiority considered *tolerable* may no longer be a meaningful value ( $R_{BW}$ ) in relation to the observed control response. By contrast, the percentage lower bound can always be used for hypothesis testing, and will typically be a relevant threshold for non-inferiority, regardless of the magnitude of observed *positive* response in the control group. Further, when conducting hypothesis tests for multiple related outcome variables, the application of the same *high fractional percentage* of the standard as a general criterion for testing non-inferiority will enhance the credibility of the analysis. It avoids pre-specifying different (and sometimes arbitrary) delta values for each outcome variable to define clinical tolerance and allows for a consistent interpretation of the study.

At completion of the study, the observed means  $\bar{X}_{et}$  and  $\bar{X}_{st}$  would be used to produce an estimate of  $R_{True}$  ( $\hat{R} = \bar{X}_{et} / \bar{X}_{st}$ ), and to construct the following test statistic:

$$(\bar{X}_{et} - R_{LB}\bar{X}_{st}) / [S^2(1 + R_{LB}^2)]^{1/2} \quad (7)$$

which, if larger than the single-sided critical constant  $t_{\alpha, n-2}$ , would allow us to reject  $H_0: \bar{X}_{et} / \bar{X}_{st} \leq R_{LB}$  and infer that the experimental therapy is 'at least as good as' the standard therapy by an amount exceeding ( $R_{LB} \times 100$ ) per cent.

#### Example (continued)

Mean blood pressure reductions at the conclusion of the study were observed to be very similar in the experimental and standard therapy groups (12.0 and 13.2 mmHg, respectively), with  $s = 8$  and  $n = 142$  per group. Here, from (7)

$$t_{282} = (12.0 - (0.8)13.2) / [64(1 + 0.8^2)]^{1/2} \approx (12.0 - 10.56) / 0.86 = 1.67$$

The test statistic 1.67 would cause the rejection of the null hypothesis of inferiority, indicating that the experimental therapy provides over 80 per cent of the anti-hypertensive effect of the standard drug ( $p < 0.05$ , single-sided). Strictly speaking, the null hypothesis defines clinical inferiority to include  $R_{LB}$ , the boundary condition for *clinical tolerance* (to be consistent with Blackwelder's formulation). Rejecting the null hypothesis implies that the experimental therapy has an effect above the lower limit,  $R_{LB}$ .

The approximate  $100(1 - \alpha)$  per cent two-sided symmetric confidence interval for  $R$  has been derived as

$$\hat{R} \pm z [S^2(1 + \hat{R}^2) / n(\bar{X}_{st})^2]^{1/2} \quad (9)$$

Because  $\hat{R}$  is included in the estimate of its own variance, these limits will be wider than those generated under the null hypothesis (using  $R_{LB}$  to construct the variance of  $\hat{R}$ ). This explains why the use of the confidence interval (centred by  $\hat{R}$ ) to exclude  $R_{LB}$  will be *less efficient* than testing  $H_0: R_{True} \leq R_{LB}$  to reject clinical inferiority.

It has been shown, as well, under typical conditions for use of the 'at least as good as' criterion (that is, when  $R_{LB} < 1$  and  $R_{LB} < R_{True}$ ), that the hypothesis test based on the ratio format ( $H_0: R_{True} \leq R_{LB}$ ) will be *more powerful* than Blackwelder's test of  $H_0: \bar{X}_{st} - \bar{X}_{et} \geq R_{BW}$

to detect any given alternative hypothesis contained in  $H_1: \mu_{et} = \mu_{st} \geq R_{LB}$  or, equivalently,  $\mu_{st} - \mu_{et} \leq R_{LB}$ . The increased efficiency is a result of smaller SEs for the corresponding contrasts ( $R_{LB} \leq 1$ ) in the ratio-formatted hypotheses defined in terms of their Blackwelder equivalents.

When  $R_{True} = 1:0$ , it has been shown that these comparisons in efficiency also apply to tests of conventional superiority (where  $[\mu_{et} - \mu_{st}]_{CSD} = CSD$ ). This provides a simple way to derive the effect size, or *clinically significant difference*, that would be detectable in a *superiority trial* for the same total sample size required to establish a non-inferiority claim with the type I and type II errors held constant. As part of the trial planning process, the clinical and marketing staff could then visualize trial outcomes and examine trade-offs with use of 'at least as good as' and superiority designs, when patient resources are limited and the merits of the new therapy versus the standard are under consideration.

Finally, in addition to sample size and power considerations, the interpretation of active control trials without a concurrent placebo control group is further complicated by the need to prove that the experimental treatment would have outperformed a placebo, had one been included in the trial. The procedures described in this paper do not eliminate problems establishing the efficacy of the experimental drug in comparison to a *hypothetical placebo*. Nevertheless, the ratio-based hypothesis test of non-inferiority offers a useful and often more efficient alternative to Blackwelder's approach when the objective is to prove that the experimental therapy produces an acceptably high percentage of the standard therapy's effect.

APPENDIX: NON-INFERIORITY TRIALS: THE 'AT LEAST AS GOOD AS' CRITERION

*A1. Sample size formulation using ratio estimators in the 'at least as good as' model*

The sample size equation below is expressed in terms of a continuous response measure, with higher values denoting greater improvement. If smaller metrics represent improvement, the inequalities would be reversed in the hypotheses below and referenced to  $R_{Upper Bound}$  instead.

*Hypotheses*

$$\text{Null: } R_{True} \leq R_{Lower Bound} \quad \text{Alternative: } R_{True} \geq R_{Lower Bound}$$

where  $R_{True} = \mu_{et} / \mu_{st}$ .

*Re-parameterization*

In order to obtain an optimal solution (Paulson [27]), we reparameterize the ratio-formatted hypotheses to

$$\text{Null: } \mu_{et} - R_{Lower Bound} \mu_{st} \leq 0 \quad \text{Alternative: } \mu_{et} - R_{Lower Bound} \mu_{st} \geq 0$$

*Distributions*

Operationally then, the null ( $H_0$ ) is centred at zero, the alternative ( $H_1$ ) at  $\mu_{et} - R_{Lower Bound} \mu_{st}$ , with common  $\text{var}(\bar{x}_{et} - R_{LB} \bar{x}_{st}) = \sigma^2(1 + R_{LB}^2)/n$  and the two distributions asymptotically normally distributed as

$$Z_{H_0} = [\bar{x}_{et} - R_{LB} \bar{x}_{st}] / \{ \sigma^2(1 + R_{LB}^2)/n \}^{1/2} \tag{A1}$$

and

$$Z_{H_1} = [(\bar{x}_{et} - R_{LB}\bar{x}_{st}) - (e_{et} - R_{LB}e_{st})] \sqrt{(1 + R_{LB}^2)\eta}^{1-\alpha} \tag{A2}$$

If  $Z_{H_0}$  and  $Z_{H_1}$  are each solved in terms  $\bar{x}_{et} - R_{LB}\bar{x}_{st}$ , and equated as

$$Z_{H_0} \sqrt{(1 + R_{LB}^2)\eta}^{1-\alpha} = Z_{H_1} \sqrt{(1 + R_{LB}^2)\eta}^{1-\alpha} + [e_{et} - R_{LB}e_{st}] \tag{A3}$$

(A3) results. Transposition and factoring in (A3) gives

$$[Z_{H_0} - Z_{H_1}] \sqrt{(1 + R_{LB}^2)\eta}^{1-\alpha} = [e_{et} - R_{LB}e_{st}] \tag{A4}$$

Squaring (A4), factoring  $[e_{et} - R_{LB}e_{st}]$  as  $e_{st} [R_T - R_{LB}]$ , for  $R_T = e_{et}/e_{st}$ , while setting  $(CV)^2 = \eta^2 = \frac{\eta^2}{st}$  and solving for  $n$  yields

$$n_{per\ group} = [(CV)^2(z_{1-\alpha} - z)^2(1 + R_{LB}^2)] / (R_T - R_{LB})^2 \tag{A5}$$

where now  $z_{1-\alpha}$  and  $z$  replace  $Z_{H_0}$  and  $Z_{H_1}$ , respectively, as the normal deviates producing the chosen single-sided probabilities under the operationally specified distributions.

*A2. Ratio format-Blackwelder relative efficiencies*

*Identities*

Blackwelder's  $n_{per\ group}$  in continuous variate form can be shown to be

$$n_{BW} = 2 \frac{(z_{1-\alpha} - z)^2}{(standard - experimental - BW)^2} \tag{A6}$$

where  $BW$  is Blackwelder's designation for clinical tolerance (see Blackwelder [1]).

Ratio format  $n_{per\ group}$  is

$$n_{RF} = \frac{2[(z_{1-\alpha} - z)^2(1 + R_{LB}^2)]}{\frac{2}{standard}(R_T - R_{LB})^2} \tag{A7}$$

(from (A5)) where as seen before,  $(CV)^2 = \eta^2 = \frac{\eta^2}{standard}$ .

With  $BW$  defined in terms of a ratio percentage of  $\frac{standard}{standard}$  as

$$BW = (1 - R_{LB})e_{st} \tag{A8}$$

( $e_{st} > 0$ ); where  $R_{LB} \in ]0, 1[$ ,  $(R_T - R_{LB})$  in (A7) would be

$$(e_{et}/e_{st}) - [(e_{st} - BW)/e_{st}] = (e_{et} - e_{st} + BW)/e_{st} \tag{A9}$$

where  $R_T = e_{et}/e_{st}$ . Hence, with

$$(R_T - R_{LB})^2 = (e_{et} - e_{st} + BW)^2 / e_{st}^2 \tag{A10}$$

the denominator seen in (A7)

$$\frac{2}{st}(R_T - R_{LB})^2 = (e_{et} - e_{st} + BW)^2 \tag{A11}$$

Since  $(e_{st} - e_{et} - BW)^2$  and  $(e_{et} - e_{st} + BW)^2$  from (A6) and (A11) are identically equivalent, both can be set equal to  $\Delta^2$ , as

$$(e_{et} - e_{st} + BW)^2 = (e_{st} - e_{et} - BW)^2 = \Delta^2 \tag{A12}$$

*Relative efficiency*

The ratio of  $n_{RF} = n_{BW}$  would then equal (A7) = (A6) with  $\Delta^2$  substituted for each denominator from the identities established in (A11) and (A12), resulting in the efficiency ratio (in effect, the ratio of the squared non-centrality parameters):

$$EF = [(z_{-} - z)^2(1 + R_{LB}^2)] = 2(z_{-} - z)^2 = (1 + R_{LB}^2) = 2 \quad (A13)$$

after both  $\Delta^2$  and  $\Delta^2$  cancel.

The resulting efficiency ratio EF seen in (A13) indicates that when  $R_{LB} \geq 1:0$ , then

$$n_{RF} = n_{BW} \geq 1:0$$

always, for the restriction imposed, or, under this restriction on EF,  $n_{RF}$  will always be smaller than  $n_{BW}$ , and thus the efficiency always greater.

## ACKNOWLEDGEMENTS

The authors are grateful to Professors James Pickands III and Abba Krieger, University of Pennsylvania, David Hoberman, FDA, Mitchell Kotler, Colgate-Palmolive Co., and Howard Proskin, Howard Proskin & Associates, for their reviews and valuable discussions of this paper. We are especially grateful to the editor and two referees for their careful reviews and suggestions that led to a substantial improvement in the exposition. The research of the first author was supported, in part, by the Colgate-Palmolive Co.

## REFERENCES

1. Blackwelder WC. 'Proving the null hypothesis' in clinical trials. *Controlled Clinical Trials* 1982; **3**:345–353.
2. Blackwelder WC, Chang MA. Sample size graphs: 'for proving the null hypothesis'. *Controlled Clinical Trials* 1984; **5**:97–105.
3. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports* 1978; **62**:1037–1040.
4. Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Science* 1972; **61**:1340–1341.
5. Metzler CM. Bioavailability—a problem equivalence. *Biometrics* 1974; **30**:309–317.
6. Westlake WJ. The use of balanced incomplete block designs in comparative bioavailability trials. *Biometrics* 1974; **30**:319–327.
7. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. *Biometrics* 1976; **32**:741–744.
8. Westlake WJ. Statistical aspects of comparative bioavailability trials. *Biometrics* 1979; **35**:273–280.
9. Spriet A, Beiler D. When can 'non-significantly different' treatments be considered as 'equivalent?' *British Journal of Clinical Pharmacology* 1979; **7**:623–624.
10. Rodda BE, Davis RL. Determining the probability of an important difference in bioavailability. *Clinical Pharmacological Therapy* 1980; **28**:247–252.
11. Patel HI, Gupta GD. A problem of equivalence in clinical trials. *Presented at the Meetings of the American Statistical Association and the Biometric Society* 1981; Richmond, VA (March).
12. Selwyn MR, Dempster AP, Hall NR. A Bayesian approach to bioequivalence for the 2×2 changeover design. *Biometrics* 1981; **37**:11–21.
13. Schuirmann D. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of coverage bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**:657–680.
14. Munk A. An improvement on commonly used tests in bioequivalence assessment. *Biometrics* 1993; **50**:884–886.
15. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests, and equivalence confidence sets (with discussion). *Statistical Science* 1996; **11**:283–319.
16. Brown LD, Hwang JTG, Munk A. An unbiased test for the bioequivalence problem. *Annals of Statistics* 1997; **26**:2345–2367.
17. Metzler CM. Sample sizes for bioequivalence studies. *Statistics in Medicine* 1991; **10**:961–970.

18. Anderson S, Hauck WW. Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 1990; **18**:259–273.
19. Hwang J, Wang W. The validity of the test of individual equivalence ratios. *Biometrika* 1997; **84**:893–900.
20. Wang W. On testing of individual bioequivalence. *Journal of the American Statistical Association* 1999; **94**: 880–887.
21. Wang W DasGupta A, Hwang JTG. Statistical tests for multivariate bioequivalence. Technical report, Cornell University, 1997.
22. Munk A, Pflugger R. 1 – equivariant confidence rules for convex alternatives are  $\alpha$ -level tests—with applications to the multivariate assessment of bioequivalence. *Journal of the American Statistical Association* 1999; **94**:1311–1319.
23. Lehmann EL. *Testing Statistical Hypotheses*. Wiley: New York, 1959.
24. Holmgren EB. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* 1999; **9**:651–659.
25. Fieller EC. The distribution of the index in a normal bivariate population. *Biometrics* 1932; **24**:428–440.
26. Miller RG. *Beyond ANOVA, Basics of Applied Statistics*. Wiley: New York, 1986.
27. Paulson E. A note on the estimation of some mean values for a bivariate distribution. *Annals of Mathematical Statistics* 1942; **13**:440–445.
28. Mood A, Graybill FA, Boes DC. *Introduction to the Theory of Statistics*, 3rd edn. McGraw-Hill: New York 1974; 176–181.