STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1115–1130
Published online 28 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2476

Non-inferiority trials: the '*at least as good as*' criterion with dichotomous data

Larry L. Laster[1,*,†] Mary F. Johnson[2,‡] and Mitchell L. Kotler[3,§]

[1]*School of Veterinary Medicine, University of Pennsylvania, 3900 Delancey St., Philadelphia, PA 19104, U.S.A.*
[2]*PharmaNet, Inc., 504 Carnegie Center, Princeton, NJ 08540, U.S.A.*
[3]*GlaxoSmithKline, Consumer Healthcare, 1500 Littleton Rd., Parsippany, NJ 07054, U.S.A.*

SUMMARY

The '*at least as good as*' criterion, introduced by Laster and Johnson for a continuous response variate, is developed here for applications with dichotomous data. This approach is *adaptive* in nature, as the margin of non-inferiority is not taken as a fixed difference; it varies as a function of the positive control response. When the non-inferiority margin is referenced as a high fraction of the positive control response, the procedure is seen to be *uniformly* more efficient than the fixed margin approach, yielding smaller sample sizes when sizing non-inferiority trials under identically specified conditions. Extending this method to proportions is straightforward, but highlights special considerations in the design of non-inferiority trials *versus* superiority trials, including potential trade-offs in statistical efficiency and interpretability. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: non-inferiority trials; '*at least as good as*'; statistical efficiency; adaptive *versus* fixed margins

INTRODUCTION

The general framework for non-inferiority testing initially proposed by Blackwelder [1] is now well established and has been explored by a number of authors [2–4]. Briefly, the test requires that the effect of a new therapy be no more than a fixed amount, 'delta' (δ_{BW}), worse than the effect of an active control. Recent extensions include 'adaptive' testing methods that allow the non-inferiority margin to vary as a function of the active control response [5–7]. In particular, the method proposed by Laster and Johnson [5] describes the statistical advantage of expressing the margin as a *percentage* of the control response, referred to as the

[*] Correspondence to: Larry L. Laster, School of Veterinary Medicine, University of Pennsylvania, 3900 Delancey St., Philadelphia, PA 19104, U.S.A.
[†] E-mail: larryl@vet.upenn.edu
[‡] E-mail: mjohnson@pharmanet.com
[§] E-mail: mitchell.l.kotler@gsk.com

Received 7 February 2005
Accepted 7 November 2005

'*at least as good as*' criterion, based on the mean of a continuous response variable. We now develop this method for application to dichotomous data and, in the process, examine both statistical and practical aspects of its use. Extending this method to proportions is straightforward, but highlights special considerations in the design of non-inferiority trials *versus* superiority trials, including potential trade-offs in statistical efficiency and interpretability.

Blackwelder [1] introduced a single-sided null hypothesis for clinical inferiority to be rejected in favour of the *at least as good as* hypothesis, defined in terms of a dichotomous outcome variable, with *higher proportions* denoting greater *success*, as

$$H_0\colon \Pi_{\mathrm{st}} - \Pi_{\mathrm{et}} \geqslant \delta_{\mathrm{BW}} \quad \textit{versus} \quad H_1\colon \Pi_{\mathrm{st}} - \Pi_{\mathrm{et}} < \delta_{\mathrm{BW}} \tag{1}$$

where Π_{st} and Π_{et} are *success* proportions with standard and experimental therapies and δ_{BW} represents the clinical tolerance selected. A trial designed in this framework would be successful if the outcome with the test therapy (et) were no worse than the outcome with the active control (st), by some clinically tolerable amount, δ_{BW}, usually envisioned as a small fractional part of the effect attributed to the active control.

Blackwelder assumes that Π_{st} (the success rate of a standard therapy) is far enough from 0 or 1, *given sufficiently large sample sizes*, that its estimate p_{st} is distributed approximately normal. Under H_0, with δ_{BW} a small amount, Π_{et} should also be approximately normal justifying the test statistic

$$z = (p_{\mathrm{st}} - p_{\mathrm{et}} - \delta_{\mathrm{BW}})/s \tag{2}$$

with

$$s^2 = [p_{\mathrm{st}}(1 - p_{\mathrm{st}})/n_{\mathrm{st}} + p_{\mathrm{et}}(1 - p_{\mathrm{et}})/n_{\mathrm{et}}] \tag{3}$$

where p_{st} and p_{et} are simply the observed experimental proportions.

It should be noted that Blackwelder's large sample-based choice for s^2, when dealing with the *null distribution* of the difference between two independently distributed proportions, was not without controversy. See References [8, 9]. We feel the unpooled version (3) is consistent with the conditions imposed by non-inferiority testing. Farrington and Manning [2] examined several comparative forms for s^2 for use with dichotomous data in non-inferiority trials and suggest that maximum likelihood estimates would be more accurate in establishing the null variance. Phillips [6] argues that with large samples, asymptotic arguments should suffice to allow the use of observed proportions and presents non-inferiority trial formulae on this basis. *In essence*, Phillips uses the Blackwelder (unpooled) s^2 when comparing the difference in proportions with a constant (δ_{BW}) for clinical tolerance.

In this paper, as in Reference [5], Blackwelder's approach of testing Π_{et} against $\Pi_{\mathrm{st}} - \delta_{\mathrm{BW}}$ will be recast as a test of Π_{et} against *a lower bound, a high percentage or fraction* (R_{LB}) *of* Π_{st} ($R_{\mathrm{LB}} < 1$). In this format, the null and alternative hypotheses take the form

$$H_0\colon \Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} \leqslant 0 \quad \text{and} \quad H_1\colon \Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} > 0 \tag{4}$$

where higher proportions denote greater *success*.

The relationship between Π_{et} and Π_{st} may be characterized by their ratio as $R_{\mathrm{True}} = \Pi_{\mathrm{et}}/\Pi_{\mathrm{st}}$. This parameter ($R_{\mathrm{T}}$) is useful for planning non-inferiority trials as it allows one to quantify the effectiveness of the new therapy in direct relation to expected efficacy of the active comparator.

In this format, the null and alternative hypotheses would have the form

$$H_0\colon R_{\mathrm{T}} \leqslant R_{\mathrm{LB}} \quad \textit{versus} \quad H_1\colon R_{\mathrm{T}} > R_{\mathrm{LB}} \tag{5}$$

HYPOTHESIS TESTING

Blackwelder's approach in (1) is identically equivalent to the high-fraction approach in (4) when $\Pi_{\mathrm{st}} - \delta_{\mathrm{BW}} = R_{\mathrm{LB}}\Pi_{\mathrm{st}}$, or

$$\delta_{\mathrm{BW}} = (1 - R_{\mathrm{LB}})\Pi_{\mathrm{st}} \tag{6}$$

Note, as in Reference [1], to justify the *as least as good as* application, δ_{BW} must be positive, and thus $R_{\mathrm{LB}} < 1$ *as defined here*. As Blackwelder points out, δ_{BW} could in theory be zero or negative, thus $R_{\mathrm{LB}} \geqslant 1$, but these cases are uncommon for non-inferiority testing. More typically, when δ_{BW} is selected as a small part of Π_{st}, $R_{\mathrm{LB}}\Pi_{\mathrm{st}}$ is the complementary larger part of it.

Our large sample formulations for the test statistics considered in this paper will rely on the experimentally observed proportions as in References [1, 6]. Unconditionally based exact tests for non-inferiority are available and are discussed below. The sample-based contrasts and their variances for both approaches (Blackwelder and the high-fraction approach) are as follows.

For Blackwelder,

$$p_{\mathrm{st}} - p_{\mathrm{et}} - \delta_{\mathrm{BW}} \tag{7}$$

with

$$\mathrm{Var}(p_{\mathrm{st}} - p_{\mathrm{et}} - \delta_{\mathrm{BW}}) = [\Pi_{\mathrm{st}}(1 - \Pi_{\mathrm{st}})/n_{\mathrm{st}} + \Pi_{\mathrm{et}}(1 - \Pi_{\mathrm{et}})/n_{\mathrm{et}}] \tag{8}$$

by independence. For the high-fraction lower-bound approach

$$p_{\mathrm{et}} - R_{\mathrm{LB}}\, p_{\mathrm{st}} \tag{9}$$

with

$$\mathrm{Var}(p_{\mathrm{et}} - R_{\mathrm{LB}}\, p_{\mathrm{st}}) = [\Pi_{\mathrm{et}}(1 - \Pi_{\mathrm{et}})/n_{\mathrm{et}} + \Pi_{\mathrm{st}}(1 - \Pi_{\mathrm{st}})R_{\mathrm{LB}}^2/n_{\mathrm{st}}] \tag{10}$$

Therefore, when $R_{\mathrm{LB}} < 1$,

$$\mathrm{Var}(p_{\mathrm{st}} - p_{\mathrm{et}} - \delta_{\mathrm{BW}}) > \mathrm{Var}(p_{\mathrm{et}} - R_{\mathrm{LB}}\, p_{\mathrm{st}}) \tag{11}$$

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.

In this form, the resultant test statistic is asymptotically normal as

$$(p_{\mathrm{et}} - R_{\mathrm{LB}}\, p_{\mathrm{st}})/[p_{\mathrm{et}}(1 - p_{\mathrm{et}})/n_{\mathrm{et}} + p_{\mathrm{st}}(1 - p_{\mathrm{st}})R_{\mathrm{LB}}^2/n_{\mathrm{st}}]^{1/2} \tag{12}$$

where here, *higher proportions denote greater success*. Phillips [6] confirms the form of this test statistic produced for an *adaptive test* for non-inferiority. The *adaptive nature* of Laster and Johnson's test for continuous data [5] as derived here for proportions, and Phillips' test [6], arises from the fact that the margin for non-inferiority varies as a function of the positive control response.

As noted by Phillips [6], similar results are obtained using a confidence interval based on the same $\mathrm{Var}(p_{\mathrm{et}} - R_{\mathrm{LB}}p_{\mathrm{st}})$ as

$$p_{\mathrm{et}} - R_{\mathrm{LB}}p_{\mathrm{st}} - Z_{(1-\alpha)}[p_{\mathrm{et}}(1-p_{\mathrm{et}})/n_{\mathrm{et}} + p_{\mathrm{st}}(1-p_{\mathrm{st}})R_{\mathrm{LB}}^2/n_{\mathrm{st}}]^{1/2} > 0 \tag{12a}$$

to reject H_0: $\Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} \leqslant 0$.

If a confidence interval procedure were applied to exclude δ_{BW} based on the unpooled $\mathrm{Var}(p_{\mathrm{st}} - p_{\mathrm{et}} - \delta_{\mathrm{BW}})$ seen in (8), it would not produce the same results, generally, relative to the confidence interval seen in (12a) using $\mathrm{Var}(p_{\mathrm{et}} - R_{\mathrm{LB}}p_{\mathrm{st}})$, for it would have the wrong size and be generally inefficient (given $R_{\mathrm{LB}} < 1$). A similar phenomenon was pointed out in Reference [5] using continuous data. The inefficiency of Blackwelder's hypothesis test in (11), in similar circumstances, also pertains to the associated confidence interval procedure.

SAMPLE SIZE REQUIREMENTS

For studies large enough to justify the normal approximation, the variance of $(p_{\mathrm{et}} - p_{\mathrm{st}}R_{\mathrm{LB}})$ shown in (10) leads to

$$n_{\text{per group}} = [(Z_{1-\alpha} - Z_\beta)^2(\Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}}) + \Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}})R_{\mathrm{LB}}^2)]/(\Pi_{\mathrm{et}} - \Pi_{\mathrm{st}}R_{\mathrm{LB}})^2 \tag{13a}$$

or

$$n_{\text{per group}} = [(Z_{1-\alpha} - Z_\beta)^2(\Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}}) + \Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}})R_{\mathrm{LB}}^2)]/\Pi_{\mathrm{st}}^2(R_{\mathrm{T}} - R_{\mathrm{LB}})^2 \tag{13b}$$

where Π_{st} and Π_{et} are the *success proportions* with standard and experimental therapies, $R_{\mathrm{T}} = \Pi_{\mathrm{et}}/\Pi_{\mathrm{st}}$, R_{LB} the lower bound (high fraction, $R_{\mathrm{LB}} < R_{\mathrm{T}}$) with $Z_{1-\alpha}$ and Z_β the normal deviates, producing the chosen single-sided probabilities under the operationally specified hypotheses

$$H_0\colon\ \Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} \leqslant 0 \quad \text{and} \quad H_1\colon\ \Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} > 0 \tag{4}$$

Phillips [6] confirms the form of (13a) as a *special case* of a more general sample size equation produced for an *adaptive test* for non-inferiority.

A *special case* occurs as well, when $\Pi_{\mathrm{et}} = \Pi_{\mathrm{st}} = \Pi$ or $R_{\mathrm{T}} = 1$, where (13b) can then be rewritten as

$$n_{\text{per group}} = [(Z_{1-\alpha} - Z_\beta)^2(1-\Pi)(1+R_{\mathrm{LB}}^2)]/\Pi(1-R_{\mathrm{LB}})^2 \tag{14}$$

The above formulae rely on asymptotic arguments in using the *usual* normal approximations. Phillips [6] argues that studies that typically enrol 100–200 or more patients per treatment group, with *success rates* below 95 per cent, should hold up reasonably well, so that the size of the test is near the nominal α-level.

COMPARATIVE EFFICIENCY

With continuous data in Reference [5], an efficiency ratio was derived to compare the sample size requirements for the high-fraction and Blackwelder approaches. The high-fraction approach was found to be uniformly *more efficient* than Blackwelder's method when $R_{\mathrm{LB}} < 1$, yielding smaller sample sizes for non-inferiority trials under commonly specified conditions. With a dichotomous outcome variable, two versions of the efficiency ratio result from a comparison of sample sizes under different assumptions for R_{T} ($\mathrm{EF}_{\Pi\mathrm{a}}$ and $\mathrm{EF}_{\Pi\mathrm{b}}$). In the *general case* with no restriction on R_{T},

$$\mathrm{EF}_{\Pi\mathrm{a}} = [\Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}}) + \Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}})R_{\mathrm{LB}}^2]/[\Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}}) + \Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}})] \tag{15a}$$

whereas in the *special case* of $R_{\mathrm{T}} = 1$ (or $\Pi_{\mathrm{et}} = \Pi_{\mathrm{st}} = \Pi$),

$$\mathrm{EF}_{\Pi\mathrm{b}} = [1 + R_{\mathrm{LB}}^2]/2 \tag{15b}$$

This form (15b) for the efficiency ratio is identical to the version produced for continuous data in Reference [5]. In either case, the relative efficiency of the high-fractioned test found with continuous data remains *uniformly* (when $R_{\mathrm{LB}} < 1$) in the case of proportions, as indicated above by (11).

DEALING WITH PROPORTIONS OF FAILURE

Note that in the original work [5] with continuous data, whenever *smaller values* denote *improvement*, the ratio R_{T} may be inverted (as $\mu_{\mathrm{st}}/\mu_{\mathrm{et}}$) for testing against a lower bound, to maintain the advantage of improved efficiency. We will make the same suggestion here. When dealing with failure data or the like, simply redefine R_{T} as

$$R'_{\mathrm{T}} = \Pi_{\mathrm{st}}/\Pi_{\mathrm{et}}$$

This will allow the continued use of the lower bound $R_{\mathrm{LB}}(< 1)$ to insure the increased efficiency whenever smaller values denote improvement. When this inverted definition for R_{T} is used, changes will be required in the statements involving hypotheses, test statistics, and efficiency equations. Sample size formulations, with the exception of $R'_{\mathrm{T}} = R_{\mathrm{T}} = 1$, will require changes as well.

The null and alternative hypotheses would then be rewritten as

$$H_0\colon\ \Pi_{\mathrm{st}} \leqslant R_{\mathrm{LB}}\Pi_{\mathrm{et}} \quad \textit{versus} \quad H_1\colon\ \Pi_{\mathrm{st}} > R_{\mathrm{LB}}\Pi_{\mathrm{et}} \tag{16}$$

when lower proportions denote *greater success*. The test statistic in like fashion, would be

$$(p_{\mathrm{st}} - R_{\mathrm{LB}}\,p_{\mathrm{et}})/[p_{\mathrm{st}}(1-p_{\mathrm{st}})/n_{\mathrm{st}} + p_{\mathrm{et}}(1-p_{\mathrm{et}})R_{\mathrm{LB}}^2/n_{\mathrm{et}}]^{1/2} \tag{17}$$

Sample size equations would be modified ($R_{\mathrm{T}} \neq 1$) to

$$n_{\text{per group}} = [(Z_{1-\alpha} - Z_{\beta})^2(\Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}}) + \Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}})R_{\mathrm{LB}}^2)]/(\Pi_{\mathrm{st}} - \Pi_{\mathrm{et}}R_{\mathrm{LB}})^2 \tag{18a}$$

or

$$n_{\text{per group}} = [(Z_{1-\alpha} - Z_{\beta})^2(\Pi_{\mathrm{st}}(1-\Pi_{\mathrm{st}}) + \Pi_{\mathrm{et}}(1-\Pi_{\mathrm{et}})R_{\mathrm{LB}}^2)]/\Pi_{\mathrm{et}}^2(R'_{\mathrm{T}} - R_{\mathrm{LB}})^2 \tag{18b}$$

with $R'_{\mathrm{T}} = \Pi_{\mathrm{st}}/\Pi_{\mathrm{et}}$. Similar substitutions would be required *to right* the efficiency equations as well. Note that with failure data as described here, Blackwelder equivalence is referenced to the experimental therapy as $\delta_{\mathrm{BW}} = (1 - R_{\mathrm{LB}})\Pi_{\mathrm{et}}$.

Tables I and II display corresponding sample size (*per group*) requirements for both *success* and *failure* data (respectively), when testing lower bounds on non-inferiority (R_{LB}) of up to 95 per cent '*as good as*' (with 80 per cent power; one-tailed $\alpha = 0.05$). To demonstrate sample sizes with similar *critical regions* for the lower-valued proportions in either set of tables, an R_{LB} of 50 per cent is included. The savings in sample size associated with the high-fractioned approach *versus* Blackwelder's method (up to 39 per cent for $R_{\mathrm{LB}} = 0.75$ to 0.95, and up to 69 per cent for R_{LB} of 0.50, under the conditions examined in Tables I and II) makes it an attractive option for planning non-inferiority studies, and, as illustrated in the next section, adds a new dimension to the study planning process.

STUDY DESIGN SCENARIOS

We illustrate the use of this method for planning a study when various hypothesis-testing strategies are under consideration. Suppose a new dental gel is to be investigated for use as a topical anaesthetic during periodontal scaling and subgingival curettage. Pain severity will be assessed on a 5-point scale, with control of pain dichotomized as either success (scores of 0 or 1) *versus* failure (scores of 3, 4 or 5). The sponsor believes that the new product is at least as effective as a standard, marketed agent in controlling pain, with a *faster* acting effect. In previous placebo-controlled studies, the marketed product demonstrated a 70 per cent success rate in patients with advanced periodontitis, on the basis of the same pain assessment scale. The sponsor will employ a randomized, double-blind, parallel-group design to compare the new product to placebo as well as an active control (the standard, marketed anaesthetic) in the same target population. To establish efficacy, the study must demonstrate that the new treatment is superior to placebo and not inferior to the active control. To fully validate the outcome of the study, response to the active control must also be superior to placebo. *Weighing both cost and regulatory concerns, we evaluate sample size requirements under a number of different, but equally plausible assumptions and hypothesis-testing frameworks.*

Non-inferiority trials

Prior studies in this periodontal maintenance setting indicate that a difference in success rates of 16 per cent (treatment-70 per cent *versus* placebo-54 per cent) is a clinically meaningful effect. To detect this difference and demonstrate *superiority* of each active treatment over placebo, with two-tailed $\alpha = 0.05$ and $\beta = 0.2$ (80 per cent power), the required sample size per group would be

$$n_{\text{per group}} \approx 141$$

using equation (13a) with $R_{\mathrm{LB}} = 1$ (which yields the usual sample size formula for superiority based on the variance in (3) above).

Note, a continuity correction is not used here. Furthermore, adjustment of the α-level is unnecessary for all three hypotheses proposed must be rejected to meet the requirements of the study (see Reference [10]).

For *non-inferiority*, we set $R_{\mathrm{LB}} = 0.80$ to determine if the new product is at least 80 per cent as effective as the standard, keeping $\alpha = 0.05$ (one sided in this instance) and $\beta = 0.2$. Using equation (14) with $\Pi_{\mathrm{et}} = \Pi_{\mathrm{st}} = 0.7$ (for $R_{\mathrm{T}} = 1$)

$$n_{\text{per group}} = [(1.645 + 0.84)^2(1 - 0.7)(1 + 0.8^2)]/0.7(1 - 0.8)^2$$
$$\approx 109 \text{ (see Table I)}$$

This sample size for testing non-inferiority is smaller than that required for testing superiority ($n = 141$), even though a smaller difference is sought (0.14 for non-inferiority *versus* 0.16 for superiority). Two reasons explain the difference in sample size requirements: use of the single-sided rejection region and a smaller variance. By comparison, Blackwelder's $n_{\text{per group}} = 132$, for the same non-inferiority margin (again 0.14 or 0.2(0.7)), with a single-tailed rejection region as well (see Table I). *Clearly*, if the study were sized for a test of superiority, it would have more than adequate power to establish non-inferiority under the assumed conditions (once again, no adjustment for multiple comparisons has been made).

Note in particular that, if $R_{\mathrm{LB}} = 0.8$ were considered too lenient, higher limits could be considered, but sample sizes will soar. This is illustrated in the following table (from Table I) using the high-fraction (HF) and Blackwelder (BW) approach to non-inferiority testing, under identical conditions (with $R_{\mathrm{T}} = 1$). *Sample sizes are per group*:

R_{LB}	Sample size (HF)	Sample size (BW)
0.85	203	236
0.90	480	530
0.95	2016	2120

Non-inferiority versus Superiority

There is often limited evidence, or only a theoretical basis, for assuming that the new product will out-perform the active control, and the magnitude of the difference is usually uncertain. This debate often gives rise to the following question: *What effect size in superiority could be detected with identical type 1 error* (*adjusted for two-tailed hypothesis testing*) *and type 2 error*, *with sample sizes similar to those just determined for a non-inferiority test*, *if the new therapy were*, *in fact*, *better than the active control* ($R_{\mathrm{T}} > 1$)?

Returning to the above example, suppose the sponsor suspects that the new dental gel is more effective than the active control, to some marginal degree. What would be the case for testing superior performance based on an assumed 10 per cent or even a 20 per cent improvement in response (e.g. $R_{\mathrm{T}} = 1.1$ or 1.2), thus efficacies of about 77 per cent or 84 per cent, respectively? Again, with two-tailed $\alpha = 0.05$ and 80 per cent power, the required sample sizes per group (based on the superiority form of equation (13a) with $R_{\mathrm{LB}} = 1$) would be

$$n_{(R_{\mathrm{T}}=1.1)} = 619 \quad \text{and} \quad n_{(R_{\mathrm{T}}=1.2)} = 138$$

Table I. Comparative sample size solutions* for non-inferiority tests with *proportions of success* using the high-fraction and Blackwelder approaches for various values of R_T (Π_{et}/Π_{st}), R_{LB} and Π_{st} with $1-\beta = 0.8$ and one-tailed $\alpha = 0.05$.

		$R_T = \Pi_{et}/\Pi_{st}$																	
	R_{LB}	0.80		0.85		0.90		0.95		1.00		1.05		1.10		1.15		1.20	
$\Pi_{st} = 0.10$	0.50	660	(1124)	506	(847)	403	(664)	331	(537)	278	(445)	238	(376)	207	(323)	182	(281)	162	(247)
	0.75	30721	(40459)	7938	(10373)	3642	(4723)	2111	(2720)	1391	(1781)	993	(1264)	750	(948)	589	(741)	477	(597)
	0.80			33479	(41491)	8625	(10628)	3945	(4835)	2281	(2782)	1499	(1820)	1068	(1291)	804	(968)	631	(756)
	0.85					36335	(42511)	9336	(10880)	4260	(4946)	2458	(2844)	1612	(1859)	1146	(1317)	861	(987)
	0.90							39290	(43519)	10071	(11129)	4585	(5055)	2640	(2904)	1728	(1897)	1226	(1344)
	0.95									42344	(44514)	10832	(11374)	4922	(5163)	2829	(2964)	1848	(1935)
$\Pi_{st} = 0.20$	0.50	300	(506)	229	(380)	181	(297)	148	(240)	124	(198)	105	(167)	91	(142)	79	(123)	70	(108)
	0.75	13874	(18201)	3572	(4654)	1632	(2113)	942	(1213)	618	(791)	439	(560)	330	(418)	258	(326)	208	(261)
	0.80			15055	(18616)	3864	(4754)	1761	(2156)	1014	(1237)	664	(806)	471	(569)	353	(425)	275	(331)
	0.85					16272	(19018)	4165	(4852)	1893	(2198)	1088	(1259)	710	(820)	503	(579)	376	(432)
	0.90							17528	(19407)	4476	(4946)	2030	(2239)	1164	(1281)	758	(834)	536	(588)
	0.95									18820	(19784)	4796	(5037)	2171	(2278)	1242	(1303)	808	(847)
$\Pi_{st} = 0.30$	0.50	179	(300)	136	(224)	107	(175)	87	(140)	72	(115)	61	(97)	52	(82)	45	(71)	40	(62)
	0.75	8258	(10782)	2116	(2748)	962	(1243)	553	(711)	361	(462)	255	(325)	190	(242)	148	(187)	118	(149)
	0.80			8913	(10991)	2277	(2797)	1032	(1263)	591	(721)	385	(468)	271	(329)	202	(244)	157	(189)
	0.85					9585	(11186)	2442	(2842)	1104	(1282)	631	(731)	410	(474)	288	(333)	214	(247)
	0.90							10273	(11370)	2611	(2885)	1178	(1300)	672	(740)	435	(479)	306	(336)
	0.95									10978	(11541)	2784	(2925)	1254	(1316)	714	(749)	462	(484)
$\Pi_{st} = 0.40$	0.50	119	(196)	90	(146)	70	(114)	56	(91)	46	(74)	39	(62)	33	(52)	28	(45)	24	(39)
	0.75	5450	(7073)	1389	(1794)	628	(808)	358	(459)	232	(297)	163	(208)	120	(153)	93	(118)	73	(93)
	0.80			5843	(7178)	1484	(1818)	668	(817)	380	(464)	246	(299)	172	(209)	127	(154)	97	(118)
	0.85					6241	(7271)	1580	(1838)	710	(824)	403	(467)	260	(301)	181	(210)	133	(154)
	0.90							6646	(7351)	1679	(1855)	752	(831)	426	(470)	274	(302)	191	(210)
	0.95									7057	(7419)	1778	(1869)	795	(835)	449	(472)	288	(303)

$\Pi_{st} = 0.50$	0.50	83	(135)	62	(100)	48	(77)	38	(61)	31	(49)	25	(41)	21	(34)	18	(29)	15	(25)
	0.75	3765	(4847)	952	(1223)	427	(547)	241	(309)	155	(198)	107	(137)	78	(100)	60	(76)	46	(60)
	0.80			4000	(4890)	1008	(1230)	450	(549)	253	(309)	162	(198)	112	(137)	82	(100)	62	(76)
	0.85					4235	(4921)	1063	(1235)	473	(550)	266	(309)	169	(197)	117	(136)	85	(99)
	0.90							4470	(4940)	1119	(1237)	497	(549)	278	(308)	177	(196)	122	(135)
	0.95									4705	(4946)	1175	(1235)	520	(547)	291	(306)	184	(194)
$\Pi_{st} = 0.60$	0.50	59	(93)	43	(69)	33	(52)	26	(41)	21	(33)	17	(27)	14	(22)	11	(18)	9	(15)
	0.75	2642	(3363)	661	(841)	293	(373)	163	(208)	103	(132)	70	(90)	50	(65)	37	(49)	29	(37)
	0.80			2772	(3365)	690	(839)	304	(370)	169	(206)	106	(130)	72	(89)	52	(64)	38	(47)
	0.85					2898	(3355)	719	(833)	316	(366)	175	(203)	109	(128)	74	(87)	53	(62)
	0.90							3019	(3332)	746	(824)	326	(361)	180	(199)	112	(125)	76	(84)
	0.95									3137	(3297)	772	(812)	337	(354)	185	(195)	115	(121)
$\Pi_{st} = 0.70$	0.50	42	(64)	30	(46)	23	(35)	17	(27)	13	(21)	10	(17)	8	(14)	6	(11)	5	(9)
	0.75	1840	(2303)	453	(569)	197	(248)	108	(137)	66	(85)	44	(57)	30	(40)	22	(29)	16	(21)
	0.80			1895	(2276)	464	(559)	200	(243)	109	(132)	66	(82)	44	(54)	30	(38)	21	(27)
	0.85					1942	(2236)	473	(546)	203	(236)	109	(128)	66	(78)	43	(51)	29	(35)
	0.90							1983	(2184)	480	(530)	205	(227)	110	(122)	66	(74)	43	(48)
	0.95									2016	(2120)	485	(511)	206	(217)	109	(116)	65	(70)
$\Pi_{st} = 0.80$	0.50	29	(42)	20	(30)	15	(22)	11	(16)	8	(12)	6	(9)	4	(7)				
	0.75	1238	(1509)	297	(365)	125	(155)	66	(83)	39	(49)	24	(32)	15	(21)	10	(14)	6	(9)
	0.80			1237	(1459)	294	(349)	122	(147)	63	(77)	37	(46)	22	(29)	14	(18)	9	(12)
	0.85					1226	(1397)	288	(331)	118	(137)	60	(71)	34	(41)	20	(25)	12	(16)
	0.90							1206	(1323)	280	(309)	113	(126)	57	(64)	31	(36)	18	(21)
	0.95									1176	(1237)	269	(284)	107	(114)	53	(56)	28	(31)
$\Pi_{st} = 0.90$	0.50	19	(25)	13	(17)	8	(12)	6	(8)										
	0.75	770	(890)	176	(206)	69	(83)	33	(41)	17	(22)	9	(12)	4	(6)				
	0.80			725	(824)	161	(186)	62	(73)	28	(34)	13	(17)	6	(8)				
	0.85					668	(745)	144	(163)	53	(61)	22	(27)	9	(12)				
	0.90							601	(653)	124	(137)	42	(48)	16	(19)				
	0.95									523	(550)	102	(108)	31	(34)				

*Values seen in ( ) are sample size solutions for the Blackwelder equivalent hypothesis test to detect $\delta_{BW} = (1 - R_{LB})\Pi_{st}$ *as defined here.*

Note: Due to the distributional properties of the *smaller sample size projections* indicated here, they should be considered only approximate.

Table II. Comparative sample size solutions* for non-inferiority tests with *proportions of failure* using the high-fraction and Blackwelder approaches for various values of R'_{T} $(\Pi_{\mathrm{st}}/\Pi_{\mathrm{et}})$, R_{LB} and Π_{st} with $1-\beta=0.8$ and one-tailed $\alpha=0.05$.

		$R'_{\mathrm{T}} = \Pi_{\mathrm{st}}/\Pi_{\mathrm{et}}$																	
	R_{LB}	0.80		0.85		0.90		0.95		1.00		1.05		1.10		1.15		1.20	
$\Pi_{\mathrm{st}}=0.10$	0.50	516	(877)	423	(707)	359	(591)	313	(508)	278	(445)	251	(397)	230	(359)	213	(328)	198	(302)
	0.75	23982	(31556)	6628	(8657)	3240	(4201)	1994	(2569)	1391	(1781)	1049	(1334)	834	(1054)	688	(866)	585	(732)
	0.80			27951	(34628)	7673	(9453)	3727	(4568)	2281	(2782)	1583	(1921)	1188	(1435)	940	(1131)	773	(926)
	0.85					32322	(37813)	8819	(10277)	4260	(4946)	2595	(3002)	1792	(2066)	1339	(1539)	1055	(1209)
	0.90							37114	(41108)	10071	(11129)	4841	(5337)	2935	(3229)	2019	(2216)	1502	(1646)
	0.95									42344	(44514)	11435	(12008)	5472	(5740)	3304	(3463)	2264	(2370)
$\Pi_{\mathrm{st}}=0.20$	0.50	227	(382)	187	(310)	159	(260)	139	(225)	124	(198)	112	(177)	102	(160)	95	(147)	88	(136)
	0.75	10504	(13750)	2917	(3796)	1431	(1852)	884	(1138)	618	(791)	467	(595)	372	(471)	308	(388)	262	(329)
	0.80			12291	(15184)	3388	(4167)	1651	(2022)	1014	(1237)	705	(857)	530	(642)	420	(507)	346	(416)
	0.85					14266	(16668)	3907	(4550)	1893	(2198)	1156	(1339)	800	(924)	599	(690)	473	(543)
	0.90							16439	(18201)	4476	(4946)	2158	(2380)	1311	(1444)	904	(993)	674	(739)
	0.95									18820	(19784)	5098	(5354)	2446	(2566)	1480	(1552)	1016	(1064)
$\Pi_{\mathrm{st}}=0.30$	0.50	131	(217)	108	(178)	92	(150)	81	(130)	72	(115)	65	(104)	60	(94)	56	(87)	52	(80)
	0.75	6012	(7815)	1680	(2176)	828	(1069)	514	(660)	361	(462)	273	(348)	218	(277)	181	(229)	154	(194)
	0.80			7071	(8703)	1960	(2405)	960	(1174)	591	(721)	413	(502)	311	(377)	247	(299)	204	(246)
	0.85					8248	(9620)	2270	(2641)	1104	(1282)	677	(784)	470	(543)	353	(407)	279	(321)
	0.90							9548	(10566)	2611	(2885)	1263	(1394)	770	(849)	532	(586)	398	(437)
	0.95									10978	(11541)	2985	(3136)	1437	(1509)	872	(915)	600	(629)
$\Pi_{\mathrm{st}}=0.40$	0.50	83	(135)	69	(111)	59	(95)	52	(83)	46	(74)	42	(67)	39	(61)	36	(56)	34	(52)
	0.75	3765	(4847)	1061	(1366)	527	(677)	329	(422)	232	(297)	176	(225)	141	(180)	117	(149)	100	(127)
	0.80			4461	(5462)	1246	(1524)	614	(750)	380	(464)	266	(324)	202	(245)	161	(195)	133	(161)
	0.85					5238	(6096)	1451	(1687)	710	(824)	437	(507)	305	(353)	229	(265)	182	(210)
	0.90							6102	(6748)	1679	(1855)	816	(901)	500	(551)	346	(382)	260	(286)
	0.95									7057	(7419)	1929	(2027)	933	(980)	568	(596)	392	(412)

$\Pi_{st} = 0.50$	0.50	54	(85)	45	(72)	39	(62)	34	(55)	31	(49)	28	(45)	26	(41)	24	(38)	23	(36)
	0.75	2417	(3067)	690	(879)	346	(442)	218	(279)	155	(198)	118	(151)	95	(122)	79	(101)	68	(87)
	0.80			2895	(3518)	817	(995)	406	(495)	253	(309)	179	(218)	136	(166)	109	(132)	90	(110)
	0.85					3433	(3982)	960	(1114)	473	(550)	293	(340)	205	(238)	155	(180)	124	(143)
	0.90							4035	(4458)	1119	(1237)	548	(605)	337	(372)	235	(259)	177	(195)
	0.95									4705	(4946)	1295	(1362)	630	(662)	386	(405)	267	(281)
$\Pi_{st} = 0.60$	0.50	35	(52)	30	(45)	26	(40)	23	(36)	21	(33)	19	(30)	17	(28)	16	(26)	15	(25)
	0.75	1519	(1879)	443	(555)	226	(286)	144	(183)	103	(132)	79	(102)	64	(83)	54	(69)	46	(60)
	0.80			1851	(2222)	532	(643)	268	(326)	169	(206)	120	(147)	92	(113)	74	(91)	62	(76)
	0.85					2229	(2572)	633	(733)	316	(366)	197	(230)	139	(162)	106	(124)	85	(99)
	0.90							2656	(2931)	746	(824)	369	(408)	229	(253)	161	(178)	122	(135)
	0.95									3137	(3297)	873	(918)	428	(451)	264	(278)	184	(194)
$\Pi_{st} = 0.70$	0.50	21	(29)	18	(26)	16	(24)	15	(23)	13	(21)	12	(20)	11	(19)	11	(18)	10	(17)
	0.75	877	(1032)	266	(324)	140	(174)	91	(115)	66	(85)	52	(67)	42	(55)	36	(47)	31	(41)
	0.80			1105	(1296)	328	(391)	169	(204)	109	(132)	78	(96)	61	(75)	49	(61)	42	(51)
	0.85					1369	(1565)	399	(460)	203	(236)	129	(150)	92	(108)	71	(83)	57	(67)
	0.90							1672	(1840)	480	(530)	241	(267)	152	(168)	108	(120)	82	(91)
	0.95									2016	(2120)	571	(601)	284	(300)	177	(187)	125	(132)
$\Pi_{st} = 0.80$	0.50			10	(12)	9	(13)	8	(13)	8	(12)	7	(12)	7	(12)	6	(11)	6	(11)
	0.75	396	(396)	133	(150)	75	(90)	51	(64)	39	(49)	31	(40)	26	(34)	22	(30)	20	(26)
	0.80			546	(601)	175	(202)	95	(114)	63	(77)	47	(58)	37	(47)	31	(39)	26	(33)
	0.85					724	(810)	223	(255)	118	(137)	77	(91)	57	(67)	44	(53)	36	(43)
	0.90							934	(1022)	280	(309)	145	(162)	94	(105)	68	(76)	53	(59)
	0.95									1176	(1237)	345	(364)	176	(186)	112	(119)	80	(85)
$\Pi_{st} = 0.90$	0.50																		
	0.75							20	(24)	17	(22)	15	(20)	13	(18)	12	(16)	11	(15)
	0.80							37	(43)	28	(34)	23	(29)	19	(25)	16	(21)	14	(19)
	0.85							87	(96)	53	(61)	38	(45)	29	(35)	24	(29)	20	(25)
	0.90							359	(385)	124	(137)	71	(79)	49	(55)	37	(42)	30	(34)
	0.95									523	(550)	169	(179)	92	(98)	61	(66)	46	(49)

*Values seen in ( ) are sample size solutions for the Blackwelder equivalent hypothesis test to detect $\delta_{BW} = (1 - R_{LB})\Pi_{et}$ *as defined here.*

Note: Due to the distributional properties of the *smaller sample size projections* indicated here, they should be considered only approximate.

in order to detect differences in success rates of 7 or 14 per cent (respectively). In this instance, when the new product offers a relatively small benefit ($R_T = 1.1$) over standard therapy, the sample size adequate for testing non-inferiority provides insufficient power for a test of superiority. *By contrast*, Table I indicates that $n_{\text{per group}} = 110$ would be sufficient to demonstrate that the new dental gel is *at least* '90 per cent *as good as*' (not superior to) the active comparator if, in fact, $R_T = 1.1$. With $R_T = 1.2$, $n_{\text{per group}} = 65$ would suffice to make a '95 per cent *as good as*' claim. Again, the sample size of $n = 141$ per group required for testing superiority against placebo (with $R_T = 1.0$) would fulfil power requirements for these other hypothesis tests as well.

The *upshot* of these considerations is clear based on the assumptions $R_T = 1.1$ or 1.2. Raise the sample size to establish the superiority claim ($n \approx 619$ or 138 per group, respectively), *or* instead, keep costs down and design a non-inferiority trial ($n = 110$ or 65 per group) to establish that the new dental gel is *at least* 90 or 95 per cent (respectively) *as good as* the active control product. If the $R_T = 1$ assumption is considered *more realistic*, the 80 per cent *as good as* design, with $n = 141$ per group, would cover all of the above with the exception of the superiority case for $R_T = 1.1$.

Typically, the choice depends on financial considerations and the degree of *acceptable risk*. A relatively safe strategy is to *size the study for superiority while proposing a high-fractioned non-inferiority trial* (*at 90 per cent or 95 per cent*). Even if the study failed to demonstrate superiority, it could still support a claim that the new product is '*at least as good as*' the active control. This ability is built into the non-inferiority paradigm. It could be argued that the strategy generates a potential multiplicity concern, allowing two ways to succeed in claiming significance: superiority or non-inferiority. However, Morikawa and Yoshida [10] demonstrate, on the basis of closed hypothesis-testing procedures (CTP), that no adjustment of the α-level is required. That is, if superiority were to be established, non-inferiority is contained in that result; thus by CTP, no adjustment to the type 1 error is required even if only non-inferiority is obtained.

INTERPRETING STUDY RESULTS

To perform the non-inferiority test at study completion, the observed proportions p_{et} and p_{st} are used to construct the following test statistic

$$(p_{\text{et}} - R_{\text{LB}}\, p_{\text{st}})/[p_{\text{et}}(1 - p_{\text{et}})/n_{\text{et}} + p_{\text{st}}(1 - p_{\text{st}})R_{\text{LB}}^2/n_{\text{st}}]^{1/2} \tag{12}$$

which, if larger than the single-sided critical constant Z_α, allows us to reject H_0: $\Pi_{\text{et}} - R_{\text{LB}}\Pi_{\text{st}} \leqslant 0$ and infer that the experimental therapy is *at least as good as* the standard therapy by an amount exceeding ($R_{\text{LB}} \times 100$) per cent.

We illustrate the use of this test in the context of the above example. Let us say the sponsor sizes the study to detect superiority of each active treatment over placebo, and to perform the non-inferiority test based on $R_T = 1$, with a prescribed lower bound of $R_{\text{LB}} = 0.8$ (80 per cent *as good as*). This approach yields a sample size of 141 per group (or 423 total for the three-arm trial), since the larger number required for superiority testing supersedes the smaller number needed to evaluate non-inferiority.

It should be noted that in those instances where no superiority justifications are required, thus only the two-arm demonstration of non-inferiority, the smaller sample size would suffice.

In the present example, the two-arm non-inferiority comparison would yield considerably smaller sizes of 109 per group.

Study results turn out as follows:

Groups	p_{success} (per cent)
New dental gel	69
Active control	72
Placebo	54

Non-inferiority test (high fractioned equation (12))

$$Z = (0.69 - 0.8(0.72))/[0.69(1-0.69)/141 + 0.72(1-0.72)0.8^2)/141]^{1/2} \approx 2.31$$

($p < 0.01$ single sided)

Superiority test (active control *versus* placebo)

$$Z = (0.72 - 0.54)/[0.72(1-0.72)/141 + 0.54(1-0.54)/141]^{1/2} \approx 5.54$$

($p < 0.0000$ two sided)

Superiority test (new dental gel *versus* placebo)

$$Z = (0.69 - 0.54)/[0.69(1-0.69)/141 + 0.54(1-0.54)/141]^{1/2} \approx 2.62$$

($p < 0.009$ two sided)

The trial succeeds in establishing non-inferiority based on 80 per cent *as good as*, and in both tests of superiority. Note, as mentioned above, no control of the joint α-level is required.

A different scenario

Now consider results if the new drug performed considerably better than the active control. That is, suppose the success rates were 87 per cent in the new drug group, 69 per cent with active control, and 54 per cent in the placebo group.

Group	p_{success} (per cent)
New dental gel	87
Active control	69
Placebo	54

In this scenario, the new drug is statistically significantly more effective than the active control ($p < 0.0001$, two sided). Based on CTP as discussed above, the claim of superiority is justified, despite the fact that the trial was designed with a non-inferiority objective. *Clearly*, non-inferiority is subsumed under the alternative hypothesis for superiority. The superiority of the active control over placebo ($p<0.0001$, two sided) further validates the efficacy of the new product.

Once again, it could be more cost-effective to plan the study as a non-inferiority trial, assuming that the test product offers a small but realistic advantage over the active control. This conservative approach allows for a broader range of successful trial outcomes, spanning both non-inferiority and superiority claims. Given the possible scenarios noted above, it would be prudent to size a study for superiority testing, while proposing a high-fraction non-inferiority margin for the primary hypothesis test. Such an approach has minimal associated risk, with the maximum chance for a successful marketing claim.

DISCUSSION

The *at least as good as* criterion previously developed for continuous data has similar advantages in the case of binomial endpoints. It has been shown under typical conditions for use in non-inferiority trials (i.e. where $R_{\mathrm{LB}}<R_{\mathrm{T}}$ and $R_{\mathrm{LB}}<1$), that the hypothesis test based on the high-fraction format H_0: $\Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} \leqslant 0$ is more powerful than Blackwelder's test of H_0: $\Pi_{\mathrm{st}} - \Pi_{\mathrm{et}} \geqslant \delta_{\mathrm{BW}}$ to detect any given alternative hypothesis contained in H_1: $\Pi_{\mathrm{et}} - R_{\mathrm{LB}}\Pi_{\mathrm{st}} > 0$ or, equivalently, $\Pi_{\mathrm{st}} - \Pi_{\mathrm{et}} < \delta_{\mathrm{BW}}$. The increased efficiency is a result of smaller SE's for contrasts defined by the high-fraction hypothesis test ($R_{\mathrm{LB}}<1$) compared to their Blackwelder equivalents.

We have also suggested a strategy to evaluate the effect size that can be detected in a superiority trial of the same sample size required to support non-inferiority claims, with type 1 and type 2 errors held constant. As illustrated, study planners can explore a range of study scenarios and examine trade-offs with the use of *at least as good as* and superiority designs, when patient resources are limited and the merits of the new therapy *versus* standard are uncertain.

Although we illustrate the method using efficacy response rates, the high-fraction approach to trial planning is particularly useful for non-inferiority tests involving safety outcomes, including event rates for toxicity. Oncology and critical care studies are attractive settings for this procedure, given the high mortality typical of such studies and the need to examine clinical benefits of new therapies on morbidity or quality of life. For instance, a drug may not be expected to improve survival but could prolong the duration of tumour response or delay the need for narcotics to control pain. The objective of such studies may be to assure that the new therapy has no adverse impact on mortality (e.g. survival of patients administered the new therapy should be at least 95 per cent of survival with standard care), while the new therapy is shown to offer a significant advantage in terms of other important clinical outcomes.

The selection of a non-inferiority margin that is '*adaptive*' to the control group response offers many advantages. First, study planners 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 must achieve, than to select an absolute value for clinical tolerance (Blackwelder's δ_{BW}). Expressing non-inferiority as a high numerical fraction of the expected active control response yields a more efficient testing procedure, and thus, savings in sample size. 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 (δ_{BW}) in relation to the observed control response. By contrast, the percentage lower bound (R_{LB}) 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. *In this sense, the method is adaptive* as seen in Reference [6]. Further, when conducting hypothesis tests for multiple related outcome variables, the application of the same *high fraction* 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) δ-values for each outcome variable to define clinical tolerance and allows for a consistent interpretation of the study.

Phillips [6] addresses the value of the adaptive lower bound R_{LB} by indicating its analogy with δ_{BW} as '*fixed in advance by the experimenter*', and thus not dependent on the observed success rates. He points out, further, that his work has indicated no discontinuities in the size or power functions for these adaptive methods.

Regardless of the approach to non-inferiority testing, there is potential loss of power if the control group response is not predicted accurately. It is important to review past performance of the active control treatment in previous placebo-controlled trials in order to assess the constancy and reproducibility of its effect relative to placebo. The '*constancy assumption*' can be critical to the interpretation of efficacy in non-inferiority trials, as it would require that the size of the active control effect be bounded by its effect size in historical placebo-controlled trials. See the ICH guidelines [11] and Hung *et al.* [12] for more in-depth coverage of this topic.

The scenarios presented in this paper underscore the advantage of nominally designating *high-fractioned* non-inferiority objectives for clinical trials that are, *in fact*, adequately powered to demonstrate clinical superiority. Insofar as the results support non-inferiority, at a minimum, the trial will be successful. However, this approach also allows for a claim of superiority if treatment effects emerge as anticipated. Clearly, this approach should only be taken if the new product is believed to be superior to the active control. It would absurd to plan a non-inferiority study with R_{T} greater than 1 (*speciously*), purely for the purpose of decreasing the sample size, as this would lead to gross under-powering of the study.

A NOTE ON EXACT AND CONTINUITY-CORRECTED TESTS OF NON-INFERIORITY

Exact tests are usually considered for small samples, when asymptotic normality may be in question. Farrington and Manning [2] remind us that an exact test of null hypotheses for proportions with *non-zero differences* (or *non-unity relative risks*) does not exist, thus providing no absolute reference to compare different methods. To be clear, there is in fact no permutation-based exact test *conditional* on the actual observations in an experiment when non-zero centred null hypotheses are involved (or non-unity relative risks). See Reference [13]. Chan [14] has discussed *unconditional exact tests of non-inferiority* for smaller data sets using the unconditional Z statistic described by Farrington and Manning [2].

Since non-inferiority testing with proportions conditionally has no exact test basis with non-zero centred null hypotheses and Chan's exact unconditional approach [14] does not require the continuity adjustment, such corrections were not applied in these examples. No mention or use of the adjustment is found in Reference [6]. For a comprehensive review of the necessity of such adjustments with approximate formulae, see Reference [15].

ACKNOWLEDGEMENTS

The authors are grateful to the editor and two referees for their careful reviews and suggestions that led to a substantial improvement in the exposition of the paper.

REFERENCES

1. Blackwelder WC. 'Proving the null hypothesis' in clinical trials. *Controlled Clinical Trials* 1982; **3**:345–353.
2. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypotheses of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* 1990; **9**:1447–1454.
3. Machin D, Campbell M, Fayers P, Pinol A. Sample Size Tables for Clinical Studies. Blackwell Science: Oxford, 1997.
4. 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.
5. Laster LL, Johnson MF. Non-inferiority trials: 'the at least as good as' criterion. *Statistics in Medicine* 2003; **22**:187–200.
6. Phillips KF. A new test of non-inferiority for anti-infective trials. *Statistics in Medicine* 2003; **22**:201–212.
7. Rohmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**:1703–1714.
8. Eberhardt KR, Fligner MA. A comparison of two tests for equality of two proportions. *American Statistician* 1977; **31**:151–155.
9. Fleiss JL. Statistical Methods for Rates and Proportions. Wiley: New York, 1981.
10. Morikawa T, Yoshida M. A useful testing strategy in Phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 1995; **5**(3):297–306.
11. *International Conference on Harmonization*: *Statistical Principles for Clinical Trials* (*ICH E-9*). Food and Drug Administration, DHHS, 1998.
12. Hung HMJ, Wang S-J, Tsong Y, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213–226.
13. Barndorff-Nielsen. Nonformation. *Biometrika* 1976; **63**:567–571.
14. Chan ISF. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* 1998; **17**:1403–1413.
15. Haviland MG. Yates' correction for continuity and the analysis of 2×2 contingency tables. With comments. *Statistics in Medicine* 1990; **9**:363–384.