


11-1-2010

Notes on Hypothesis Testing under a Single-Stage Design in Phase II Trial

Kung-Jong Lui

San Diego State University, kjl@rohan.sdsu.edu

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Lui, Kung-Jong (2010) "Notes on Hypothesis Testing under a Single-Stage Design in Phase II Trial," *Journal of Modern Applied Statistical Methods*: Vol. 9: Iss. 2, Article 7.

Available at: <http://digitalcommons.wayne.edu/jmasm/vol9/iss2/7>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

Notes on Hypothesis Testing under a Single-Stage Design in Phase II Trial

Kung-Jong Lui
San Diego State University,
San Diego, CA USA

A primary objective of a phase II trial is to determine future development is warranted for a new treatment based on whether it has sufficient activity against a specified type of tumor. Limitations exist in the commonly-used hypothesis setting and the standard test procedure for a phase II trial. This study reformats the hypothesis setting to mirror the clinical decision process in practice. Under the proposed hypothesis setting, the critical points and the minimum required sample size for a desired power of finding a superior treatment at a given α -level are presented. An example is provided to illustrate how the power of finding a superior treatment by accounting for a secondary endpoint may be improved without inflating the given Type I error.

Key words: Phase II trial, Type I error, power, union-intersection test, sample size, equivalence.

Introduction

One of the primary objectives in a phase II trial for a new anti-cancer treatment is to make a preliminary determination on whether the treatment has sufficient activity or benefits against a specified type of tumor to warrant its further development. Based on subjective knowledge, researchers commonly choose two response rates in advance p_0 and p_1 (where $0 < p_0 < p_1 < 1$) for the uninteresting and desirable levels, respectively. Test hypotheses: $H_0: p \leq p_0$ versus $H_a: p \geq p_1$ (Simon, 1989; Lin, Allred & Andrews, 2008; Lu, Jin & Lamborn, 2005) are considered using p_1 to determine the minimum required sample size for a desired power $1 - \beta$ of rejecting $H_0: p \leq p_0$ at a nominal α -level when $p = p_1$. This hypothesis setting can cause clinicians to misinterpret their findings that rejecting the null hypothesis $H_0: p \leq p_0$ is equivalent to supporting the alternative hypothesis $H_a: p \geq p_1$ and vice versa (Storer, 1992).

Despite employing a large sample size to meet a desired power, the probability of excluding a potentially interesting treatment from further consideration can still be large. To illustrate the above points, for example, consider testing $H_0: p \leq 0.30$ versus $H_a: p \geq 0.50$ at the 0.05 level. When using the common sample size calculation formula for a desired 90% power of rejecting $H_0: p \leq 0.30$ for $p = 0.50$ at the 0.05-level, the minimum required sample size is determined to be 49 patients.

Suppose that ($x =$) 20 patients respond among these ($n =$) 49 patients (i.e., the sample proportion response $\hat{p} = 20/49 = 0.408$). Using these data, the p-value for testing $H_0: p \leq 0.30$ is 0.049 (on the basis of normal approximation) and thereby, H_0 is rejected at the 0.05 level. Note that because $\hat{p} = 20/49 (= 0.41)$ is less than 0.50, there is no evidence that the underlying response rate p is larger than 0.50. Conversely, there is statistically significant evidence, given $\hat{p} = 20/49$, to indicate that the underlying response rate p is less than the desirable level 0.50 at the 10% level for testing

$$H_0: p \geq 0.50$$

versus

$$H_a: p < 0.50.$$

Kung-Jong Lui is a Professor in the Department of Mathematics and Statistics. Email him at: kjl@rohan.sdsu.edu.

Furthermore, when a treatment has the response rate $p = 0.35$ (which is larger than the uninteresting level $p = 0.30$) and is of potential interest, it can be shown that the probability of terminating this treatment for further consideration by not rejecting $H_0: p \leq 0.30$ is approximately 80%.

The above concerns and criticisms are partially due to the fact that the complement of $\{p|p \leq p_0\}$ is not the set $\{p|p \geq p_1\}$ and there is no explicit instruction about what should be done when the underlying response rate p falls in the borderline region $\{p|p_0 < p < p_1\}$. This motivates the recent development of a design with three outcomes, including an outcome allowable to account for other factors, including toxicity, cost or convenience, when making a decision (Storer, 1992; Sargent, Chan & Goldberg, 2001; Hong & Wang, 2007). One intuitive and logical justification of this practice is that - if the response rate of a new treatment was not much different from that of the standard treatment - it would be reasonable to recommend the new treatment for further study if the new treatment was less toxic, cheaper and/or easier to administer.

Treating both $H_0: \{p|p \leq p_0\}$ and $H_a: \{p|p \geq p_1\}$ as two separate competing null hypotheses, Storer (1992) proposed a three-outcome design to accommodate the situation in which one might reject neither H_0 nor H_a and he suggested sample size calculation based on $P(X \geq r_u | H_0) \leq \alpha$, $P(X \leq r_l | H_a) \leq \beta$, and $P(\text{rejecting } H_i | p_m) \leq \gamma$ for $i = 0, a$, where r_u and r_l are minimum and the maximum critical points satisfying the above probability constrains and where $p_m \approx (p_0 + p_1) / 2$.

On the basis of Simon's setting (1989) and the normal approximation for the binomial distribution, Sargent, Chan & Goldberg (2001) proposed a three-outcome test procedure with an inconclusive region in which neither H_0 nor H_a were rejected and they discussed sample size calculation for given errors of α and β , and the minimum probabilities of concluding correctly. Hong & Wang (2007) further extended sample size calculation to

accommodate a two-sample randomized comparative trial. In fact, the design suggested by Sargent, Chan & Goldberg (2001) can be expressed in terms of Storer's setting (1989) by treating H_0 and H_a as two competing null hypotheses in the following:

- (1) testing $H_0: p \leq p_0$ (versus $p > p_0$) at α -level, and rejecting H_0 when $X \geq r_u$ where r_u is the minimum point satisfying $P(X \geq r_u | p = p_0) \leq \alpha$;
- (2) testing $H_a: p \geq p_1$ (versus $p < p_1$) at β -level, and rejecting H_a when $X \leq r_l$, where r_l is the maximum point satisfying $P(X \leq r_l | p = p_0) \leq \beta$.

The inconclusive region then simply corresponds to the set of sample points $\{r_l < X < r_u\}$. Based on the normal approximation, it can be shown that the inconclusive region consists of

$$\{X | np_1 - Z_\beta \sqrt{np_1(1-p_1)} - 0.5 < X < np_0 + Z_\alpha \sqrt{np_0(1-p_0)} + 0.5\},$$

where Z_α is the upper $100(\alpha)^{\text{th}}$ percentile of the standard normal distribution. Note that this inconclusive region is a function of errors α , β , and the sample size, which are all operating parameters of the statistical test procedure rather than the biological characteristics of patient response to treatments. Various choices of α , β , or the sample size can lead to obtain different inconclusive regions despite that the underlying p_1 and p_0 are fixed. This is not appealing because the inconclusive region should represent the values falling in the borderline between the uninteresting and desirable levels and should be related to the biological aspects. Furthermore, it is possible that both $H_0: p \leq p_0$ and $H_a: p \geq p_1$ may be rejected in the design proposed by Sargent, Chan & Goldberg (2001); in this case, the above inconclusive region will no longer exist. This

can occur even when the sample size is moderate and both α and β errors are controlled.

To clarify this point, consider the above example of testing $H_0: p \leq 0.30$ versus $H_a: p \geq 0.50$. Given ($x =$) 20 patients with response among ($n =$) 49 patients, $H_0: p \leq 0.30$ can be rejected at $\alpha = 0.05$ level and $H_a: p \geq 0.50$ would be rejected at $\beta = 0.10$ level. When choosing $\alpha = 0.05$ and $\beta = 0.10$, by definition $r_u < r_l$ in this case and the inconclusive region does not exist. There is no discussion on what action to take when both $H_0: p \leq p_0$ and $H_a: p \geq p_1$ are rejected in the three-outcome design as proposed previously (Storer, 1992; Sargent, et al., 2001; Hong & Wang, 2007).

When determining in practice whether a new treatment warrants further study at the end of a phase II trial the decision is almost always based on multiple risk/benefit considerations rather than the testing result of a single primary endpoint, especially when no clear decision can be derived from the testing result. In other words, unless the response rate of the new treatment can be shown to be different from that of the standard treatment by a magnitude of clinical importance, relevant factors are incorporated into the determination of whether the new treatment should be studied further. Thus, it is desirable to design a test procedure that can mirror the clinical decision process in reality.

To avoid distracting readers' attention from the main focus of this article, discussion is restricted to a single-stage design. Under the proposed setting, the critical points and the minimum required sample size for a desired power of finding a superior treatment in a variety of situations are presented. Furthermore, using an idea suggested by Lin, Allred and Andrews (2008) and Lu, Jin and Lamborn (2005), an example is included to illustrate how the power of detecting a superior treatment may be improved by considering a secondary endpoint without inflating the given Type I error. Finally, another alternative procedure is considered and its difference, advantage, and

disadvantage are noted and compared with the proposed procedure.

Notation and Hypothesis Testing

Consider a phase II trial in which a random sample of size n patients is taken from a studied population and assigned to receive a new treatment under study. Suppose that x out of n patients are obtained with objective (or primary) response. Let p_0 denote the objective response rate determined from the historical data for the standard treatment. Let δ denote the level of difference such that, if the objective response rate p is larger than $p_u = p_0 + \delta$, the new treatment is regarded as superior to the standard treatment and hence is warranted for further study.

Similarly, if the objective response rate p is less than $p_l = p_0 - \delta$, the new treatment is regarded as inferior to the standard treatment and is terminated from further investigation. Recall that in the standard setting, statistical significance against $H_0: p \leq p_0$ does not provide information on how large the difference $p - p_0$ is between the new and standard treatments. By contrast, statistical significance evidence to support that $p > p_0 + \delta$ (i.e., the new treatment is larger than the standard treatment by a magnitude δ of clinical significance) will provide better evidence. Conversely, when statistically significant evidence exists that the new treatment is inferior to the standard treatment (i.e., $p < p_0 - \delta$), the new treatment may be excluded from further consideration for ethical reasons. This occurrence will not be known unless the data against the hypothesis $p \geq p_0 - \delta$ is examined. Thus, despite the fact that the main interest in a phase II trial is to find a potentially promising treatment, the critical region may also include the sample points to test the hypothesis $p \geq p_0 - \delta$. However, the calculation of sample size required for power of detecting a given $p (< p_0 - \delta)$ is of no practical interest. Defining $p_l = p_0 - \delta$ and $p_u = p_0 + \delta$, therefore, the hypotheses considered in testing are:

$$H_0: p_l \leq p \leq p_u \tag{1}$$

versus

$$H_a: p > p_u \text{ or } p < p_l.$$

$H_0: p_l \leq p \leq p_u$ will be rejected at the α -level if $x \geq x_u(\alpha_1)$ or $x \leq x_l(\alpha_2)$, where $\alpha = \alpha_1 + \alpha_2$, $x_u(\alpha_1)$ is the minimum point such that

$$P(X \geq x_u(\alpha_1) | p_u) = \sum_{x=x_u(\alpha_1)}^n \binom{n}{x} p_u^x (1-p_u)^{n-x} \leq \alpha_1, \tag{2}$$

and $x_l(\alpha_2)$ is the maximum point such that

$$P(X \leq x_l(\alpha_2) | p_l) = \sum_{x=0}^{x_l(\alpha_2)} \binom{n}{x} p_l^x (1-p_l)^{n-x} \leq \alpha_2. \tag{3}$$

Note that the hypothesis setting (1) is simply a switch between the null and alternative hypotheses when testing equivalence (Dunnett & Gent, 1997; Westlake, 1979; Liu & Weng, 1995; Liu & Chow, 1992; Hauck & Anderson, 1984; Lui, 1997a, 1997b; Lui & Cumberland, 2001a, 2001b). Note also that the above test procedure for (1) is a union-intersection test (Casella & Berger, 1990). When making an error in recommending an ineffective or harmful treatment for phase III trial is considered more serious than making an error of missing a potentially interesting treatment, an investigator may wish to choose $\alpha_1 \leq \alpha_2$.

For a given true value $p \in \{p | p > p_u\}$, the power is equal to

$$\Phi(n, p, \alpha_1, \alpha_2, \delta) = P(X \leq x_l(\alpha_2) | p) + P(X \geq x_u(\alpha_1) | p). \tag{4}$$

Thus, given p , α_1 , α_2 , and δ , a trial-and-error procedure can be applied to determine the critical points: $x_l(\alpha_2)$ and $x_u(\alpha_1)$, as well as the minimum required sample size n for a desired power $1 - \beta$ based on (4) such that

$$\Phi(n, p, \alpha_1, \alpha_2, \delta) \geq 1 - \beta. \tag{5}$$

Sample Size Determination and Critical Points

Programs were written in SAS (1990) to find the minimum required sample size n satisfying equation (5). For illustration purposes, $\delta = 2.5\%$ was arbitrarily chosen for the following discussion. Table 1 summarizes the critical points $x_u(\alpha_1)$, $x_l(\alpha_2)$, and the minimum required sample size n for $\alpha_1 = \alpha_2 = 0.10$ calculated from $\Phi(n, p, \alpha_1, \alpha_2) \geq 1 - \beta$ (5) for a desired power $1 - \beta = 0.80, 0.90$ in testing

$$H_0: p_l \leq p \leq p_u$$

versus

$$H_a: p > p_u \text{ or } p < p_l,$$

where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta = 2.5\%$; $p_0 = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$; and $p = p_0 + 0.15, p_0 + 0.20$.

For example, consider testing

$$H_0: 0.325 \leq p \leq 0.375 \text{ (i.e., } p_0 = 0.35)$$

versus

$$H_a: p > 0.375 \text{ or } p < 0.325$$

at levels of $\alpha_1 = \alpha_2 = 0.10$. If the desired power for rejecting H_0 when the underlying objective response rate p equals 0.50 is 80%, for example, based on equation (5), 77 patients would be required. Furthermore, Table 1 shows that if $(x_u(\alpha_1)) = 35$ or more patients are obtained with an objective response out of the 77 patients, then the new treatment would be recommended for further study.

On the other hand, if 19 or less patients are obtained with objective responses, the new treatment would be terminated from further consideration. Finally, if the number of patients with objective responses falls between 20 and 34, other factors would be considered to determine whether the experimental treatment warrants further study. Table 2 summarizes the corresponding critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$

and the minimum required sample size n for $\alpha_1 = 0.05$ and $\alpha_2 = 0.15$ in the same configurations as those considered in Table 1.

Discussion

Multiple factors are almost always accounted for at the end of a phase II trial to determine whether a new treatment warrants further study unless there is a clear cut decision in the testing results. The test procedure proposed herein has the advantage of resembling the actual clinical decision process more closely than the standard test procedure. By contrast, in Simon’s setting, the determination of a new treatment for further study may completely depend on the testing result of a single primary point, but this may not be the case in practice. Furthermore, in the three-outcome design, the inconclusive region depends on the operating characteristics, such as errors α , β , and the sample size, of a test procedure. Thus, the inconclusive region can change or may not even exist for different given values of these parameters even when the underlying objective response rate is fixed. For this reason the inconclusive region is defined here in terms of biological equivalence. Based on the proposed hypothesis setting (1), it is possible to control both the errors of recommending a non-superior treatment and of terminating a non-inferior treatment to be less than a given error-level.

When there is no statistical evidence against the hypothesis $H_0: p \in [p_l, p_u]$ based on the primary endpoint, a reasonable and appealing action can be to consider a secondary endpoint to improve power. For example, in traditional phase II trials, the total response (TR) rate, the sum of the complete response (CR) rate and the partial response (PR) rate, is often used as the objective (or primary) response rate p . Because CR is generally rare for many tumors, even a small increase in the number of CRs can be important in evaluation of the efficacy of a treatment. Thus, clinicians will welcome a decision rule that accepts a new treatment for further study based on an improved CR rate even when the treatment does not achieve the desirable objective response rate of TR (Lin, Allred & Andrews, 2008; Lu, Jin & Lamborn, 2005).

Table 1: The critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$ and the minimum required sample size n calculated from $\Phi(n, p, \alpha_1, \alpha_2) \geq 1 - \beta$ in equation (5) for a desired power $1 - \beta = 0.80, 0.90$ in testing $H_0: p_l \leq p \leq p_u$ versus $H_a: p > p_u$ or $p < p_l$ where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta = 2.5\%$; $p_0 = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$; $p = p_0 + 0.15, p_0 + 0.20$; $\alpha_1 = 0.10$ and $\alpha_2 = 0.10$.

p_0	p	n	$x_l(\alpha_2)$	$x_u(\alpha_1)$
$1 - \beta = 0.80$				
0.15	0.30	51	2	13
	0.35	31	1	9
0.25	0.40	68	10	24
	0.45	36	4	14
0.35	0.50	77	19	35
	0.55	41	9	20
0.45	0.60	77	26	43
	0.65	37	11	22
0.55	0.70	73	32	48
	0.75	36	14	25
0.65	0.80	59	31	45
	0.85	30	14	24
0.75	0.90	39	24	34
	0.95	16	8	15
$1 - \beta = 0.90$				
0.15	0.30	79	5	19
	0.35	45	2	12
0.25	0.40	94	15	32
	0.45	52	7	19
0.35	0.50	109	28	48
	0.55	53	12	25
0.45	0.60	105	37	57
	0.65	54	17	31
0.55	0.70	101	46	65
	0.75	50	21	34
0.65	0.80	83	45	62
	0.85	41	21	32
0.75	0.90	61	39	52
	0.95	22	12	20

Table 2: The critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$ and the minimum required sample size n calculated from $\Phi(n, p, \alpha_1, \alpha_2) \geq 1 - \beta$ in equation (5) for a desired power $1 - \beta = 0.80, 0.90$ in testing $H_0: p_l \leq p \leq p_u$ versus $H_a: p > p_u$ or $p < p_l$, where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta = 2.5\%$; $p_0 = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$; $p = p_0 + 0.15, p_0 + 0.20$; $\alpha_1 = 0.05$ and $\alpha_2 = 0.15$.

p_0	p	n	$x_l(\alpha_2)$	$x_u(\alpha_1)$
$1 - \beta = 0.80$				
0.15	0.30	73	5	19
	0.35	41	2	12
0.25	0.40	92	16	33
	0.45	48	7	19
0.35	0.50	102	27	47
	0.55	50	12	25
0.45	0.60	103	38	58
	0.65	53	18	32
0.55	0.70	95	44	63
	0.75	48	21	34
0.65	0.80	81	45	62
	0.85	41	21	33
0.75	0.90	56	36	49
	0.95	26	15	24
$1 - \beta = 0.90$				
0.15	0.30	102	8	25
	0.35	55	3	15
0.25	0.40	121	21	42
	0.45	66	10	25
0.35	0.50	136	38	61
	0.55	71	18	34
0.45	0.60	140	52	77
	0.65	72	25	42
0.55	0.70	129	61	84
	0.75	64	28	44
0.65	0.80	110	62	83
	0.85	53	28	42
0.75	0.90	78	51	67
	0.95	32	20	29

When studying the efficacy of a treatment for brain tumors the TR rate can be small as well. In this case, the objective response can be stabilization disease (SD) progression for six months after post-treatment initiation, while the secondary endpoint can be either CR or PR. For both of the above examples, a critical region may be found based on the objective and secondary responses such that if the objective response rate cannot be used to decide whether a new treatment warrants further study, an opportunity may still exist to justify the acceptance of the new treatment based on its secondary response rate subject to the originally given α_1 error. To illustrate this point, consider the example for patients with glioblastomas. On the basis of the standard for the North American Brain Tumor Consortium (NABTC), interest lies in determining whether the objective response rate of SD increases from $p_0 = 0.15$ to $p = 0.35$ (Lu, Jin & Lamborn, 2005). Thus, testing

$$H_0: 0.125 \leq p \leq 0.175 \text{ (with } \delta = 2.5\%)$$

versus

$$H_a: p > 0.175 \text{ or } p < 0.125$$

is considered. From equation (5), the minimum required number of patients is determined to be 31 patients for a desired power of 80% when $p = 0.35$ at ($\alpha_1 = \alpha_2 =$) 0.10-level and the corresponding critical points $x_l(\alpha_2)$ and $x_u(\alpha_1)$ are 1 and 9, respectively (Table 1).

When no evidence exists to claim the experimental treatment to be superior (i.e., $p > 0.175$) to the standard treatment based on the objective response rate of SD, for example, the experimental treatment may be still determined to warrant further study. This could occur if the secondary response rate, p_s , that the tumor shrinkage is sufficient to be regarded as either CR or PR for a 6-month interval is larger than 0.05.

Let x_s denote the number of patients with the secondary response among 31 patients. While keeping the above critical point $x_u(\alpha_1)$ for the objective response of SD, SAS programs are written to search for the secondary endpoint

for the critical point x_{CS} , which is the minimum point x_s such that the probability $P(X \geq 9 \text{ or } X_s \geq x_s | p_u = 0.175, p_s = 0.05) \leq 0.10$. The critical point, x_{CS} , is 5 if an observation $(x, x_s) = (8, 6)$ is obtained. Although the number ($x = 8$) of patients with the objective response of SD is not ≥ 9 , the experimental treatment may be recommended for further development because the number of ($x_s = 6$) patients with the secondary response is above the critical point ($x_{CS} = 5$). In fact, the joint power for given values p and p_s based on the trinomial distribution can also be calculated:

$$P(X \geq 9 \text{ or } X_s \geq 5 | p, p_s) = \sum_i \sum_j 1_{\{i+j \geq 9 \text{ or } i \geq 5\}} \frac{31!}{i! j! (n-i-j)!} \times p_s^i (p - p_s)^j (1-p)^{(31-i-j)} \quad (6)$$

where the indicator function, $1_{\{condition\}}$, equals 1 if the condition in braces is true, and equals 0 otherwise.

For example, when $p = 0.35$ and $p_s = 0.20$, the joint power obtained from (6) ≈ 0.88 , which is larger than the original desired actual power $P(X \geq 9 | p = 0.35) \approx 0.81$ exclusively based on the objective response by approximately 7%. Note that because the binomial distribution is discrete, the true Type I error $P(X \geq 9 | p_u = 0.175)$ based on the objective response is actually equal to 0.079, which is less than the nominal ($\alpha_1 = 0.10$) level. This is the reason why the critical region can be expanded from $\{X \geq 9\}$ to $\{X \geq 9 \text{ or } X_s \geq 5\}$ to increase power without the necessity of inflating the given α_1 error. Conaway & Petroni (1995) proposed methods for designing group sequential phase II trials with two binary endpoints.

Conaway & Petroni (1995) also focused discussion on the situation in which a new treatment is recommended for further study when the new treatment has both a high response and lower toxicity. By contrast, consider the situation in which the new

treatment is recommended for further study if the new treatment has either a high objective response rate or a high secondary response rate. Thus, Conaway & Petroni's results cannot be applicable to the situations discussed here.

It may be shown that

$$P(X \geq x | p) (= \sum_{X=x}^n \binom{n}{X} p^x (1-p)^{n-x}) \leq \alpha^*$$

if and only if the $100(1 - \alpha^*)\%$ lower confidence limit (LCL) (one-sided), given by $x / (x + (n - x + 1) F_{2(n-x+1), 2x, \alpha^*})$, falls above the underlying response rate p , where $F_{2(n-x+1), 2x, \alpha^*}$

is the upper $100(\alpha^*)^{\text{th}}$ percentile of the central F-distribution with degrees of freedom $2(n-x+1)$ and $2x$, respectively (Casella & Berger, 1990; Lui, 2004). Similarly, it can be shown that $P(X \leq x | p) \leq \alpha^*$ if and only if the $100(1 - \alpha^*)\%$ upper confidence limit (UCL) (one-sided), given by

$$\frac{\{(x+1) F_{2(x+1), 2(n-x), \alpha^*}\}}{\{(n-x) + (x+1) F_{2(x+1), 2(n-x), \alpha^*}\}}$$

falls below p . Thus, the hypothesis setting and test procedure defined in (1-3) is equivalent to the decision procedure defined as follows: when the UCL with $\alpha^* = \alpha_2$ falls below $p_l (= p_0 - \delta)$, the new treatment is terminated; when the LCL with $\alpha^* = \alpha_1$ falls above $p_u (= p_0 + \delta)$, the new treatment warrants further consideration; when neither of the above conditions hold relevant factors are accounted for in the final decision. Compared with hypothesis testing, the use of confidence intervals to present the testing results may shed light on the magnitude of the difference between the two treatments under comparison.

Rather than excluding a new treatment from further consideration when it is shown to be inferior (i.e., $p < p_0 - \delta$) to the standard treatment in the procedure proposed, an alternative procedure can be considered by including a new treatment into further

consideration only when it is shown to be non-inferior to the latter (i.e., $p > p_0 - \delta$). That is, the following design may be employed: (1) for the LCL with a given $\alpha^* = \alpha_1$ falling below $p_0 - \delta$, the new treatment is excluded from further consideration; (2) for the LCL with $\alpha^* = \alpha_1$ falling into $[p_0 - \delta, p_0 + \delta]$, accounting for other factors; and (3) for the LCL falling above $p_0 + \delta$, the new treatment is recommended for further study. To avoid missing a potentially useful treatment when a new treatment for a specified type of cancer is hard to find, the hypothesis setting and test procedure (1-3) described herein may be employed to terminate a new treatment only when it is shown to be inferior to the standard treatment. To alleviate the concern of including an inferior treatment for phase III trials, a large value for α_2 may be chosen (e.g., 0.15) in (3); on the other hand, when new experimental treatments are easier to find, the alternative decision procedure, including only those treatments shown to be non-inferior to the standard treatment for further consideration, can be of potential use.

In summary, limitations in the commonly-used hypothesis setting and the recently proposed three-outcome design have been described. The hypothesis testing has been reformatted and a test procedure proposed to more closely resemble the clinical decision process. The minimum required sample size for a desired power of finding a superior treatment at a given α -level has been presented and the corresponding critical points in a variety of situations provided. Discussion and an example were used to illustrate how power may be improved by accounting for the secondary endpoint without inflating the given Type I error in the proposed test procedure. Also included was a discussion on an alternative procedure and for which situations in which this procedure can be of use. The findings and the discussion should be helpful for clinicians when exploring a new treatment in a phase II trial.

References

- Casella, G., & Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury.
- Conaway, M. R., & Petroni, G. R. (1995). Bivariate sequential designs for phase II trials. *Biometrics*, *51*, 656-664.
- Dunnett, C. W., & Gent, N. (1997). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*, *33*, 593-602.
- Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*, 83-91.
- Hong, S., & Wang, Y. (2007). A three-outcome design for randomized comparative phase II clinical trials. *Statistics in Medicine*, *26*, 3525-3534.
- Lin, X., Allred, R., & Andrews, G. (2008). A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharmaceutical Statistics*, *7*, 88-92.
- Liu, J.-P., & Chow, S.-C. (1992). Sample size determination for the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *20*, 101-104.
- Liu, J.-P., & Weng, W.-S. (1995). Bias of two one-sided tests procedures in assessment of bioequivalence. *Statistics in Medicine*, *14*, 853-861.
- Lu, Y., Jin, H., & Lamborn, K. R. (2005). A design of phase II cancer trials using total and complete response endpoints. *Statistics in Medicine*, *24*, 3155-3170.
- Lui, K.-J. (1997a). Sample size determination for repeated measurements in bioequivalence test. *Journal of Pharmacokinetics and Biopharmaceutics*, *25*, 507-513.
- Lui, K.-J. (1997b). Exact equivalence test for risk ratio and its sample size determination under inverse sampling. *Statistics in Medicine*, *16*, 1777-1786.
- Lui, K.-J. (2004). *Statistical Estimation of Epidemiological Risk*. New York: Wiley.

Lui, K.-J., & Cumberland, W. G. (2001a). A test procedure of equivalence in ordinal data with matched-pairs. *Biometrical Journal*, *43*, 977-983.

Lui, K.-J., & Cumberland, W. G. (2001b). Sample size determination for equivalence test using rate ratio of sensitivity and specificity in paired-sample data. *Controlled Clinical Trials*, *22*, 373-389.

Sargent, D. J., Chan, V., & Goldberg, R. M. (2001). A three-outcome design for phase II clinical trials. *Controlled Clinical Trials*, *22*, 117-125.

SAS Institute Inc. (1990). *SAS Language Version 6*, 1st edition. Cary, North Carolina SAS Institute.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, *10*, 1-10.

Storer, B. E. (1992). A class of phase II designs with three possible outcomes. *Biometrics*, *48*, 55-60.

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics*, *35*, 273-280.