ELSEVIER

Controlled Clinical Trials 23 (2002) 2–14

Controlled Clinical Trials

Choosing an equivalence limit for noninferiority or equivalence studies

Brian L. Wiens, Ph.D.*

Department of Biostatistics, Amgen Inc.,

Manuscript received February 20, 2001; manuscript accepted September 24, 2001

Abstract

Studies that compare treatments with the purpose of demonstrating that the treatments are similar require an a priori definition of an equivalence limit, how different the treatments can be before the difference is of concern. Defining such an equivalence limit is one of the most difficult aspects of planning the study. Three principles are proposed for setting such limits, depending on the objective of the study: a putative placebo calculation, an approach based on clinically important differences, and methods based on statistical properties. All methods will be useful for many studies, but the study objective should determine the final choice of an equivalence limit. The statistician must play an integral role in determining the final equivalence limit. Advice is offered for helping the statistician participate in the decision on the equivalence limits.

Keywords: Equivalence limit; Active control; Role of statistician; Putative placebo

Introduction

Clinical trials are commonly conducted to show equivalence or noninferiority of an investigational treatment compared to an active control or another treatment. Equivalence trials have the objective of showing that an investigational treatment and a comparator treatment have similar effects (e.g., similar population means). Noninferiority trials have the objective of showing that an investigational treatment has an effect that is either better than or not much worse than a comparative therapy. There are many arguments both for and against doing active control equivalence or noninferiority trials rather than simple placebo controlled trials but the current state of practice is to accept such studies as necessary in many cases [1–5].

* Corresponding author: Brian L. Wiens, M/S 24-3-C, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA. Tel.: 805-447-8904; fax: 805-499-5860.

E-mail address: bwiens@amgen.com

0197-2456/02/$—see front matter
PII: S0197-2456(01)00196-9

The standard testing scheme for comparing an investigational treatment to placebo involves hypotheses of the form

$$H_0: \mu_{test} - \mu_{pbo} = 0$$
$$H_1: \mu_{test} - \mu_{pbo} \neq 0$$

where μ_{test} and μ_{pbo} refer to the population means of the investigational treatment and placebo, respectively. Blackwelder suggested writing the null and alternative hypotheses for a noninferiority trial as

$$H_0: \mu_{ac} - \mu_{test} \geq \delta_0$$
$$H_1: \mu_{ac} - \mu_{test} < \delta_0$$

where μ_{ac} refers to the population mean of the active control and δ_0 is an "equivalence limit" (also called equivalence margin or zone of indifference) [6]. This assumes that larger values of the mean are better. An appropriate test statistic is

$$z_{equiv} = \frac{\bar{x}_{ac} - \bar{x}_{test} - \delta_0}{SE(\bar{x}_{ac} - \bar{x}_{test})}.$$

This test statistic is asymptotically normal, or with smaller sample sizes can be compared to a *t* distribution critical value under certain assumptions about the distributions. For an equivalence trial, a one-sided test procedure can be used twice [7]. Note that other formulations of equivalence hypotheses have been proposed [8].

Many papers have appeared in the literature discussing methods for testing these hypotheses. In most cases, it is assumed that the equivalence limit has been chosen appropriately. With the entire testing process dependent on choosing an appropriate equivalence limit, such an assumption ignores a vital part of the testing process. If δ_0 is too large, then rejecting the null hypothesis in favor of the alternative is meaningless. If δ_0 is too small, then the power of the test will be dramatically reduced. Choosing a smaller value for δ_0 is a conservative strategy but can waste resources. Discussions on choosing an equivalence limit, either with general discussion of such studies or in a specific study or therapeutic area, are numerous, even though few of these articles have a primary focus on choosing the equivalence limit [5,9–21].

The International Conference on Harmonisation (ICH) E9 guidance has the following statement [22]:

> An equivalence margin should be specified in the protocol; this margin is the largest difference that can be judged as being clinically acceptable and should be smaller than differences observed in superiority trials of the active comparator. For the active control equivalence trial, both the upper and the lower equivalence margins are needed, while only the lower margin is needed for the active control non-inferiority trial. The choice of equivalence margins should be justified clinically.

The ICH E10 guidance has the following statement [23]:

> The margin chosen for a noninferiority trial cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with

placebo in the setting of the planned trial, but may be smaller based on clinical judgment.

The Food and Drug Administration (FDA) guidance for industry in developing drugs, biologics, or devices for rheumatoid arthritis states that the equivalence limit "represents a consensus, in that particular circumstance and for that particular claim, on what small potential difference can be considered clinically insignificant, to allow the treatments to be considered clinically equivalent" [24]. Another view is that the equivalence limit should be "small enough to be of little consequence and well within range of background variability" [25].

These various statements can generally be classified as denoting an equivalence limit based on one or more of the following three criteria:

1. A value that is small enough to conclude an effect of the test treatment compared to placebo.
2. The smallest value that would represent a clinically meaningful difference, or the largest value that would represent a clinically meaningless difference.
3. A value that is small compared to background variability or has other good statistical properties.

However, there seems to be no consensus as to the appropriate method of choosing an equivalence limit. The Committee on Proprietary Medicinal Products of the European Agency for the Evaluation of Medicinal Products recently issued a concept paper calling for more discussion on defining an equivalence limit [26].

In this paper three strategies for selecting an appropriate value of δ_0 for equivalence or noninferiority testing are discussed. The equivalence limit in the hypotheses will be designated δ_0 and the other values will be designated δ_{pbo}, δ_{clin}, and δ_{stat} as alternative ways of defending or deriving δ_0. The first method will be choosing a value δ_{pbo} such that if the difference $\mu_{ac}-\mu_{test}$ is less than δ_{pbo} it can be concluded that $\mu_{test}>\mu_{pbo}$, demonstrating effectiveness without a direct comparison to placebo. The second method will be choosing a value δ_{clin} such that concluding the difference $\mu_{ac}-\mu_{test}$ is less than δ_{clin} implies that the difference is unlikely to be of practical importance. This conclusion would result in an interpretation that the active control and the investigational treatment are similar enough so that neither would have a preferable outcome. The third method will be choosing a value δ_{stat} such that concluding that the difference $\mu_{ac}-\mu_{test}$ is less than δ_{stat} results in some desirable statistical or mathematical properties. While previous guidance noted above has recognized several criteria for an equivalence limit [22–25], it is proposed in this paper that the various criteria should be first considered individually rather than in combination. Further study of the properties is provided for some practical interpretations.

Some very simple models and assumptions are introduced in the following section to help explain some of the ideas. These simple models are not meant to be taken literally but are meant to help the reader understand some of the ideas that are presented. Most situations encountered in clinical trials will be more complex, but the ideas from this article carry through.

Strategies for selecting an equivalence limit

Three strategies for choosing an equivalence limit will be discussed: the putative placebo strategy, the clinical importance strategy, and statistical strategies.

Putative placebo

A value of δ_{pbo} should be small enough that a conclusion that μ_{ac} and μ_{test} differ by less than δ_{pbo} is tantamount to concluding that μ_{test} is larger than μ_{pbo} by at least some amount (or, $p<0.050$ in a placebo-controlled trial). One common method for choosing δ_{pbo} is as follows. The difference $\mu_{ac}-\mu_{pbo}$ can be estimated from historical studies with a 95% confidence interval, although Hauck and Anderson proposed that a different confidence level (68–90%) might be appropriate when calculating the confidence interval for $\mu_{ac}-\mu_{pbo}$ [16]. The lower bound of this confidence interval — call it ϵ — can be considered a conservative estimate of the true difference: the true difference is probably larger than ϵ. To control for model misspecification, some fraction of this amount, $\phi\epsilon$, can be used, where $0<\phi<1$. (A common value of ϕ is 0.50.) The objective in the noninferiority study becomes showing that $\mu_{ac}-\mu_{test}$ is less than $\delta_{pbo}=\phi\epsilon$.

This analysis strategy assumes that the expected difference in treatment effects is identical across various studies. If the observed mean response on treatment j in study i is $\bar{x}_j^{(i)}$, then $E\{\bar{x}_1^{(i)}-\bar{x}_2^{(i)}\}$ will be the same in every study that compares those two treatments, even when the expected responses are not always the same.

Using the confidence interval approach to show that μ_{test} is superior to placebo, it must be shown that the upper bound of the confidence interval for $\mu_{ac}-\mu_{test}$ is less than $\delta_{pbo}=\phi\epsilon$. To again simplify, assume that only one study comparing the active control to placebo has been conducted (or that one confidence interval has been calculated combining all previous studies). The difference $\mu_{test}-\mu_{pbo}$ is estimated by $(\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)})-(\bar{x}_{ac}^{(2)}-\bar{x}_{test}^{(2)})$. If the constant differences assumption holds, this quantity estimates the effect of the test treatment compared to placebo. If all effects are fixed effects, the standard error (SE) of this estimate is $2\sigma n^{-1/2}$. Then $\delta_{pbo}=\phi\epsilon=\phi[\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)}-z_\alpha(SE)]$. To conclude equivalence, the inequality that must be satisfied in the active-control trial is

$$\bar{x}_{ac}^{(2)}-\bar{x}_{test}^{(2)}+z_\alpha SE<\phi[\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)}-z_\alpha(SE)]$$

which corresponds to showing that

$$t_1(x)=[\phi(\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)})-(\bar{x}_{ac}^{(2)}-\bar{x}_{test}^{(2)})]/(1+\phi)SE>z_\alpha. \quad (1)$$

Note that this is a simplification, as it assumes equal standard errors in both studies, but it is useful for purposes of gaining an understanding of the test procedure.

Consider an alternative strategy based on linear models theory [27]. Under the constant differences assumption, an unbiased estimate of the difference between the investigational treatment and placebo is $(\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)})-(\bar{x}_{ac}^{(2)}-\bar{x}_{text}^{(2)})$. The null hypothesis that $\mu_{test}=\mu_{pbo}$ can be rejected if

$$t_2(x)=[(\bar{x}_{ac}^{(1)}-\bar{x}_{pbo}^{(1)})-(\bar{x}_{ac}^{(2)}-\bar{x}_{test}^{(2)})]/SE>z_\alpha. \quad (2)$$

If it can be shown that $t_1(x)<t_2(x)$ in all cases, then concluding equivalence via the putative placebo method in inequality (1) would imply concluding equivalence via the linear

models method in inequality (2). The numerator of $t_1(x)$ will always be smaller than the numerator of $t_2(x)$ if there is a positive effect of the active control, but the denominator of $t_1(x)$ may be smaller than that of $t_2(x)$ when $\phi < \sqrt{2} + 1$. The test $t_1(x)$, with the intermediate step of comparing the active control to placebo, also requires that the active control be better than placebo, whereas blind application of $t_2(x)$ may be used to conclude superiority of the test treatment over placebo even when the active control is worse than placebo. It can be shown that

$$\frac{\partial}{\partial\phi} t_1(x) > 0,$$

or smaller values of ϕ will result in smaller values of $t_1(x)$, when $\bar{x}_{ac}^{(1)} - \bar{x}_{pbo}^{(1)} > \bar{x}_{test}^{(2)} - \bar{x}_{ac}^{(2)}$ (see Appendix). This condition will be realized in the situation of interest: the active control is better than placebo, and the test treatment is not markedly better than the active control.

The test statistic $t_1(x)$ is of a common form, with a point estimate in the numerator and an error term in the denominator. The test statistic could be testing the hypotheses

$$H_0: \mu_{ac} - \mu_{test} \geq \phi[\mu_{ac} - \mu_{pbo}] \text{v}$$

$$H_1: \mu_{ac} - \mu_{test} < \phi[\mu_{ac} - \mu_{pbo}].$$

If $\mu_{ac} - \mu_{pbo}$ is positive (which it will be if the active control has been previously shown to be superior to placebo) then the alternative hypothesis implies that $\mu_{test} - \mu_{pbo} > 0$.

The appropriate error term should be $\text{sqrt}(\phi^2\sigma^2 n^{-1} + \phi^2\sigma^2 n^{-1} + \sigma^2 n^{-1} + \sigma^2 n^{-1}) = \text{sqrt}(1 + \phi^2)\text{SE}$, rather than the denominator in $t_1(x)$, which is $(1 + \phi)\text{SE}$. So the appropriate test statistic of the hypotheses is

$$t_3(x) = \frac{\phi[\bar{x}_{ac}^{(1)} - \bar{x}_{pbo}^{(1)}] - [\bar{x}_{ac}^{(2)} - \bar{x}_{test}^{(2)}]}{\text{sqrt}(1 + \phi^2)\text{SE}} = \frac{1 + \phi}{\text{sqrt}(1 + \phi^2)} t_1(x).$$

Since $|t_3(x)| > |t_1(x)|$ for positive ϕ and the same critical values are used, concluding a treatment effect with $t_1(x)$ implies that a treatment effect will also be concluded with $t_3(x)$, so again the use of test statistic $t_1(x)$ is conservative. Thus the use of test statistic $t_1(x)$, while not directly supported by statistical theory, is conservative in many ways, and the factor ϕ provides for some robustness against departures from the constant differences model.

If treatment by study interaction exists, then this discussion becomes more complicated. Assuming interaction in the model results in a biased estimate of the effect of test treatment after the active control study is finished. Further, the assumption of a treatment by study interaction is probably not testable as a practical matter since so few studies will be run comparing any pair of treatments. This is an inherent problem in noninferiority studies. The value ϕ can be used as an ad hoc control against treatment by study interaction, with a smaller ϕ used when there is more concern.

Now consider a slightly different issue. Not only must it be concluded that the effect of the test treatment is larger than the effect of placebo in general, but it must be concluded that the test treatment "could have been distinguished from placebo *in that trial*" [2]. Applying $t_1(x)$ mechanically will allow a problem, namely assay sensitivity [10]. One cannot automatically use this procedure in an equivalence trial unless there is some assurance that when equivalence is concluded, the test treatment will be better than placebo would have been had

placebo been used in that study. With the confidence interval method inherent in $t_1(x)$, the confidence interval for the true difference will estimate the average treatment effect of the control versus placebo across studies. This is different than estimating the treatment effect of the control versus placebo in a given study. To evaluate the comparative effect of the test treatment versus placebo, one can consider $\bar{x}_{pbo}^{(2)}$ which is the mean treatment effect of placebo that would have been observed in the equivalence/noninferiority study. This suggests that

$$P\left[\frac{\bar{x}_{\text{test}}^{(2)} - \bar{x}_{\text{pbo}}^{(2)}}{\text{SE}} > z_\alpha \,\middle|\, \text{equivalence is concluded}\right]$$

should be evaluated, which will require distributional assumptions on $\bar{x}_{\text{pbo}}^{(2)}$ or $\bar{x}_{\text{ac}} - \bar{x}_{\text{pbo}}$. This can be formalized by constructing a tolerance interval for the true difference between placebo and the active control rather than a confidence interval. Tolerance intervals may not be a practical solution as in many cases the lower bound of the tolerance interval will be negative due to having few historical placebo-controlled studies.

In summary, the strategy of showing superiority over placebo by showing equivalence to an active control and setting $\delta_{\text{pbo}} = \phi[\bar{x}_{\text{ac}} - \bar{x}_{\text{pbo}} + z_\alpha \text{sqrt}(2\sigma^2)]$ is a conservative strategy if the constant differences model holds. The chance of concluding that an ineffective drug is effective is much smaller than the advertised type I error rate. If the constant differences model does not hold, it can be either more or less likely to conclude that an ineffective drug is superior to placebo. Setting ϕ less than one, when the constant differences model does not hold, is an ad hoc strategy to use what may be thought of as inefficiency in variance estimation to avoid a high false-positive rate [8]. The value of ϕ should be in some manner based on the variability of observed treatment effects in previous studies.

Clinical importance

Testing for equivalence or noninferiority with an equivalence limit that is "clinically important" is a common strategy. In this section some of the properties of δ_{clin} will be discussed and compared to strategies for choosing δ_{pbo}.

The value δ_{clin} can be defined in a number of ways. It is usually thought of as the smallest value that would present an important difference or a value (e.g., the largest value) that would not present an important difference. If $|\mu_{\text{ac}} - \mu_{\text{test}}| > \delta_{\text{clin}}$ it may be that there is no preference between the test treatment and the active control. By "no preference" one may mean either a patient new to therapy could begin either treatment or (a possibly different idea) that a patient currently on one treatment could be switched to the other treatment. A value half that of the "undisputed clinical importance" may be an acceptable value for δ_{clin} [11].

It must be obvious that the concept of δ_{clin} is imprecise at best and indefinable at worst. Each physician, and probably each patient, will have a different concept of an important difference in a health-care setting. This limits the usefulness of δ_{clin}. If an equivalence limit is chosen based on a clinically meaningful difference, it is important to present not only the hypothesis test results but also the confidence interval estimate of the true difference when discussing the study, with a note that each reader must decide individually whether the difference is meaningful.

It is also important to define δ_{clin} in the context of the hypothesis testing. If the null and alternative hypotheses are based on whether $|\mu_{ac} - \mu_{test}| > \delta_{clin}$, then δ_{clin} is a difference in population means. This can be very different than defining δ_{clin} to be a difference in the two observations or the theoretical difference in response of one individual to two different treatments. (In the next section such a criterion for equivalence will be discussed in more detail.) A change of δ_{clin} in the average treatment effect implies a change of δ_{clin} in the expected value of the response of an individual when changing treatment for that individual. However, if there is a treatment by individual interaction, then a difference of δ_{clin} in the expected value of one individual study subject is not necessarily a result of $|\mu_{ac} - \mu_{test}| = \delta_{clin}$. It is often assumed that a subject who responds well to one treatment will also respond well to the other, but this assumption is rarely tested (or testable, except in a crossover study). Just as treatment by study interaction poses problems for δ_{pbo}, treatment by subject interaction poses problems for δ_{clin}.

The quantity ϕ introduced in the previous section also has interpretation for δ_{clin}. Letting $\delta_{clin} = \phi\epsilon$ (some proportion of the effect of the active control versus placebo) requires that a new therapy retains some effect of the active control. If not used for robustness against departures from the constant differences assumption, $\phi < 1$ is protecting some clinically important benefit and not preventing a false conclusion that a test treatment is better than placebo.

The "number needed to treat" (NNT) has been proposed for interpretation of clinical trials [28]. The NNT is the inverse of the absolute risk reduction and estimates the number of subjects that would need to be exposed to the new therapy instead of the standard to see one less event (e.g., one less adverse event or one less death). Such a number could also be used in setting up an equivalence limit when the outcome is binomial. The inverse of NNT could serve as an upper bound for δ_{clin}. NNT might be smaller in a rare disease, supporting a larger value of δ_{clin}. NNT may also be smaller for specialists than for physicians in general practice; the context must be considered before δ_{clin} is determined.

Some have argued that the approach of setting a clinically meaningful difference is inappropriate, as any difference in certain endpoints such as death will be meaningful [18]. Others have commented that the smallest treatment difference of clinical importance may be quite large if the new treatment offers improvements in administration, adverse events, or costs [29]. This apparent disagreement is mainly a difference in semantics, as there is agreement that an acceptable reduction in efficacy is best viewed in terms of cost-benefit analysis of efficacy with safety or administrative differences. It is likely that any difference in death or irreversible morbidity that will not be noticed by the physician or patient (such as one in a thousand or one in a million) will be small enough that a sample size attainable in a clinical trial will not have sufficient power to conclude equivalence. When differences between treatments are not known before the study commences (such as when there may be a difference in safety profile rather than a difference in ease of administration), it may not be possible to define δ_{clin} a priori. The potential biases of defining δ_{clin} after the study should be weighed against the cost and inconvenience of better understanding the differences by completing a pilot study before the definitive equivalence/noninferiority study. A practical alternative would be to pick a value of δ_{clin} a priori with the understanding that it is appropriate only if the safety profiles (or other information in the study) do not show an unanticipated result.

Statistical models

In this section some ideas on setting the equivalence limits to take advantage of statistical properties are discussed. Two measures of closeness are presented as examples of measures other than the standard $\mu_1 - \mu_2$ to study equivalence. The methods can be used by the statistician and client to explore the consequences of choosing various equivalence limits or to brainstorm on equivalence limits.

Let δ_{stat} be the equivalence limit chosen for statistical properties. One idea is to set $\delta_{stat} = k\sigma$. This has been used in an example to demonstrate an equivalence method without comment on the appropriateness of doing so [30]. A study with higher than expected variability would not have an adverse impact on the test if $\delta_{stat} = k\hat{\sigma}$, an obvious choice when σ is not known. The value $|\mu_{ac} - \mu_{test}|/\sigma$ is known as the standardized increment or the treatment effect and can be used to help decide what differences are clinically important [31], although cautions presented by Lenth on the standardized increment should be noted [32].

A value for δ_{stat} can be considered with various measurements of important differences between observations such as a nonparametric analysis of equivalence based on $P(|X_i - Y_j| > \theta)$ where X_i is a value from treatment 1 and Y_j is a value from treatment 2 [33]. The number of times that pairs of observed values differ by more than θ is the basis for a test of this measure. Assume that data from both treatments are normally distributed. $P(|X_i - Y_j| > \theta)$ under the null can be calculated from the distributional assumptions. Let $(X, Y) \sim N(\underline{\mu}, \Sigma)$. If $\underline{\mu} = (\mu_X, \mu_Y), \mu_X - \mu_Y = k\sigma$, and $\Sigma = \sigma^2 I$, then P

$$(|X_i - Y_j| > \theta) = P(X_i - Y_j > \theta) + P(X_i - Y_j < \theta) =$$

$$1 - \Phi\left(\frac{\theta - k\sigma}{\sigma\sqrt{2}}\right) + \Phi\left(\frac{-\theta - k\sigma}{\sigma\sqrt{2}}\right)$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. Consider $k=1$ and $\theta = 2\sigma$ as an example. $P(|X_i - Y_j| > \theta) = 1 + - \Phi(-\sqrt{2/2})\Phi(3\sqrt{2/2}) = 0.257$. If $\mu_{ac} - \mu_{test} = \sigma$($H_o$ is true with $k = 1$) then $P(|X_i - Y_j| > 2\sigma | H_o) = 0.257$. If $k=1/3$, the equivalence limit is $\mu_{ac} - \mu_{test} = \sigma/3$, and $P(|X_i - Y_j| > 2\sigma Ho) = 0.169$. This can be compared to $P(|X_i - Y_j| > 2\sigma | \mu_{ac} = \mu_{test}) = 0.157$, very close to the value when $k=1/3$. The approach can also be adapted for noninferiority studies and is statistically interpretable even if crossover study designs cannot be employed. Hauck, Hyslop, and Anderson proposed a similar metric, $P(X_i > Y_j)$ [34].

Heyse and Stine proposed another criterion for equivalence, the proportion of similar responses (PSR) [35]. This is essentially the area that is under both density curves (assuming continuous data), or $PSR = \int_{-\infty}^{\infty} \min[f_i(x)]dx$ with $0 \le PSR \le 1$. A large value of PSR is indicative of a similarity in distributions. This is another tool to study the effect of various equivalence limits. As above, let $\delta_{stat} = k\sigma$, with $k=0$, 1/3, or 1, with data normally distributed. For $k=0$, PSR=1. For $k=1/3$, PSR=0.868. For $k=1$, PSR=0.617. Such a criterion is not appropriate for noninferiority without some adjustments.

The above interpretations of these approaches make distributional assumptions that often are not appropriate — a normal distribution is a stronger assumption than the induction of the central limit theorem. Both methods were proposed for nonparametric settings, and they can be used as such. (The normal assumption was for ease of presentation.) Choosing single val-

ues of k, θ, and λ or of PSR will also be difficult, perhaps more difficult than choosing δ_{clin}. Clinicians are trained to be more interested in differences in individual patients than differences in group means, so specifying an important difference in individual responses along with corresponding probabilities might be easier for the clinician than specifying an important difference in means. The prepared statistician will play an important role in helping the clinician understand the implications of choosing various values of δ_0 by elucidating the resulting statistical properties.

Other choices

Other rules and guidelines have been proposed for determining an equivalence limit. In pharmacokinetic and bioavailability studies, a ratio of means between 0.80 and 1.25 is often used for determining bioequivalence. In anti-infective studies, the U.S. FDA Division of Anti-Infective Drug Products proposed an adaptive equivalence limit, in which the acceptable difference in response rates is either 10, 15, or 20 percentage points depending on the observed cure rates in the study [36–38]. These equivalence limits seem to be based on reasoning other than what is discussed in this paper. Although these limits have advantages (ease of implementation, reasonable and consistent sample size requirements, etc.), it is difficult to justify them in general as appropriate compared to a placebo effect, or as clinically important, without further information.

Comparisons, contrasts, and comments

Three strategies for determining equivalence limits have been presented. In this section the three will be compared and contrasted.

Necessity of a clear objective

It is imperative that the objective of the study be known and stated prior to searching for an appropriate value of δ_0. That objective will be to show that the test treatment is superior to placebo, to show that the active control is not much better than the test treatment, or to show that there is not an important difference between the test treatment and the active control.

If a value of δ_{pbo} can be found, even crudely, it will be an upper bound for a noninferiority limit, whatever the objective of the study. Such a value of δ_0 is inherently one-sided in nature: it makes sense to use δ_{pbo} as an upper bound on $\mu_{ac}-\mu_{test}$ for testing noninferiority but not as a lower bound for testing equivalence. Finding a value of δ_{pbo} requires that the difference between the active control and placebo is well documented. If the only objective of the trial is to show that the investigational treatment is active, then δ_{pbo} may be used as the equivalence limit, even if $\delta_{pbo}>\delta_{clin}$.

If a value of δ_{pbo} cannot be found, there may be a temptation to continue the study with δ_{clin}. In most situations this should not be done, since it is possible to accept as noninferior a

treatment that is only as good as or even worse than placebo. However, there may be special cases when it is possible. Consider comparing a treatment to placebo for purposes of evaluating an important safety endpoint. It would be absurd to try to define a noninferiority limit based on the difference between the control and placebo, since the control is placebo. If it is desired to show that the treatment is not clinically worse than placebo for incidence of the safety endpoint, choosing the noninferiority limit based on δ_{clin} in the absence of δ_{pbo} may be appropriate. Another example is consistency lots for new vaccine products, in which three or more investigational lots are compared to each other to confirm consistency of the manufacturing process [39–42]. In this case, all of the treatments are investigational, so the concept of δ_{pbo} is not appropriate. These are two imperfect counterexamples to the argument that one can never choose a noninferiority limit when δ_{pbo} is not evaluable, as each is a departure from the usual active control paradigm. Planning a study of noninferiority or equivalence based on δ_{clin} in the absence of knowledge about the effect of an active control versus placebo will rarely be appropriate, and the burden of proof will be on the study sponsor to defend such an equivalence limit.

Simply showing that $\mu_{test} - \mu_{ac} < \delta_{pbo}$ is not sufficient to show that the two treatments are interchangeable when the objective is to rule out clinically important differences. Input from the clinician on an important difference will be valuable in determining equivalence. Such input is often looked upon skeptically by the statistician, as it involves biases and judgments. However, it will provide information on how the treatment must perform before it will be readily used by the medical community. In registrational clinical trials, it is probably more important for marketing purposes than for registration purposes.

When there are multiple objectives, multiple values of δ may be chosen before the study begins. Such planning might be appropriate when one value is chosen in the manner of δ_{pbo} for purposes of confirming efficacy and supporting registration and another value is chosen in the manner of δ_{clin} for purposes of a favorable package circular or for publications to support marketing objectives. The type I error rate can be controlled with a simple modification of a testing procedure that has been proposed for conditional superiority testing in noninferiority studies [43,44].

The statistical approach to equivalence limits is valuable for exploring various equivalence limits or brainstorming on possibilities. The approach of δ_{stat} will also be useful to critique a value of δ_{pbo} or δ_{clin} that has been proposed.

Role of the statistician

The role of the statistician in selecting a value of δ_0 has sometimes been minimized, even in the statistical literature. Schuirmann stated that the equivalence limits should be made “by the experts in the fields of biopharmaceutics and medicine (not by the statistician!)” [7]. Chuang-Stein wrote that “the choice of (δ) should be based on clinical and not statistical considerations” [45]. For selecting a value of δ_0 in the manner of δ_{clin}, this may be at least partially correct. However, the statistician must be involved in setting δ_{pbo}, which will be a bound for δ_0. If the clinical “expert” picks a value of δ_{clin} that is large enough to allow equivalence with placebo (or worse than placebo), the statistician must be ready to advise against such a limit. The statistician can also use δ_{stat} to assist the client in choosing a value of δ_{clin}.

Summary

In few equivalence or noninferiority studies will there be one obvious choice for an equivalence limit. The intent of this paper is not to supply a strategy that will result in a single choice of an equivalence limit. Rather, the intent of this paper is to provide several options. The statistician and clinician, in consultation with regulatory authorities for registration studies, should collaborate to find an equivalence limit during the planning phase of the study. (See NG [46] for one recent opinion from a regulatory reviewer.) When determining an appropriate value for an equivalence limit in a given study, each of the three strategies discussed should be considered. Depending on the objective of the study, any or all may be of interest and of help in choosing the value to be used in the study. Interpretations of the various strategies should be presented to the client for input and feedback, and the final equivalence limit can then be selected jointly by the statistician and client. This process will not be quick or painless, but will result in an improved understanding of the chosen equivalence limit by all parties, and is a better plan than an instant or routine one-size-fits-all approach to choosing δ_0. The strategies discussed in this paper are an attempt to make an inherently subjective process a little more objective.

Acknowledgments

The author thanks William Blackwelder, Paul Flyer, the editor, and two anonymous referees for providing helpful criticism and review of this paper.

References

[1] Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med 1994;331:394–398.
[2] Temple RJ. When are clinical trials of a given agent vs. placebo no longer appropriate or feasible? Control Clin Trials 1997;18:613–620.
[3] Temple RJ, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments, Part 1: ethical and scientific issues. Ann Intern Med 2000;133:455–463.
[4] Ellenberg SS, Temple RJ. Placebo-controlled trials and active-control trials in the evaluation of new treatments, Part 2: practical issues and specific cases. Ann Intern Med 2000;133:464–470.
[5] Simon R. Are placebo-controlled clinical trials ethical or needed when alternative treatment exists? Ann Intern Med 2000;133:474–475.
[6] Blackwelder WC. “Proving the null hypothesis” in clinical trials. Control Clin Trials 1982;3:345–353.
[7] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 1987;15:657–680.
[8] Holmgren EB. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. J Biopharm Stat 2000;9:651–659.
[9] Windeler J, Trampisch H-J. Recommendations concerning studies on therapeutic equivalence. Drug Inf J 1996;30:195–200.
[10] Temple R. Problems in interpreting active control equivalence trials. Accountability in Research 1996;4:267–275.
[11] Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996;313:36–39.
[12] Rosenkranz G. Can we reduce the dose of a vaccine? Control Clin Trials 1997;18:43–53.
[13] Stenijans VW, Neuhäuser M, Hummel T, et al. Asthma management: the challenge of equivalence. Int J Clin Ther 1998;36:117–125.
[14] Califf RM. A perspective on the regulation of the evaluation of antithrombotic drugs. Am J Cardiol 1998;82(8B):25P–35P.
[15] Ebbutt AF, Frith L. Practical issues in equivalence trials. Stat Med 1998;17:1691–1701.
[16] Hauck WW, Anderson S. Some issues in the design and analysis of equivalence trials. Drug Inf J 1999;33:109–118.
[17] Hwang IK, Morikawa T. Design issues in noninferiority/equivalence trials. Drug Inf J 1999;33:1205–1218.
[18] Siegel JP. Equivalence and noninferiority trials. Am Heart J 2000;139:S166–170.
[19] Källén A, Larsson P. On the definition of therapeutic equivalence. Drug Inf J 2000;34:349–354.

[20] Gould AL. Sample sizes for event rate equivalence trials using prior information. Stat Med 1993;12:2009–2323.
[21] Garbe E, Rohmel J, Gundert-Remy U. Clinical and statistical issues in therapeutic equivalence trials. Eur J Clin Pharmacol 1993;45:1–7.
[22] International Conference on Harmonisation. E9: Guidance on statistical principles for clinical trials. Federal Register 63(179), September 16, 1998.
[23] International Conference on Harmonisation. E10: Choice of control group in clinical trials. Federal Register 64(185), September 24, 1999.
[24] Guidance for industry: clinical development programs for drugs, devices, and biological products for the treatment of rheumatoid arthritis. Food and Drug Administration, Washington, DC, February 1999.
[25] Meyerson LJ, Wiens BL, LaVange LM, Koutsoukos AD. Understanding clinical trials: quality control in oncology clinical trials. In: Allegra CJ, Kramer BS, editors. Hematology/Oncology Clinics of North America. Philadelpha: WB Saunders, 2000.
[26] Concept paper on the development of a committee for proprietary medicinal products (CPMP) points to consider on biostatistical/methodological issues arising from recent CPMP discussions on licensing applications: choice of delta. Committee on Proprietary Medicinal Products, London, 23 September 1999.
[27] Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. Drug Inf J 2001;35:435–449.
[28] Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–454.
[29] Fleming TR. Design and interpretation of equivalence trials. Am Heart J 2000;139:S171–S176.
[30] Giani G, Straβberger K. Testing and selecting for equivalence with respect to a control. J Am Stat Assoc 1994;89:320–329.
[31] Feinstein AR. Indexes of contrast and quantitative significance for comparisons of two groups. Stat Med 1999;18:2557–25581.
[32] Lenth RV. Some practical guidelines for effective sample size determination. The American Statistician 2001;55:187–193.
[33] Ponnapalli RM, Dubey SD. Nonparametric tests for equivalence. ASA 1991 Proceedings of the Biopharmaceutical Section. 1991. p. 102–105.
[34] Hauck WW, Hyslop T, Anderson S. Generalized treatment effects for clinical trials. Stat Med 2000;19:887–899.
[35] Heyse JF, Stine R. Use of the overlapping coefficient for measuring the similarity of treatments. ASA 2000 Proceedings of the Biopharmaceutical Section. 2000, p. 29–32.
[36] Points to consider: clinical development and labeling of anti-infective drug products. Food and Drug Administration Division of Anti Infective Drug Products, Washington, DC, October 26, 1992.
[37] Bristol DR. Determining equivalence and the impact of sample size in anti-infective studies: a point to consider. J Biopharm Stat 1996;6:319–326.
[38] Wiens BL, Iglewicz B. Testing noninferiority of response rates for regulatory filings using transformations. Drug Inf J 2001;35:1165–1171.
[39] Wiens BL, Iglewicz B. On testing equivalence of three populations. J Biopharm Stat 1999;9:465–483.
[40] Wiens BL, Iglewicz B. Design and analysis of three treatment equivalence trials. Control Clin Trials 2000;21:127–137.
[41] Ng T-H. Equivalence testing with three or more treatment groups. American Statistical Association 2000 Proceeding of the Biopharmaceutical Section. 2000, p. 156–160.
[42] Wiens BL, Heyse JF, Matthews H. Similarity of three treatments, with application to vaccine development. ASA 1996 Proceedings of the Biopharmaceutical Section. 1996, p. 203–206.
[43] Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. Stat Med 1996;15:1729–1738.
[44] Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. J Biopharm Stat 1995;5:297–306.
[45] Chuang-Stein C. Clinical equivalence—a clarification. Drug Inf J 1999;33:1189–1194.
[46] Ng T-H. Choice of delta in equivalence testing. Drug Inf J 2001;35:1517–1527.

Appendix

$$\frac{\partial}{\partial\phi}t_1(x) = \{(1+\phi)[\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}]\sqrt{2\sigma^2/n} -$$
$$[\phi(\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}) - (\bar{x}_{\mathrm{ac}}^{(2)} - \bar{x}_{\mathrm{test}}^{(2)})]\sqrt{2\sigma^2/n}\}/[(1+\phi)(\sqrt{2\sigma^2/n})]^2 > 0$$

$$\Leftrightarrow$$

$$(1+\phi)[\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}]\sqrt{2\sigma^2/n} -$$
$$[\phi(\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}) - (\bar{x}_{\mathrm{ac}}^{(2)} - \bar{x}_{\mathrm{test}}^{(2)})]\sqrt{2\sigma^2/n} > 0 \text{ since denominator} > 0$$

$$\Leftrightarrow$$

$$\sqrt{2\sigma^2/n}\{(1+\phi)[\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}] - [\phi(\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}) - (\bar{x}_{\mathrm{ac}}^{(2)} - \bar{x}_{\mathrm{test}}^{(2)})]\} > 0$$

$$\Leftrightarrow$$

$$(\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)}) + (\bar{x}_{\mathrm{ac}}^{(2)} - \bar{x}_{\mathrm{test}}^{(2)}) > 0$$

$$\Leftrightarrow$$

$$\bar{x}_{\mathrm{ac}}^{(1)} - \bar{x}_{\mathrm{pbo}}^{(1)} > \bar{x}_{\mathrm{test}}^{(2)} - \bar{x}_{\mathrm{ac}}^{(2)} .$$