Clinical Trials

The general concepts of an equivalence trial, applied to ASSENT-2, a large-scale mortality study comparing two fibrinolytic agents in acute myocardial infarction

E. Lesaffre¹, E. Bluhmki², F. Wang-Clow³, S. Berioli², T. Danays², N. L. Fox⁴ and F. Van de Werf⁴

¹Biostatistical Centre, University of Leuven, Belgium; ²Boehringer Ingelheim Pharma KG, Ingelheim, Germany; ³Genentech Inc., San Francisco, USA; ⁴Department of Cardiology, University of Leuven, Belgium

Introduction

About a decade ago clinical trials were almost exclusively ‘superiority’ trials. This situation has changed, and nowadays ‘equivalence’ trials have gained popularity. Loosely speaking, the intention of an equivalence trial is to show that two treatments have about the same therapeutic effect. This is in contrast to a superiority trial, where the purpose is to show superiority of the new active treatment over placebo or over the standard active treatment. Thus, the goals set for an equivalence trial are less ambitious than the goals set for a superiority trial. Why then the recent interest in equivalence trials?

For new drugs within a particular pharmacological class it is becoming increasingly difficult to improve upon existing treatments. In particular, the introduction of thrombolytics for acute myocardial infarction patients has significantly decreased the 30-day mortality over the last two decades from about 10% to 6.3% in GUSTO-1[1]. A further reduction in mortality is therefore not very likely. While one might strive for new thrombolytic treatment to be superior to existing treatment, it is more realistic to hope for an equally good performance regarding mortality and to focus on other possible benefits of the new treatment, such as: side effects, cost, ease of administration. There may be several reasons why the equivalence of two treatments is of scientific and economic interest.

The meaning of a non-significant result in a superiority trial

It is useful to take academic examples to clarify ideas. In all examples two thrombolytic treatments, A and B, are compared for 30-day mortality in the treatment of acute myocardial infarction patients. In the first example, it is assumed that in a superiority study the 30-day mortality rate for A is mA=6% and for B mB=8%. The observed mortality rates will be denoted as mA (mB), while for the true mortality rates $\pi_A$ ($\pi_B$) will be used. Suppose the $P$ value obtained from a statistical test (say a chi-square test) yields a significant $P$ value (<0.05). In that case one concludes that treatment A is probably better for treating acute infarction patients than treatment B, when evaluated with 30-day mortality, i.e. that probably $\pi_A<\pi_B$. When for the same mortality rates a $P$ value $\geq 0.05$ is obtained, one concludes that the study does not give evidence for a real difference in treatment effects.

In a second study, 8% 30-day mortality is obtained in both arms of the study. In that case the $P$ value approximately equals 1, and the result is, of course, not significant (at $\alpha=0.05$). However, one cannot infer that treatments A and B have equal 30-day mortality rates. Indeed, when the study involves 50 patients in each arm, 8% mortality simply means that for each treatment four patients died within 30 days after randomization. Clearly, with such a small study population the equal results could have been obtained by chance, and we would not claim that the two treatments are equal in performance in future patients. Hence we would not dare claim that $\pi_A=\pi_B$. This uncertainty is numerically summarized by the 95% confidence interval (CI) of the difference in 30-day mortality rates, given by $[-11\% , 11\%]$. In general, the 95% CI of the difference of two mortality rates expresses the region of uncertainty of the true difference in mortality rates. In other words,

Key Words: Alteplase, equivalence, mortality, myocardial infarction, non-inferiority, tenecteplase.

Revision submitted 25 May 2000, and accepted 7 June 2000.

Correspondence: Emmanuel Lesaffre, Biostatistical Centre, U.Z. St. Rafael, Kapucijnenvoer 35, 3000 Leuven, Belgium.
the 95% CI expresses our belief in what the true difference in mortality rates might be, i.e. it expresses our belief in $\pi_A - \pi_B$.

With the above study of 50 patients per arm, the 95% CI of $[-11\%, 11\%]$ simply states that with 0.95 probability, treatment A might yield up to an 11% higher or an 11% lower 30-day mortality rate than treatment B. This clearly represents a large uncertainty. With 5000 patients per arm and the same mortality rates, the 95% CI for the true difference in 30-day mortality rates equals $[-1\%, 1\%]$, which is clearly much smaller. Hence, the larger the study the less uncertainty we have about the true difference in 30-day mortality rates. The latter study indicates that the two treatments are close in performance since there is still uncertainty about the true difference.

How many patients do we need to reduce the above uncertainty to zero? This is equivalent to asking how many patients do we need to produce a 95% CI of zero. To answer this question, let us look at a study with 50,000 patients per arm. With the same 30-day mortality rates of 8%, the 95% CI for the difference in mortality rates becomes $[-0.3\%, 0.3\%]$. Apparently there is still uncertainty about the true difference. In fact, there will always be uncertainty, irrespective of the sample size. Indeed, one can never conclude from a statistical study that two treatments are equally good. One can only show that they are close in performance. In other words, a negative study testing for superiority, i.e. a study with a non-significant result, never implies (in a statistical sense) that the two treatments are equally good in performance.

**Testing equivalence**

Since we can only show that two treatments are close in performance we need to define ‘close’. Therefore we must define a ‘region of therapeutic equivalence’. Most often this region is an interval. Here we need to define an interval of differences in 30-day mortality rates, which we value as therapeutically equivalent. Clinical reasoning might dictate an interval $[-1\%, 1\%]$. In that case we claim that the two treatments, A and B, are therapeutically equivalent when they differ from each other in 30-day mortality rate by at most 1% (in absolute value). Since we never know in practice what the true mortality rates are, we need a rule to claim in practice therapeutic equivalence. This rule must be based on the mortality rates found in the study and must also take into account the uncertainty with which the true rates are estimated. Thus we require, in practice, that the 95% CI for the true difference lies (entirely) in the region of therapeutic equivalence before we call the two treatments (therapeutically) equivalent. Good statistical practice requires that this interval of equivalence be fixed in advance.

The definition of therapeutic equivalence is derived from the associated statistical test for equivalence (on a two-sided 0.05 significance level). It can be shown that this test consists of determining the 95% CI for $\pi_A - \pi_B$ and evaluating whether this CI lies entirely in the interval of therapeutic equivalence. For example, suppose that the interval of equivalence is a priori defined as $[-1\%, 1\%]$. Further, suppose that a study is based on 5000 patients per arm, and that each treatment yields a 30-day mortality rate of 8%. Then the 95% CI for the true difference is given as $[-1\%, 1\%]$. Since this interval exceeds the region of therapeutic equivalence, the two treatments cannot be called therapeutically equivalent, even when observing the same mortality rates. This is because there is too much uncertainty about the true difference in mortality rates. Namely, this true difference might be above 1% or below −1%. On the other hand, when the same results were obtained on 50,000 patients per arm, the 95% CI was given by $[-0.3\%, 0.3\%]$, and since this interval lies entirely in the region of therapeutic equivalence, the two treatments will then be called therapeutically equivalent.

To summarize, when the 95% CI of the difference in 30-day mortality rates lies entirely in the region of therapeutic equivalence, the two treatments will be regarded as (therapeutically) equivalent. This corresponds to a significant $P$ value ($<0.05$) for an equivalence test, implying the equivalence of the two treatments. When $P \geq 0.05$, the result is not-significant and this signifies, for an equivalence test, that the two treatments cannot be declared equivalent. Compare this interpretation with a $P$ value $<0.05$ for a superiority trial. In that case the study does not provide evidence for a difference between the two treatments in their 30-day mortality rates. Observe that we have formulated this non-significant result as ‘the study does not give evidence for . . .’. In other words we actually say that the study was not powerful enough to observe a difference, which is not the same as claiming that there is evidence that the two treatments are close in performance. Finally, observe also that the $P$ value for a superiority trial is calculated differently from that for an equivalence trial.

To summarize, in terms of a 95% CI, a significant $P$ value ($<0.05$) is obtained from an equivalence trial when the 95% CI of the difference in mortality rates lies entirely in the region of equivalence. On the other hand, for a superiority trial a $P$ value $<0.05$ implies that the zero difference is not contained in this confidence interval.

**Testing non-inferiority**

Let us suppose that $[-1\%, 1\%]$ is the region of therapeutic equivalence in the difference in 30-day mortality rates between treatment A and treatment B, namely for the true difference $\pi_A - \pi_B$. When the observed difference for $\pi_A - \pi_B$ equals $-3\%$, treatment A has a lower observed 30-day mortality rate. While this is an excellent result for treatment A, with the above definition of the
region of equivalence the treatment can never be called equivalent to B with such an outcome, whatever the sample size. Thus, even when this result is significant in a superiority trial, when set up as an equivalence trial, this study would be called ‘negative’. But clearly, one would not like to miss this better result for treatment A. Adjusting the definition of equivalence could solve this problem.

When the previous interval of equivalence \([-1\%, 1]\%\) is replaced by \([-100\%, 1]\%\) the new definition of equivalence admits all negative differences for true 30-day mortality rates. In other words, with the new region of equivalence, treatment A is said to be ‘equivalent to treatment B’ when it has a much lower 30-day mortality rate than B or is worse than B by 1% at most. In statistical terms, this type of equivalence is deemed to be ‘non-inferior’. In our case, this means that when the true difference \(\pi_A-\pi_B\) lies in the interval \([-100\%, 1]\%\) treatment A will be called ‘not inferior to treatment B’. When \(-100\%\) is not a possible difference, it will have to be replaced by its minimum possible value, which we will denote as \(-\min\%\).

Similar to equivalence, we can only claim non-inferiority of treatment A over treatment B when the 95% CI for \(\pi_A-\pi_B\) lies entirely in the interval of non-inferiority, e.g. in \([-\min\%, 1]\%\). For this type of region, it is customary to take the 95% CI as one-sided, i.e. open-ended at the left-hand side (or at the right-hand side when positive differences are aimed at). Namely, the left end of the one-sided 95% CI is minimally \(-\min\%\). But, one-sided CIs are not very popular in the medical literature and that is why the one-sided 95% CI is replaced by a two-sided 90% CI, which has the same right end-point (which is the only relevant end-point). For example, when the observed 30-day mortality rates are 6% for treatment A and 7% for treatment B with 8000 patients per arm, the one-sided 95% CI is \([-\min\%, 0.36\%]\) and the two-sided 90% CI is equal to \([-1.64\%, -0.36\%]\). Hence, the relevant boundary value of the confidence interval is \(-0.36\%\), which is the same for both intervals.

The region of non-inferiority for the ASSENT-2 study

Many so-called equivalence studies are in fact non-inferiority trials, as is the ASSENT-2 trial\[3\]. When the region of non-inferiority is \([-\min\%, 1]\%\), Fig. 1 is obtained. The interpretation of this figure goes as follows: choose a value for \(\pi_{\text{alteplase}}\) (abbreviated as \(\pi_A\)), say 7%, then draw a vertical line and observe where this line intersects with the line ‘DIFF=1%’. Clearly, this intersection will, in this case, be at 8% mortality with tenecteplase. All values for \(\pi_{\text{tenecteplase}}\) (abbreviated as \(\pi_B\)) below this 8% belong to the region of non-inferiority. Thus, we allow that tenecteplase has, at most, a 1% higher 30-day mortality rate than alteplase in order to conclude that it is not inferior to alteplase (of course in practice the two-sided 90% CI also needs to be taken into account). The above region of non-inferiority becomes quite loose for low mortality rates. For instance, suppose (in the unlikely event) that the 30-day mortality rate for alteplase equals 1%, then the region of non-inferiority allows a doubling of this mortality rate for tenecteplase, which is clearly unacceptable. That is why Fig. 1 is replaced by Fig. 2 for the definition of the region of non-inferiority in the ASSENT-2 study\[3\]. For mortality rates higher than or equal to 7.2%, the 30-day mortality rate obtained from the GUSTO III study\[3\], the absolute criterion of 1% still applies. But, for lower mortality rates a relative criterion defines the region of non-inferiority. When the relative risk \(\pi_A/\pi_B\) is lower than 1/14, tenecteplase is declared therapeutically not inferior to alteplase. For example, when the true 30-day
mortality rates are 5.9% for tenecteplase and 5% for alteplase, then tenecteplase would be called therapeutically not inferior to alteplase according to Fig. 1, because the difference is less than 1%. However, the relative risk equals 1.18, higher than 1.14 and thus, according to the ASSENT-2 definition, tenecteplase can not be called therapeutically equivalent (not-inferior) to alteplase. For a 5% mortality rate for alteplase, the ASSENT-2 definition of non-inferiority implies that the mortality rate of tenecteplase cannot be higher than 5% × 1.14 = 5.7%. Of course, as repeatedly stated, we do not know these ‘true’ mortality rates and therefore we claim that tenecteplase is not inferior to alteplase when the two-sided 90% CI of their true difference lies entirely in the region of non-inferiority, as exhibited in Fig. 2.

Observe that at 7.2% the absolute criterion and the relative criterion coincide since 8.2%/7.2% = 1.14.

Finally, in addition to INJECT[13], COBALT[14] and COMPASS[15], the ASSENT-2 trial is a leading example on how to design a non-inferiority study with regard to 30-day mortality when comparing two fibrinolytic agents. It reflects a broad scientific consensus, especially on the definition of the particular shape of the non-inferiority region. However, the definition of therapeutic non-inferiority regarding 30-day mortality is not the same for the above trials. For ASSENT-2, showing non-inferiority for tenecteplase vs alteplase implied showing that the 95% one-sided CI for the absolute difference and for the relative risk between tenecteplase and accelerated alteplase did not exceed 1% or 14%, respectively, whichever is the smallest. The rationale for the 1% absolute difference limit was inspired by the GUSTO-I trial, in which a statistically significant and therapeutically relevant 1% difference in favour of accelerated alteplase over SK was proven. The rationale for the 14% relative risk (RR) limit was derived from the Fibrinolitics Therapy Trialists meta-analysis[7].

The results obtained in the ASSENT-2 study[2] were, for 30-day mortality, 6.16% for tenecteplase and 6.18% for alteplase. This corresponds to a relative risk of 0.997 with a two-sided 90% confidence interval of [0.904, 1.011]. Since the mortality rates fell below the threshold value of 7.2%, the relative part of the equivalence criterion applies. Hence, according to our above defined equivalence region it was concluded that tenecteplase was not inferior to alteplase with respect to 30-day mortality. The results of the ASSENT-II study are summarized in Table 1. In total, three analyses were performed. The primary analysis of the ASSENT-2 study was based on an adjusted rate[9], corrected for the five most important risk factors[13], but the conclusion was the same as for the unadjusted analysis, which corresponds to the comparison of the mortality rates as above. The third analysis, logistic regression, also takes the five most important covariates into account and was added as an exploratory analysis because of the popularity of the logistic model when dealing with covariates. Again non-inferiority could be concluded from this analysis.

**Discussion**

Rapid restoration of patency of the infarction-related artery is the key mechanism of action for preserving myocardial tissue and improving survival. This understanding has led to the development of new plasminogen activators that induce faster and more complete reperfusion and are associated with a reduced risk of bleeding complications. However, it is becoming increasingly difficult to improve upon the performance of well established treatments, such as accelerated alteplase. As a consequence, increasingly more equivalence (and also non-inferiority) trials in this field have been conducted. Unfortunately, the principles that govern the design and analysis of equivalence trials are not well understood[9-12].

There are two fundamental methodological features distinguishing an equivalence trial from a superiority trial. The first feature relates to the statistical methodology and the second to the lack of an internal check of validity in equivalence trials. Both features are linked to a solid and stringent definition of the region of equivalence.

The output of an equivalence study is considerably more difficult to understand than that of a superiority trial. The results have to be interpreted in opposite ways. Furthermore, it is not trivial to compare and appreciate the results of two equivalence trials when their region of equivalence is differently defined. For example, when the ASSENT-2 study claims equivalence of tenecteplase to alteplase and another study also claims equivalence of another new fibrinolytic to alteplase, using a different, possibly looser definition of equivalence than in the ASSENT-2 study, then one cannot automatically conclude that both studies have reached the same

<p>| Table 1 | Results on 30-day mortality of ASSENT-II study. The P-value corresponds to the non-inferiority test with region of non-inferiority as indicated in Fig. 2 |
|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>Tenecteplase (%)</th>
<th>Alteplase (%)</th>
<th>Absolute difference</th>
<th>Relative risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary analysis</td>
<td>6.179</td>
<td>6.151</td>
<td>0.028 (– 0.554, 0.609)</td>
<td>0.0059</td>
</tr>
<tr>
<td>Unadjusted analysis</td>
<td>6.160</td>
<td>6.176</td>
<td>– 0.016 (– 0.624, 0.592)</td>
<td>0.0060</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>6.089</td>
<td>6.140</td>
<td>– 0.051 (– 0.623, 0.522)</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

*Eur Heart J, Vol. 22, issue 11, June 2001*
conclusion. Moreover, it is not possible to claim equivalence of the three treatments. The same is true for superiority trials, but there is more room for alternative interpretations in equivalence trials because of the absence of a gold standard in the choice of a region of equivalence in contrast to the generally accepted clinically and statistically significant difference in superiority trials.

Finally, it is possible to test a superiority hypothesis in equivalence or non-inferiority trials as well, without being penalized in terms of the significance level. The superiority and non-inferiority can be tested using an hypothesis region described by Fleming[13] or by using hierarchical hypothesis testing, as described by Koch and Gansky[14]. Also in the ASSENT-2 trial, a hierarchical testing procedure was proposed in the statistical analysis plan. This implies that if non-inferiority is proven it can be tested whether tenecteplase is superior to alteplase as regards 30-day mortality, or not. Because the results for both treatments were so close it is obvious that superiority of tenecteplase over alteplase could not be claimed.

The authors would like to thank Gary Koch, PhD, and Robert Califf, MD, for their helpful comments on the ASSENT-2 protocol.

References