ELSEVIER

Journal of Hepatology 46 (2007) 947–954

Journal of Hepatology

www.elsevier.com/locate/jhep

Review

Methodology of superiority vs. equivalence trials and non-inferiority trials

Erik Christensen*

Clinic of Internal Medicine I, Bispebjerg University Hospital, Bispebjerg Bakke 23, DK-2400 Copenhagen NV, Copenhagen, Denmark

The randomized clinical trial (RCT) is generally accepted as the best method of comparing effects of therapies. Most often the aim of an RCT is to show that a new therapy is superior to an established therapy or placebo, i.e. they are planned and performed as superiority trials. Sometimes the aim of an RCT is just to show that a new therapy is not superior but equivalent to or not inferior to an established therapy, i.e. they are planned and performed as equivalence trials or non-inferiority trials. Since the types of trials have different aims, they differ significantly in various methodological aspects. The awareness of the methodological differences is generally quite limited. This paper reviews the methodology of these types of trials with special reference to differences in respect to planning, performance, analysis and reporting of the trial. In this context the relevant basal statistical concepts are reviewed. Some of the important points are illustrated by examples.

1. Introduction

The randomized clinical trial (RCT) is generally accepted as the best method of comparing effects of therapies [1,2]. Most often the aim of an RCT is to show that a new therapy is superior to an established therapy or placebo, i.e. they are planned and performed as superiority trials. Sometimes the aim of an RCT is just to show that a new therapy is not superior but equivalent to or not inferior to an established therapy, i.e. they are planned and performed as equivalence trials or non-inferiority trials [3]. Since these types of trials have different aims, they differ significantly in various methodological aspects [4]. The awareness of the methodological differences is generally quite limited. For example it is a rather common belief that failure of finding a significant difference between therapies in a superiority trial implies that the therapies have the same effect or are equivalent [5–10]. However, such a conclusion is not correct because of a considerable risk of overlooking a clinically relevant effect due to insufficient sample size.

The purpose of this paper is to review the methodology of the different types of trials, with special reference to differences in respect to planning, performance, analysis and reporting of the trial. In this context the relevant basal statistical concepts will be reviewed. Some of the important points will be illustrated by examples.

2. Superiority trials

2.1. Sample size estimation and power of an RCT

An important aspect in the planning of any RCT is to estimate the number of patients necessary i.e. the sample size. The various types of trials differ in this respect [1,2,11]. A superiority trial aims to demonstrate the superiority of a new therapy compared to an established therapy or placebo. The following description applies to a superiority trial. The features, by which an equivalence or a non-inferiority trial differ, will be described later.

* Tel.: +45 3531 2854; fax: +45 3531 3556.
E-mail address: ec05@bbh.hosp.dk

0168-8278/$32.00
doi:10.1016/j.jhep.2007.02.015

To estimate the sample size one needs to consider some important aspects described in the following.

By how much should the new therapy be better than the reference therapy? This extra effect of the new compared to the reference therapy is called the Least Relevant Difference or the Clinical Significance. It is often denoted by the Greek letter Δ (Fig. 1).

By how much would the difference in effect between the two groups be influenced by random factors? Like any other biological measurement a treatment effect is subject to a considerable "random" variation, which needs to be determined and taken into account. The magnitude of the variation is described in statistical terms by the standard deviation S or the variance S^2 (see Fig. 1c). The variance of the effect variable would need to be obtained from a pilot study or from previously published similar studies. The trial should demonstrate as precisely as possible the true difference in effect between the treatments. However, because of the random variation the final result of the trial may deviate from the true difference and give erroneous results. If for example the null hypothesis H_0 of no difference were true, it could be still that the trial in some cases would show a difference. This type of error – the type 1 error ("false positive") (Fig. 1) – would have the consequence of introducing an ineffective therapy. If on the other hand the alternative hypothesis H_Δ of the difference being Δ were true, the trial could in some cases fail to show a difference. This type of error – the type 2 error ("false negative") (Fig. 1) – would have the consequence of rejecting an effective therapy.

Thus one needs to specify how large risks of type 1 and type 2 errors would be acceptable for the trial. Ideally the type 1 and type 2 error risks should be near zero, but this would need extremely large trials. Limited resources and patient numbers make it necessary to accept some small risk of type 1 and 2 errors.

Most often the type 1 error risk α would be specified to 5%. In this paper, α means the type 1 error risk in one direction i.e. either up or down from H_0 i.e. $\alpha = 5\%$. However, in many situations one would be interested in detecting both beneficial and harmful effects of the new therapy compared to the control therapy, i.e. one would be interested in "two-sided" testing for a difference in both "upward" and "downward" direction (Fig. 1). Hence we would instead specifiy the type 1 error risk to be 2α (i.e. $\alpha_{\text{upwards}} + \alpha_{\text{downwards}}$), i.e. $2\alpha = 5\%$.

The type 2 error risk β would normally be specified to 10-20%. Since a given value of Δ is always either above or below zero (H_0), the type 2 error risk β is always one-sided. The smaller β, the larger the complementary probability $1 - \beta$ of accepting H_Δ when it is in fact true. $1 - \beta$ is called the power of the trial because it states the probability of finding Δ if this difference truly exists.

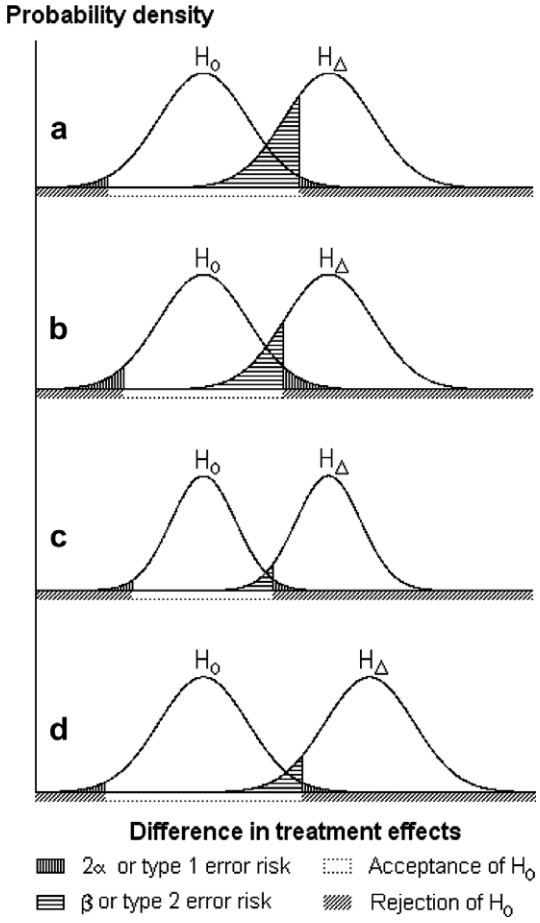


Fig. 1. Illustration of factors influencing the sample size of a trial. The effect difference found in a trial will be subject to random variation. The variation is illustrated by bell-shaped normal distribution curves for a difference of zero corresponding to the null hypothesis (H_0) and for a difference of Δ corresponding to the alternative hypothesis (H_Δ), respectively. Defined areas under the curves indicate the probability of a given difference being compatible with H_0 or H_Δ, respectively. If the difference lies near H_0, one would accept H_0. The farther the difference would be from H_0, the less probable H_0 would be. If the probability of H_0 becomes very small (less than the specified type 1 error) risk 2α (being α in either tail of the curve) one would reject H_0. The sample distribution curves show some overlap. A large overlap will result in considerable risk of interpretation error, in particular the type 2 error risk may be substantial as indicated in the figure. An important issue would be to reduce the type 2 error risk β (and increase the power $1 - \beta$) to a reasonable level. Three ways of doing that are shown in (b–d), a being a reference situation. (b) Isolated increase of 2α will decrease β and increase power. Conversely, isolated decrease of 2α will increase β and decrease power. (c) Isolated narrowing of the sample distribution curves – by increasing sample size $2N$ and/or decreasing variance of the difference S^2 – will decrease β and increase power. Conversely, isolated widening of the sample distribution curves – by decreasing sample size and/or increasing variance of the difference – will increase β and decrease power. (d) Isolated increase of Δ – larger therapeutic effect – will decrease β and increase power. Conversely, isolated decrease of Δ – smaller therapeutic effect – will increase β and decrease power.

From given values of Δ, S^2, α and β the needed number (N) of patients in each group can be estimated using this relatively simple general formula:

$$N = (Z_{2\alpha} + Z_{\beta})^2 \times S^2/\Delta^2,$$

where $Z_{2\alpha}$ and Z_{β} are the standardized normal deviates corresponding to the levels of the defined values of 2α (Table 1, left), and β (Table 1, right), respectively. If for some reason one wants to test for difference in only one direction (''one-sided'' testing) one should replace $Z_{2\alpha}$ with Z_{α} in the formula and apply the right side of Table 1. The formula is approximate, but it gives in most cases a good estimation of the necessary number of patients. For a trial with two parallel groups of equal size the total sample size will be $2N$.

The values used for 2α, β and Δ should be decided by the researcher, not by the statistician. The values chosen should take into account the disease, its stage, the effectiveness and side effects of the control therapy and an estimate of how much extra effect may be reasonably expected by the new therapy.

If for example the disease is rather benign with a relatively good prognosis and the new therapy is more expensive and may have more side effects than a rather effective control therapy, one should specify a relatively larger Δ and β and a smaller 2α, because the new therapy would only be interesting if it is markedly better than the control therapy.

If on the other hand the disease is aggressive, the new therapy is less expensive or may have less side effects than a not very effective control therapy, one should specify a relatively smaller Δ and β and a larger 2α, because the new therapy would be interesting even if it is only slightly better than the control therapy.

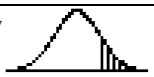
As mentioned above 2α would normally be specified to 5% or 0.05, but one may justify values of 0.10 or 0.01 in certain situations as mentioned above. β would normally be specified to 0.10–0.20, but in special situations a higher or lower value may be justified. Δ should be decided on clinical grounds as the least relevant therapeutic gain of the new therapy considering the disease and its prognosis, the efficacy of the control therapy and what may reasonably be expected of the new therapy. Preliminary data from pilot studies or historical observational data can be guidelines for the choice of Δ. Even if it may be tempting to specify a relatively large Δ as fewer patients will then be needed, Δ should never be specified larger than what is biologically reasonable. It will always be unethical to perform trials with unrealistic aims. Fig. 1 illustrates the effects on the type 2 error risk β and hence also on the power $(1 - \beta)$ of changing 2α, N, S^2 and Δ. Thus β will be decreased and the power $1 - \beta$ will be increased if 2α is increased (Fig. 1b), if the sample size is increased (Fig. 1c), and if Δ is increased (Fig. 1d).

The estimated sample size should be increased in proportion to the expected loss of patients during follow-up due to drop-outs and withdrawals.

2.2. The confidence interval

An important concept indicating the confidence of the result obtained in an RCT is the width of the confidence interval of the difference D in effect between the therapies investigated [1,2]. The narrower the confidence

Table 1
Abbreviated table of the standardized normal distribution (adapted for this paper)

Two-sided probability 		One-sided probability 			
$Z_{2\alpha}$	2α	Z_{α} or Z_{β}	α or β	Z_{α} or Z_{β}	α or β
3.72	0.0002	3.72	0.0001	0.00	0.50
3.29	0.001	3.29	0.0005	−0.13	0.55
3.09	0.002	3.09	0.001	−0.25	0.60
2.58	0.01	2.58	0.005	−0.39	0.65
2.33	0.02	2.33	0.010	−0.52	0.70
1.96	0.05	1.96	0.025	−0.67	0.75
1.64	0.1	1.64	0.05	−0.84	0.80
1.28	0.2	1.28	0.10	−1.04	0.85
1.04	0.3	1.04	0.15	−1.28	0.90
0.84	0.4	0.84	0.20	−1.64	0.95
0.67	0.5	0.67	0.25	−1.96	0.975
0.52	0.6	0.52	0.30	−2.33	0.990
0.39	0.7	0.39	0.35	−2.58	0.995
0.25	0.8	0.25	0.40	−3.09	0.999
0.13	0.9	0.13	0.45	−3.29	0.9995
0.00	1.0	0.00	0.50	−3.72	0.9999

Note. The total area under the normal distribution curve is one. The area under a given part of the curve gives the probability of an observation being in that part. The y-axis indicates the ''probability density'', which is highest in the middle of the curve and decreases in either direction toward the tails of the curve. The normal distribution is symmetric, i.e. the probability from Z to plus infinity (right side of the table) is the same as from $-Z$ to $-\infty$. The right side of the table gives the one-sided probability from a given Z-value on the x-axis to $+\infty$. The left side of the table gives the two-sided probability as the sum of the probability from a given positive Z-value to $+\infty$ and the probability from the corresponding negative Z-value to $-\infty$.

interval would be, the more reliable the result would be. In general the width of the confidence interval is determined by the sample size. A large sample size would result in a narrow confidence interval. Normally the 95% confidence interval would be estimated. The 95% confidence interval is the interval, which would on average include the true difference in 95 out of 100 similar studies. This is illustrated in Fig. 2 where 100 trial samples of the same size have been randomly drawn from the same population. It is important to note that in 5 of the 100 samples the 95% confidence interval of the difference in effect *D* does not include the true difference found in the population. When the sorted confidence intervals are aligned to their middle (Fig. 2c), the variation in relation to the true value in the population becomes even clearer. If simulation is carried out on an even greater scale, the likelihood distribution of the true difference in the population, given the results from a certain trial sample, will follow a normal distribution like that presented in Fig. 3 [2]. It is seen that the likelihood of the true difference in the population is maximum at the difference *D* found in the sample and that it decreases with higher and lower values. The figure also illustrates the 95% confidence interval, which is the interval that includes the middle 95% of the total likelihood area under the normal curve. This area can be calculated from the difference *D* and its standard error SED. To be surer that the true difference is included in the confidence interval, one may calculate a 99% confidence interval, which would be wider, since it should include the middle 99% of the total likelihood area.

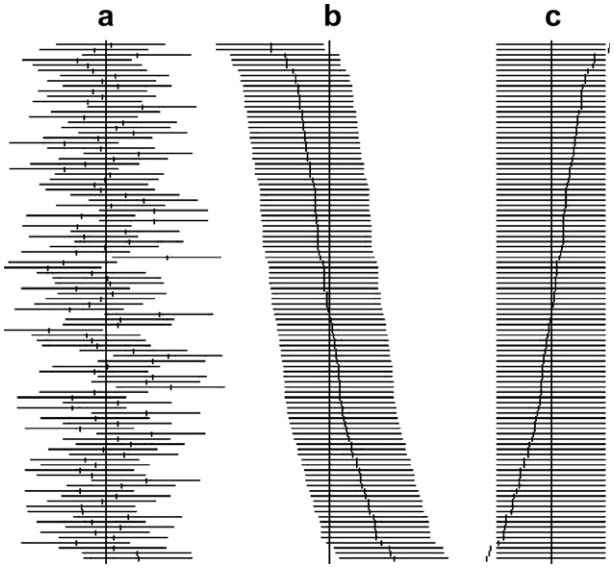


Fig. 2. Illustration of the variation of confidence limits in random samples (computer simulation). (a) ninety-five percent confidence intervals in 100 random samples of same size from the same population aligned according to the true value in the population. In 5 of the samples the 95% confidence interval does not include the true value found in the population. (b) The same confidence intervals are here sorted according to their values. (c) When the sorted confidence intervals are aligned to their middle, their variation in relation to the true value in the population is again clearly seen. This presentation corresponds to how investigators would see the world. They investigate samples in order to extrapolate the findings to the population. However, the potential imprecision of extrapolating from a sample to the population is apparent – especially if the confidence interval is wide. Thus keeping confidence intervals rather narrow is important. This would mean relatively large trials.

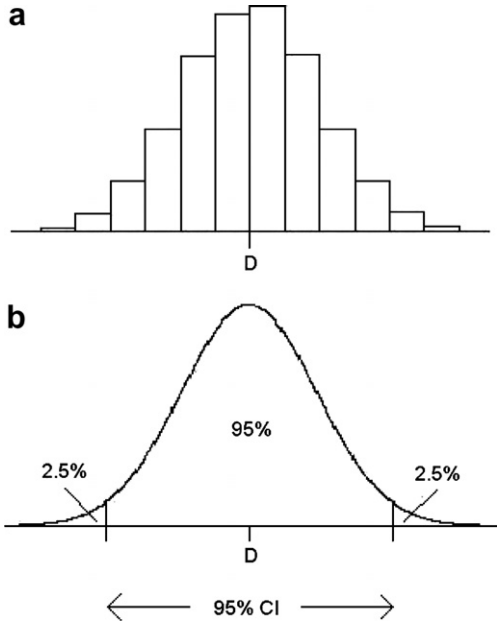


Fig. 3. (a) Histogram showing the distribution of the true difference in the population in relation to the difference *D* found in the trial sample (computer simulation of 10,000 samples). (b) The normally distributed likelihood curve of the true difference in the population in relation to the difference *D* found in a trial sample. The 95% confidence interval (CI) is shown.

2.3. The type 2 error risk of having overlooked a difference Δ

If the 95% confidence interval of *D* includes zero, then there is no significant difference in effect between the two therapies. However, this does not mean that one can conclude that the effects of the therapies are the same. There may still be a true difference in effect between the therapies, which the RCT has just not been able to detect e.g. because of insufficient sample size and power. The risk of having overlooked a certain difference in effect of Δ between the therapies is the type 2 error risk β. In some cases this risk may be substantial. Example 1 gives an illustration of this.

Example 1. In naïve cases of chronic hepatitis C genotype 1 pegylated interferon plus ribavirin for 3 months induce sustained virologic response in about 40%. One wishes to test if a new therapeutic regimen can increase the sustained response in this type of patients to

60% with a power $(1-\beta)$ of 80%. The type 1 error risk (2α) should be 5%. One needs to estimate the number of patients necessary for this trial. For comparison of proportions like in this trial, the variance of the difference (S^2) is equal to $p_1(1-p_1)+\ p_2(1-p_2)$, where p_1 and p_2 are the proportions with response in the compared groups. So we have:

$$2\alpha = 0.05 \Rightarrow Z_{2\alpha} = 1.96. \qquad \beta = 0.20 \Rightarrow Z_\beta = 0.84$$
$$p_1 = 0.4 \qquad p_2 = 0.6 \qquad \Delta = 0.2.$$

Using $N = (Z_{2\alpha} + Z_\beta)^2 \times p_1(1-\ p_1) + p_2(1-p_2)/\Delta^2$ one gets:

$$N = (1.96 + 0.84)^2 \times (0.4 \times 0.6 + 0.6 \times 0.4)/0.2^2$$
$$= 7.84 \times 0.48/0.04 = 94.$$

Therefore the necessary number of patients $(2N)$ would be 188 patients.

However, due to various difficulties only 120 patients (60 in each group) of this kind could be recruited. By solving the general sample size formula according to Z_β one obtains:

$$Z_\beta = \frac{\sqrt{N}}{S} \times \Delta - Z_{2\alpha}.$$

Using this formula, the power of the trial with the reduced number of patients can be estimated as follows:

$$Z_\beta = \frac{\sqrt{60}}{\sqrt{0.48}} \times 0.2 - Z_{2\alpha} \qquad Z_\beta = 7.75/0.69 \times 0.2 - 1.96 = 0.29$$

Using Table 1 (right part) with interpolation β becomes 0.39. Thus with this limited number of patients, the power $1-\beta$ is now only 0.61 or 61% (a reduction from 80%). This markedly reduced power seriously diminishes the chances of demonstrating a significant treatment effect. A post hoc power calculation like this can only be used to explain why a superiority trial is inconclusive; it can never be used to support a negative result of a superiority trial.

The result of the trial was as follows: sustained virologic response was found in 26 of 60 (0.43 or 43%) in the control group and in 35/60 (0.58 or 58%) in the new therapy group. The difference D is 0.15 or 15%, but it is not statistically significant ($p > 0.10$). A simple approximate formula for the standard error of the difference is:

$$\text{SED} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$
$$= \sqrt{0.43 \times 0.57/60 + 0.58 \times 0.42/60} = 0.09$$

The 95% confidence interval for D is $D \pm Z_{2\alpha} \times \text{SED} = 0.15 \pm 1.96 \times 0.09$ or -0.026 to 0.326 (-2.6% to 32.6%), which is rather wide, as it includes both zero and Δ. The type 2 error risk of overlooking an effect of 20% (corresponding to Δ) can be estimated as follows: $Z_\beta = (\Delta - D)/\text{SED} = (0.20–0.15)/0.09 = 0.55$. Using Table 1 (right part) with interpolation β becomes 0.29. Thus the risk of having overlooked an effect of 20% is 29%. This is a consequence of the smaller number of patients included and the reduced power of the trial. The situation corresponds to that illustrated in Fig. 4. As seen from this figure the result of a negative RCT like this does not rule out that the true difference may be Δ, since the type 2 error risk β of having overlooked an effect of Δ is substantial.

Fig. 4. Illustration of the type 2 error risk β in an RCT showing a difference D in effect, which is not significant, since zero (0) difference lies between the lower (L) and upper (U) 95% confidence limits. The type 2 error risk of having overlooked an effect of Δ is substantial.

3. Equivalence trials

The purpose of an equivalence trial is to establish identical effects of the therapies being compared [12–17,15]. Complete equivalent effects would mean a Δ-value of zero. As seen from the formula for estimation of the sample size (see above) this would mean division by zero, which is not possible. Dividing by a very small Δ-value would result in an unrealistic large estimated sample size. Therefore, as a manageable compromise, the aim of an equivalence trial would be to determine if the difference in effects between two therapies lies within a specified small interval $-\Delta$ to $+\Delta$.

An equivalence trial would be relevant if the new therapy is simpler, associated with fewer side-effects or less expensive, even if it is not expected to have a larger therapeutic effect than the control therapy.

It is crucial to specify a relevant size of Δ [14,17]. This is not simple. One should aim at limiting as much as possible the acceptance of a new therapy, which is inferior to the control therapy. Therefore Δ should be specified rather small and in any case smaller than the smallest value that would represent a clinically meaningful difference. As a crude general rule Δ should be specified to no more than half the value which may be used in a superiority trial [13]. Equivalence between the therapies would be demonstrated if the confidence interval

for the difference in effect between the therapies turns out to lie entirely between $-\Delta$ and $+\Delta$ [13]. Fig. 5 illustrates the conclusions that can be drawn from the position of the confidence limits for the difference in effect found in the performed trial.

In the equivalence trial the roles of the null and alternative hypotheses are reversed. In the equivalence trial the relevant null hypothesis is that a difference of at least Δ exists, and the aim of the trial is to disprove this in favor of the alternative hypothesis that no difference exists [13]. Even if this situation is like a mirror image of the situation for the superiority trial, it turns out that the method for sample size estimation is similar in the two types of trial, although Δ has different meanings in the superiority and equivalence trials.

Example 2. In the same patients as described in Example 1 one wishes to test in an RCT the therapeutic equivalence of the current regimen of pegylated interferon plus ribavirin (giving a sustained response in 40%) and another new inexpensive therapeutic regimen having less side-effects.

One needs to estimate the number of patients necessary for this trial. The power $(1-\beta)$ of the trial should be 80%. The type 1 error risk (2α) should be 5%. The therapies would be considered equivalent if the confidence interval for the difference in proportion with sustained response falls entirely within the interval $\pm 0.10\%$ or $\pm 10\%$. Thus Δ is specified to 0.10. So we have:

$$2\alpha = 0.05 \Rightarrow Z_{2\alpha} = 1.96. \qquad \beta = 0.20 \Rightarrow Z_{\beta} = 0.84$$
$$p_1 = 0.4 \qquad p_2 = 0.4 \qquad \Delta = 0.10.$$

Using the same expression for the variance of the difference (S^2) as in Example 1 this result is obtained:

$$N = (1.96 + 0.84)^2 \times (0.4 \times 0.6 + 0.4 \times 0.6)/0.1^2$$
$$= 7.84 \times 0.48/0.01 = 376.$$

Therefore the necessary number of patients ($2N$) would be 752 patients.

The trial was conducted and the result of the trial was as follows: Sustained virologic response was found in 145 of 372 (0.39 or 39%) in the control group and in 156/380 (0.41 or 41%) in the new therapy group. The difference D is 0.02 or 2%, but it is not statistically significant ($p > 0.50$). The standard error of the difference is:

$$\mathrm{SED} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$
$$= \sqrt{0.39 \times 0.61/372 + 0.41 \times 0.59/380} = 0.036$$

The 95% confidence interval for D is $D \pm Z_{2\alpha} \times \mathrm{SED} = 0.02 \pm 1.96 \times 0.036$ or -0.050 to 0.091 (-5.0% to 9.1%). Since this confidence interval lies completely within the specified interval for Δ from -0.1 to $+0.1$, the effects of the two therapies can be considered equivalent. The situation corresponds to B or C in Fig. 5.

Like in this example the necessary sample size in an equivalence trial will often be at least 4× that of a corresponding superiority trial. Therefore the necessary resources will be larger.

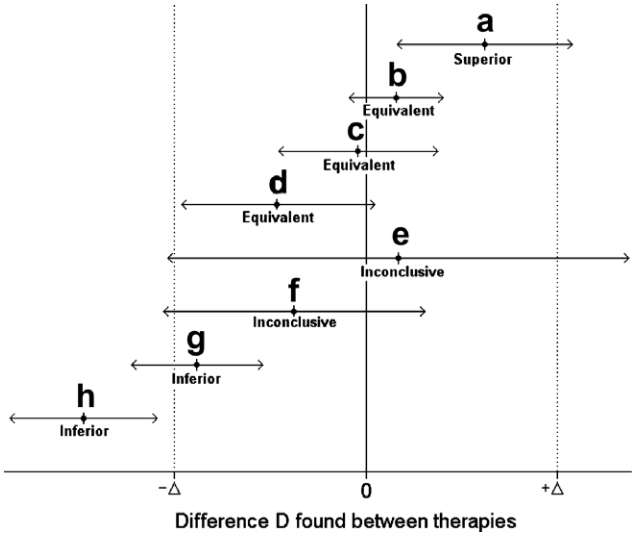


Fig. 5. Examples of observed treatment differences (new therapy – control therapy) with 95% confidence intervals and conclusions to be drawn. (a) The new therapy is significantly better than the control therapy. However, the magnitude of the effect may be clinically unimportant. (b–d) The therapies can be considered having equivalent effects. (e–f) Result inconclusive. (g) The new therapy is significantly worse than the control therapy, but the magnitude of the difference may be clinically unimportant. (h) The new therapy is significantly worse than the control therapy.

4. Non-inferiority trials

The non-inferiority trial, which is related to the equivalence trial, aims not at showing equivalence but only at showing that the new therapy is no worse than the reference therapy. Thus the non-inferiority trial is designed to demonstrate that the difference in effect (new therapy–control therapy) should be no less than $-\Delta$. Non-inferiority of the new therapy would then be demonstrated if the lower confidence limit for the difference in effect between the therapies turns out to lie above $-\Delta$. The position of the upper confidence limit is not of primary interest. Thus the non-inferiority trial is designed as a one-sided trial. For that reason the necessary number of patients would be less than for a corresponding equivalence trial as illustrated in the following example.

Example 3. We want to conduct the trial described in Example 2 not as an equivalence trial but as a non-

inferiority trial. Thus the trial should be one-sided instead of the two-sided equivalence trial. The only difference would be that one should use Z_α instead of $Z_{2\alpha}$. For $\alpha = 0.05$ one gets $Z_\alpha = 1.64$ (Table 1, right side). Thus we obtain:

$$N = (1.64 + 0.84)^2 \times (0.4 \times 0.6 + 0.4 \times 0.6)/0.1^2$$
$$= 6,15 \times 0.48/0.01 = 295.$$

Therefore the necessary number of patients ($2N$) would be 590 patients.

The trial was conducted and the result of the trial was as follows: Sustained virologic response was found in 114 of 292 (0.39 or 39%) in the control group and in 125/298 (0.42 or 42%) in the new therapy group. The difference D is 0.03 or 3%, but it is not statistically significant ($p > 0.50$). The standard error of the difference is:

$$\text{SED} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$
$$= \sqrt{0.39 \times 0.61/292 + 0.42 \times 0.58/298} = 0.040$$

The lower one-sided 95% confidence limit would be $D - Z_\alpha \times \text{SED} = 0.03 - \quad 1.64 \times 0.040 = -0.036 \quad (-3.6\%)$. Since the lower confidence limit lies above the specified limit for Δ of -0.1, the effect of the new therapy is not inferior to the control therapy. If the two-sided 95% confidence interval (which is recommended by some even for the non-inferiority trial [18]) is being estimated, one obtains $D \pm Z_{2\alpha} \times \text{SED} = 0.03 \pm 1.96 \times 0.040$ or -0.048 to 0.108 (-4.8% to 10.8%). The lower confidence limit still lies above -0.1, but the upper confidence limit lies above 0.1 (the upper limit for equivalence – see Example 2). Therefore the new therapy may be slightly better than the control therapy. The type 2 error risk of having overlooked an effect of 0.1 or 10% could be estimated as follows: $Z_\beta = (\Delta - \text{D})/\text{SED} = (0.10 \quad -0.03)/0.04 = 1.75$. Using Table 1 (right part) with interpolation β becomes 0.04, a rather small risk.

5. Other factors

Since the aim of an equivalence or non-inferiority trial is to establish equivalence between the therapies or non-inferiority of the new therapy, there is not the same incentive to remove factors likely to obscure any difference between the treatments as in a superiority trial. Thus in some cases finding of equivalence may be due to trial deficiencies like small sample size, lack of double blinding, lack of concealed random allocation, incorrect doses of drugs, effects of concomitant medicine or spontaneous recovery of patients without medical intervention [19].

An equivalence or non-inferiority trial should mirror as closely as possible the methods used in previous superiority trials assessing the effect of the control therapy versus placebo. In particular it is important that the inclusion and exclusion criteria, which define the patient population, the blinding, the randomization, the dosing schedule of the standard treatment, the use of concomitant medication and other interventions, the primary response variable and its schedule of measurements, are the same as in the preceding superiority trials, which have evaluated the reference therapy being used in the comparison. In addition one should pay attention to patient compliance, the response during any run in period, and the scale of patient losses and the reasons for them. These should not be different from previous superiority trials.

6. Analysis: both "intention to treat" and "per protocol"

An important point in the analysis of equivalence and non-inferiority trials concerns whether to use an "intention to treat" or a "per protocol" analysis. In a superiority trial, where the aim is to decide if two treatments are different, an intention to treat analysis is generally conservative: the inclusion of protocol violators and withdrawals will usually tend to make the results from the two treatment groups more similar. However, for an equivalence or non-inferiority trial this effect is no longer conservative: any blurring of the difference between the treatment groups will increase the chance of finding equivalence or non-inferiority.

A per protocol analysis compares patients according to the treatment actually received and includes only those patients who satisfied the entry criteria and properly followed the protocol. In a superiority trial this approach may tend to enhance any difference between the treatments rather than diminishing it, because uninformative "noise" is removed. In an equivalence or non-inferiority trial both types of analysis should be performed and equivalence or non-inferiority can only be established if both analyses support it. To ensure the best possible quality of the analysis it is important to collect complete follow-up data on all randomized patients as per protocol, irrespective of whether they are subsequently found to have failed entry criteria, withdraw from trial medication prematurely, or violate the protocol in some other way [20]. Such a rigid approach to data collection allows maximum flexibility during later analysis and hence provides a more robust basis for decisions.

The most common problem in reported equivalence or non-inferiority studies is that they are planned and analyzed as if they were superiority trials and that the lack of a statistically significant difference is taken as proof of equivalence [7–10]. Thus there seems to be a need for a better knowledge of how equivalence and non-inferiority studies should be planned, performed, analyzed and reported.

7. Ensuring a high quality

A recent paper reported on the quality of reporting of published equivalence trials [21]. They found that some trials had been planned as superiority trials but were reported as if they had been equivalence trials after failure to demonstrate superiority, since they did not include an equivalence margin. They also found that one-third of the reports which included a sample size calculation had omitted elements needed to reproduce it; one third of the reports described a confidence interval whose size was not in accordance with the type 1 error risk used in the sample size calculation; and half the reports that used statistical tests did not take the margins into account. In addition, only 20% of the trials surveyed provided the 4 necessary basic requirements: equivalence margin defined, sample size calculation taking this margin into account, both intention-to-treat and per-protocol analyses, and confidence interval for the result. Only 4% of the trials gave a justification, which is essential, for the margin used.

An extension concerning equivalence and non-inferiority trials of the CONSORT statement about publication of RCTs [22–24] has been suggested [18]. This includes description of the rationale for adopting an equivalence or non-inferiority design, how study hypotheses were incorporated into the design, choice of participants, interventions (especially the reference treatment), and outcomes, statistical methods, including sample size calculation and how the design affects interpretation and conclusions [18].

8. Summary

Clinicians should always remember that a negative result in a superiority trial never would prove that the investigated therapies are equivalent; Most often there may be a large risk of type 2 error (false negative result). Equivalence and non-inferiority trials demand high standards to provide reliable results. Clinicians should especially bear in mind that equivalence margins are often far too large to be clinically meaningful and that a claim of equivalence may be misleading if a trial has not been conducted to an appropriately high standard. Furthermore, clinicians should be somewhat skeptical of trials that fail to include the basic reporting requirements including definition and justification of the equivalence margin, calculation of sample size taking this margin into account, presentation of both intention-to-treat and per-protocol analyses, and providing confidence intervals for the results.

Equivalence and non-inferiority trials are indicated in certain areas. If the necessary strict adherence to the specific methodology is followed, such trials may provide important new knowledge.

References

[1] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.
[2] Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell; 1994.
[3] Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 1978;62:1037–1040.
[4] Fleiss JL. General design issues in efficacy, equivalency and superiority trials. J Periodontal Res 1992;27:306–313.
[5] Garrett AD. Therapeutic equivalence: fallacies and falsification. Stat Med 2003;22:741–762.
[6] Blackwelder WC. "Proving the null hypothesis" in clinical trials. Control Clin Trials 1982;3:345–353.
[7] Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence?. Ann Intern Med 2000;132:715–722.
[8] Costa LJ, Xavier ACG, del Giglio A. Negative results in cancer clinical trials – equivalence or poor accrual?. Control Clin Trials 2004;25:525–533.
[9] Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? Arch Surg 2001;136:796–800.
[10] Detsky AS, Sackett DL. When was a negative clinical trial big enough? how many patients you needed depends on what you found. Arch Intern Med 1985;145:709–712.
[11] Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122–124.
[12] Djulbegovic B, Clarke M. Scientific and ethical issues in equivalence trials. JAMA 2001;285:1206–1208.
[13] Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996;313:36–39.
[14] Lange S, Freitag G. Choice of delta: requirements and reality – results of a systematic review. Biomed J 2005;47:12–27.
[15] Durrleman S, Simon R. Planning and monitoring of equivalence studies. Biometrics 1990;46:329–336.
[16] Ebbutt AF, Frith L. Practical issues in equivalence trials. Stat Med 1998;17:1691–1701.
[17] Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. Control Clin Trials 2002;23:2–14.
[18] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW. CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295:1152–1160.
[19] Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457–2465.
[20] Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–46.
[21] Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA 2006;295:1147–1151.
[22] Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA 1996;276:637–639.
[23] Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191–1194.
[24] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–694.