Can we trust equivalence and non-inferiority trials?

Giovanni Casazza · Monica Solbiati ·
Gruppo di Autoformazione Metodologica (GrAM)

Received: 28 February 2013 / Accepted: 26 March 2013 / Published online: 12 April 2013
© SIMI 2013

The trial by Rickard and colleagues [1] was designed to assess the incidence of phlebitis in a group of patients who had intravenous catheters replaced when clinically indicated in comparison with a group of patients with catheters removed every third day. The aim of the study was not to demonstrate that one of the two strategies was better than the other, but that patients who had intravenous catheters replaced when clinically indicated would have equivalent rates of phlebitis, and no difference in other complications, but reduced costs and number of catheter insertions, compared with patients with catheters removed every third day. For this reason, the study was designed as an equivalence trial rather than a superiority trial.

Usually, randomized clinical trials (RCT) are designed and conducted to prove that a new treatment (that we will call A) is better than the standard treatment or placebo (B) (superiority trial). With a significant result, we can conclude that A is significantly better than B and we can replace the old with the new treatment in clinical practice. However, in case of a non-significant result, we cannot state that the two treatments are similar or that A is worse: absence of evidence is not evidence of absence. Sometimes a non-significant result may be due to lack of statistical power. If it has already been proved that the new treatment

(A) reduces either harms or costs, then it should be considered for use in clinical practice if it is as effective as the old standard therapy (B). Unfortunately, since exact equality is usually impossible to be proved, a new RCT could be designed to answer two possible clinical questions:

1. “is treatment A as effective as treatment B?”
2. “is treatment A not significantly less effective than treatment B?”

An equivalence trial is needed to answer the first question, while the non-inferiority design is the right choice for the second one. Sometimes the choice of showing equivalence or non-inferiority rather than superiority can be appropriate (and also methodologically adequate) because it may be unethical to conduct a placebo-controlled trial [2]. Usually, a non-inferiority trial could be conducted when a treatment of proven efficacy already exists and the new treatment is not expected to be more effective than the existing one but has other advantages (e.g., it is cheaper, easier to provide or has fewer side effects).

Nevertheless, some authors believe that also in the above reported situation equivalence and non-inferiority trials may be unethical for several reasons. First, because in some cases the supposed extra advantages of the new treatment (fewer side effects) are not definitely proven. Then the new treatment, proven to be non-inferior, might actually be not superior to placebo. For these reasons, it may be unethical to randomize patients to a treatment (the new one) that is expected to be no more effective than the standard treatment and that could also be less safe. Moreover, these studies are usually preferred by drug producers just because they can reduce research costs. [3]. Anyway, these types of trials are usually accepted by regulatory authorities for drugs approval.

G. Casazza (✉)
Dipartimento di Scienze Biomediche e Cliniche L. Sacco,
Università degli Studi di Milano, Milan, Italy
e-mail: giovanni.casazza@unimi.it

M. Solbiati
Unità Operativa di Medicina Interna II, Dipartimento di Scienze
Cliniche L. Sacco, Ospedale L. Sacco, Università degli Studi di
Milano, Via GB Grassi 74, 20157 Milan, Italy
e-mail: monica.solbiati@gmail.com

Although equivalence and non-inferiority trials answer two different questions, from a statistical and methodological stand point they share their main characteristics. They both should prove that two treatments are not significantly different; that is, that the confidence intervals of the measures of efficacy (odds ratio, relative risk or risk ratio) lie between a bound of equivalence ($-\Delta$ to $+\Delta$) or below a non-inferiority limit (Δ). In particular, equivalence trials are designed to prove that treatment A is neither better nor worse than treatment B, while non-inferiority trials are intended to show that treatment A is not worse than B (it might be better or less effective by an amount less than Δ).

Figure 1 resumes different significant results of RCT.

In a superiority trial (I), the results are reported as a measure of effectiveness (risk difference, risk ratio, odds ratio or hazard ratio) with its 95 % confidence interval (CI, the continuous horizontal line). The superiority of the new treatment is claimed when a pre-specified level of statistical significance is reached (usually $p < 0.05$), that is to say, when the 95 % CI lies totally under the threshold of indifference (risk difference = 0, risk ratio, odds ratio or hazard ratio = 1). In an equivalence trial (II), the difference between the two treatments is reported in terms of risk difference and the equivalence (or non-equivalence) between the treatments is assessed comparing the observed 95 % CI with the pre-specified interval of equivalence ($-\Delta$ to $+\Delta$). The two treatments can be declared equivalent if the entire 95 % CI of the risk difference lies between the lower and upper bounds of equivalence. Finally, in a non-inferiority trial (III), the 95 % CI of the risk difference is compared with a pre-specified non-inferiority limit: if the 95 % CI of the risk difference crosses the non-inferiority limit, then the new treatment cannot be declared non-inferior to the existing one. On the contrary, if the 95 % CI

of the risk difference does not cross the non-inferiority limit, the new treatment is declared as “non-inferior” than the existing one. This means that the new treatment is similar to the existing one. Actually, it means that the new treatment is not too worse (by less than the pre-specified bound) than the standard one. In some circumstances, depending on the magnitude of the 95 % CI of the risk difference, this might also mean that treatment A is better than B [4, 5].

For a better understanding, the possible results of a non-inferiority trial are illustrated in Fig. 2.

In cases I and II, the upper limit of the 95 % CI does not cross the non-inferiority bound (Δ): the new treatment (A) is non-inferior to the existing one (B). In particular, in case I, treatment A is even better than B: the new treatment can be declared better than the standard. In case II, the two treatments do not differ in terms of efficacy (the new treatment is non-inferior). It is clear that moving Δ toward the threshold of “No difference” (more conservative hypothesis), the conclusion of the trial in case II would be different. In cases III and IV, the upper limit of the 95 % CI goes beyond the non-inferiority bound: A is not non-inferior to B. In particular, in case IV, since even the lower limit of the CI is higher than the non-inferiority bound the new treatment can be claimed worse than the standard. From the above examples, the pivotal role of the 95 % CI for a full understanding of the results of non-inferiority and equivalence is clear.

Moreover, the choice of the margin can affect the results’ interpretation. Indeed, while the proof of superiority is somewhat objective, the proof of non-inferiority (equivalence) is strictly dependent on the definition of similarity (choice of the margins). In the analysis of a superiority trial, once a cut-off for significance is chosen (usually $p < 0.05$), the same numerical finding leads to the

Fig. 1 Different significant results of randomized clinical trials. The horizontal bars represent the 95 % CIs. The gray area is the zone of superiority (I), equivalence (II) or non-inferiority (III)

Fig. 2 Possible results of a non-inferiority trial. The horizontal bars represent the 95 % CIs. The gray area is the zone of non-inferiority

Fig. 3 Schematic results of the trial by Rickard and coworkers. The horizontal bar represents the 95 % CI. The gray area is the zone of equivalence

same conclusion. On the other hand, the numerical finding of a non-inferiority trial may lead to different conclusions. An hypothetical observed difference of 5 % (95 % CI 1–9 %) could be considered not relevant by some clinicians (non-inferiority limit set, for instance, at 10 %, non-inferiority proven) and could be considered relevant (non-inferiority limit set, for instance, at 6 %, non-inferiority not proven) by some others. Non-inferiority claim strongly depends on subjective (hopefully a priori) assumptions.

While in superiority trials the sample size depends on the incidence of events in the control group and on the minimum clinically relevant reduction, the number of patients to enroll in non-inferiority (or equivalence) trials depends on the incidence of events in the control group and on the magnitude of the chosen non-inferiority (equivalence) margin. Actually, the number of patients to enroll decreases as the non-inferiority limit (equivalence interval) increases. Namely, if you wish to demonstrate that the efficacy of the two treatments is “highly similar” (setting narrow non-inferiority margins) you need to enroll a large number of patients. On the contrary, if you believe it is enough to prove that the two treatments are “quite similar” (setting wide non-inferiority margins) then the number of patients needed becomes relatively small [6]. Since a wide

non-inferiority limit could be chosen to enroll a smaller number of patients, it’s very important, in the interpretation of the results of such trials, to assess the width of the non-inferiority bounds in the specific clinical context. Actually, we should accept the conclusions of the study only if the non-inferiority limit seems really clinically reasonable.

Going back to the example above, Rickard and colleagues enrolled about 3,000 patients to detect equivalence between the two strategies at an incidence of 4 % of phlebitis in the control group, with an equivalence margin of 3 % (significance level 0.05, power 95 %). Phlebitis occurred in 114 of 1,593 (7 %) patients in the clinically indicated group and in 114 of 1,690 (7 %) patients in the routine replacement group, with an absolute risk difference of 0.41 % (95 % CI –1.33 % to 2.15 %). Since the CI lied between –3 % and +3 % (the pre-specified equivalence margin) (Fig. 3), the authors concluded that the two treatments were equivalent [1].

Bottom line for clinicians

When reading an equivalence or non-inferiority trial, we should first ask ourselves what the study rationale is

because we are going to choose a new treatment also if it could be not too worse (not so inferior that it would cause concern) than the standard one provided that it has other proven advantages (e.g. it is cheaper, easier to provide or has fewer side effects).

The critical issue in interpreting equivalence and non-inferiority trials is the choice of an acceptable (clinically sensible) threshold of equivalence or non-inferiority. This is the maximum allowable excess of outcome events arising from the new treatment compared with the standard one. When designing non-inferiority trials, investigators set their own thresholds, therefore we should make sure that they choose a reasonable bound so that a positive result indicates a true absence of a significant difference between two treatments. Moreover, a too wide non-inferiority margin could be the dishonest attempt of having a positive trial enrolling a smaller sample of patients.

Finally, the results of equivalence and non-inferiority trials should always be interpreted looking at confidence intervals rather than at the point estimate of efficacy.

Conflict of interest None.

References

1. Rickard CM, Webster J, Wallis MC et al (2012) Routine versus clinically indicated replacement of peripheral intravenous catheters: a randomised controlled equivalence trial. *Lancet* 380:1066–1074
2. Pocock SJ (2003) The pros and cons of noninferiority trials. *Fundam Clin Pharmacol* 17:483–490
3. Garattini S, Bertelé V (2007) Non-inferiority trials are unethical because they disregard patients' interests. *Lancet* 370:1875–1877
4. Mulla SM, Scott IA, Jackevicius CA et al (2012) How to use a noninferiority trial. *JAMA* 308(24):2605–2611
5. Piaggio G, Elbourne DR, Pocock SJ et al (2012) Reporting of noninferiority and equivalence randomized trials. *JAMA* 308(24):2594–2604
6. Snapinn SM (2000) Noninferiority trials. *Curr Control Trials Cardiovasc Med* 1:19–21