Point of view

The interpretation of lack of evidence of a difference in efficacy: equivalence trials and the treatment of fungal infections

The demonstration of treatment efficacy by a prospective randomised placebo controlled trial has become established in the medical literature. In situations where effective therapy already exists, the introduction of newer therapeutic agents using placebo controlled trials is controversial and comparisons with “standard therapy” are frequently undertaken using so-called equivalence or non-inferiority trials. The use of such trials obviously presupposes the established efficacy of a therapy, but the formulation of a “standard therapy” in the critical care setting has been difficult, as opposed to, say, the practice of cardiology. For instance, the Fibrinolytic Therapy Trialists’ Collaborative Group reviewed reports of fibrinolytic and standard therapy for myocardial infarction (ST segment elevation and/or bundle-branch block with randomisation within 6 hours of symptom onset) in 58600 patients and demonstrated an overall absolute 35-day mortality reduction of -1.84% (95% CI: -2.34% to -1.35%). Thus fibrinolytic therapy, in particular streptokinase, has become a standard therapy and it is “no longer ethical to withhold... (such therapy)... from patients...”. The usual (i.e. placebo) controlled trial is a superiority trial, where the aim is to rule out treatment equality by rejection of the null hypothesis that the two treatments are the same. However, the converse proposition does not hold; that the failure to reject the null hypothesis (the “negative” clinical trial) establishes equivalence. An illustration of this was the report of a clinical trial comparing trimethoprim-sulphamethoxazole and pentamidine in the treatment of Pneumocystis carinii pneumonia. Forty patients were enrolled and no difference was seen in 21-day mortality rates (p = 0.18) or other indices of improvement or of toxicities. The trialists concluded that the two study treatment arms were “probably of equal effectiveness”. As Polis and Blackwelder noted, apropos the question of study sample size and β error, “With additional patients, this study may have contributed more toward the resolution of this issue... (therapy of P. carinii pneumonia)... . However, it does not demonstrate that trimethoprim-sulphamethoxazole and pentamidine are equally effective; failure to show a significant difference ... is not at all the same as showing equivalence.

The equivalence trial reverses the logic of the superiority trial; the null hypothesis is instead that of a specified difference (δ) between the experimental therapy and an active control. Thus if μ is the “true” treatment difference (experimental vs control therapy and μ is positive when the experimental is superior to standard therapy),

i) in a superiority trial the test is μ = 0 vs μ ≠ 0 at the 5% level (or rather, μ ≤ 0 vs μ ≥ 0 at the 2.5% level)

ii) in an equivalence trial, the purpose is to demonstrate minimal differences (i.e. experimental therapy vs standard) in either direction. Therefore, μ ≤ -δ or μ ≥ δ is tested (two-sided) against -δ < μ < δ. Alternatively, a pair of one-sided hypotheses are tested: H1, μ ≤ -δ vs μ > -δ and H2, μ ≥ δ vs μ < δ (both one-sided hypotheses need to be rejected). True equivalence trials are usually bio-equivalence trials.

iii) in a non-inferiority trial, the purpose is to demonstrate that the experimental therapy is not substantially worse than active-control. Therefore, μ ≤ -δ is tested against μ > -δ. The test is one-sided at a significance level (usually 0.05). Such testing may be subject to the known limitations of hypothesis testing in general. Alternatively, a 100(1-2α) percent two-sided confidence interval for the treatment difference is computed and if the lower bound of the CI is < -δ, non-inferiority can be claimed. Whether this be at the 90% or 95% is a point of some dispute, although recent regulatory recommendations suggest α = 0.025 for one-sided testing of non-inferiority. It is also noted that the strategy of using a two sided 90% CI for a one-sided 5% test assumes that the 90% CI is equal-tailed (each end of the interval excludes 5%). In the clinical literature equivalence is often used synonymously with non-inferiority and, unless specified otherwise, this review will conform to this practice.

In figure 1, trials 1-4 show the above as hypothetical trials with point estimates and CI (95% for trials 1 & 2 and 90% CI for trials 3 & 4). In trial 1, placebo vs standard drug, the lower 95% CI approximates, in this scenario, the value of δ which is set at 20%. Trial 2 shows a successful superiority trial with lower 95% CI above zero. Trial 3 is an equivalence trial showing upper and lower 90% CI within ±δ. In trial 4, a non-inferiority trial, the lower 90% CI is < -δ (but the upper 95% CI is > +δ).
Figure 1. The vertical axis indicates the absolute risk response success (experimental – standard, as %), such that + ve values reflect efficacy of the experimental therapy and – ve values reflect efficacy of standard therapy. Diamonds represent point estimate with 90% CI, unless indicated. The horizontal axis indicates trial type. Trials 1 - 4 are hypothetical examples of: 1. placebo/standard therapy, 2. superiority trial (successful, 95% CI), 3. equivalence trial (successful), 4. non-inferiority trial (successful). Trials 6 - 9 (Phillips et al, reference 69): 6. PP analysis, 7. PP analysis with 95% CI, 8. ITT analysis, 9. ITT analysis with 95% CI. Trials 11 - 14 (Rex et al, reference 67): 11. PP analysis, 12. PP analysis with 95% CI, 13. ITT analysis, 14. ITT analysis with 95% CI.

ISSUES IN EQUIVALENCE TRIALS

1. Assumptions made when conducting equivalence trials
   a. assay sensitivity: that the active-control would have been superior to a placebo if such had been employed in the current trial. That is, an equivalence trial requires the consideration of “…information external to the trial”.
   b. sensitivity to drug effects: the ability of well designed trials to reliably demonstrate active-control drug effect (with respect to placebo)
      If the above two assumptions have not been met, then the interpretation of equivalence trials can be problematic. This was demonstrated by Tramer et al, in their recent review of anti-emetics; in particular, the efficacy of ondansetron. They concluded that where no gold standard treatment existed and event rates (in this case, of emesis) varied widely “…trial designs without placebo controls are unlikely to yield sensible results”.
   c. constancy assumption, that the historical treatment difference is preserved in the current trial. This may be difficult to sustain given changes in medical practice and the effect of different patient populations.

2. Intention-to-treat (ITT) vs per-protocol (PP) analysis: In superiority trials ITT is the preferred analysis compared with PP. Such is not the case with equivalence trials where PP analysis is the more conservative and ITT tends to make treatment arms appear similar, although this will depend upon the pattern of patient drop out and treatment assignment. Both types of analyses should be
presented, but this strategy must take into consideration the reduced patient number in a PP analysis when initial sample size calculations are made.

3. **Biocreep:** whereby a slightly inferior treatment becomes the active control for the next generation of equivalence trials and active controls become little different from placebos.\(^{31}\)

4. **Conduct of the trial:** poor trial conduct in an equivalence trial will widen CI of the observed treatment effect and make the declaration of equivalence more difficult,\(^{23,24,32}\) whereas in a superiority trial there will be a tendency to a null result which may be mistakenly claimed as indicating equivalence.

5. **Sample size requirements:** for equivalence studies sample size requirements are variably increased above similar superiority trials; on average about 10%.\(^1\) Formulas for such calculations are provided in numerous articles\(^{16,18,21,26}\) and specialised software is available.\(^{33,34}\)

6. **The determination of \(\delta\) (or non-inferiority margin):** this may be formally defined as the largest acceptable clinical difference in treatment efficacy (experimental therapy vs standard) or, in the reverse, as a difference in patient status with an effect size \(\leq \delta\) that is non-detectable.\(^{17}\) Thus \(\delta\) is different (and usually smaller)\(^{35}\) from the differences in proportions between two treatments \((\pi_1 - \pi_2)\) used in routine sample size calculations for superiority trials.\(^{36,16}\)

Recommendations for the calculation of \(\delta\) have been numerous and are found, not surprisingly, in the biopharmaceutical literature,\(^{22,37-39}\) but the regulatory literature has been somewhat circumspect in prescribing \(\delta\) a priori.\(^{15,40}\) From a statistical perspective, \(\delta\) has been defined as a certain fraction of:

(a) the treatment effect of control drug vs placebo (for example, 0.2 - 0.5) or,

(b) the lower 95% CI of this treatment effect (for example, 0.5) derived from a meta-analysis or large trial.\(^{23,38,41,42}\)

In the anti-infective drug testing domain, the Food and Drug Administration (FDA) in the United States of America had informally provided a so-called step-down function of \(\delta\) that reflected the observed response rates in the equivalence study. For response rates (one or both arms) of at least 90%, \(\geq 80\% \text{ but } < 90\%\), and \(\geq 70\% \text{ but } < 80\%\), \(\delta\) was suggested to be 10, 15 and 20% respectively.\(^{41}\) The recent removal of this step-down function and the use of a more conservative \(\delta\) (unofficially 10%\(^{15}\)) has provoked comment regarding the unavoidable and large increment in trial size consequent upon this decision.\(^{31}\)

In cardiology, where standard care in the treatment of acute myocardial infarction has been effectively established, equivalence trials have used streptokinase as the standard and \(\delta\) has been set at a much lower level. In the INJECT trial,\(^{48}\) which compared reteplase with streptokinase using 35 day mortality as end-point, equivalence was established if the (upper) CI of mortality difference excluded the possibility that reteplase mortality was \(> 1\%\) worse than streptokinase mortality.\(^{45}\) (Note here and subsequently, the algebraic reversal when a positive treatment difference, i.e. \(\mu\), indicates worse outcome). The difference was in fact 0.5% with two-sided 90% CI for the difference: \(-1.7\% \text{ to } 0.71\%\) (two-sided 95% CI: \(-1.96\% \text{ to } 0.98\%\)).

A different perspective may be taken of the attempt to infer equivalence from a large negative superiority trial.\(^{39}\) GUSTO III,\(^ {46}\) was designed as a superiority trial (15059 patients enrolled) to detect a 20% difference in 30-day post myocardial infarction mortality, comparing double-bolus reteplase relative to an accelerated infusion of alteplase. The mortality for the reteplase arm was 7.47% and that of alteplase 7.24, an absolute mortality difference of 0.23% (two-sided 95% CI: \(-0.65\% \text{ to } 1.11\%\)). This would, as the trialists noted “exceed a definition of equivalence requiring a difference of less than 1%”,\(^ {46,47}\) but they did observe, parenthetically, that the INJECT trial \(^{48}\) had used 90% CI to establish equivalence. On this basis, equivalence would have been established in GUSTO III (two-sided 90% CI: \(-0.51\% \text{ to } 0.98\%\)).

What is an appropriate equivalence margin is further illustrated by a consideration of the COBALT equivalence trial,\(^ {48}\) where 7169 patients were enrolled to compare 30-day post myocardial infarction mortalities, weight adjusted accelerated alteplase versus double bolus alteplase. On the basis of a 0.4% lower 95% CI for the absolute difference of accelerate infusion of alteplase vs streptokinase in the GUSTO I trial,\(^ {49}\) equivalence was defined in the COBALT trial if the upper boundary of a one-sided 95% CI of the difference in mortality did not exceed 0.4%. The absolute mortality difference of 0.44% with two-sided 90% CI: \(-0.57 \text{ to } 1.49\%\) thus failed to sustain equivalence. The differences in these approaches provoked editorial comment by Ware and Antmann,\(^ {50}\) who also noted the consequences of the calculation of sample size when based upon the assumption of unequal mortality rates in the two arms. In the COBALT trial it was assumed that 30-day mortality rates would be 6.3% in accelerated alteplase and 5.4% with double bolus alteplase and the trialists calculated an initial equivalence sample size of 4029 per group (although, by our calculations...
it was 4039, using the software package PASS 2002
34). However, the sample size required for
approximately equal mortality rates of 7.5% (the
range of the two arms reported in the COBALT trial)
was identified by Ware and Antman as about 30 000
in each group and the power of the COBALT study
as effectively 0.16 (our calculations: 53 634 in each
group and power 0.158 ). Ware and Antman further
suggested a absolute difference of 1.5% as a
reasonable compromise for equivalence studies of
this type, which, with 80% power, would require
3832 in each arm, a not impossible task for
cardiology trials.
7. Testing for non-inferiority and superiority: within the
same trial it is possible to test sequentially for non-
inferiority and superiority,51 although there are
inherent problems in this strategy.52,53 At the least,
trial methodology statements must pre-specify these
analyses; the direction of testing should be:
(a) initial demonstration of non-inferiority and,
(b) subsequent testing for superiority (the reverse is
problematic in interpretation, if superiority is
shown in ITT analysis, but non-inferiority is not
demonstrated in PP analysis); type I error must be
preserved and the potential problems of
different/unequal patient populations (PP
analysis for non-inferiority and ITT analysis for
superiority) must be addressed, for example, imputation of missing values.23
Declaring (formal) non-inferiority when the primary
trial methodology of superiority has been
unsuccessful “…should be looked upon with healthy
skepticism” …30 that is, it has the status of a post-hoc
analysis.

Overviews of equivalence trials
Two recent papers have assessed the performance of
trials where clinical or therapeutic equivalence had been
affirmed. Greene et al,55 studied 88 reports from 1992 to
1996 claiming equivalence; Δ was formally set in only
23% of reports, in 67% equivalence was declared after a
failed test of superiority, the sample size was calculated
in advance in only 33%, and in 25% of reports n was ≤
20 per group. Of interest, Δ ranged from 0 to 76% for
proportionate differences. McCalister et al,25 reviewing
4 recent hypertensive trials which appeared to show
equivalence between treatment arms, found a lack of
fulfillment in all of these trials of the 6 additional
features that were defined as distinguishing superiority
from equivalence trials; that the active control was
previously shown to be effective, similarity of (current)
patients and outcome variables to those of original
trials, optimal application of regimens, appropriate
analysis, pre-specified Δ and adequate sample size.

CRITICAL CARE IMPLICATIONS

Large trials with small treatment effect margins
In this context, it is if interest to look again at the
protocols of the ANZICS Clinical Trials Group SAFE
trial,56 which will compare saline and albumin solutions
for resuscitation. A total of 7000 patients are to be
enrolled to detect a ≥3% absolute mortality difference
between treatment groups based upon an assumed 15%
control mortality and β error 0.1. The magnitude of this
mortality difference was derived from the lower 95% CI
of the estimated treatment effect (albumin versus non-
albunin use) from the 1998 Cochrane Injury Group
Albumin Reviewers paper.55 As the SAFE trial (a
superiority trial by its methodology description) is of
large size with a relatively small treatment difference
targeted, the question may be posed: if the null
hypothesis of no treatment effect is not rejected, are we
able to conclude equivalence between the two
regimens? Despite the cautions above, some support for
this scenario is provided by Ng who argues that “If the
sample size is such that the …β error …at some Δ is
sufficiently small (e.g. < 0.05) we can conclude that the
two treatments are Δ-equivalent”.39 However, two
restrictions may apply to this proposition: the small β
error and a robust Δ. From the 1998 Cochrane paper, the
lower 95% CI of treatment effect with a random effects
estimator was 0.016 (our calculations are based upon
the “metan” routine,58 using StataTM software59). Thus
possible values of Δ would be 3% (lower 95% CI of
fixed effects meta-analysis), 0.016 (lower 95% CI of
random effects meta-analysis), 0.015 (50% of lower
95% CI of fixed effects meta-analysis). Total sample
sizes under these scenarios are variably in excess of that
of the SAFE trial.

Treatment of fungal infections
Therapy for mycotic infections in the intensive care
unit has been recently transformed by the introduction
of a group of drugs which have been marketed as
“equivalent” to the enduring yardstick, amphotericin
B.58 Since its introduction in 1957, amphotericin B has
been the standard (and until recently, the only)
antifungal therapy,60 despite never being compared with
placebo and only two randomised trials of its action
reported before the more recent comparisons with
fluconazole.60 The effect of no specific therapy for
mycotic infections (e.g. candidaemia) is difficult to
estimate, but the prospective observational study (n =
427) of Nguyen et al,55 suggests mortality rates of 27%
with and 74% without antifungal therapy.

Amphotericin B and fluconazole
There are 5 randomised trials comparing amphoteri-
Amphotericin and new-generation antifungals

The “high” incidence of toxicity, especially renal, associated with amphotericin B has prompted reassessments of the first-line position of this drug in mycotic infections.1,7 Two recent large equivalence trials of Caspofungin, an echinocandin (total n = 687, δ = 0.1)13 and liposomal amphotericin B (total n = 239, δ = 0.2)14 versus amphotericin B have suggested comparable efficacy of these two drugs with respect to amphotericin B and a reduction in toxicities. A similarly large superiority trial of Amphotericin B colloidal-dispersion versus the parent compound concluded “comparable efficacy,”15 but noted increased non-renal infusion toxicities with colloidal-dispersion amphotericin, a point re-iterated by Drew et al in correspondence over the liposomal amphotericin B trial.16 What is pertinent in this series of trials is the high cost of the group of amphotericin comparators; that is, the trade-off of the extra costs of toxicities (especially renal) versus drug costs. Cagnoni et al,77 undertook a pharmacoeconomic analysis of the liposomal amphotericin B trial above and found that the hospital acquisition cost of the drug was critical in determining the break-even point, but it is noted that only 60% of patients were evaluated and “costs” were inferred from billing data, a strategy which has been previously criticized.18 O’Connell et al undertook a similar cost-effectiveness study of the use of liposomal amphotericin B in 2002 and estimated the cost per additional life saved to be £ 23,819 (= S$US 57, 574).

Conclusions

Claims in the literature as to the demonstration of “equivalence” must be subjected to careful scrutiny. The particular methodology of equivalence trials is of critical importance with respect to the conclusions that may be inferred, especially as these trials require the concurrent assessment of appropriate “external information”. Extension of conclusions beyond the “equivalence” hypothesis must also be formally assessed.
REFERENCES

11. Sprit A, Beiler, D. When can 'non significantly different' treatments be considered as 'equivalent'? Brit J Clin Pharmacol 1979;7:623-624.
37. Holmgren EB. Establishing equivalence by showing that the effect of the active control over placebo is maintained. J Biopharm Stat 1999;9:651-659.
41. Blackwelder WC. Showing a treatment is good because it is not bad: when does "noninferiority" imply effectiveness? Control Clin Trials 2002;23:52-54.
45. Hampton JR. Mega-trials and equivalence trials: experience from the INJECT study. European Heart Journal 1996;17(Suppl E);28-34.
57. Staet Statical Software, Version 8.0, 2003. Stata Corporation, College Station, TX


