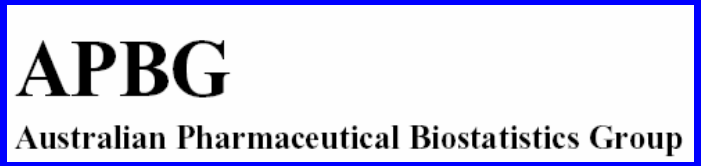
Indirect Comparisons Workshop

Designs Issues

Dr. Philip McCloud
Roche Products Pty Ltd

30 March 2007



Introduction

- In conducting economic evaluations the preferred approach is a head-to-head comparison of the two drugs
- For several reasons such head-to-head comparisons will often not be available
  - several comparators available; we cannot test against all
  - the best comparator will change over time
  - the best comparator may not be approved in all markets
- For these reasons an Indirect Comparison of two drugs via a common placebo/comparator may be necessary

Precision is out of Control

- Indirect Comparisons are not rigorously designed experiments; for example compare the design process for a non-inferiority study
- The sample size of a non-inferiority clinical trial can only be calculated once five quantities are set:
  - the level of significance; usually 0.025
  - the power; usually 80%
  - an estimate of the inherent random variation of the primary efficacy variable
  - the expected treatment difference; usually set to zero
  - the non-inferiority margin

Precision is out of Control

- Similar for difference or superiority studies
- Through such a process the sponsor commences the trial with a good chance of success provided the two drugs are identical

Precision is out of control

- The use of an adjusted indirect comparison to assess the non-inferiority of two treatments is not based on such rigorous statistical foundations. The design properties of an adjusted indirect comparison are as follows:
  - The studies being compared were never designed for such a purpose; many shall be superiority or difference studies
  - The non-inferiority margin will probably not have been pre-specified
  - The inherent random variation of the adjusted indirect comparison will be much larger than that of the original studies
  - The sample size of the studies was not calculated with the intention of conducting an adjusted indirect comparison

Precision is out of control

- Therefore the power of the adjusted indirect comparison shall be low unless an unacceptably large limit of non-inferiority is chosen. Therefore the probability of a successful outcome is unacceptably low.

Testing Non-inferiority

- Non-inferiority clinical trials generally demand a large sample size
- For example, in 30 day mortality trials of acute MI the American Cardiology Society has set the non-inferiority margin to be 1%, which requires a sample size of 14,128 in each group

Testing Non-inferiority

- Example: Suppose two products are from the same class of drug. Product A was registered before product B. Suppose that the registration trials compared A and B to a common comparator/placebo. Suppose the number of patients was 1,000 in each group, and that the population failure rates for A and B were 10% compared to 20% for the placebo groups.
- The assumptions set A and B to be equivalent.
- The use of the fixed-effects analysis with only two-studies is not an endorsement of the approach

Testing non-inferiority

- Let P_A denote the proportion of failures for Treatment A, similarly P_B, and P_P
- Let RR_A, RR_B respectively denote the relative risks of A and B compared to placebo from the respective trials, where RR_A = risk of failure in A divided by the risk of failure in Placebo; therefore the smaller RR_A the better

Testing non-inferiority

- Product B will be inferior to A if the population or true RR_B is greater than RR_A; suppose the non-inferiority margin is set to 1.30
- $H_0 : RR_B / RR_A \geq 1.30$ vs $H_1 : RR_B / RR_A < 1.30$
- The non-inferiority margin of 1.30 corresponds to a risk difference of 3%

SSAI
Statistical Society of Australia Inc.

APBG
Australian Pharmaceutical Biostatistics Group

Testing non-inferiority

- The 95% CI for the $\log(RR_B / RR_A)$ equals:

  $\log(RR_B / RR_A) \pm 1.96 \times se(\log(RR_B / RR_A))$
- Similarly the 95% CI for the upper limit of the 95% CI of $\log(RR_B/RR_A)$ equals:

  $(\log(RR_B / RR_A), \log(RR_B / RR_A) + 2 \times 1.96 \times se(\log(RR_B / RR_A)))$

SSAI
Statistical Society of Australia Inc.

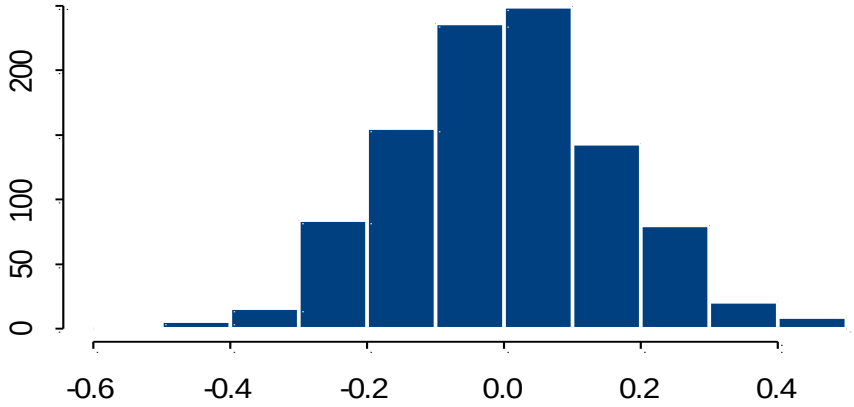
APBG
Australian Pharmaceutical Biostatistics Group

Testing non-inferiority

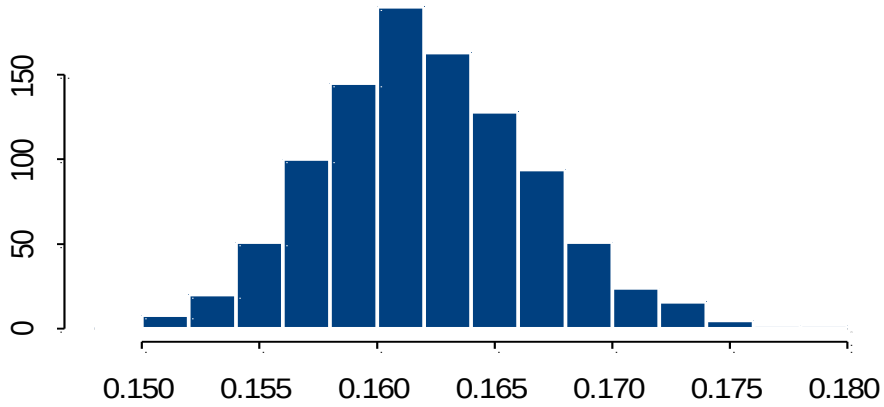
- A theoretical calculation shows that the expected 2.5th and 97.5th percentiles of the distribution of the upper limit of the 95% equal 1.00, and 1.88 respectively
- The probability of observing an upper limit greater than 1.30 equals 0.63. That is we have a low probability of success (=0.37).

Simulation for Response Rates

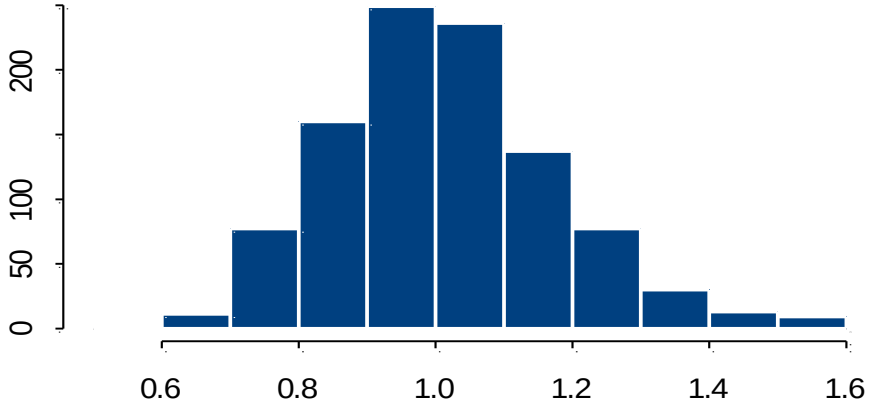
Logarithm of Ratio of Relative Risk



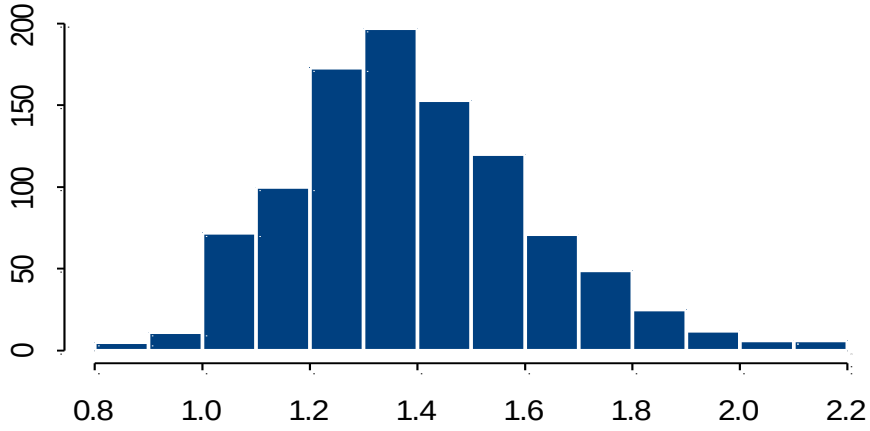
Standard error of log of Ratio



Ratio of Relative Risks



Upper 95% CI



Testing a difference

- Example: The lower the response rate the better. Active treatments A and B have population response rates of 30%, and the placebo response rate equals 40%. Clearly A and B are identical.
- The first study compared B to placebo in a difference study: level of sig = 0.05, two-sided test, power = 80%, then sample size was 360.
- In 1,000 simulations 789 significant

Testing a difference

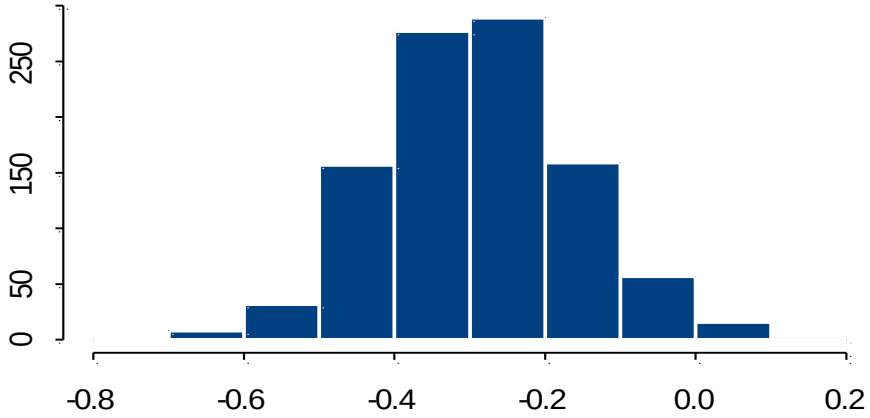
- Example contd: The second non-inferiority study was to demonstrate that A was non-inferior compared to B. The settings were: level of sig was 0.025 (one-sided), non-inferiority margin was 6%, power 80%, then n per group was 920
- In 1,000 simulations 808 significant; that is the phase III study/program was successful 80% of the time

Testing a difference

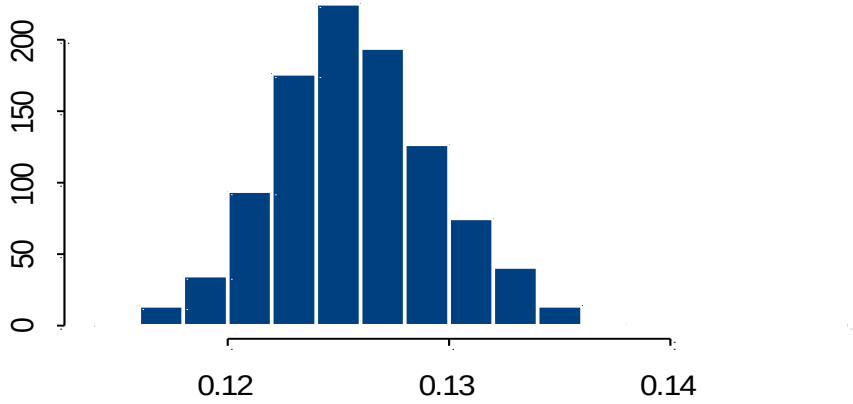
- Example contd: Asked to demonstrate that A is superior to placebo with an adjusted indirect comparison
- In 1,000 simulations A versus placebo was significant 632 occasions
- Therefore lose roughly 25% of the significant non-inferiority studies

Simulation for Response Rates - Difference Study

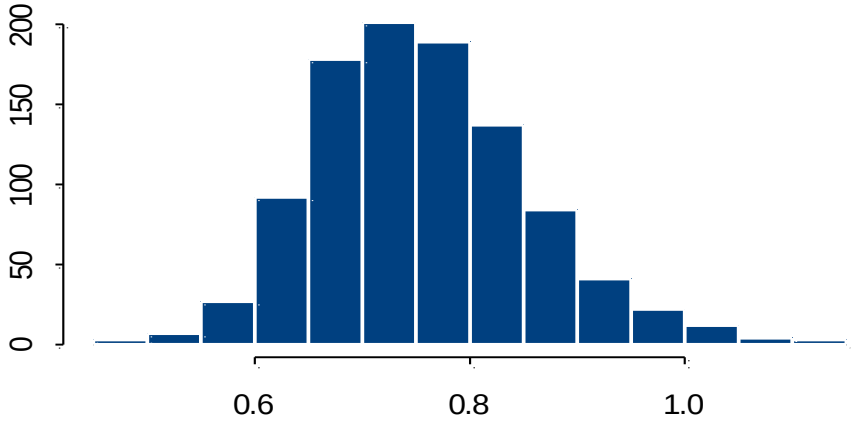
Logarithm of Ratio of Relative Risk



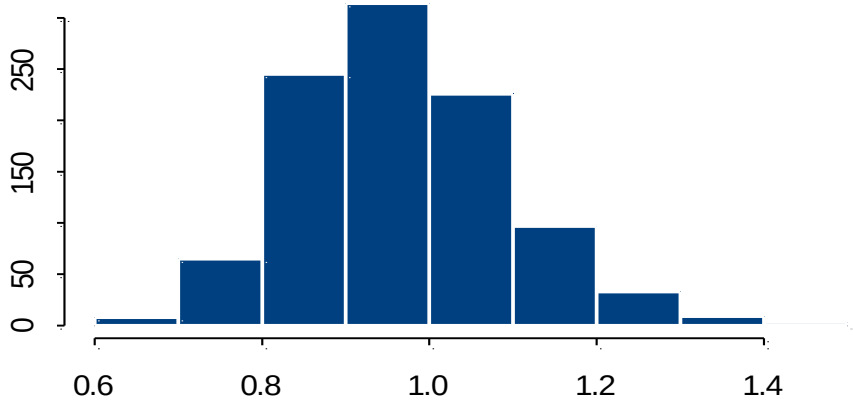
Standard error of log of Ratio



Ratio of Relative Risks



Upper 95% CI



Testing a difference

- Example contd: What if the null hypothesis was true, and A was the same as placebo?
- In 1,000 simulations A was always inferior to B; that is zero significant results in 1,000 simulations

Appropriate Replication

- Replication is fundamental to design
- Increasing replication increases the precision of the treatment effect
- Replication must be appropriate for the treatment comparison
- The following erroneous interpretation of replication produced embarrassing results

Appropriate Replication

- Example: Hermann & Bischof (1986) Experimental Brain Research
- Twenty-two zebra finches of both sexes were evaluated for the effects of monocular deprivation on neuron size and total volume
- Ten birds were monocularly deprived on the first or second day of life until they were sacrifice
- The remaining 12 birds were brought up normally and were used as controls.

Appropriate Replication

Treatment Group	Days before Sacrifice	No. of finches
Monocularly deprived	20	4
Monocularly deprived	40	3
Monocularly deprived	>100	3
Control	20	4
Control	40	4
Control	>100	4

Appropriate Replication

- For the cell measurements of the nucleus rotundus the outlines of 100 neurons showing visible nucleus were drawn at a magnification of 800x. Hence, each finch had 100 neurons drawn.
- The cross-sectional areas of the neurons were determined as a measure of size. From these measurements means, medians and standard deviations were computed and statistical comparison between the cells sizes was performed using a two-tailed *t*-test.

Appropriate Replication

- Then we find the following comment in the results section.
- *"With 40 days of monocular closure, cells in the deprived nucleus rotundus are only little affected, as the mean cell size is only 4.9% larger in normal birds than in the deprived rotundus (184.2 vs 175.1 μ m2), and this difference is only weakly significant ($p<0.02$, $t = 2.11$, d.f. = 792)."*

Appropriate Replication

- Where did all the d.f. come from? The study reports only 7 birds being sacrificed at 40 days, which gives a maximum of 6 d.f.
- The most likely explanation for the excessive d.f. is each neuron (100 for each finch) was counted as an independent data point, which is clearly incorrect. The investigators sensed there was something wrong, because they were trying to explain away the statistically significant difference, which was a small clinical difference. Where have they gone wrong?

Appropriate Replication

- What the investigators have done is use an incorrect level of replication to test the treatment effect.
- The SED of their *t*-test is based on neuron-to-neuron variation.
- The treatments were not applied to the individual neuron, but to the bird.
- Therefore the appropriate level of replication to test the treatment effect against is bird-to-bird variation.

Appropriate Replication

- Adapted from Mead (1988, section 6.4): to use within study variance to make adjusted indirect comparisons, it is necessary to assume:
  - That there are no intrinsic differences between studies which might influence patient response
  - That the variation between patients, both within and between studies, is essentially because of between patient variation; that is other sources of variation are negligible

Appropriate Replication

- Adapted from Mead (1988, section 6.4) continued:
  - there is no systematic variation between patients within a study which is induced by the treatment or the effects of the treatment
- The crucial requirement is that sponsor or evaluator can believe that the variation between patients is so dominated by patient-to-patient variation that other sources of variation can be ignored.

Appropriate Replication

- If such an assumption is credible, then the sponsor or evaluator may feel justified in using within study variation to test the adjusted indirect comparison
- Note, however, that in making this assumption the standard errors based on within study variation are not based on appropriate replication; we are no longer protected from systematic bias through proper randomisation and genuine replication
- The sponsor is seeking an act of faith from the evaluator in accepting that the within study variation between patients is the same as between study variation between patients
- The sponsor must make the assertion explicitly and has no statistical basis for arguing against the rejection of the assertion.

Conclusion

- Indirect Comparisons are not rigorously designed clinical trials
- The power of Indirect Comparisons will often be low; therefore success is likely to be luck, and failure bad luck
- The interpretation of indirect comparisons is problematic
- There is no statistical basis to justify a fixed effects analysis
- The estimate of the variance cannot be justified on the basis of random allocation
- The estimate of the variance from the random effects analysis cannot be regarded as reliable unless the number of observations is greater than 12, 20, 30