



Subgroup Analysis of Randomised Trials: Anticipating Chance Variation During Study Design

Ian Marschner
Macquarie University

A decorative graphic in the top-left corner consisting of a blue square with a white circular shape partially overlapping it.

Subgroup Analysis of Clinical Trials

- Analysis of a subset of subjects defined according to a baseline characteristic
- Can provide information on whether the efficacy of a treatment depends on individual characteristics
- Can be misleading due to chance variation
- Requires an assessment of the effect of chance variation and multiple comparisons



Multinational Trials

- Faster and more efficient clinical trials through pooling of resources
- Greater generalisability through broad assessment of treatments in different regions and countries
- Reduction in the need for replication of research results in different populations
- **Subgroup Analysis**: Multinational trials are often difficult to interpret when treatment effects appear to differ between regions or countries



Treatment Effect Differences

- Differences in treatment effect in multinational trials can arise from various sources, e.g.
 - Ethnic differences
 - Cultural differences
 - Treatment administration differences

- Differences in treatment effect can mean treatments may be
 - Beneficial or harmful for some ethnicities but not others
 - Ineffective in some cultural contexts but not others
 - Able to be effectively administered in some hospitals but not others



Chance Variation

- While there may be many plausible explanations for variation in treatment effects, none can be entertained unless “chance” can be ruled out as an explanation

- Evidence of treatment effect differences over and above chance variation comes only from a test of heterogeneity across regions/subgroups

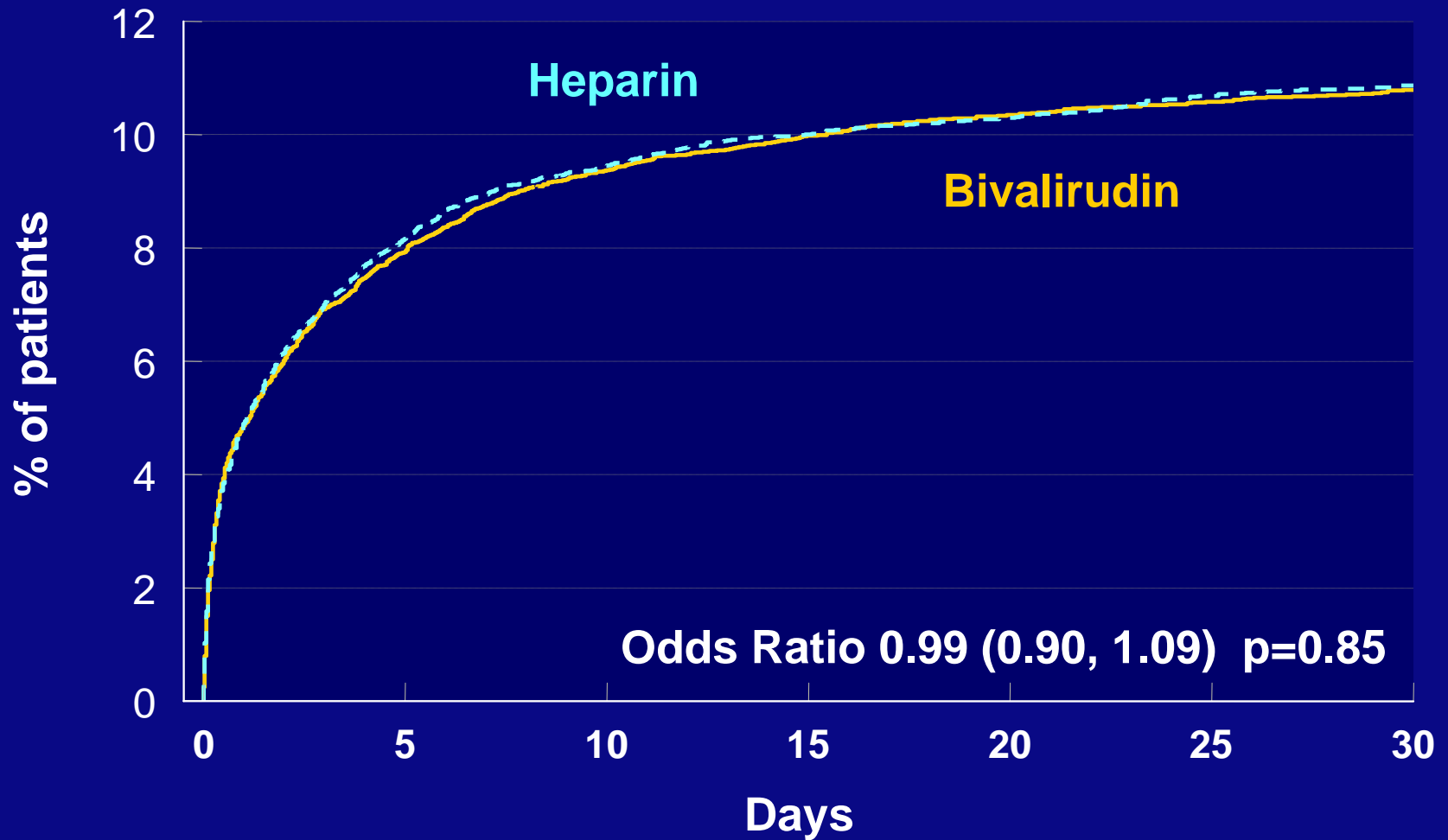
- Separate tests of treatment effect in different subgroups/regions do not provide evidence of treatment effect differences
 - Underpowered
 - Multiple comparisons



Case Study: HERO-2 Trial

- Thrombin-specific anticoagulation with **bivalirudin versus heparin** in patients receiving fibrinolytic therapy for acute myocardial infarction
- Lancet, **358**, 2001:1855-1863
- 17,073 patients with acute MI (heart attack)
- 539 hospitals in 46 countries
- Primary Outcome:
 - 30 day mortality

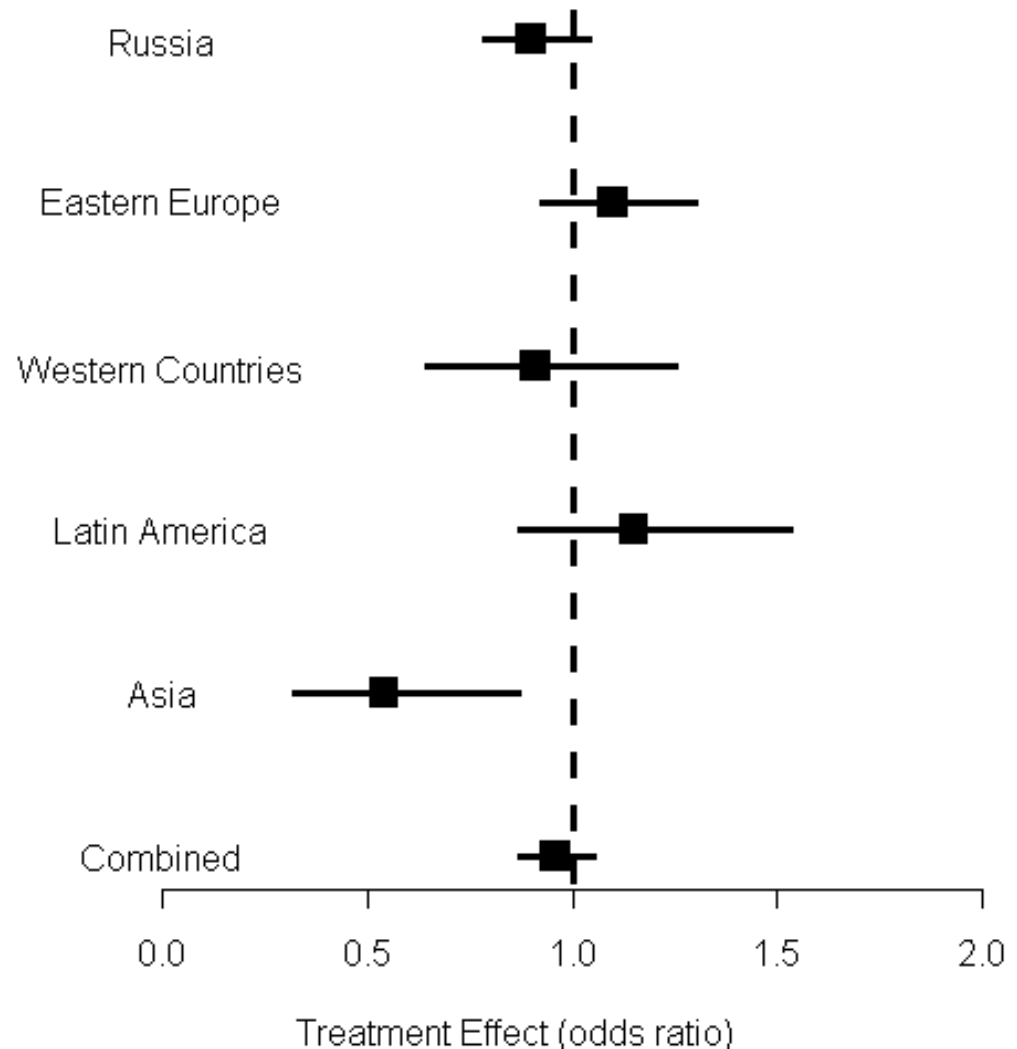
Mortality



Regional Differences

- Overall insignificant treatment effect and no significant test of heterogeneity (at 5% level)
- A significant treatment effect in a particular region does not provide evidence of treatment effectiveness in that region (see result for Asia)

Heterogeneity $P=0.063$





Estimating Treatment Effect

If there is

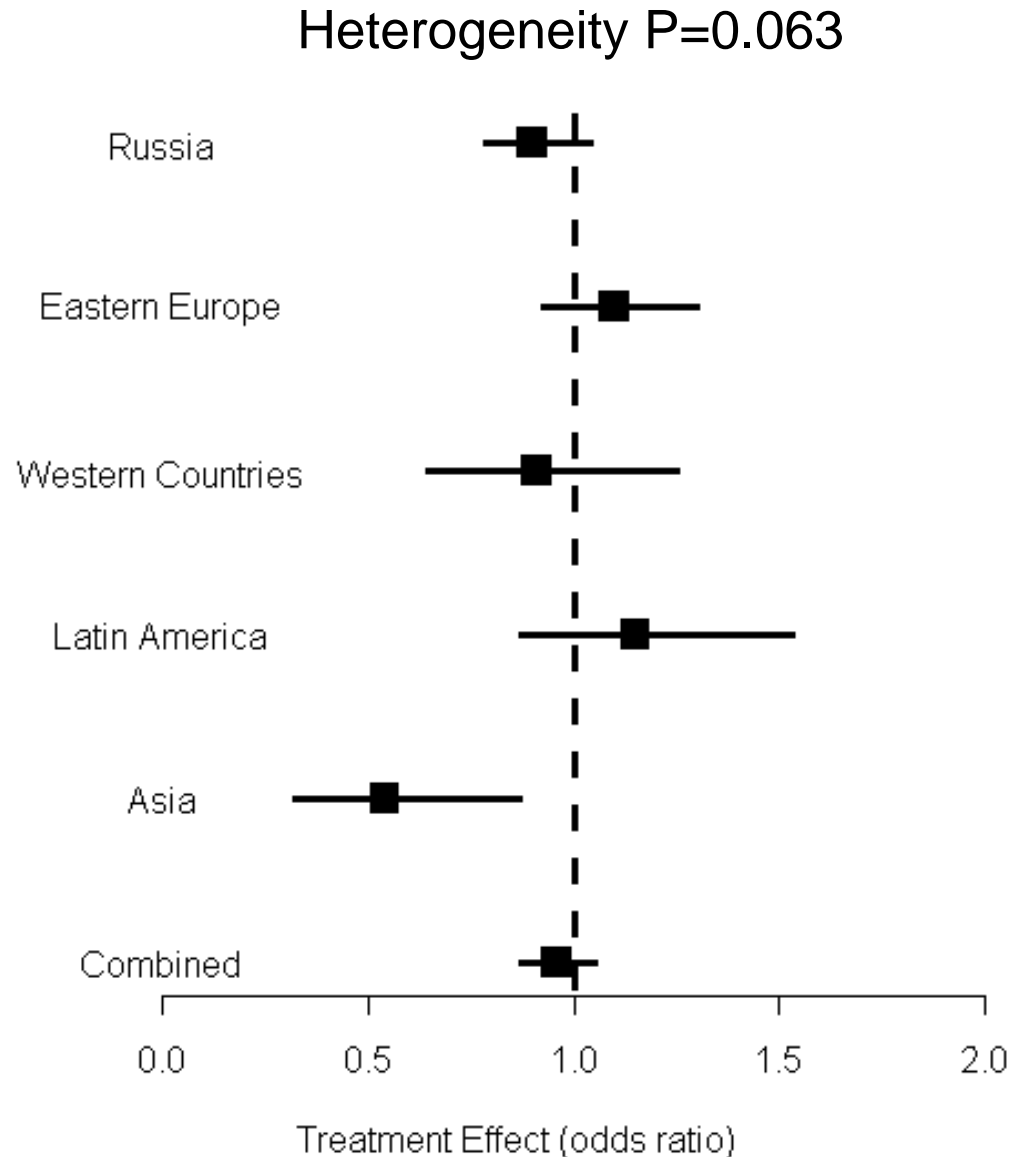
- *no statistical evidence for a difference between regions/subgroups and*
- *no expected mechanism by which such a difference is likely to exist*

Then

- *the most appropriate estimate of the treatment effect in individual regions/subgroups is the **combined overall treatment effect** from the clinical trial.*

Treatment effect estimation

- The most appropriate estimate of mortality treatment effect in Asia is a statistically insignificant odds ratio of **0.96** (combined analysis) not a statistically significant odds ratio of **0.54**





Case Study: *MERIT-HF Trial*

- Beta-blocker treatment in heart failure
- Lancet, 1999, 353:2001-2007
- American Heart Journal, 2001, 142:502-511

- 3,991 patients with chronic heart failure
- Randomised to beta-blocker or placebo

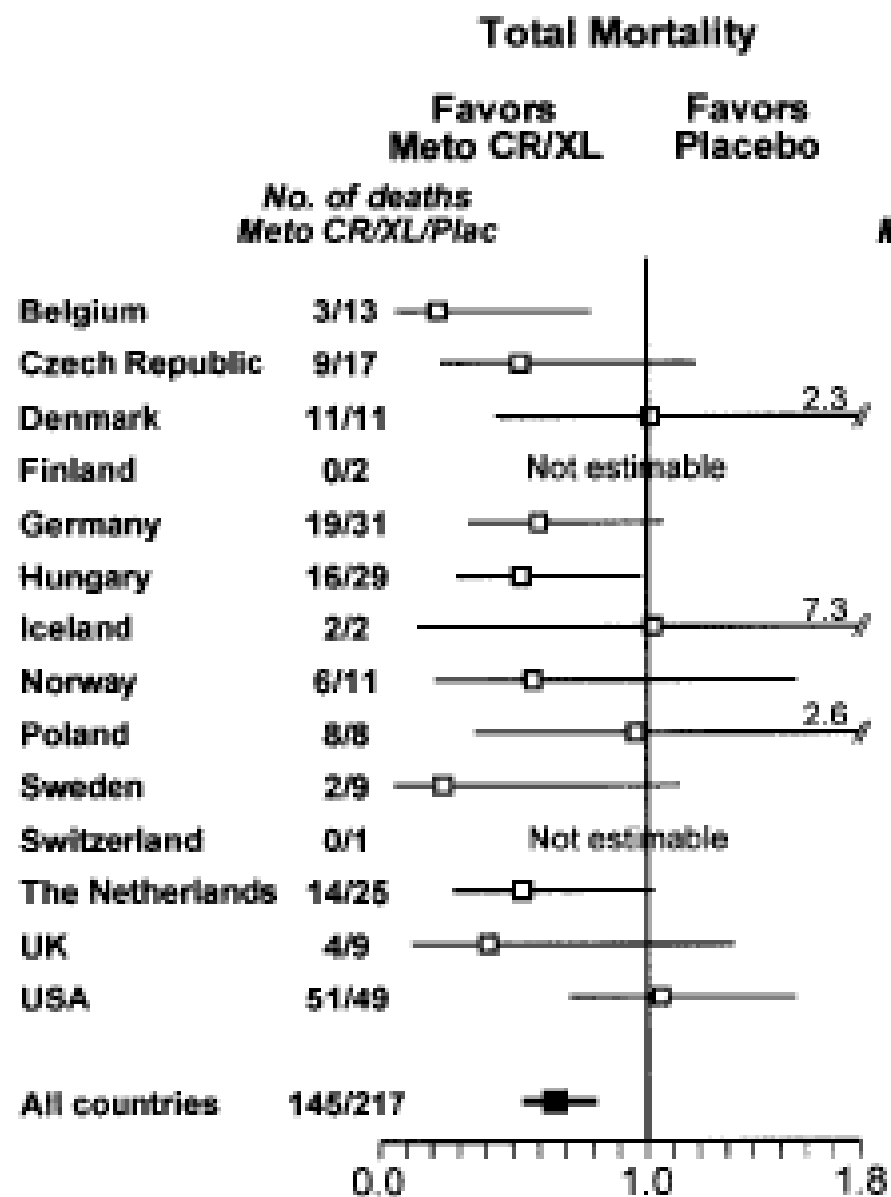
- 313 hospitals in 14 countries


- Primary Outcome:
 - Mortality

Case Study: MERIT-HF Trial



Country	No. randomized	
	Meto CR/XL	Plac
Belgium	68	66
Czech Republic	123	124
Denmark	141	150
Finland	20	14
Germany	252	247
Hungary	211	212
Iceland	19	22
Norway	97	105
Poland	102	102
Sweden	39	46
Switzerland	21	21
The Netherlands	278	270
United Kingdom	87	83
United States	532	539
All countries	1990	2001





Analysis Results – MERIT-HF Trial

- Overall study results:
 - Hazard ratio: 0.66, $P=0.00009$
 - Differences between countries test: $P=0.22$
 - Conclusion: beta-blockers reduce mortality in heart failure

- Post-hoc regional subgroup analysis conducted by FDA:
 - US hazard ratio:
 - 1.05 (95% CI: 0.71 – 1.56)
 - Other countries combined hazard ratio:
 - 0.55 (95% CI: 0.43 – 0.70)
 - US vs. Other: heterogeneity test $P=0.003$
 - Conclusion: mortality benefit not demonstrated in US



Anticipating Chance Variation

- Despite a general understanding of subgroup analysis principles there is still a temptation to over-interpret “surprising” variation in treatment effects across subgroups
- Particularly important for regulatory authorities looking at results for a single country from a multinational trial
- Study design process should include an assessment of the anticipated extent of chance variation in treatment effects and this should be documented in the analysis plan or design paper
- Calibrate expectations, reduce surprises, head off over-interpretation



Measures of Expected Variation

- Expected range of treatment effects
 - *Particularly the expected minimum treatment effect*

- Probability of an “extreme” event
 - *Particularly the chance of at least one group favouring the control when the treatment is beneficial*

- These quantities can be studied by treating the collection of subgroup specific treatment effects as a normally distributed sample (possibly heteroscedastic)

- Order statistics (and their distributions) from such a sample can be used to assess the above quantities

Assumptions

- Two arm study with 1:1 randomisation
- Total of N subjects in R subgroups of size n_i
- Normally distributed endpoint; mean δ , variance σ^2
- D and D_i : overall and subgroup specific treatment differences
- N is chosen for adequate power of the overall comparison

$$N = \frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

- Subgroup sizes n_i may be equal or unequal
- Equal: D_i distributed $N(\delta, s^2 \delta^2)$ where $s^2 = R(z_{1-\alpha/2} + z_{1-\beta})^{-2}$.
- Unequal: D_i distributed $N(\delta, s_i^2 \delta^2)$ where $s_i^2 = p_i^{-1} (z_{1-\alpha/2} + z_{1-\beta})^{-2}$.

Expected Range – Equal Subgroups

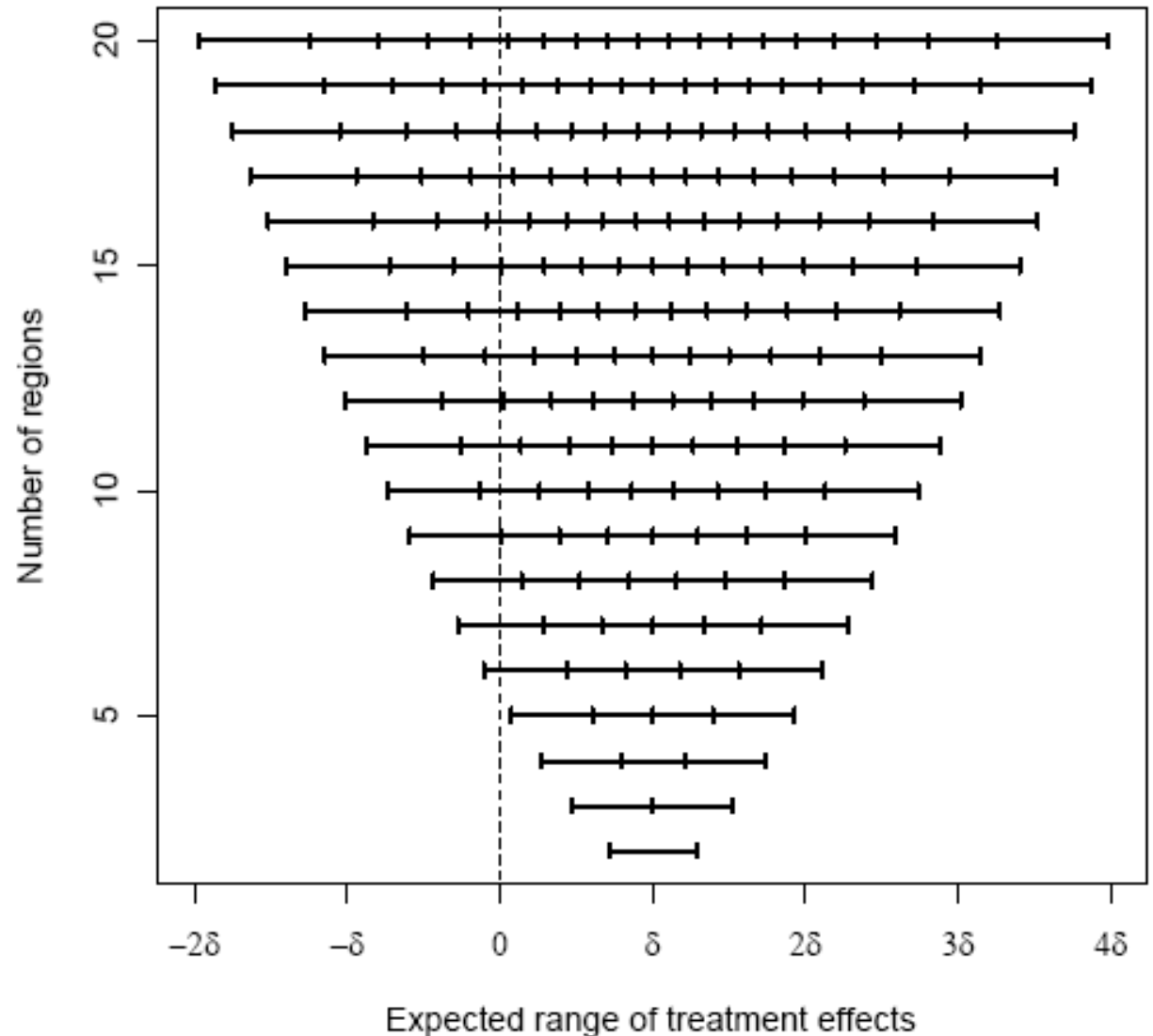
- If subgroup sizes are equal then the expected range can be determined from normal scores

$$e_j(\delta) = E(D_{(j)}) = \delta(1 + O_j^{(R)}s) = \delta e_j(1)$$

- If N is chosen based on power considerations then the expected range is independent of σ and N .
- Expected range can be expressed as a multiple of δ without specifying δ

Expected Range – 80% power, 5% level

- With more than 5 regions we should expect the smallest to favour the control
- With more than 10 regions we should expect multiple treatments to favour the control
- With 10 – 20 regions we should expect the regional treatment effects to range down to between $-\delta$ to -2δ



Expected Range – Unequal Subgroups

- Previous results may underestimate the expected range of treatment effects because subgroups (regions) will normally have unequal sample sizes which increases the variability
- If subgroup sizes are unequal then the smallest treatment effect will satisfy:

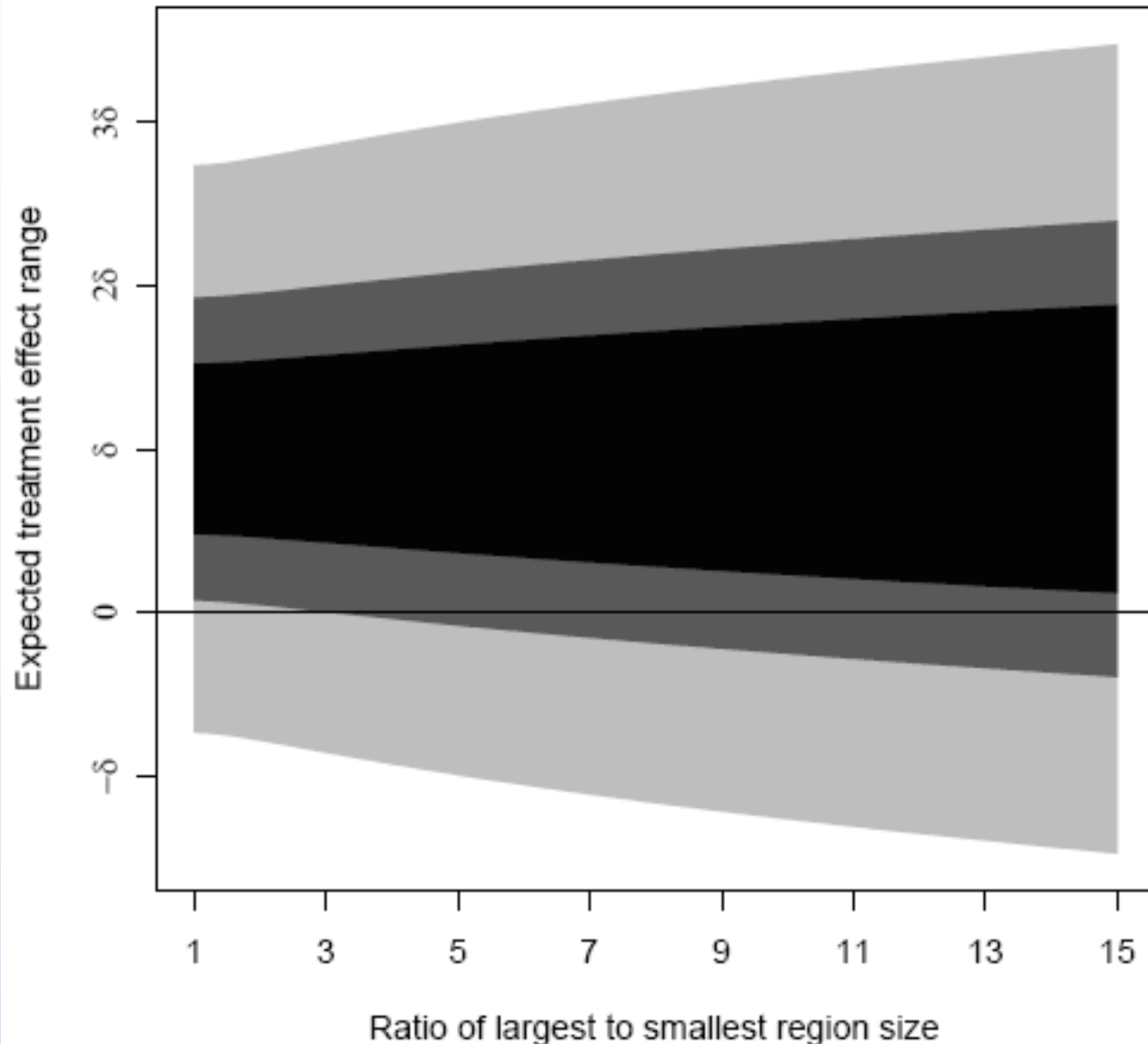
$$F_{\min}(x) = \Pr(D_{(1)} \leq x) = 1 - \prod_{i=1}^R \{1 - \Phi(s_i^{-1}[x\delta^{-1} - 1])\}$$

$$e_{\min}(\delta) = E(D_{(1)}) = \delta \int_{-\infty}^{\infty} y \sum_{i=1}^R s_i^{-1} \phi(s_i^{-1}[y - 1]) \prod_{\substack{j=1 \\ j \neq i}}^R \{1 - \Phi(s_j^{-1}[y - 1])\} dy = \delta e_{\min}(1)$$

- Similar expressions can be determined for the maximum effect
- If N is chosen based on power considerations then the expected range is independent of σ and N .
- Expected range can be expressed as a multiple of δ without specifying δ

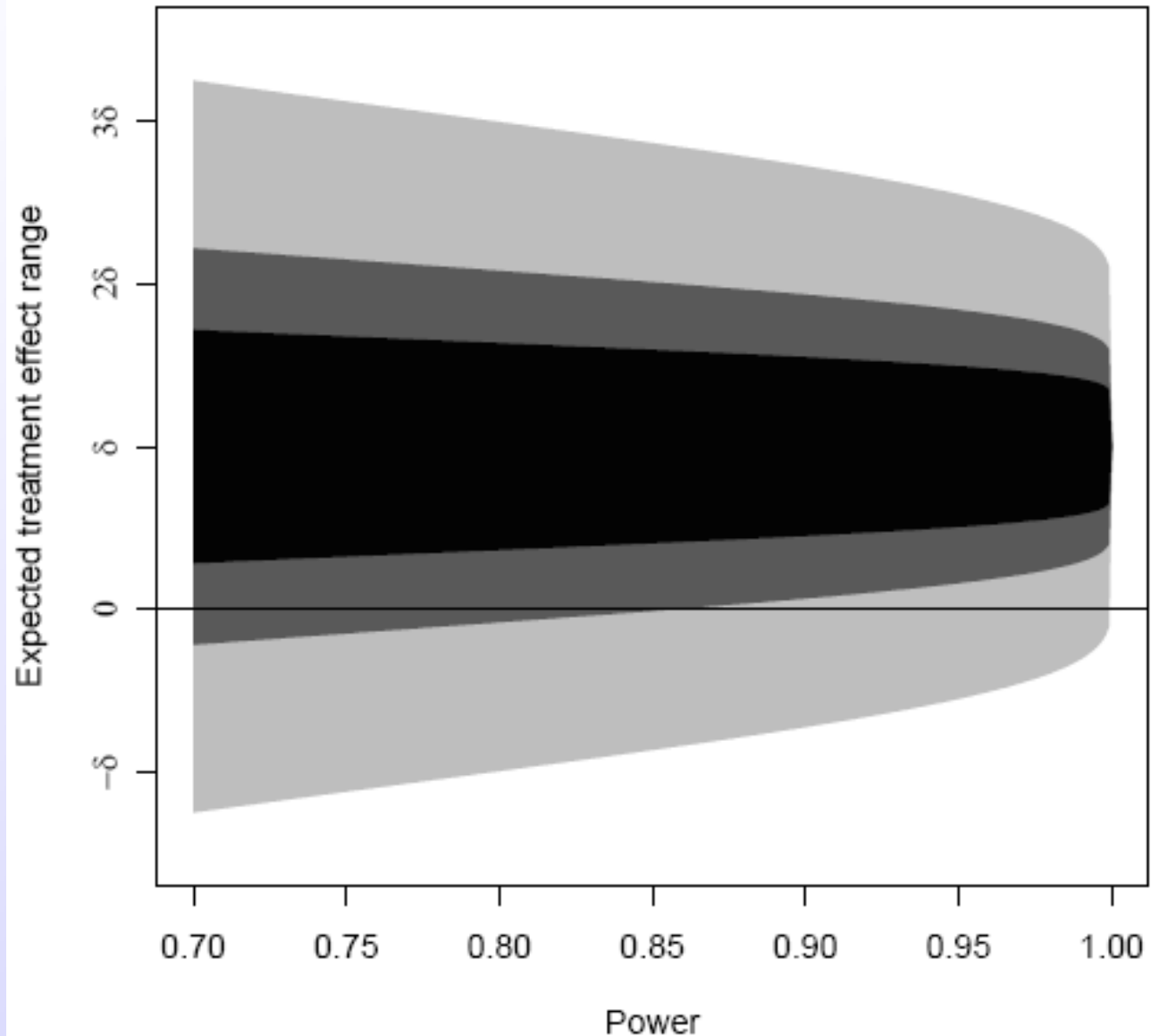
Expected Range – Unequal Subgroups

- Light: $R=10$
 - Medium: $R=5$
 - Dark: $R=3$
- Does variation in regional sample sizes increase expected range of treatment effects?
- Yes, a little bit. Not hugely important



Effect of Increasing the Study's Power

- Light: $R=10$
 - Medium: $R=5$
 - Dark: $R=3$
- Does over-powering a study protect against treatment effect variation?
- No, not with a moderate to large number of regions



Chance of favouring the control

- If a treatment is beneficial and the overall sample size has been determined based on power considerations then the probability of one or more subgroups having an observed treatment effect that favours the control arm is:

$$P_0 = F_{\min}(0) = 1 - \prod_{i=1}^R \{1 - \Phi(-s_i^{-1})\}$$

$$\text{where } s_i^2 = p_i^{-1} (z_{1-\alpha/2} + z_{1-\beta})^{-2}$$

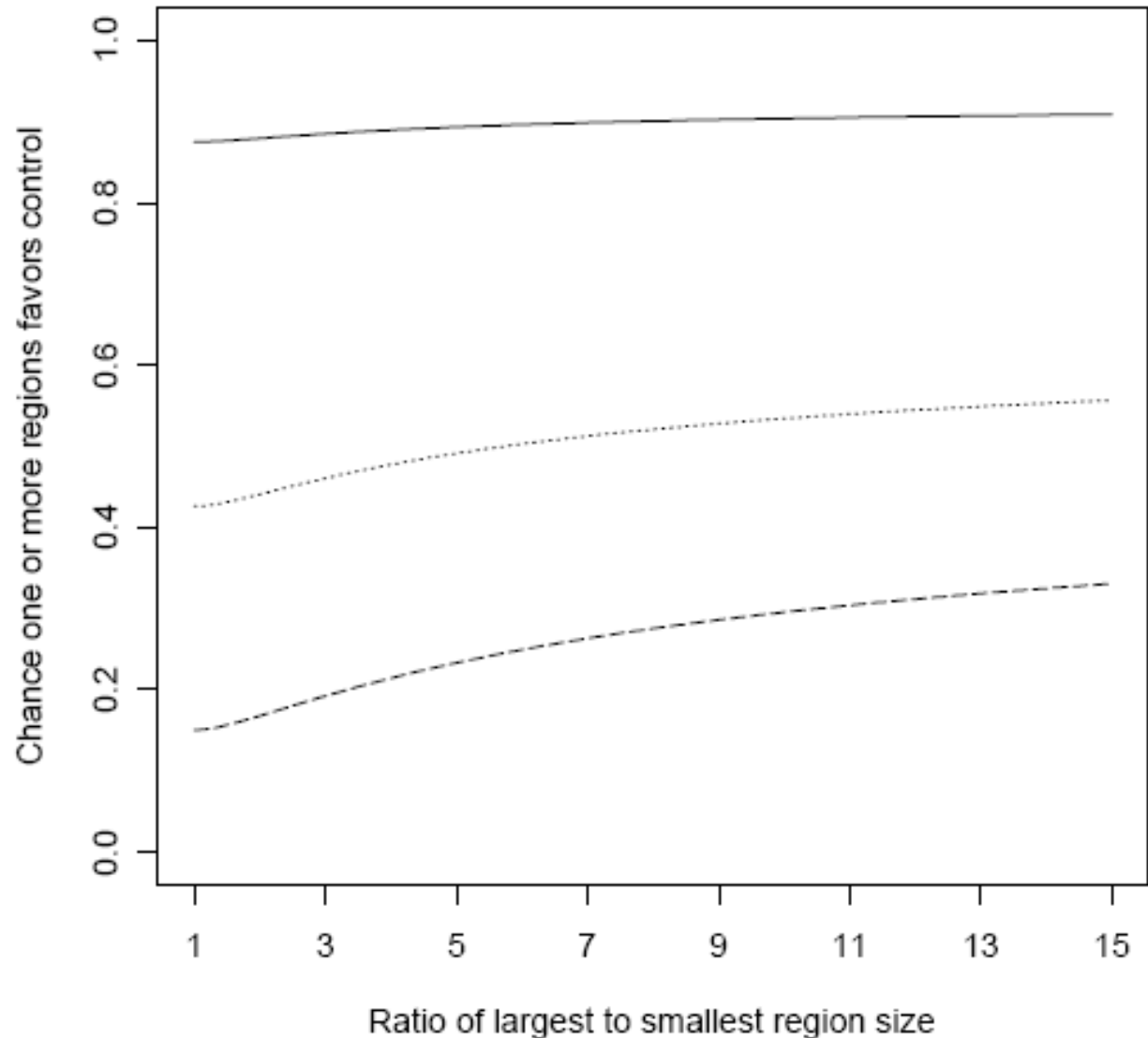
- Very general: independent of σ , N and δ .

Chance of favouring the control

- Solid: R=10
- Dotted: R=5
- Dashed: R=3

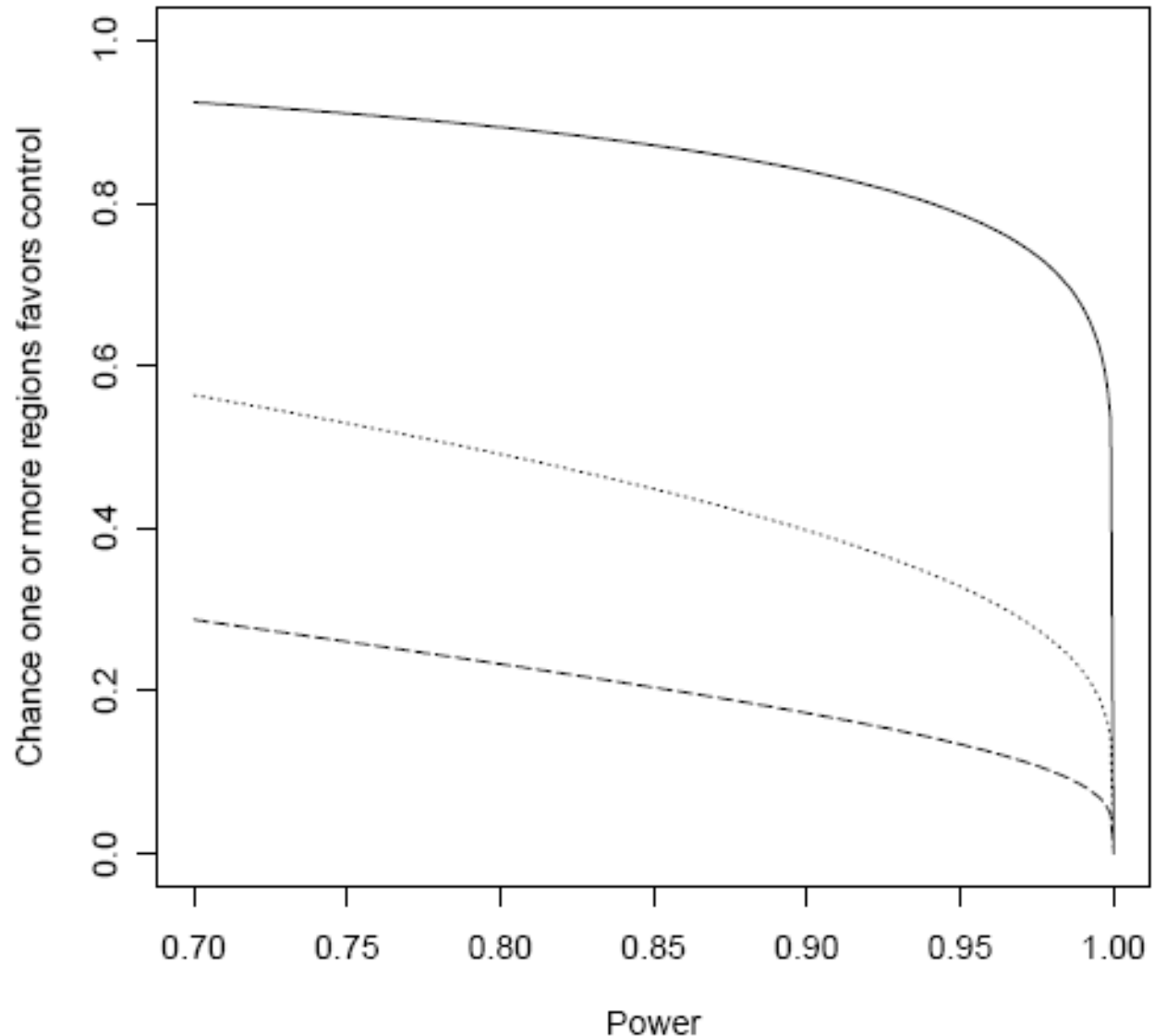
- Chance ranges from low to 50-50 to very high

- Variation in regional sample sizes has little effect



Effect of increasing the study's power

- Solid: $R=10$
 - Dotted: $R=5$
 - Dashed: $R=3$
- Does over-powering a study protect against one or more regions seeming to favour the control?
- No, hardly at all (e.g. compare 80% to 95% power)





Key points

- Purely by chance, the observed experimental treatment effect in different regions or subgroups can be expected to range from beneficial to apparently harmful
- It should not be surprising if the experimental treatment seems to favour the control arm in one or more subgroups/regions, even if the overall study shows significant benefit
- Increasing the power of a study to high levels (e.g. 95%) does not provide protection against this phenomenon
- When we design studies we need to prepare ourselves for the extent of likely variation in treatment effects, rather than waiting until the end of the study and being surprised by the apparent variation



Example documentation – 10 regions

*Even if the experimental treatment is beneficial, it will not be surprising if chance variation leads to one or more individual regions having treatment effect estimates that appear to show no benefit or even harm. In particular, assuming a beneficial experimental treatment effect of δ as used in the power calculations, and assuming that the treatment works homogeneously across all regions, **we can expect the regions to have treatment effect estimates that range from a difference on the order of δ in favour of the control arm to a difference on the order of 3δ in favour of the experimental arm. Furthermore there is a probability of greater than 85% that at least one region will have a treatment effect estimate that favours the control arm.** These calculations indicate that even if the experimental treatment is homogeneously beneficial across all regions, we can expect substantial chance variation between regions. Accordingly, unless there is a significant test of heterogeneity of region-specific treatment effects, apparent differences in treatment effects between regions will be attributed to chance variation rather than to genuine differences between the regions.*



Chance variation in treatment effects

- Seemingly large variation in the observed treatment effects in different countries can arise even though there is no underlying “true” difference between subgroups
- The methods suggested here are a complement to formal interaction/heterogeneity testing which is sometimes ignored in favour of intuition
- Only compare individual treatment effects in different subgroup/regions when a test of heterogeneity has shown evidence of true underlying differences in the treatment effects



Inflated Significance level

- Another reason why interaction tests are sometimes ignored is their low power. Sometimes people try to increase the power by increasing α
- Use of an unconventionally high significance level (e.g. 10% or 20%) to conduct tests of heterogeneity between regions/subgroups may be misleading: simply replaces one problem with another
- Evidence that treatment effects differ between subgroups is only convincing if based on a conventional significance level (e.g. 5%).



Explaining treatment differences

- Even when there is a significant test of heterogeneity, the potential explanations for treatment effect differences should be examined carefully
- Subgroup analysis principles generally require a plausible biological mechanism, in addition to statistical significance, before treatment heterogeneity can be concluded
- Even when significant treatment effect differences do emerge, they may not be caused by the treatment having different efficacy in different regions/subgroups



Summary of Principles

- Subgroup analysis principles should be used to assess subgroup differences, including the use of formal statistical tests of heterogeneity at conventional significance levels
- Seemingly large differences between subgroups can arise purely by chance
- Separate tests of treatment effect in different subgroups do not provide evidence of treatment effect differences
- In the absence of statistically significant heterogeneity, the most appropriate treatment effect estimate in any region/subgroup is the combined overall estimate from the multinational or multicenter trial
- Consistent with subgroup analysis principles, plausible biological or other mechanisms are generally required for a definitive conclusion of heterogeneity
- **Multinational trials can benefit from documentation of expected inter-region differences in the analysis plan and/or design paper**