Methods for exploring treatment effect heterogeneity in global clinical trials

Ian Marschner

(joint work with Manjula Schou)

Macquarie University & CTC

Multinational Trials

- Faster and more efficient clinical trials through pooling of resources
- Greater generalizability through broad assessment of treatments in different regions and countries
- Reduction in the need for replication of research results in different populations
- Multinational trials can be 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

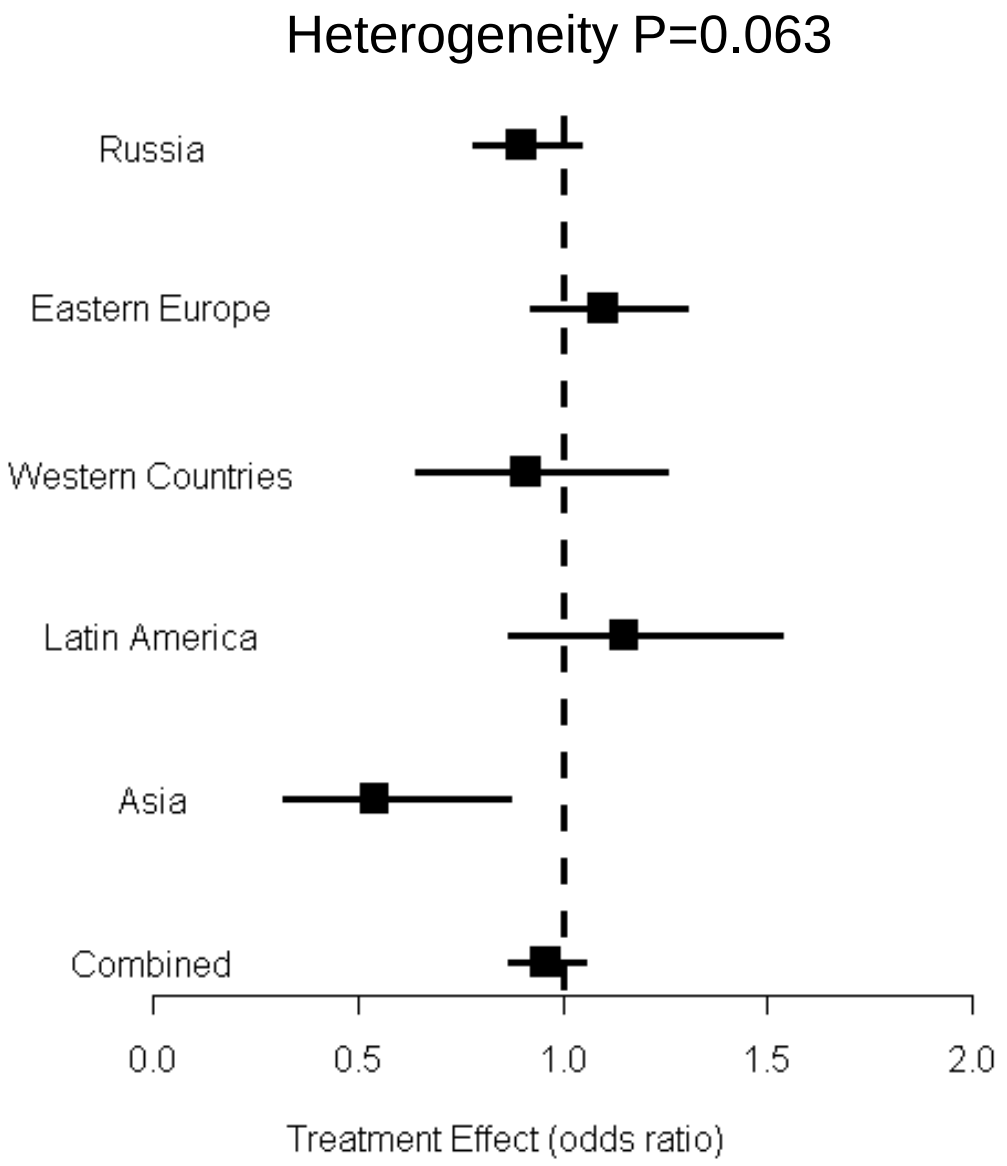
- While there may be many plausible explanations for variation in treatment effects, none can be entertained unless “chance” can be ruled out as an explanation
- Evidence of treatment effect differences over and above chance variation comes only from a test of heterogeneity across regions/subgroups
- Separate tests of treatment effect in different regions do not provide evidence of treatment effect differences

Example 1: HERO-2 Trial

- Thrombin-specific anticoagulation with **bivalirudin versus heparin** in patients receiving fibrinolytic therapy for acute myocardial infarction
- Lancet 2001; 358:1855-1863
- 17,073 patients with acute MI (heart attack)
- 539 hospitals in 46 countries
- Primary Outcome:
  - 30 day 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
- The most appropriate estimate of mortality treatment effect in Asia is a statistically insignificant odds ratio of 0.96



Example 2: MERIT-HF Trial

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

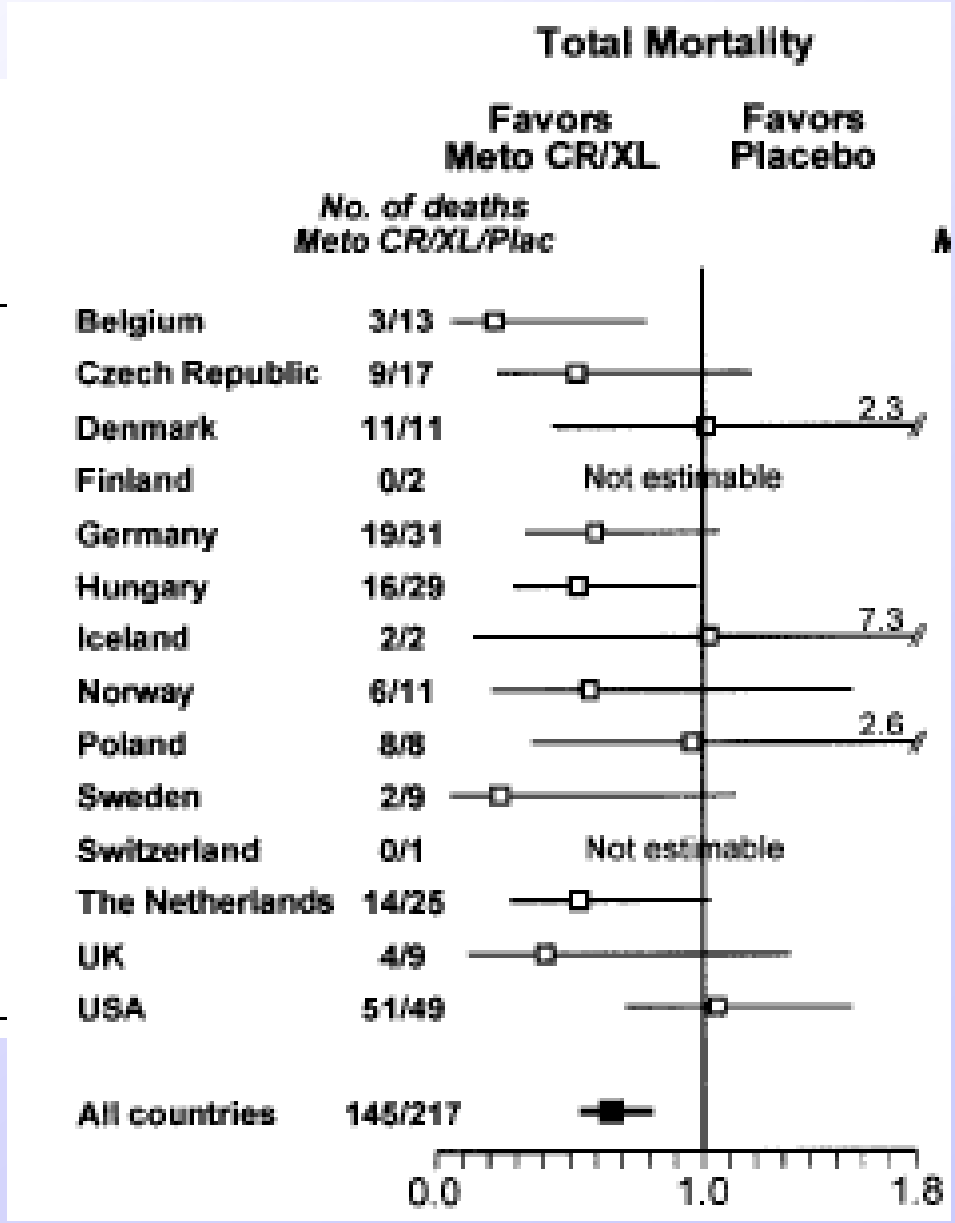
- 3,991 patients with chronic heart failure
- Randomised to beta-blocker or placebo

- 313 hospitals in 14 countries

- Primary Outcome:
  - Mortality

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

Prospective meta-analysis

- An alternative to a multinational clinical trial is a prospective meta-analysis (PMA) of separate studies run under the same (or similar) protocol in different countries/regions
- Prospectively plan to combine study results in a meta-analysis
- Like multinational studies, PMAs can be difficult to interpret if there is variation in treatment effects between studies

Example 3: Oxygen Use in neonates

- Does lower versus higher O_2 targeting from birth in extremely preterm infants improve Retinopathy of Prematurity (ROP) rates (i.e. visual impairment outcomes) without increasing mortality by more than 3%?

Option 1: do a single trial large enough to detect a 3% difference (90% power, 5% sig level), minimum N=5737

Option 2: meta-analyse results from several trials designed to address the same question prospectively

Example 3: Oxygen Use in neonates

- Single study too big for any single funding agency
- Prospective meta-analysis planned of individual patient data from 6 similar worldwide trials

COUNTRY	STUDY	N
USA	SUPPORT	1310
USA	POST	1220
Australia	BOOST II	1200
New Zealand	BOOST NZ	320
UK	BOOST II UK	1200
Canada	COT	1200

First study to report
(NEJM, May 2010)

Example 3: Oxygen Use in neonates

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Target Ranges of Oxygen Saturation in Extremely Preterm Infants

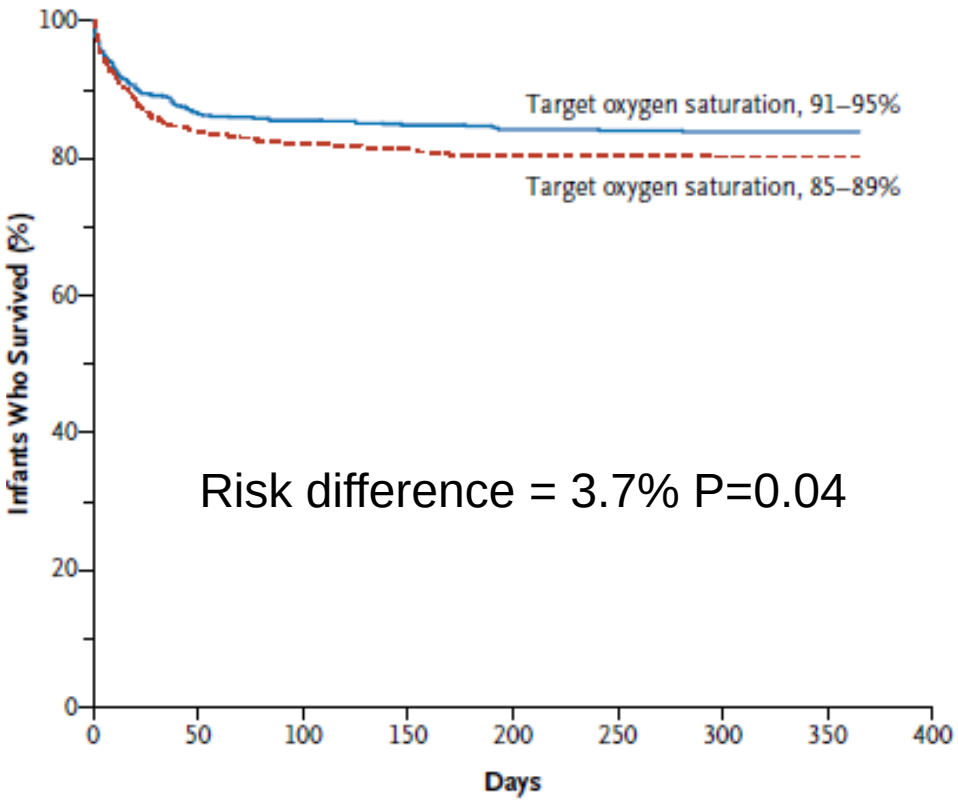
SUPPORT Study Group of the Eunice Kennedy Shriver NICHD Neonatal Research Network*

The NEW ENGLAND JOURNAL of MEDICINE

EDITORIAL

CPAP and Low Oxygen Saturation for Very Preterm Babies?

Colin J. Morley, M.D.



Risk difference = 3.7% P=0.04

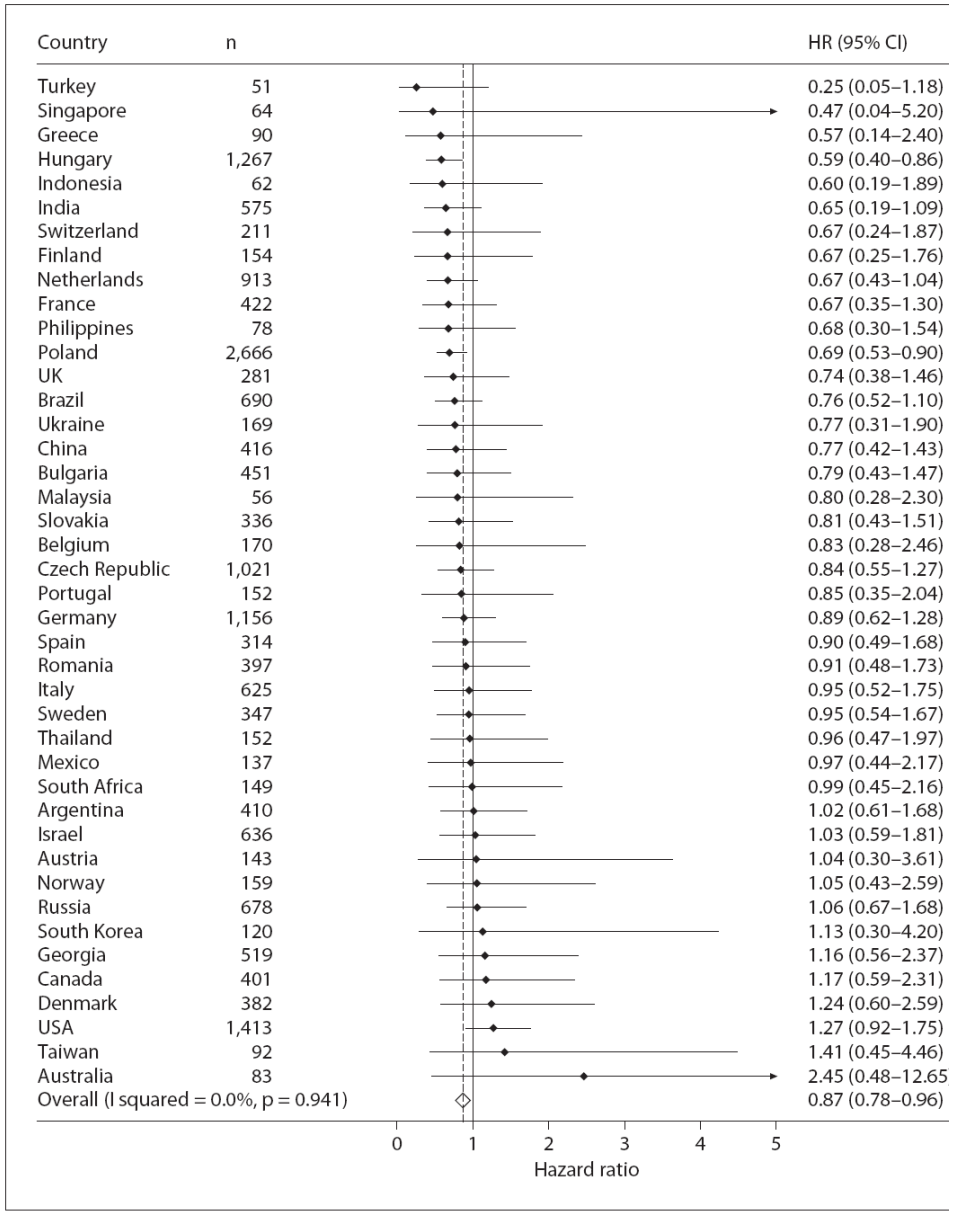
CONCLUSIONS

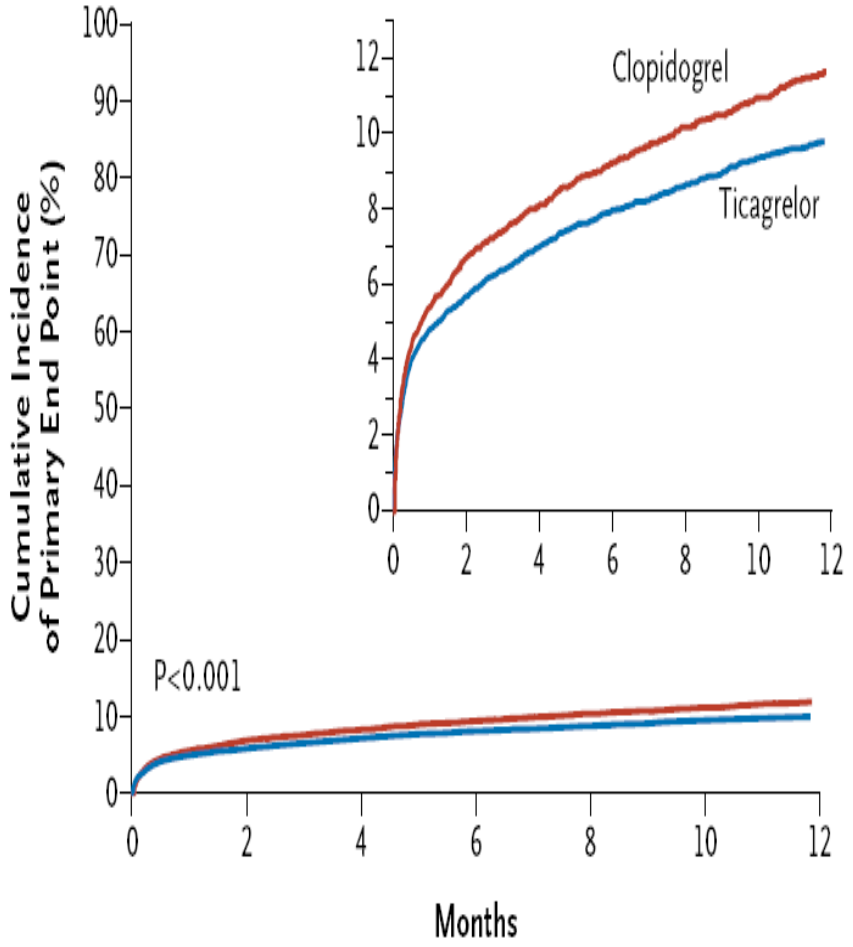
A lower target range of oxygenation (85 to 89%), as compared with a higher range (91 to 95%), did not significantly decrease the composite outcome of severe retinopathy or death, but it resulted in an increase in mortality and a substantial decrease in severe retinopathy among survivors. The increase in mortality is a major concern, since a lower target range of oxygen saturation is increasingly being advocated to prevent retinopathy of prematurity. (ClinicalTrials.gov number, NCT00233324.)

Example 4: PLATO Trial

- Ticagrelor versus clopidogrel for the prevention of cardiovascular events
- NEJM 2009; 361:1045-1057
- 18,624 patients with acute coronary syndromes
- 42 countries
- Primary Outcome:
  - Death from vascular causes, MI, stroke

PLATO Trial


Country
n
HR (95% CI)
Turkey 51 0.25 (0.05–1.18)
Singapore 64 0.47 (0.04–5.20)
Greece 90 0.57 (0.14–2.40)
Hungary 1,267 0.59 (0.40–0.86)
Indonesia 62 0.60 (0.19–1.89)
India 575 0.65 (0.19–1.09)
Switzerland 211 0.67 (0.24–1.87)
Finland 154 0.67 (0.25–1.76)
Netherlands 913 0.67 (0.43–1.04)
France 422 0.67 (0.35–1.30)
Philippines 78 0.68 (0.30–1.54)
Poland 2,666 0.69 (0.53–0.90)
UK 281 0.74 (0.38–1.46)
Brazil 690 0.76 (0.52–1.10)
Ukraine 169 0.77 (0.31–1.90)
China 416 0.77 (0.42–1.43)
Bulgaria 451 0.79 (0.43–1.47)
Malaysia 56 0.80 (0.28–2.30)
Slovakia 336 0.81 (0.43–1.51)
Belgium 170 0.83 (0.28–2.46)
Czech Republic 1,021 0.84 (0.55–1.27)
Portugal 152 0.85 (0.35–2.04)
Germany 1,156 0.89 (0.62–1.28)
Spain 314 0.90 (0.49–1.68)
Romania 397 0.91 (0.48–1.73)
Italy 625 0.95 (0.52–1.75)
Sweden 347 0.95 (0.54–1.67)
Thailand 152 0.96 (0.47–1.97)
Mexico 137 0.97 (0.44–2.17)
South Africa 149 0.99 (0.45–2.16)
Argentina 410 1.02 (0.61–1.68)
Israel 636 1.03 (0.59–1.81)
Austria 143 1.04 (0.30–3.61)
Norway 159 1.05 (0.43–2.59)
Russia 678 1.06 (0.67–1.68)
South Korea 120 1.13 (0.30–4.20)
Georgia 519 1.16 (0.56–2.37)
Canada 401 1.17 (0.59–2.31)
Denmark 382 1.24 (0.60–2.59)
USA 1,413 1.27 (0.92–1.75)
Taiwan 92 1.41 (0.45–4.46)
Australia 83 2.45 (0.48–12.65
Overall (I squared = 0.0%, p = 0.941) 0.87 (0.78–0.96)
0 1 2 3 4 5
Hazard ratio


Cumulative Incidence of Primary End Point (%)
Clopidogrel
Ticagrelor
P<0.001
Months
No. at Risk
Ticagrelor 9333 8628 8460 8219 6743 5161 4147
Clopidogrel 9291 8521 8362 8124 6650 5096 4047

Analysis Results – PLATO Trial

- Overall study results:
  - Hazard ratio: 0.84, P<0.001
  - Conclusion: ticagrelor is superior to clopidogrel for reduce cardiovascular events
  - But ... differences between 5 regions test: P=0.045

    (*one of 33 subgroup analyses*)

- Post-hoc regional subgroup analysis conducted by FDA:
  - US hazard ratio: 1.27 (95% CI: 0.92 – 1.75)
  - Conclusion: mortality benefit not demonstrated in US
  - Many “non-chance” explanations proposed
    - e.g. monitoring in US done by CRO whereas in other countries it was done by company

Chance Variation

- Despite a general understanding of subgroup analysis principles there is still a temptation to over-interpret “surprising” variation in treatment effects across regions
- Particularly important for regulatory authorities looking at results for a single country from a multinational trial
- We consider a range of graphical summaries of treatment effect variation to supplement a formal heterogeneity test
- Calibrate expectations, head off over-interpretation

Measures of Expected Variation

- Expected range of regional treatment effects
  (particularly the expected minimum treatment effect)
- Generalisation: Expected distribution of subgroup specific treatment effects
- Probability of “inconsistency” *(particularly the chance of at least one region favouring the control when the treatment is beneficial)*
- Generalisation*:* Distribution of the number of subgroups that show an inconsistent treatment effects

Measures of Expected Variation

- These quantities can be studied by treating the collection of region-specific treatment effects as a normally distributed sample
- The theory of Order Statistics can be used to assess the above quantities
- Graphical comparison of the **expected** behaviour of subgroup-specific treatment effects with the **observed** behaviour of subgroup-specific treatment effects can supplement information available from a formal test of heterogeneity

Assumptions

- Two arm study with 1:1 randomisation
- Total of *N* subjects in *R* regions of size n_i
- Normally distributed endpoint; mean δ, variance σ^2
  (actually all that is needed is a normally distributed test statistic)
- *D* and D_i : overall and region-specific treatment differences
- *N* is chosen for adequate power of the overall comparison
- Equal n_i : D_i distributed $N(\delta, s^2\delta^2)$ where $s^2 = R(z_{1-\alpha/2} + z_{1-\beta})^{-2}$.
- Unequal n_i : 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 the region sizes are equal then the expected range can be determined from the "normal scores" $\left\{O_j^{(R)};\, j=1,\ldots,R\right\}$ for a sample of size R

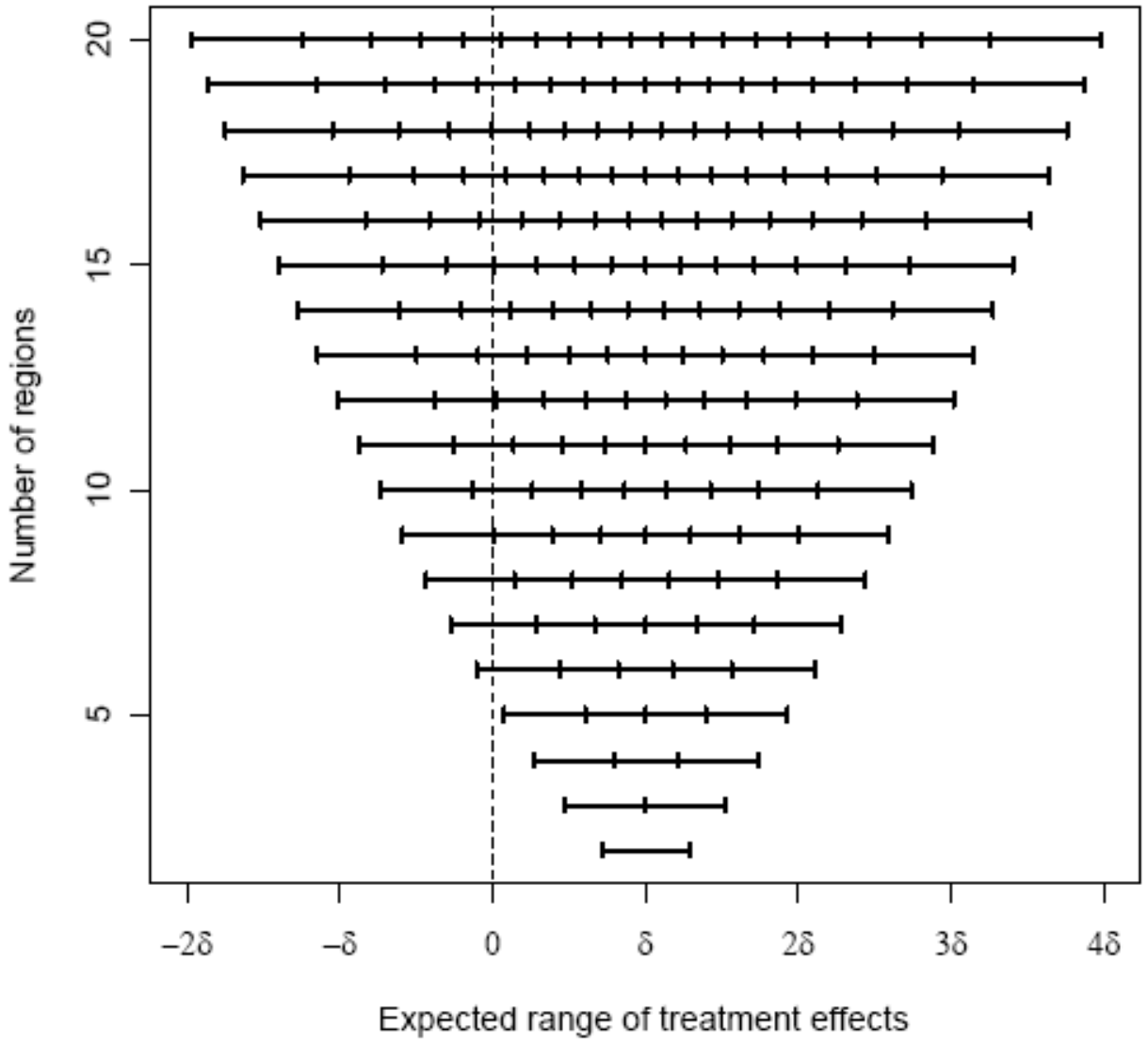
$$e_j(\delta) = E\left(D_{(j)}\right) = \delta\left(1 + O_j^{(R)} s\right) = \delta e_j(1)$$

- 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 regions will usually have unequal sample sizes which increases the variability
- If region sizes are unequal then the smallest treatment effect will satisfy:

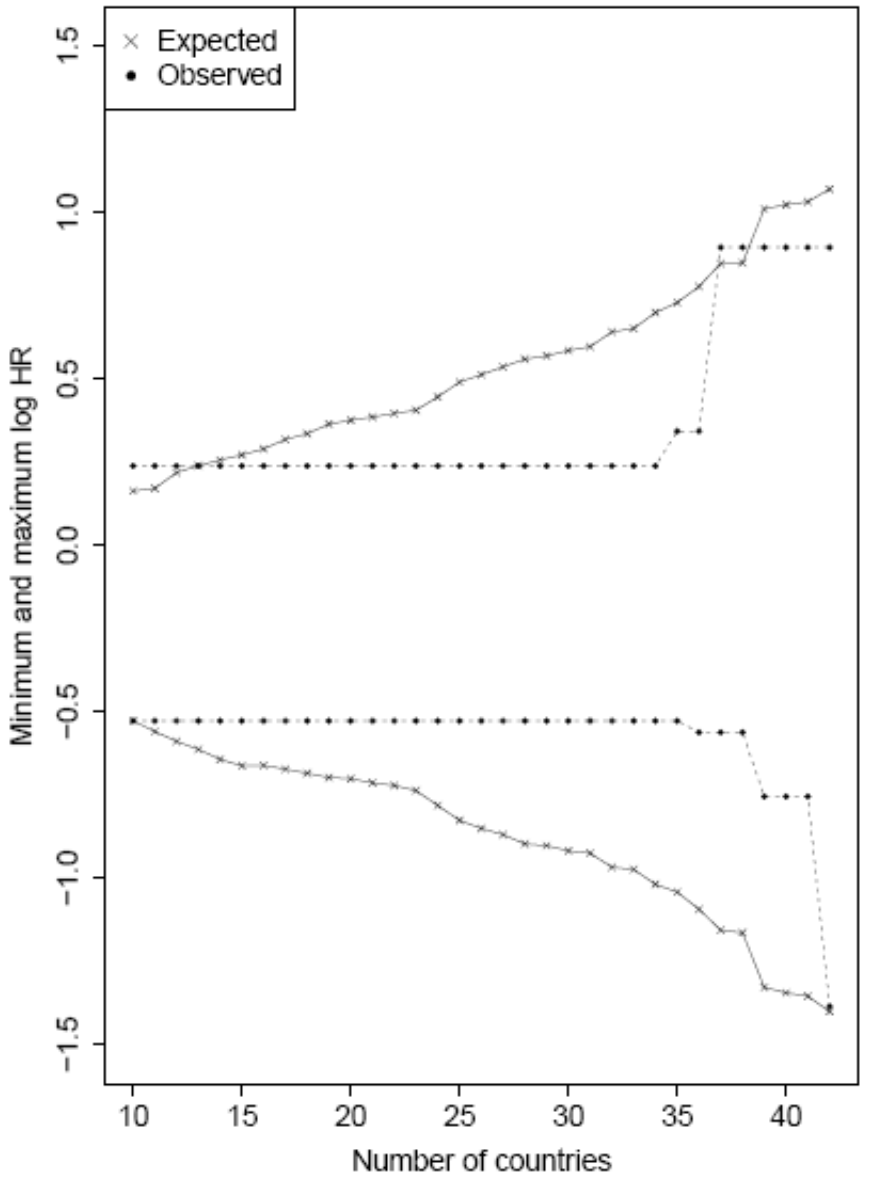
$$F_{\min}(x) = \Pr\left(D_{(1)} \le x\right) = 1 - \prod_{i=1}^{R} \left\{1 - \Phi\left(s_i^{-1}[x\delta^{-1} - 1]\right)\right\}$$

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

- 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

Example 4: PLATO study expected range

- The observed range of treatment effects in the PLATO study is not larger than the expected range due to chance, regardless of whether all studies, or just the largest studies are considered



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 regions having an observed treatment effect that favours the control arm is:

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

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

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

Generalisations

1. Rather than just comparing the observed range of treatment effects with the expected range one can compare the observed range with the sampling distribution of the range

2. Rather than just considering the range of subgroup-specific treatment effects one can consider the observed distribution of all treatment effects, and compare this with the expected distribution of treatment effects

3. Rather than looking just at the probability of **any** inconsistent treatment effects, one can look at the probability of observing as many inconsistent effects as were observed

Implementation

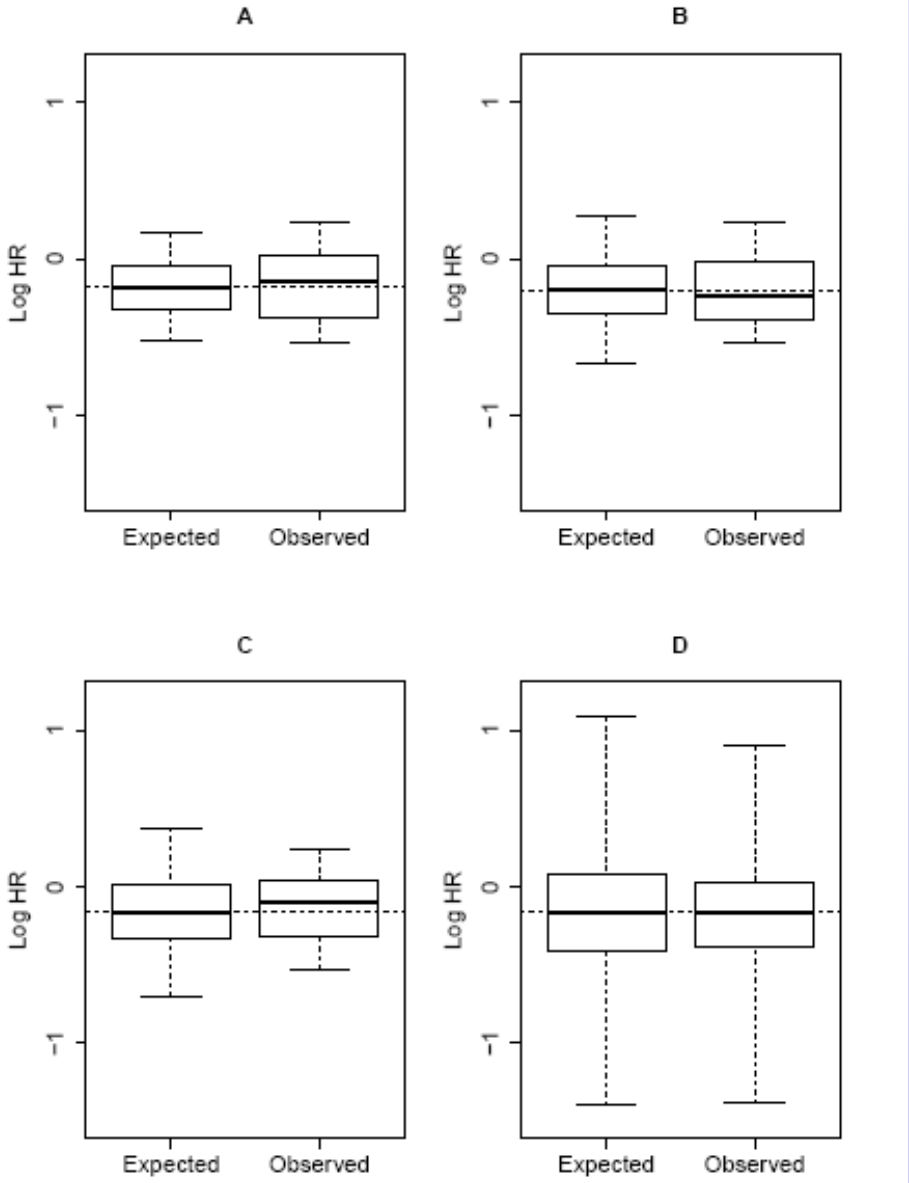
- The above generalisations can be determined based on the theory of order statistics for heteroscedastic normal samples
- The formulas are formidable but are manageable computationally either by direct theoretical computations or through simulation
- An R package is currently under development to produce graphical assessments of subgroup variation based on the above measures

Example 4: PLATO Study

- A: largest 10 countries
- B: largest 15 countries
- C: largest 20 countries
- D: largest 42 countries (all)

- The distribution of observed treatment effects across countries is consistent with the expected distribution

- If anything the observed spread of treatment effects is less than expected



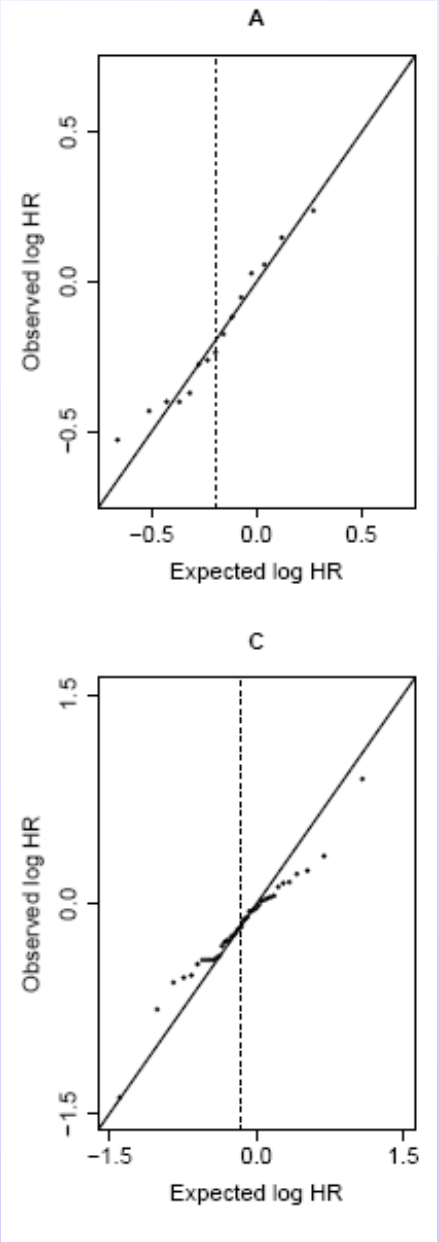
Example 4: PLATO Study

- A: largest 10 countries
- C: largest 42 countries (all)

- A Q-Q type plot is another way to display the same information

- No evidence of wider variation than expected

- If anything less variation than expected

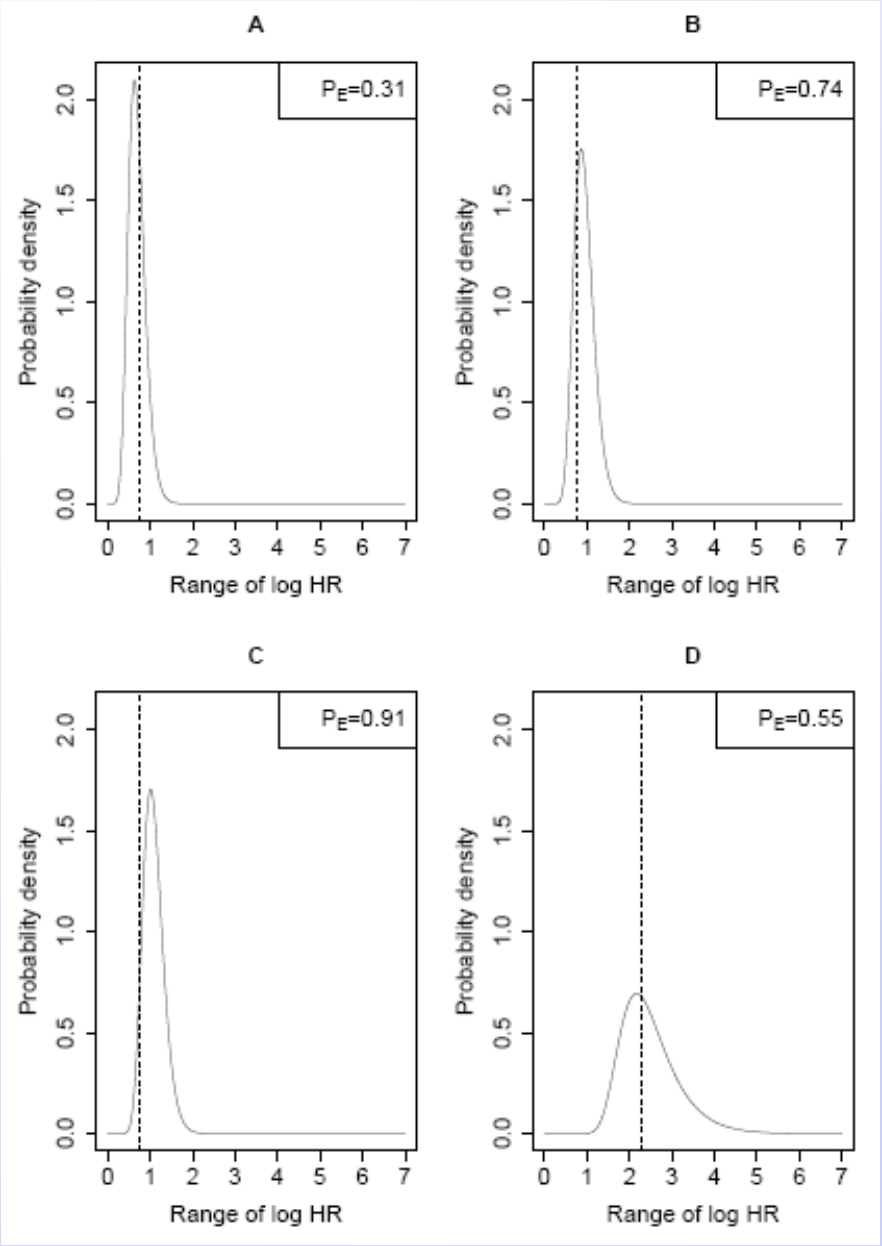


Example 4: PLATO Study

- A: largest 10 countries
- B: largest 15 countries
- C: largest 20 countries
- D: largest 42 countries (all)

- The observed treatment effect range is consistent with the sampling distribution assuming treatment homogeneity

- P_E is probability of obtaining a range at least as extreme as the one observed

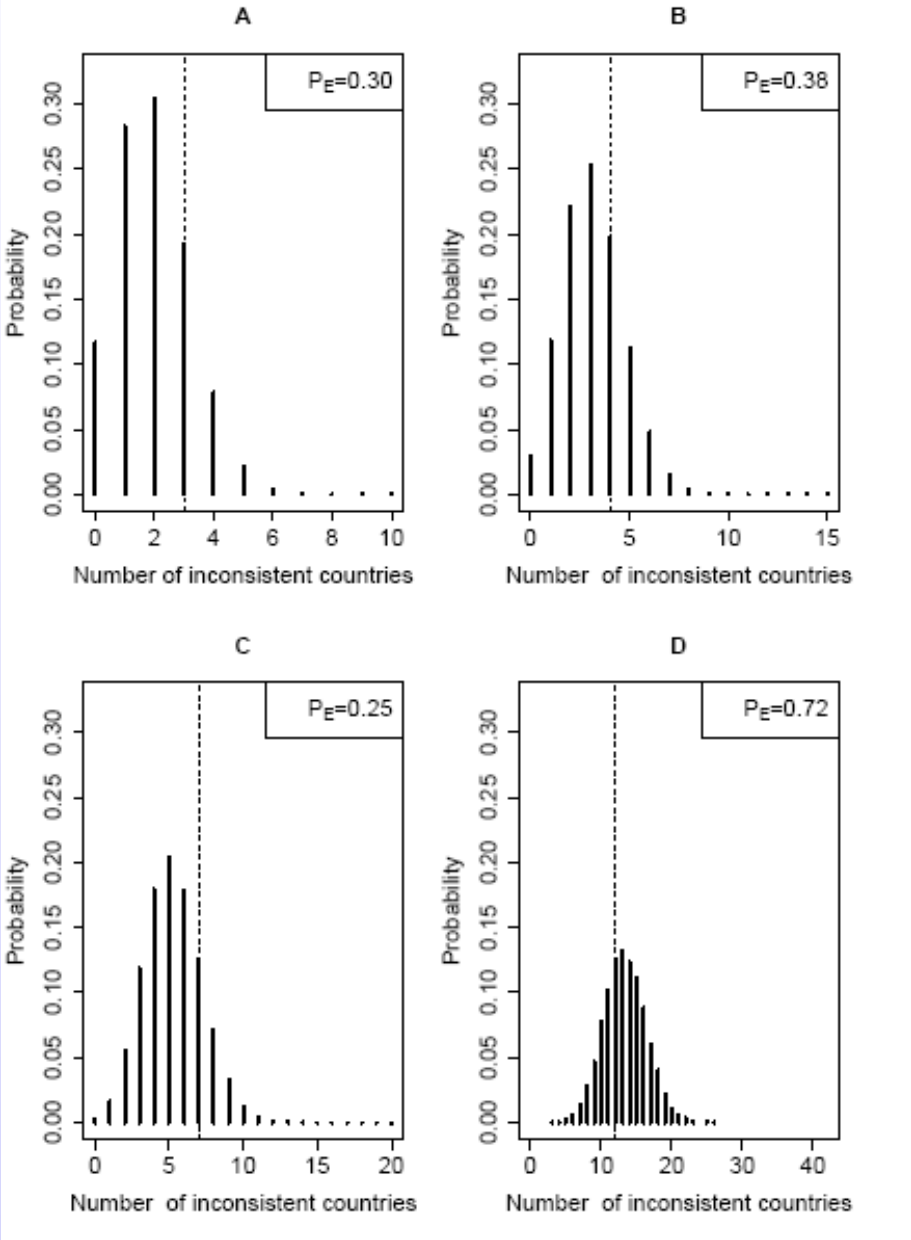


Example 4: PLATO Study

- A: largest 10 countries
- B: largest 15 countries
- C: largest 20 countries
- D: largest 42 countries (all)

- The observed number of inconsistent countries is consistent with what would be expected due to chance

- P_E is probability of obtaining a number of inconsistent countries at least as extreme as the one observed



Conclusions

- Purely by chance, the observed experimental treatment effect in different regions is often expected to range from beneficial to apparently harmful
- The range of observed treatment effects can be wider than intuitively expected and should not be over-interpretted
- It is usually not surprising if the treatment effect seems to favour the control arm in one or more regions, even when the overall study shows a significant benefit
- Graphical assessment of the expected and observed variation in subgroup-specific treatment effects can be a useful supplement to a formal test of heterogeneity