Heterogeneity of Treatment Effects: True Effect or an Illusion

John R. Cook
Merck Research Labs
West Point, PA

ISPOR 15th Annual International Meeting
Atlanta, Georgia
May 17, 2010

Heterogeneity as Interaction

• Heterogeneity of treatment effect manifests itself statistically as an interaction
  – Without heterogeneity, treatment effect is characterized by overall mean
  – With heterogeneity, average treatment effect differs between subgroups

• Basic types of interaction
  – Quantitative (differ in magnitude)
  – Qualitative (differ in direction)

Quantitative Interaction
Same Direction of Effect

Group 1

A

B

Treatment

Group 2

Qualitative Interaction
Different Direction of Effect

Group 1

A

B

Treatment

Group 2

Subgroup Analyses

• Often investigate whether the overall treatment effect differs by one or more subgroup
  – Survey of 50 RCTs from 4 medical journals (Assmann)
  – 2 out of 3 reported at least one subgroup analysis

• Subgroup analyses are appropriate when
  – Test whether treatment effect differs between subgroups (interaction test)
  – Subgroup analyses are pre-specified
  – Subgroups are sufficiently large
  – Statistical analysis appropriately conservative

Issue: Inappropriate Analysis

• Test treatment effect within each subgroup
  – What is the treatment effect for a given subgroup? Does it differ from the null?

• Such tests address the wrong question
  – What do we conclude if one subgroup treatment difference is significant and another is not?
  – What do we conclude if both subgroup-specific treatment differences are significant?
  – Don’t know if the subgroups differ from each other

• Need to perform formal test of interaction
Interpreting Within Subgroup Analyses

One (blue) subgroup with significant treatment effect

Both subgroups with significant treatment effect

VS.

VS.

Do subgroups differ? Need an interaction test

Issue: Multiplicity

- With any test, there is a risk of “finding” a significant result when none exists (type I error, \( \alpha \))
  - May attribute clinical importance to one subgroup or withhold treatment to another inappropriately
  - Are you born under Libra or Gemini?
    - Analysis of ISIS-2 trial: aspirin ineffective vs placebo (Rothwell 2005)
  - In practice, multiple subgroup analyses are often conducted
    - Risk of a false positive can be very high
      - Risk: 40% with 10 tests using \( \alpha = .05 \)

Issue: Lack of Power

- Tests of interaction typically have low statistical power
  - Studies are designed to test overall significance
  - Sample size inadequate – particularly problematic if there are many subgroups
  - Simulation by Brookes (2004): trial with 80% power had 29% power to detect an interaction
  - Consequently, may be tempted to lower the threshold for significance (i.e., set \( \alpha = .20 \))
    - Easier to identify a true subgroup effect
    - But, false positive rate may be substantial

Issue: Risk of Misinterpretation

- Tendency to report “interesting” subgroups
  - Post-hoc emphasis on the subgroup of interest can lead to exaggerated claims
  - “Uninteresting” subgroup analyses may not be reported
    - Particularly problematic if post-hoc subgroup analyses are conducted, but not reported
    - Cannot properly adjust for multiplicity
- Unanticipated results should be confirmed
Designing for Heterogeneity

- Conduct trial specifically to confirm prior observation
  - Pre-specify direction and magnitude of anticipated subgroup effects (biologic/clinical rationale)
  - Consider stratification by important baseline characteristic(s) suspected to influence outcome
  - Sample size should provide sufficient power
- Cautions
  - Study may over-power for overall treatment effect
  - So may require adaptive design to minimize exposure
- Key: Confirmation thru replication

True Effect or an Illusion

In practice, it is difficult to determine whether an observed subgroup difference is real or simply reflects noise in the data.

References