Analysis in RCTs: Subgroup and Interim Analyses

Office of AIDS Research, NIH
ICSSC, FHI
Goa, India,
September 2009

Kenneth F. Schulz, PhD
As discussed, I will address more multiplicity problems
Subgroup analyses

- Effect by age categories, for example
  - Characteristics at baseline
- Have specious appeal
- Sound logical and intuitive . . . Even fun
- Multiplicity and naivety lead to interpretational missteps
- Subgroup effects may be illusory
Discarded Subgroup Analyses (p>0.05)

Potential Subgroup Analyses
Subgroup analyses

- Test enough subgroups – probably a false-positive result by chance alone
- Worse yet, investigators might just report the statistically significant results
  - Distorts the medical literature
- In general, we discourage subgroup analyses
Inappropriate subgroup analyses a problem

- Investigators persist after many warnings
- Statistical tests within every subgroup
- Multiple subgroups and multiple outcomes – a profusion of tests
- Seeking positive subgroup effects in the absence of an overall effect may fuel the epidemic of subgroup analyses
Effect of new versus standard antibiotic on febrile morbidity in four age strata and overall

<table>
<thead>
<tr>
<th>Strata</th>
<th>Rate Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 20-24 years</td>
<td>1.4 (0.6-3.2)</td>
</tr>
<tr>
<td>Age 25-29 years</td>
<td>1.2 (0.4-3.1)</td>
</tr>
<tr>
<td>Age 30-34 years</td>
<td>0.3 (0.1-0.9)</td>
</tr>
<tr>
<td>Age 35-39 years</td>
<td>1.1 (0.6-1.4)</td>
</tr>
<tr>
<td>Total</td>
<td>0.9 (0.6-1.4)</td>
</tr>
</tbody>
</table>

Test for statistical interaction yields $p=0.103$
Effect of new versus standard antibiotic on febrile morbidity in four age strata and overall

<table>
<thead>
<tr>
<th>Strata</th>
<th>Rate Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 20-24 years</td>
<td>1.4 (0.6-3.2)</td>
</tr>
<tr>
<td>Age 25-29 years</td>
<td>1.2 (0.4-3.1)</td>
</tr>
<tr>
<td>Age 30-34 years</td>
<td>0.3 (0.1-0.9)</td>
</tr>
<tr>
<td>Age 35-39 years</td>
<td>1.1 (0.6-1.4)</td>
</tr>
<tr>
<td>Total</td>
<td>0.9 (0.6-1.4)</td>
</tr>
</tbody>
</table>

Test for statistical interaction yields p=0.103
Proper analysis of antibiotic trial

- Focus on the overall analysis
- If the subgroup analysis is relevant, use a test of statistical interaction
- Improper – stratum-specific testing
  - Might highlight the 30-34 age results
  - An example of superfluous subgroup salvage of an otherwise indeterminate trial
RCT of Aspirin and Death

- CTSU trial (Peto) of 17,187 patients
- Overall, aspirin prevented death
  - 23% reduction (15%-30%), p<0.00001
- Lancet editors wanted about 40 subgroup analyses
- Initially, Peto et al declined
  - Absurd in any case, particularly 40
Agreed to subgroup analyses under one condition

- Reluctantly agreed, as long as they could add their own absurd subgroups
- Illustrate their unreliability
- Recoded all 17, 187 patients by astrological sign
- Data-dredged by subgroup
Aspirin Does not Prevent Death in Gemini or Libra

- Born under the sign of Gemini or Libra, 9% increase, SD 13; NS
- Born under other astrological signs, strikingly beneficial, 28% reduction, SD 5; p<0.00001
- You’re sunk if you’re born in the wrong month
- Negotiated with Lancet
An Illustration of Post Hoc Lucidity

- Data dredging bolstered by post hoc lucidity
- Call from Hendricks
- On a plane sitting next to Nancy Reagan’s astrologist
- Asked who would not benefit from aspirin
- Astrologers support the results, but anyone else?
Inserted Teaching Points in the Article

- “All these subgroup analyses should, perhaps, be taken less as evidence about who benefits than as evidence that such analyses are potentially misleading.” Peto et al.
  - Two journal pages on the absurdity of subgroup analyses in the paper
Indeterminate Trials

“The answer to a randomized controlled trial that does not confirm one’s beliefs is not the conduct of several subanalyses until one can see what one believes. Rather, the answer is to re-examine one’s beliefs carefully.”

Oei et al. BMJ 1999
Concluding thoughts on conducting subgroup analyses

- Just say “No”
- If contemplated
  - Prespecify
  - Report all subgroup analyses done
  - Use tests of interaction
- If exploratory analyses, report as such and report all exploratory analyses done
Concluding thoughts on reviewing trials with subgroup analyses

- Wary of trials that report many
- Also wary of trials that report a few
  - Selective reporting . . . ‘cherry picking’
- Prespecified or hypothesis generating?
- Discount analyses with tests within subgroups
  - Authors should use tests of interaction
- Suspicious of superfluous subgroup salvages of otherwise indeterminate trials
Interim Analyses

- Focus on statistical warnings
  - Erstwhile ‘statistical stopping rules’

- Appropriate monitoring involves more...
  - Slow accrual
  - Poor data quality
  - Poor adherence
  - Resource deficiencies
  - Unacceptable adverse effects
  - Emerging information makes trial irrelevant
  - Fraud
Many challenges of the data cause multiplicity problems

Michael Keaton - Andie MacDowell

Multiplicity.
Better living through cloning.
Interim analyses done every 6 months for 5 years. The $p$ value is shown for the comparison between the treatment group and control group.
Inflation of false-positive error rate (α)

- Intuitively, many interim analyses at p<0.05 should inflate α
- Overall α level rises with the challenges
  - α = 0.08 after two challenges
  - α = 0.11 after three challenges
  - α = 0.19 after ten challenges
- Multiplicity dictates adjustment
  - Scientific credibility depends upon it
Interim stopping levels ($p$ values) for different numbers of planned interim analyses by a group sequential design: overall $\alpha = 0.05$

<table>
<thead>
<tr>
<th>No. of planned interim analyses</th>
<th>Interim analysis</th>
<th>Pocock</th>
<th>Peto</th>
<th>O’Brien-Fleming</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0.029</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>2 (final)</td>
<td>0.029</td>
<td>0.05</td>
<td>0.048</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.022</td>
<td>0.001</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.022</td>
<td>0.001</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>3 (final)</td>
<td>0.022</td>
<td>0.05</td>
<td>0.045</td>
</tr>
</tbody>
</table>
Interim stopping levels ($p$ values) for different numbers of planned interim analyses by group sequential design: overall $\alpha = 0.05$

<table>
<thead>
<tr>
<th>No. of planned interim analyses</th>
<th>Interim analysis</th>
<th>Pocock</th>
<th>Peto</th>
<th>O'Brien-Fleming</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.018</td>
<td>0.001</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.018</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.018</td>
<td>0.001</td>
<td>0.019</td>
</tr>
<tr>
<td>4 (final)</td>
<td></td>
<td>0.018</td>
<td>0.05</td>
<td>0.043</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.016</td>
<td>0.001</td>
<td>0.00001</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.016</td>
<td>0.001</td>
<td>0.0013</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.016</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.016</td>
<td>0.001</td>
<td>0.023</td>
</tr>
<tr>
<td>5 (final)</td>
<td></td>
<td>0.016</td>
<td>0.05</td>
<td>0.041</td>
</tr>
</tbody>
</table>
Early termination and bias

- If a DMC stops a trial early based on a group sequential procedure – bias
- Random fluctuations towards greater treatment effects would more probably result in early termination
  - Than random towards lesser treatment effects
- When a trial is stopped early, readers need to grasp: effects prone to exaggeration
  - i.e., a random high
Summation on interim analyses

- Challenging the data during interim analyses at $\alpha = 0.05$ inflates the false-positive error rate
- Interim analyses dictate statistical adjustment
  - Usually a group sequential approach
- Effects are prone to exaggeration when a trial is stopped early with a group sequential approach
End