Subgroup and Post-hoc Analyses: Benefits and Challenges

Lehana Thabane, PhD
Clinical Epidemiology and Biostatistics
McMaster University
Centre for Evaluation of Medicines, St Joseph’s Healthcare
March 12, 2009

Email: thabanl@mcmaster.ca
Tel: (905) 522-1155 x3720
Objectives

• What is subgroup analysis?
• What is posthoc analysis?
• Discuss reasons for doing subgroup analyses

• Benefits of subgroup analyses
• Challenges with subgroup and post-hoc analyses
• Criteria for evaluating results of subgroup analyses
Definition

• Subgroup analysis means
  – Evaluation of treatment effect in a small group of patients defined by baseline characteristics
  – Eg gender (M, F); baseline CVD risk (L, H)
Definitions: Pre-specified or Post-hoc Subgroup Analysis

- A **pre-specified subgroup analysis** is one that is planned and documented before any examination of the data - preferably in the study protocol
  - This analysis includes specification of the end point, the baseline characteristics, the statistical method used to test for an interaction

- **Post-hoc analysis** refers to those in which the hypotheses being tested are not specified before any examination of the data
  - Hypothesis is suggested after doing the analysis
Motivations for Subgroup Analysis

• To assess the robustness of the overall results across subgroups

• To get a risk-benefit profile of treatment on different subgroups
  – Race
  – Age
  – Sex

• To explain the variation in treatment effects: Pooling results from heterogeneous populations can
  – increase the risk of missing important effects (on subgroups) - leading to meaningless conclusions based on “average” effects

• To aid in drug labeling
Regulators Perspective: FDA’s position on subgroup analysis

- **Pre-specified subgroup analysis**
  - FDA allows the sponsor to specify, before a Phase III trial, one or more subgroups among the target group for which it plans to undertake subgroup analysis.
  - **Rationale:** Minimize the risk of non-approval of drugs with significant positive effects for identifiable subgroups (False-negatives).

- **Post-hoc subgroup analysis**
  - Can be employed to justify a label that warns of negative effects for specified subgroups.
  - Because the drug’s sponsor has a financial conflict of interest, however, the FDA must conduct this analysis by itself or use an outside consultant.
  - **Rationale:** Post hoc subgroup analysis increases the risk of approving drugs that have no net beneficial effect (False-positives).

- **FDA’s uses of Subgroup Analyses (Chest 2005;127:2298-2301)**
  - to investigate the robustness and confirm the overall conclusions of a trial.
  - to assess a risk-benefit ratio for populations that are most likely to benefit from a drug.
    - if drug is superior, subgroup analysis is used to determine whether there is a consistently favorable risk/benefit ratio across subgroups.
FDA’s position and rulings on subgroup analysis
Maggioni BP et al. FDA and CPMP Rulings on Subgroup Analyses. Cardiology 2007;107:97–102

• FDA’s recommended approach
  – If interactions are anticipated or are of particular a priori interest, subgroup analysis or a statistical model including interactions should be part of analysis plan
  – Caution: subgroup or interaction analyses are often merely exploratory and should be clearly identified as such in the protocol

• Do subgroup analyses affect FDA’s drug approval?
  – Market approval of a drug is based on the overall trial results
  – No drug has so far been approved or not-approved either in the US or in the EU on the basis of subgroup analysis
  – Subgroup analysis can influence the approval or can even be required
    • It can influence the labeling of the Summary Characteristics of a Product

• Not enough planned subgroup analyses
    • 48/219 (21.9%) reported at least one planned sub-group analysis.
    • 20 (9.1%) planned sex analysis,
    • 9 (4.1%) planned a race or ethnic group analysis
    • 17 (7.8%) planned an age analysis
    • 18 (8.2%) planned study site analysis
Limitations of Subgroup Analysis

• **Subgroup analyses are notoriously unreliable**
  - JAMA 1991;266(1):93-8
  - Lancet 1987;28(8531):494-7
  - J Hypertens 1992;10(1):6-8

• **Multiplicity**
  - Performing multiple subgroup analyses can increase the risk of false-positive findings

<table>
<thead>
<tr>
<th># of tests</th>
<th>False-positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0975</td>
</tr>
<tr>
<td>3</td>
<td>0.143</td>
</tr>
<tr>
<td>5</td>
<td>0.226</td>
</tr>
</tbody>
</table>
Power for Subgroup Analysis: Detecting interaction effects

  – Trial with 80% power for the overall effect had 29% power to detect an interaction effect of the same magnitude
  – For interactions of this size to be detected with the same power as the overall effect, sample sizes should be inflated 4 times

• Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. The Lancet 2005;365(9454):176 - 186
  – If important subgroup effects are anticipated,
    • trials should either be powered to detect them reliably or
    • pooled analyses of several trials should be undertaken
Example: ISIS-2 trial

  - Mortality benefits for both interventions

- **Subgroup Analyses**
  - Patients born under Zodiac signs of Gemini and Libra had 5% higher mortality on aspirin compared to placebo
  - Patients born under other Zodiac signs had a 30% lower mortality on aspirin compared to placebo

- **No plausible explanation for the observations**
  - Illustrates the pitfall of post-hoc subgroup analyses
How to conduct a subgroup analysis

• Statistical model as an interaction term between the treatment group and the baseline variable of interest

• Remember that the presence or absence of interaction is specific to the measure of the treatment effect

• Start with the test of interaction effect
Examples of Interactions

[Graphs showing interactions between gender and dose with different responses for males and females.]
Common Mistakes about subgroup analysis

- The inability to find significant interactions is not evidence that the observed treatment effect necessarily applies to all subjects
  - Most trials are not powered for interaction tests

- Claims of subgroup statistical significance on the basis of
  - separate tests of treatment effects within each subgroup
  - the observed treatment-effect sizes within each subgroup, ignoring the uncertainty of these estimates

![95% CI Plot of A, B and (B-A)](image)
Common Mistakes about subgroup analysis (cont.)

• Claims of statistical significance (or lack of) at alpha = 0.05
  – Ignoring multiplicity

• Poor reporting of subgroup analysis results
  – No actual results on test of interaction
  – Wang et al. Reporting Subgroup Analyses in clinical Trials. NEJM 2007;357(21);2189-94

• Misinterpretation of results
  – Statistical significance is not the same as clinical significance (clinical relevance)
  – **Apply Hill’s criteria:** Hill AB. Principles on medical statistics. 9th ed. London: Lancet;1971
Criteria for evaluating results of subgroup analyses
(Oxman AD, Guyatt GH. A consumer’s guide to subgroup analyses. Ann Inter Med 1992;116(1):78-84)

• Is the subgroup difference suggested by comparisons within rather than between studies?
• Was the hypothesis stated apriori (before analysis) or after doing analysis?
• Was the subgroup effect one of the small number of effects tested?
• How large is the magnitude of the effect?
• Was the effect statistically significant?
• Is the effect consistent across studies?
• If there an indirect evidence that supports the hypothesized subgroup effect?
Another Criteria


• **The rule of chance/the strength of association**
  – Was the subgroup effect one of the small number of (subgroup/interaction) effects tested?
  – How large is the magnitude of the (interaction) effect?
  – Was the (interaction) effect statistically significant?

• **The biologic gradient**
  – eg effect vary by disease severity

• **Consistency (internal/external)**
  – Is the (interaction) effect similar across outcomes (internal)?
  – Is the (interaction) effect seen in other studies (external)?

• **Confounding**
  – Is there any other baseline factor that could explain the (interaction) effect?

• **Coherence/plausibility**
  – If there an indirect evidence that supports the hypothesized subgroup effect?
Illustrative Example: OASIS-5 Renal Function Subgroup Analysis
(Ann Intern Med 2007;147:304-10)

How to interpret the results
OASIS-5 Trial: Organization to Assess Strategies in Acute Ischemic Syndromes  
(Am Heart J 2005;150:1107)

• **Objective:** To determine whether fondaparinux was non-inferior to enoxaparin in preventing the composite of death, new myocardial infarction, and refractory ischemia at 9 days (primary), at 30 days, 180 days (secondary)

• **Design:** Phase III trial

• **Population:** Patients with non-ST-segment elevation acute coronary syndromes (ACS)

• **Main Results:**
  – Major bleeding was 2-times less frequent with fondaparinux than with enoxaparin
  – Renal dysfunction increases the risk of bleeding
Example: OASIS-5 Renal Function Subgroup Analysis
(Ann Intern Med 2007;147:304-10)

• **Design:** Four subgroups based on Glomerular Filtration rate (GFR) per 1.73 m^2
  – GFR < 58 mL/min
  – 58 < GFR < 71 mL/min
  – 71 < GFR < 86 mL/min
  – > 86 mL/min

• **Hypothesis**
  – The researchers hypothesized that the benefit of fondaparinux versus enoxaparin for patients with non-STEMI would be greatest among patients whose risk of bleeding was greater because of renal dysfunction

• **Implication of the study:**
  – In non-STEMI the net benefits of fondaparinux compared with enoxaparin are most marked among patients with GFR less than 58 ml/min
1. The rule of chance/the strength of association

- **Was the subgroup effect one of the small number of (interaction) effects tested?**
  - It was not stated apriori (ie not pre-specified in the protocol)

- **How large is the magnitude of the (interaction) effect ?**

- **Was the (interaction) effect statistically significant?**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Timing of Assessment</th>
<th>p-value for interaction test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death, MI, RI</td>
<td>9 days</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>30 days</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>180 days</td>
<td>0.85</td>
</tr>
<tr>
<td>Major bleeding</td>
<td>9 days</td>
<td><strong>0.056</strong>*</td>
</tr>
<tr>
<td></td>
<td>30 days</td>
<td><strong>0.093</strong>*</td>
</tr>
<tr>
<td></td>
<td>180 days</td>
<td>0.301</td>
</tr>
</tbody>
</table>

- p-values for death, MI, RI show **no strength** of association
- p-values for major bleeding show **some strength** of association (*)
- Remember this is just explanatory study, not powered for interaction tests
2 Biological Gradient

Does effect vary by severity of renal dysfunction?

- Benefit of Fondaparinux vs enoxaparin on risk of major bleeding decreases with decrease in GFR at 9 days
  - A similar pattern holds for 30 and 180 days, showing some biological gradient
  - A similar relationship is seen for death at 180 days

<table>
<thead>
<tr>
<th>The major bleeding on day 9</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>HR</td>
<td>95% CI</td>
</tr>
<tr>
<td>GFR &lt; 58</td>
<td>0.42</td>
<td>(0.32, 0.56)</td>
</tr>
<tr>
<td>58 &lt; GFR &lt; 71</td>
<td>0.53</td>
<td>(0.39, 0.72)</td>
</tr>
<tr>
<td>71 &lt; GFR &lt; 86</td>
<td>0.66</td>
<td>(0.46, 0.95)</td>
</tr>
<tr>
<td>GFR &gt; 86</td>
<td>0.61</td>
<td>(0.41, 0.90)</td>
</tr>
</tbody>
</table>
3 Consistency

- **Internal Consistency: Is effect similar across outcomes?**
  - The main trial shows benefit of fondaparinux vs enoxaparin on major bleeding at 9 days
  - A similar results holds between subgroups of GFR at 9, 30 and 180 days

- **External consistency: Is effect similar between studies?**
  - Similar studies suggest an association between bleeding response, use of heparin and renal dysfunction
4 Confounding

• Is there any other baseline factor that could explain the effect?
  – Key baseline characteristics of the four GFR groups seem comparable with the exception of
    • Age
    • Management with ACE inhibitors or AR blockers
    • Hospital procedure with PCI (percutaneous coronary intervention)
  – No information about the role of dose of UFH
  – Analysis did not adjust for potential confounding through multivariable analysis
    • Though minimal, confounding cannot be completely ruled out
5 Coherence/Plausibility

• If there an indirect evidence that supports the hypothesized subgroup effect?
  – Enoxaparin is associated with 2-3 times increase in bleeding when creatinine clearance is less than 0.5 mL/s (<30 mL/min per 1.73 m^2)
  – PK studies show that creatinine clearance is linked to lower bleeding because of how drugs are eliminated - through renal secretion pathways or metabolism
  – Pharmacologic properties, bioavailability of fondaparinux suggest an advantage in reducing bleeding
  – Fondaparinux is prolonged in patients with renal insufficiency
    • Major route of elimination is urinary secretion of unchanged drug
    • Note: Fondaparinux is secreted by the kidneys without previous metabolism
Criteria for evaluating results of subgroup analyses
(Oxman AD, Guyatt GH. A consumer’s guide to subgroup analyses. Ann Inter Med 1992;116(1):78-84)

• Is the subgroup difference suggested by comparisons within rather than between studies?
  – Within study

• Was the hypothesis stated apriori (before analysis) or after doing analysis?
  – Not stated apriori – adhoc
  – Abstract provides a warning about interpretation

• Was the subgroup effect one of the small number of effects tested?
  – Several outcomes – no adjustment for multiplicity
  – Not clear whether additional subgroup analyses are planned

• How large is the magnitude of the effect?
  – Substantial on major bleeding and consistent across outcomes and time points

• Was the effect statistically significant?
  – Statistically significant interaction effect on major bleeding

• Is the effect consistent across studies?
  – Yes, supported by an earlier meta-analysis

• If there an indirect evidence that supports the hypothesized subgroup effect?
  – Yes, based on PK studies
Key Messages

• It’s best to treat all subgroup analyses with some degree of skepticism
  – Clear apriori hypothesis is necessary
  – Check if subgroup was prespecified or post-hoc
  – Subgroup analyses are prone to suffer multiplicity
  – They should at best be considered exploratory or hypothesis-generating
    • Unless appropriately powered

• Report all subgroup analyses (identify if pre-specified or post-hoc)

• Subgroup analyses should be based on tests of interaction

• They should not be based on tests of significance in each of the subgroups

• Subgroup analyses are more plausible if the overall result is statistically significant
  – They explore consistency of effects across subgroups

• Subgroup analyses are less plausible if the overall result is not statistically significant

• Always exercise caution in interpreting subgroup analyses
Key References