Hierarchical testing in a group sequential design with different information times

Dong Xi, Novartis
Jiangtao Gou, Hunter College of CUNY
ASA NJ Chapter
June 14, 2019
Agenda

• Background on hierarchical testing and group sequential design
• Refined boundary with different information times
• Clinical trial application
• Other types of hierarchical testing in group sequential design
• Conclusion
Hierarchical testing

• In confirmatory trials, hierarchical testing is commonly used to control the familywise error rate at level $\alpha$

• For many oncology trials
  – Primary endpoint (PE): PFS
  – Secondary endpoint (SE): OS
  – Or the other way round (PE: OS and SE: PFS)

• First test PE at level $\alpha$

• If significant, test SE at level $\alpha$; if not significant, stop testing
Group sequential design

• Allow interim monitoring along the course of a trial
• Possible early stopping due to overwhelming benefit
  – Test the hypothesis with 50% and 100% of the planned number of patients/events
  – At any analysis, if the hypothesis is rejected, claim success
• Potentially save time to make efficacious treatment available to patients
• Due to repeated testing of the same hypothesis with accumulating data, the test has to be adjusted for Type I error control
Group sequential design

- Consider a group sequential design for testing $H_0: \theta \leq 0$ against $H_a: \theta > 0$ with an interim and a final analysis
  - E.g., $\theta = -\log HR$
- Interim analysis is planned at information time $t$ ($0 < t < 1$)
- $t$: information fraction at the interim analysis
  \[
  t = \frac{\text{information at the interim analysis}}{\text{information at the final analysis}}
  \]
- Information: inverse of the variance of the estimates
  - Normal endpoint: proportional to the sample size
  - Survival endpoint: proportional to the number of events
- E.g., a group sequential design tests the null hypothesis twice: at 50% sample size or number of events and at 100%
Group sequential design

• $Z_1$ and $Z_2$ are test statistics at the interim and final analyses, respectively
  – E.g, Log-rank test statistics with $t$ information and 100% information

• Under $H_0$, $Z_1$ and $Z_2$ follow a bivariate normal distribution with mean 0, variance 1, and correlation $\sqrt{t}$

• $H_0$ is rejected if $Z_1 \geq c_1$ or $Z_2 \geq c_2$

• We need to find boundary $c_1$ and $c_2$ such that
  \[ P(Z_1 \geq c_1 \text{ or } Z_2 \geq c_2) = 1 - P(Z_1 < c_1, Z_2 < c_2) = \alpha \]
Two group sequential designs

• O’Brien-Fleming boundary: $c_1 = c_2 / \sqrt{t}$
  $\alpha = 0.025$, $t = 0.5$, $c_1 = 2.797$, $c_2 = 1.977$ compared with $z_{1-\alpha} = 1.96$

• Pocock boundary: $c_1 = c_2$
  $\alpha = 0.025$, $t = 0.5$, $c_1 = c_2 = 2.178$ compared with $z_{1-\alpha} = 1.96$
Clinical trial example

• CheckMate 025 is a Phase 3 trial comparing nivolumab against everolimus in patients with renal-cell carcinoma (Motzer et al. 2015)
  – PE: OS
  – SE: PFS

• An interim analysis is scheduled at $t_p = 0.7$ for PE using O’Brien-Fleming boundary

At what level should we test PE and SE at interim and final?
For PE, use $\alpha$-level group sequential boundary

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Boundary (information)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PE using O’Brien-Fleming</td>
</tr>
<tr>
<td>Interim</td>
<td>2.4 ($t_p = 0.7$)</td>
</tr>
<tr>
<td>Final</td>
<td>2.008</td>
</tr>
</tbody>
</table>

- At interim with 70% PE events, reject $H_{p0}$ if $Z_{p1} \geq 2.4$
- At final, reject $H_{p0}$ if $Z_{p2} \geq 2.008$
- Since PE can only be rejected at either interim or final, there is only one chance to test SE
  - Can SE be tested at level $\alpha$ whenever PE is significant?
Literature review

• Although SE is tested only once, testing it at level $\alpha$ will inflate Type I error (Hung, Wang, O’Neil, 1997)

• Type I error inflation depends on $\rho$, the correlation between PE and SE (Tamhane, Mehta, Liu, 2010; Glimm, Maurer, Bretz, 2010)

• Test PE using the O’Brien-Fleming boundary at level $\alpha$
  – $c_1 = 2.4$ and $c_2 = 2.008$

• Type I error
  – Rejecting SE when PE is false but SE is true
  – $P(Z_{p1} \geq c_1, Z_{s1} \geq d_1) + P(Z_{p1} < c_1, Z_{p2} \geq c_2, Z_{s2} \geq d_2)$

  | Reject SE at interim | Reject SE at final |
Type I error inflation for SE tested at level \( \alpha = 0.025 \) \((d = 1.96)\) whenever PE is significant

- When \( \rho = 0 \), Type I error is controlled
- When \( \rho > 0 \), the maximum inflation increases with \( \rho \)
Solution
(Tamhane, Mehta, Liu, 2010; Glimm, Maurer, Bretz, 2010)

• Group sequential design has to be used for SE

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Boundary (information)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PE using O’Brien-Fleming</td>
</tr>
<tr>
<td>Interim</td>
<td>2.4 ((t_p = 0.7))</td>
</tr>
<tr>
<td>Final</td>
<td>2.008</td>
</tr>
</tbody>
</table>

• At interim with 70% PE events, reject \(H_{p0}\) if \(Z_{p1} \geq 2.4\)
  – If \(H_{p0}\) rejected, reject \(H_{s0}\) if \(Z_{s1} \geq 2.4\)

• If \(H_{p0}\) not rejected at interim, reject \(H_{p0}\) at final if \(Z_{p2} \geq 2.008\)
  – If \(H_{p0}\) rejected, reject \(H_{s0}\) if \(Z_{s2} \geq 2.008\)
Type I error control for SE tested at level $\alpha = 0.025$ using O’Brien-Fleming boundary

- O’Brien-Fleming boundary for PE and SE: (2.4, 2.008)
- When $\rho = 1$, Type I error achieves 0.025 under $H_{p0}$
- Usually don’t know the truth about $\rho$. Be conservative with $\rho = 1$
Group sequential design for SE can be different from the group sequential design for PE

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Boundary (information)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PE using O’Brien-Fleming</td>
</tr>
<tr>
<td>Interim</td>
<td>2.4 ($t_p = 0.7$)</td>
</tr>
<tr>
<td>Final</td>
<td>2.008</td>
</tr>
</tbody>
</table>

- O’Brien-Fleming for PE and Pocock for SE may have power advantages (Glimm, Maurer, Bretz, 2010)
SE may have a different information time from PE

• As a result of group sequential design for SE, we need to pre-specify the information time for SE at the interim
  – Information time for SE is random at the interim, depending on PE
  – But we need to give a best guess; otherwise, it would be difficult to justify any post-hoc boundary for SE after PE is significant

• SE has the same critical values of PE only if both
  – the same type of boundary is used and
  – $t_p = t_s$

• What if the information time for SE is different from the information time for PE?
Question of interest

- Past results in the literature assume that the information time is the same for PE and SE at the interim.
- However, this is unlikely to be the case for trials with time to event endpoints:
  - PFS takes less time to accumulate than OS
  - At the interim, PFS may have a larger information time.
- For a trial including non-inferiority and superiority objectives, the analysis sets may be different.
- Do we have to use a group sequential design for SE in order to control Type I error?
  - It depends on the difference of information times between PE and SE.
Type I error when information time is different

- At interim, assume $t_p = 0.5$ but $t_s = 0.2, 0.4, 0.5, 0.6, 0.8$
- PE and SE are tested using the O’Brien-Fleming boundary

- When $t_s = 0.5$, the maximum Type I error is 0.025
- When $t_s \neq 0.5$, the maximum Type I error is $< 0.025$
Why

• When the information time is different for PE and SE, the correlation structure changes

• At interim,

\[ \text{Corr}(Z_{p1}, Z_{s1}) = \rho \frac{\sqrt{\min(t_p, t_s)}}{\sqrt{\max(t_p, t_s)}} \]

• The correlation depends on how much \( t_p \) and \( t_s \) overlap
  – In the normal setting, assume \( t_p \) patients have PE measurements and \( t_s \) patients have SE measurements
  – The correlation is generated from the \( \min(t_p, t_s) \) patients who have both measurements
Refined boundary

• When the information time is different for PE and SE, we can refine the group sequential boundary for SE
  – Refined boundary: uniformly less conservative boundary

• Idea for refinement: lower the usual $\alpha$-level boundary for SE until the actual Type I error is exactly $\alpha$

• Select $\alpha$-level boundary for PE $(c_1, c_2)$

• Solve for the boundary for SE $(d_1, d_2)$ such that the Type I error is controlled exactly at level $\alpha$

$$P(Z_{p1} \geq c_1, Z_{s1} \geq d_1) + P(Z_{p1} < c_1, Z_{p2} \geq c_2, Z_{s2} \geq d_2) = \alpha$$

Reject SE at interim

Reject SE at final
Refined boundary examples

• The more different $t_p$ and $t_s$ are, the more refinement achieved (or the less conservative the boundary is)

• When $t_p = 0.5$ and $t_s = 0.2$, SE can be tested at level $\alpha$ at interim or final, whenever PE is significant
  – No group sequential adjustment is needed for SE

• When $t_p = 0.5$ and $t_s = 0.8$
  – Usual $\alpha$-level O’Brien-Fleming boundary is (2.260, 2.021)
  – Refined boundary is (2.193, 1.962)
Power gain (refined vs. usual $\alpha$-level)

- When $t_p = 0.5$ and $t_s = 0.2$, the power gain is $\sim 9\%$ if Pocock boundary for SE and very little if O'Brien-Fleming boundary for SE
- When $t_p = 0.5$ and $t_s = 0.8$, the power gain is $\sim 4\%$ if Pocock boundary for SE and $\sim 2\%$ if O'Brien-Fleming boundary for SE
Application to CheckMate 025
PE: OS and SE: PFS

• An interim analysis is scheduled at $t_p = 0.7$ for PE using O’Brien-Fleming Lan-DeMets boundary
  – Assume that $t_s = 0.8$ for SE at interim

• $\alpha$-level boundary using the spending function
  – PE: (2.4, 2.008) and SE: (2.251, 2.025)

• Solve the following equations simultaneously for $(d_1, d_2)$

\[
P(Z_{s1} \geq d_1) = \varepsilon(y, t_s = 0.8)
\]

\[
P(Z_{s1} < d_1, Z_{s2} \geq d_2) = y - \varepsilon(y, t_s = 0.8)
\]

\[
P(Z_{p1} \geq 2.438, Z_{s1} \geq d_1) + P(Z_{p1} < 2.438, Z_{p2} \geq 2, Z_{s2} \geq d_2) = 0.025
\]

• Refined boundary for SE: (2.192, 1.975)
Other types of hierarchical testing in group sequential design

Previous:
Stagewise hierarchical

Overall hierarchical

Glimm, Maurer, Bretz, 2010
Overall hierarchical

• When PE is significant at interim, test SE at its interim and final analyses, if not significant earlier
  – PE: PFS
  – SE: OS

• If the true effect on PE is very positive, PE is almost always significant

• SE is always tested at interim and final analyses, if not significant earlier

• $\alpha$-level group sequential design is required for SE to control Type I error at level $\alpha$
Other types of hierarchical testing in group sequential design

Overall hierarchical

Partial hierarchical

Glimm, Maurer, Bretz, 2010
Partial hierarchical

• When PE is not significant at interim, test both PE and SE at final analysis simultaneously
• If PE: PFS is not significant at interim, the clinical team may want to preserve a small chance to reject SE: OS, even if PFS is not significant at all
• Hierarchical at interim but not at final
• May need to split $\alpha$ between PE and SE
Partial hierarchical example
Analysis 1: PFS final and OS interim
_analysis 2: OS final

• PFS is only tested once (i.e., no PFS interim analysis)

• OS can be tested at
  – OS interim, only if PFS significant
  – OS final, regardless of PFS

• If PFS is tested at level $\alpha$, then OS can only be tested at OS interim at level $\alpha$, if PFS significant

• Any possibility to reject OS at the final analysis will inflate Type I error

• If PFS is tested at level $\alpha/2$, then OS can be tested at OS interim and final

• Refined boundary for SE is $(2.129, 2.237)$ when $t_s = 0.5$

• Less conservative than testing OS at level $\alpha/2$
  • O’Brien-Fleming $(3.183, 2.251)$
  • Pocock $(2.450, 2.450)$
Conclusions

- Different strategies to design hierarchical testing in group sequential design

- In stagewise hierarchical testing (PE: OS, SE: PFS), refinement with less conservative boundary for SE is possible when the information time is different from PE

- In overall hierarchical testing (PE: PFS, SE: OS), refinement is not needed for two stage testing
  - For more than two stages, refinement is possible (Tamhane et al., 2018)

- In partial hierarchical testing (PE:PFS, SE: OS), refinement is also possible for SE
Reference


• Tamhane, A. C., Mehta, C. R. & Liu, L. (2010), Testing a primary and a secondary endpoint in a group sequential design, Biometrics 66(4), 1174-1184

• Glimm, E., Maurer, W. & Bretz, F. (2010), Hierarchical testing of multiple endpoints in group-sequential trials, Statistics in Medicine 29(2), 219-228


• Gou, J. & Xi, D. Hierarchical testing of a primary and a secondary endpoint in a group sequential design with different information times, Statistics in Biopharmaceutical Research, https://doi.org/10.1080/19466315.2018.1546613

• R package: gsrsb
Thank you

dong.xi@novartis.com