On Adaptive Extensions of Group Sequential Trials for Clinical Investigations

Qing Liu, Ph.D.¹
qliu2@prdus.jnj.com
and
Keaven M. Anderson, Ph.D.²
keaven_anderson@merck.com

¹Johnson and Johnson Pharmaceutical Research and Development, LLC

²Merck Research Laboratories

February 22, 2008
Presentation History

- ENAR, March, 2008
- Centocor, Oncology Biostatistics Journal Club, November, 2007
- Merck Research Laboratories, Statistics Study Group, January 14, 2007
- Columbia University Biostatistics Colloquium, September 26, 2006
Outline

- Introduction and Background
- The Design Problem
- Classical Group Sequential (GS) Designs
- Limitations of Classical GS Designs
- Extended GS Designs
- Ordering of the Sample Space
- Sequential Inference and Monitoring
- Illustrative Examples
- Discussion
- References
Associated Manuscripts and Software

- Under revision following initial journal review
  - Qing Liu and Keaven M. Anderson, *Theory of Inference for Adaptively Extended Group Sequential Designs with Applications for Clinical Trials*

- gsDesign R package
  - All graphics for this presentation done with the R package gsDesign
  - Preliminary gsDesign package done as summer intern project in 2006 with Jennifer Sun and John Zhang
  - Version 1.1 now available with 30+ page manual and substantial online help
  - Possible alternative to EAST when you want flexibility or features not provided there (also free!)
  - Send me an e-mail if you are interested (comments and work on extensions welcome...)
Example 1: Fracture Prevention Study

- Women over age fifty are randomized to placebo or a treatment intended to prevent bone fracture.
- Randomization and follow-up proceed and any suspected events are adjudicated.
- Interim analyses are planned.
- If a group sequential boundary is crossed at an interim analysis, additional patient events will have occurred at the time of analysis that have not been both collected and adjudicated.
Example 1: Fracture Prevention Study

- Women over age fifty are randomized to placebo or a treatment intended to prevent bone fracture
- Randomization and follow-up proceed and any suspected events are adjudicated
- Interim analyses are planned
- If a group sequential boundary is crossed at an interim analysis, additional patient events will have occurred at the time of analysis that have not been both collected and adjudicated
- How do you do a combined analysis of the interim data that were positive plus the overrun data?
Example 2: Accelerated Approval for Oncology

- **Background**
  - An oncology drug may be approved on a conditional basis if progression-free survival is extended.
  - A definitive demonstration of a survival benefit may be required for full approval.
Example 2: Accelerated Approval for Oncology

- Background
  - An oncology drug may be approved on a conditional basis if progression-free survival is extended
  - A definitive demonstration of a survival benefit may be required for full approval

- Possible trial setup
  - Several interim analyses are planned
  - At each analysis, both survival and progression-free survival (PFS) are analyzed
Example 2: Accelerated Approval for Oncology

- **Background**
  - An oncology drug may be approved on a conditional basis if progression-free survival is extended.
  - A definitive demonstration of a survival benefit may be required for full approval.

- **Possible trial setup**
  - Several interim analyses are planned.
  - At each analysis, both survival and progression-free survival (PFS) are analyzed.

- A benefit for PFS likely to be demonstrated BEFORE a survival benefit can be demonstrated, raising two issues:
  - How do you analyze the analysis of survival so that claims of efficacy and p-values can be presented?
  - How do you incorporate the data on PFS collected *after* you have already demonstrated a benefit for this endpoint?
Background

Canner (1983)

"Decision-making in clinical trials is complicated and often protracted...no single statistical decision rule or procedure can take the place of well-reasoned consideration of all aspects of the data by a group of concerned, competent and experienced persons with a wide range of scientific backgrounds and points of view."
"Phase 3 studies ... are intended to gather the additional information about effectiveness and safety that is needed to evaluate the overall benefit-risk relationship of the drug ..."
"Phase 3 studies ... are intended to gather the additional information about effectiveness and safety that is needed to evaluate the overall benefit-risk relationship of the drug ..."

- It is ultimately a favorable benefit-risk profile of the medical product for patients that will lead to
  - Marketing approval
  - Positive public health impact
  - Commercial success for the developer
"Phase 3 studies ... are intended to gather the additional information about effectiveness and safety that is needed to evaluate the overall benefit-risk relationship of the drug ...”

- It is ultimately a favorable benefit-risk profile of the medical product for patients that will lead to
  - Marketing approval
  - Positive public health impact
  - Commercial success for the developer

- It is not unusual for a Data Monitoring Committee to recommend extending a trial after a significance boundary for the primary endpoint has been crossed in order to collect more data on secondary or safety endpoints.
Example 3: Nosocomial Pneumonia

- Standard therapy has 50% cure rate at day 14, 30% mortality at day 30
- Trial to randomize between standard and experimental therapy
- Cure at day 14 is expected to be substantially increased by experimental therapy
- Mortality is likely to affected to a lesser extent
- There may be significant side-effects with experimental therapy
Example 3: Nosocomial Pneumonia

- Standard therapy has 50% cure rate at day 14, 30% mortality at day 30
- Trial to randomize between standard and experimental therapy
- Cure at day 14 is expected to be substantially increased by experimental therapy
- Mortality is likely to be affected to a lesser extent
- There may be significant side-effects with experimental therapy
- Question: if an interim analysis shows a positive effect for the 14-day cure rate, should the trial stop?
The Design Problem

- **Hypothesis Testing**
  - $H$: $\Delta \leq 0$ against $A$: $\Delta > 0$
  - $\alpha = 0.025$ and $\beta = 0.1$ at $\Delta = \delta$
  - Upper bounds stop the trial early to declare efficacy
  - Lower bounds stop the trial early for futility

- **Applications**
  - Life threatening disease, e.g. cancer, cardiovascular disease, etc.
  - Slow enrollment with quick endpoints or time-to-event endpoints
Classical Group Sequential Designs

Asymmetric, 2-sided Group Sequential Design

![Graph showing sample size ratio relative to fixed design](image)

- Normal critical value
- Continue
- Reject H0
- Reject H1

Sample size ratio relative to fixed design

---

February 22, 2008
Liu and Anderson: Adaptive Extensions of Group Sequential Trials
Classical Group Sequential Design Setup

- $K - 1$ interim analyses and a final analysis
- Calculate $a_k < b_k$ for $k = 1, \cdots, K - 1$, $a_K = b_K$ and sample size $n_k$ for $k = 1, \cdots, K$ such that
  \[
  \alpha = \sum_{k=1}^{K} P_0 \{ Z_k \geq b_k \} \bigcap_{j=1}^{k-1} \{ a_j < Z_j < b_j \} \tag{1}
  \]
  and
  \[
  \beta = \sum_{k=1}^{K} P_\delta \{ Z_k \leq a_k \} \bigcap_{j=1}^{k-1} \{ a_j < Z_j < b_j \} \tag{2}
  \]
  where $Z_k$ are cumulative test statistics for $k = 1, \cdots, K$
Classical Group Sequential Inference and Monitoring

- At the $k$th interim analysis
  - stop to reject $H$ if $Z_k \geq b_k$,
  - stop for futility if $Z_k \leq a_k$, and
  - continue if $a_k < Z_k < b_k$
  - reject $H$ at the final analysis if $Z_K \geq b_K$

- Final Inference and Monitoring
  - Stage-wise ordering of the sample space, consisting of the stopping time and value of the test statistic (Armitage, 1957)
  - P-values and confidence intervals (Tsiatis, Rosner and Mehta, 1984)
  - Unbiased estimators (Emerson and Fleming, 1990)
  - Repeated confidence intervals (RCI) for trial monitoring (Jennison and Turnbull, 1989)
The futility boundary $a_k$ for $k = 1, \cdots, K$ may not be followed, rather it is used as a guideline

- inflation of the type I error rate
- FDA no longer accepts the significance level given by (1)

Violation of the Intent-to-Treat (ITT) principle for not being able to incorporate data beyond the interim analysis when a boundary is crossed (i.e., over-running)

- natural over-running due to additional patient enrollment as a result of delayed observations of the clinical outcomes
- adaptive extensions to address co-primary endpoints, multiple treatment comparisons, secondary efficacy endpoints or safety issues
Limitations of Classical Group Sequential Design (2)

- **Stage-wise ordering is not suitable for evidentiary evaluation of sequential data**
  - when the test statistic just reaches the boundary, the null hypothesis should be rejected at the $\alpha$-level, not at a smaller significance level according to the stage-wise ordering
  - Stagewise ordering does not provide monitoring analysis (p-values, estimates, confidence intervals)
  - Stagewise ordering does not provide final analysis when there is natural over-running or the trial is otherwise extended

- **Repeated confidence interval issues**
  - Conservative
  - Do not ensure that late results maintain conclusions
An ordering is defined for all sample paths \( \{ \tau; Z_1, Z_2, \ldots, Z_\tau \} \), where \( \tau \) is a stopping time determined by the totality of the accumulating data.
Proposed Principles of Group Sequential Inference

I An ordering is defined for all sample paths \( \{\tau; Z_1, Z_2, ..., Z_\tau\} \), where \( \tau \) is a stopping time determined by the totality of the accumulating data.

II A null hypothesis is rejected by an Extended Group Sequential test if and only if its significance boundary is crossed at or before a stopping time, or an overall p-value is less than or equal to the significance level \( \alpha \).
Proposed Principles of Group Sequential Inference

I An ordering is defined for all sample paths \( \{\tau; Z_1, Z_2, ..., Z_\tau \} \), where \( \tau \) is a stopping time determined by the totality of the accumulating data.

II A null hypothesis is rejected by an Extended Group Sequential test if and only if its significance boundary is crossed at or before a stopping time, or an overall p-value is less than or equal to the significance level \( \alpha \).

III For any \( \mu \in (0, 1) \), the group sequential design corresponding to the p-value \( p_\tau \leq \mu \) is consistent with the underlying ordering of the sample space.
Boundaries

Calculate \( a_k < b_k \) for \( k = 1, \cdots, K - 1 \), \( a_K = b_K \) and the sample size \( n_k \) for \( k = 1, \cdots, K \) such that

\[
\alpha = \sum_{k=1}^{K} P_0\{\{Z_k \geq b_k\} \cap \bigcap_{j=1}^{k-1} \{Z_j < b_j\}\}
\]

(3)

and

\[
\beta = \sum_{k=1}^{K} P_\delta\{\{Z_k \leq a_k\} \cap \bigcap_{j=1}^{k-1} \{a_j < Z_j < b_j\}\}
\]

(4)

where \( Z_k \) are cumulative test statistics for \( k = 1, \cdots, K \)
Type I Error Difference

- Classical: compute the probability of crossing the upper bound \textit{before} the lower bound is crossed
- Extended: compute the probability of \textit{ever} crossing the upper bound even if the trial is \textit{never} stopped
  - Extended group sequential design allows Type I error to be computed regardless of when a trial is stopped
Extended GS Designs: O’Brien-Fleming

1-sided O’Brien-Fleming Bounds by alpha-Level

Should p-values be the same along each line within a boundary family?
Types of Interim Decisions

- Decisions are made at an interim
  - to terminate the trial at a future specified interim analysis
  - to adjust sample size for the remaining stages
  - to continue the trial per decisions made previously,
  - to proceed the trial as originally planned by the protocol

- No need to specify how these decisions are reached but guidelines that incorporate all aspects of data are useful

**Main Theorem** Assuming that $Z_1, \cdots, Z_K$ satisfy (3), then for any stopping rule $\tau$,

$$P_0\{Z_1 \geq b_1, \cdots, Z_\tau \geq b_\tau\} \leq \alpha$$

**Bottom line:** With no lower bound you can stop at any time and maintain the ability to perform inference.
Family of Well-ordered GS Tests

- a GS test for each $\alpha \in (0, 1)$
- boundaries are well-ordered, i.e., if $\alpha' < \alpha''$ then for $k = 1, \ldots, K$
  \[ b_k(\alpha') > b_k(\alpha'') \]

Example: Wang-Tsiatis Tests

- $b_k(\alpha) = B(\alpha)(k/K)^{\rho-1/2}$
- $B(\alpha)$ is decreasing in $\alpha$
- Pocock test ($\rho = 1/2$) and O’Brien-Fleming test ($\rho = 0$)
Ordering of Sample Space: Pocock

1-sided Pocock Bounds by alpha-Level

Z

.0005
.005
.01
.025
.1
Ordering of Sample Space: Atypical

1-sided "Atypical" Spending Function by alpha-Level

Future research topic: What ordering provides "equal credibility"?
Ordering of Sample Space: KD

1–sided Kim–DeMets(4) Bounds by alpha–Level

Z

Analysis
Ordering of Sample Space

Sample Paths

\[ \omega = \{ \tau; Z_1, \ldots, Z_\tau \} \]

Smallest Significance Level

For any \( \omega \), let

\[ \hat{\mu}^{(k)} = \sup\{ \mu : Z_k \leq b_k(\mu) \} \]

for \( k = 1, \cdots, \tau \). Define

\[ p_\tau = \min\{ \hat{\mu}^{(k)} : k = 1, \cdots, \tau \}, \]

Then \( Z_k \leq b_k(p_\tau) \) for all \( k = 1, \cdots, \tau \) and \( Z_k = b_k(p_\tau) \) for at least one \( k \) in \( 1, \cdots, \tau \).

Ordering of Sample Paths

- \( \omega' \preceq \omega'' \) if and only if \( p_{\tau}' \leq p_{\tau}'' \)
- \( \omega' \preceq \omega'' \) and \( \omega'' \preceq \omega''' \) implies that \( \omega' \preceq \omega''' \)
Sequential \( p \)-values

\[ p_k = \min_{1 \leq i \leq k} \{ \hat{\mu}^{(i)} \} \]

for \( k = 1, \cdots, \tau \) are sequential \( p \)-values. In particular, \( p_\tau \) is the final \( p \)-value

**Theorem 2**

i) \( p_k \leq \alpha \) is equivalent to \( \bigcup_{i=1}^{k} \{ Z_i \geq b_i \} \)

ii) \( P_0 \{ p_k \leq \alpha \} \leq \alpha \)

iii) \( p_1 \geq p_2 \geq \cdots \geq p_\tau \)

iv) \( P_0 \{ p_\tau \leq \alpha \} \leq \alpha \)
Consider testing against $H_\delta: \Delta \leq \delta$ in favor of $A_\delta: \Delta > \delta$. Assume

- $E(Z_k) = I_k^{1/2} \Delta$ for $k = 1, \cdots, K$
- $p_k(\delta)$ for $k = 1, \cdots, \tau$ are the corresponding sequential $p$-values

Inverting the sequential $p$-values leads to sequential confidence lower bounds

$$\hat{\Delta}_k^L = \max\{Z_i/I_i^{1/2} - b_i/I_i^{1/2}: i = 1, \cdots, k\}$$

for $k = 1, \cdots, \tau$

Similarly, the sequential confidence upper bounds are given by

$$\hat{\Delta}_k^U = \min\{Z_i/I_i^{1/2} + b_i/I_i^{1/2}: i = 1, \cdots, k\}$$

for $k = 1, \cdots, \tau$
Sequential Inference

Theorem 3

i) \( \hat{\Delta}_k^L \geq 0 \) is equivalent to \( \bigcup_{i=1}^k \{ Z_i \geq b_i \} \)

ii) \( P_\Delta \{ \hat{\Delta}_k^L < \Delta \} \geq 1 - \alpha \)

iii) \( \hat{\Delta}_1^L \leq \hat{\Delta}_2^L \leq \cdots \leq \hat{\Delta}_\tau^L \)

iv) \( P_\Delta \{ \hat{\Delta}_\tau^L < \Delta \} \geq 1 - \alpha \)

Connection to RCI

- The sequential CI lower bounds are maximum cumulative RCI lower bounds

- RCI fundamentally depends on ordering of the sample space by well-ordered group sequential tests

**Median Unbiased Estimates** The confidence bounds with \( \alpha = 0.5 \) can be used to construct median unbiased sequential estimates
Illustrative Example

**Nosocomial Pneumonia (NP)**

- Current cure rate is 50% with mortality rate exceeding 30%
- New antibiotic for NP to improve cure rate by 10% (primary objective), and possibly 10% improvement of the survival rate (secondary objective)
- Arcsin transformation of proportions to apply normal approximation, with $\delta_P = 0.1424$ for the primary endpoint and $\delta_S = 0.1124$ for the secondary endpoint
- Slow enrollment and short follow-up (30 days)
- $K = 10$ analyses with $\alpha = 0.025$ and $\beta = 0.1$
**Illustrative Example 1**

Sequential $p$-values for secondary (mortality) endpoint

<table>
<thead>
<tr>
<th>$k$</th>
<th>$a_k$</th>
<th>$b_k$</th>
<th>$Z_k$</th>
<th>$p_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.998</td>
<td>4.565</td>
<td>-0.037</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>-1.169</td>
<td>3.957</td>
<td>1.697</td>
<td>1.0000</td>
</tr>
<tr>
<td>3</td>
<td>-0.584</td>
<td>3.571</td>
<td>1.593</td>
<td>1.0000</td>
</tr>
<tr>
<td>4</td>
<td>-0.099</td>
<td>3.272</td>
<td>1.679</td>
<td>1.0000</td>
</tr>
<tr>
<td>5</td>
<td>0.323</td>
<td>3.020</td>
<td>2.552</td>
<td>0.1012</td>
</tr>
<tr>
<td>6</td>
<td>0.704</td>
<td>2.796</td>
<td>2.719</td>
<td>0.0314</td>
</tr>
<tr>
<td>7</td>
<td>1.055</td>
<td>2.592</td>
<td>3.063</td>
<td>0.0061</td>
</tr>
<tr>
<td>8</td>
<td>1.380</td>
<td>2.401</td>
<td>2.917</td>
<td>0.0058</td>
</tr>
<tr>
<td>9</td>
<td>1.693</td>
<td>2.221</td>
<td>2.855</td>
<td>0.0045</td>
</tr>
<tr>
<td>10</td>
<td>2.048</td>
<td>2.048</td>
<td>3.437</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Note: primary endpoint was significant at 5th interim, but all data for primary analysis could be analyzed using EGS test through 7th interim where secondary endpoint stopped the trial.
Illustrative Example 1

P-value Isopleths and Observed Z-values

Analysis
### Illustrative Example 1

#### Sequential CI and Estimates

<table>
<thead>
<tr>
<th>$k$</th>
<th>$b_k$</th>
<th>$Z_k$</th>
<th>$\hat{\Delta}_k^L$</th>
<th>$\hat{\Delta}_k^U$</th>
<th>$\hat{\Delta}_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.565</td>
<td>-0.037</td>
<td>-0.643</td>
<td>0.546</td>
<td>-0.049</td>
</tr>
<tr>
<td>2</td>
<td>3.957</td>
<td>1.697</td>
<td>-0.208</td>
<td>0.520</td>
<td>0.156</td>
</tr>
<tr>
<td>3</td>
<td>3.571</td>
<td>1.593</td>
<td>-0.149</td>
<td>0.389</td>
<td>0.120</td>
</tr>
<tr>
<td>4</td>
<td>3.272</td>
<td>1.679</td>
<td>-0.104</td>
<td>0.323</td>
<td>0.109</td>
</tr>
<tr>
<td>5</td>
<td>3.020</td>
<td>2.552</td>
<td>-0.027</td>
<td>0.323</td>
<td>0.148</td>
</tr>
<tr>
<td>6</td>
<td>2.796</td>
<td>2.719</td>
<td>-0.004</td>
<td>0.294</td>
<td>0.145</td>
</tr>
<tr>
<td>7</td>
<td>2.592</td>
<td>3.063</td>
<td>0.0232</td>
<td>0.279</td>
<td>0.151</td>
</tr>
<tr>
<td>8</td>
<td>2.401</td>
<td>2.917</td>
<td>0.0238</td>
<td>0.246</td>
<td>0.135</td>
</tr>
<tr>
<td>9</td>
<td>2.221</td>
<td>2.855</td>
<td>0.0276</td>
<td>0.221</td>
<td>0.124</td>
</tr>
<tr>
<td>10</td>
<td>2.048</td>
<td>3.437</td>
<td>0.0574</td>
<td>0.221</td>
<td>0.142</td>
</tr>
</tbody>
</table>
Illustrative Example 1

Median Unbiased Estimate and CI by Analysis

- Estimate
- △ - Lower 95% bound
- + - Upper 95% bound
- × - True Delta
Example 2

Multiple Primary Endpoints:

- As an example of handling multiplicity, consider multiple primary endpoints or multiple treatment groups with a stepdown procedure.
- Sequential p-values for each primary endpoint can be computed at each analysis as outlined here.
- Endpoints or treatment group comparisons may become significant at different analyses.
- At each analysis, the Hochberg method can be applied since p-values never go up for a given endpoint at subsequent analyses.
- Once you have p-values, you can ignore the fact that they were generated from a group sequential design.
Discussion

- Extended GS designs are flexible for practical applications where totality of data can be incorporated to reach multiple trial objectives
- More in the paper on sample size adaptation after positive primary as well as other estimation issues.
- All inference issues are resolved
- Further developments to fulfill the needs of specific applications, e.g., multiple endpoints, multiple treatment comparisons, survival data, etc.