Adaptation of Clinical Trial Design and Inference

H.M. James Hung

Division of Biometrics I, Office of Biostatistics

OPaSS, CDER, FDA

Present in FDA/Industry Workshop, Bethesda, Maryland, September 18, 2003

Collaborators

FDA

Sue-Jane Wang

John Lawrence

Yi Tsong

Robert O'Neill

Charles Le

Industry

Lu Cui (Aventis Pharm)

Thanks are also due to Gordon Lan, Peter Bauer, Qin Liu, W.J. Shih for sharing their insights in the area of adaptive designs.

Disclaimer

The views expressed in this presentation do not represent those of the U.S. Food and Drug Administration, nor that of Aventis Pharmaceutical. Dr. Lu Cui was one of the primary investigators in our research team during his tenure with FDA.

Acknowledgments

The research was supported by FDA/CDER RSR Funds, #96-010A, #99/00-008, and #02-06.

Drug Development Program

1) Phase I - II trials (mostly small) provide limited information for planning Phase III trials
   - efficacy variables can be hard to select
   - effect sizes can be hard to postulate
   - little safety data on human
   - clinically meaningful effect maybe unavailable
2) # of Phase III trials is not fixed
   - concept of type I error rate arguably unclear per hypothesis? per trial? per indication? …..

Current Practice for Phase III trials

- control familywise type I error rate (FWE) per trial (e.g., < 5%, 2-sided) $\Rightarrow$ type I error rate per hypothesis < 5% (2-sided)
  strong control of FWE
- replicate statistical findings in at least one more independent trial

So defining 'a trial' is important in terms of controlling FWE for multiple analyses or inferences within a trial.

Ex-1. A trial following a good plan is terminated at some interim time because internal/external data strongly suggest that the plan is highly likely to fail (e.g., wrong endpoint, insufficient sample size).

1) Revise the plan and conduct a whole new trial. Only type I error for the new trial is considered (the terminated trial regarded as a pilot study).
2) Revise the plan and continue the trial. Consideration of type I error concerns both data of the terminated trial and future data.

Ex-2. Two studies are run in parallel to study a drug effect as compared to the same control on CV death/hospitalization - primary endpoint in two mostly complementary populations.
However, the primary objective of the 'program' of the two studies is all-cause mortality to be tested pooling the two studies.

What constitutes a trial?

Typical Strategies

May conduct pilot trial(s). Taking all information available and other considerations (e.g., cost), plan carefully Phase III trials (in sequence or in parallel) → aim at getting two 'positive' trials.

Discourage any interim looks unless necessary (e.g., ethical reasons) during the course of each trial.

If a trial fails, learn the lesson and may plan future trials (mostly whole new trials).

In some cases, interim data monitoring/analysis may be conducted (for ethical reasons) with an analysis plan.
Some design modification may be permitted (e.g., blinded reassessment of sample size, blinded change of methods).

Statistical analysis plan is often finalized near the trial end before unblinding data. Adjustments are often made during the course of the trial.
These generate nervousness.

Drug Development Program

Points to consider

- Phase III trials are conducted with possible modification of design specifications (e.g., sample size, analysis methods, doses, primary endpoints) based on accumulating experience / information from internal/external data.

  # of Phase III trials not fixed.

- Analysis combining data ***used to generate design modification*** and future data may compromise validity (type I error rate or bias) of commonly used tests and estimators (e.g. sample mean).

- Analysis NOT incorporating the past data that generate design change is valid and attractive because methods have good statistical properties. However, the efficiency of the entire program may be poor when lots of data are discarded from analysis for statistical inference.

- Efficiency needs to be evaluated on the basis of the entire program.

  e.g., in Ex-1, after revising the initial plan,

  continue the trial and thus combine the (pilot) data used to support design modification and future data [adaptive]

  vs.

  perform a whole new trial and thus not combine the pilot data with the future data [non-adaptive]

- Practicality and flexibility need to be a part of consideration in selection of statistical design.
- Efficiency only for efficacy assessment may be irrelevant, because evaluation of safety may require a great amount of data and comparative trials are often preferred.

Adaptation of Design/Analysis

For drug development trials, there are increasing interests in

- making better use of covariate for adjustment
- re-estimating amount of statistical information
- testing both superiority and non-inferiority in active control trials
- making change of primary endpoint
- terminating undesirable treatment arms
- making change of analysis method

What makes valid adaptation possible?

$y_{1k} \sim N(\mu_1, \sigma^2)$, $y_{2h} \sim N(\mu_2, \sigma^2)$ k, h= 1, …, n

$\Delta = (\mu_1 - \mu_2)/\sigma$ ← effect size

Test H_0: $\Delta = 0$ vs. H_1: $\Delta > 0$

with two-sample mean Z test.

Suppose that we examine data at some interim time (e.g., m subjects per group contribute data).

Information time: s = m/n

Z_s and Z_{1-s} are Z statistics applied to data before and data after the time s, respectively.

$$Z = \sqrt{s}Z_s + \sqrt{1-s}Z_{1-s}$$

$$\Pr_{H_0}(Z > C)$$

$$= \Pr_{H_0}(\sqrt{1-s}Z_{1-s} > C - \sqrt{s}Z_s)$$

$$= E_{H_0}\{\Phi(\frac{-C + \sqrt{s}Z_s}{\sqrt{1-s}})\}$$

One simple adaptation method works on the Z statistic for the data after time s, keeping the weighting factors s and (1-s) unchanged.

Replacing Z_{1-s} with any test statistic W will not change the α-level critical value C **if** **$W \mid Z_s$ follows N(0, 1) under H_0** .

Note: any 'fixed' weights (i.e., weight not depending on internal data) will maintain validity of such adaptation.

Adaptive Test U :

$$U = \sqrt{s}Z_s + \sqrt{1-s}W$$

$$\Pr_{H_0}(U > C)$$

$$= E_{H_0}\{\Phi(\frac{-C + \sqrt{s}Z_s}{\sqrt{1-s}})\}$$

$$= \Pr_{H_0}(Z > C)$$

Examples

1) Change Test Statistics

After time s, one can change Z_{1-s} to W with a distribution conditional on Z_s is N(0, 1), e.g., Lawrence (2002) for time to event analysis.

Caveat After such a change, what parameter (e.g., mean, median) to estimate for treatment effect can be unclear.

2) Interim Selection of Transformation of Covariate for Adjustment

Y (response) , X (covariate)

Best predictor from X for Y: $E(Y \mid X) = f(X; \theta) \equiv Z$

$$\overline{Y}_{aj} = \overline{Y}_j - \tilde{\beta}\,(\overline{Z}_j - \overline{\overline{Z}}), \qquad j = 1,2$$

$$Var\,(\overline{Y}_{a1} - \overline{Y}_{a2}) \approx Var\,(\overline{Y}_1 - \overline{Y}_2)[1 - \rho_{yz}^2]$$

Select f such that $Z \equiv f(X; \theta)$ maximizes the correlation between Y and Z.

Wang & Hung (2003)

At time s, model f(X; θ) w/ treatment code kept blinded. In the final analysis, use the interim-selected f and estimate θ to construct the covariate Z for adjustment.

$$\hat{\bar{Y}}_{aj} = \bar{Y}_j - \hat{\beta}\,(\hat{\bar{Z}}_j - \hat{\bar{\bar{Z}}}), \qquad j = 1,2$$

$$\hat{\bar{Y}}_{a1} - \hat{\bar{Y}}_{a2} = \bar{Y}_{a1} - \bar{Y}_{a2} + o_p(n^{-1/2})$$

$$Var\,(\hat{\bar{Y}}_{a1} - \hat{\bar{Y}}_{a2}) \approx Var\,(\bar{Y}_1 - \bar{Y}_2)[1 - \rho_{yz}^2]$$

3) Interim Sample Size Adjustment

Possibly increase total amount of statistical information (scaled to unity) from 1 to ω, at the time s, **partly** based on Z_s . Then the resulting sample mean statistic for the data after time s

$W \equiv Z_{\omega\text{-}s}$, conditional on Z_s , is N(0, 1) under H_0

Bauer & Köhne (1994) Cui, Hung, Wang (1999)
Lehmacher, Wassmer (1999) Lan (2001, 2002, 2003)
Proschan, Liu, Hunsberger (2003)

The initial sample size n is estimated based on a postulated effect size δ depending on benefit/safety/cost assessment that might not be well made before trial planning, particularly for the 1st Phase III trial (**use of conservative δ is better**).

At the time s, use accumulating data from the trial and/or external data to check if δ is reasonable. A new δ^* (judged to be worthwhile at that time) may be postulated (better NOT use only observed effect size to postulate δ^*) and re-estimate sample size.

If $\delta^* << \delta$, conduct a whole new trial. Final analysis is performed only to the data of the new trial.

If $\delta^* < \delta$, determine whether to increase sample size (this may involve subjective judgment).
If sample size is increased, adapt analysis method (how to adapt needs to be pre-specified in the protocol) if necessary.

Q: What type I error is relevant?

per the elected 'trial'?

i.e., for the new trial (Z test) only if the election is conducting a whole new trial or for the extended trial (adaptive test) only if the election is extending the existing trial?

or

associated with the mixture of Z tests for the new trial situation and adaptive test for the extended trial situation?

This simple adaptation for two-stage designs has been extended to group-sequential designs without having to modify the rejection boundaries (Cui, Hung, Wang, 1999)

This simple adaptation method can also be extended to more than one comparison, e.g.,

- change endpoint
- drop treatment arm

4) Interim Change of Endpoint

Suppose that interim data analyzed at (info) time s only to determine whether to change E1 to E2 (both are **"pre-specified"**).

Sample path: $(Z_{1s}, Z_{2s}), (Z_1, Z_2)$

$D(Z_{1s}, Z_{2s}) = 1 \Rightarrow$ Test E2 only at trial end

$= 0 \Rightarrow$ Test E1 only at trial end

Hung & Wang (2000) **Hung & O'Neill (2000)**

Test statistic (at trial end):

$$\mathbf{T = Z_2 D(Z_{1s}, Z_{2s}) + Z_1\{1- D(Z_{1s}, Z_{2s})\}}$$

Determine critical value C for T such that

$$\begin{aligned} & \Pr(T > C \mid H_0) \\ = \; & E\{ \Pr(T > C \mid Z_{1s}, Z_{2s}; H_0) \} \\ = \; & \alpha \end{aligned}$$

If sample size can increase depending on $D(Z_{1s}, Z_{2s})$, then one can adapt test statistics

$(Z_1, Z_2) \rightarrow (U_1, U_2)$ [e.g., CHW's test]

The adaptive test at trial end:

$$T^* = U_2\, D(Z_{1s}, Z_{2s}) + U_1\{1 - D(Z_{1s}, Z_{2s})\}$$

$$U_h = \sqrt{s}\, Z_{hs} + \sqrt{1-s}\, Z_{h,\omega-s} \qquad h = 1, 2$$

$$\Pr(T^* > C \mid H_0)$$
$$= \Pr(T > C \mid H_0)$$

Key to such simple valid adaptation

- adaptation is based only on past/external data, not on future data
- conditional on the past sample path, statistical inference must be valid for the future data
- how to combine the statistics of the past data and the future data does not depend on the internal sample path

Bauer & Köhne (1994, Biometrics)
Brannath, Posch & Bauer (2002, JASA)
Liu, Proschran, Pledger (2003, JASA)

Notes on Adaptive test U:

- simple implementation
  uses the initially chosen rejection boundary
- reduces to the standard fixed-sample (sequential or non-sequential) Z test if sample size is not changed
- validity (type I error rate) holds so long as 'n' modification rule does not depend on future data
  don't need a specific rule for changing n
- estimator and CI readily available and compatible with the adaptive test

- U has another look using a combination of p-values from the incremental group data

  [Lehmacher & Wassmer (1999), Brannath, Posch & Bauer (2002)]

- U may pay a big price for large sample size increase, because as such the Z statistic after the interim time s is severely down weighted (i.e., weighting with $(1\text{-}s)^{1/2}$) in U

  [Jennison & Turnbull (2002), Tsiatis & Mehta (2002)]

  But it depends on how the weighting is done and how the trial is planned.

In realistic practical applications, such simple adaptation strategy can still be competitive with fixed-sample strategy.

[Hung, Cui, Wang, Lawrence (2002), Lawrence and Hung (2003), Le and Hung (2003)]

Example

D0) Plan n to detect $\Delta=\delta$ at $\alpha=0.025$, $\beta=0.10$

Use fixed sample Z test and maintain sample size n throughout.

E1) Plan n to detect $\Delta=\delta$ at $\alpha=0.025$, $\beta=0.10$

At $t=0.5$, observe $\Delta_{0.5}$ and guess a new δ^* .

Suppose that $\delta^* = 0.6\delta$.

Determine new sample size m such that

$CP(\delta^*) = 0.90$ for the new assumed $\delta^* = 0.6\delta$

Set $n \le m \le N_{max} = 4n$

$\Delta_{0.5} \le -0.27\delta \Rightarrow$ accept H_0 and stop the trial

Use adaptive test U .

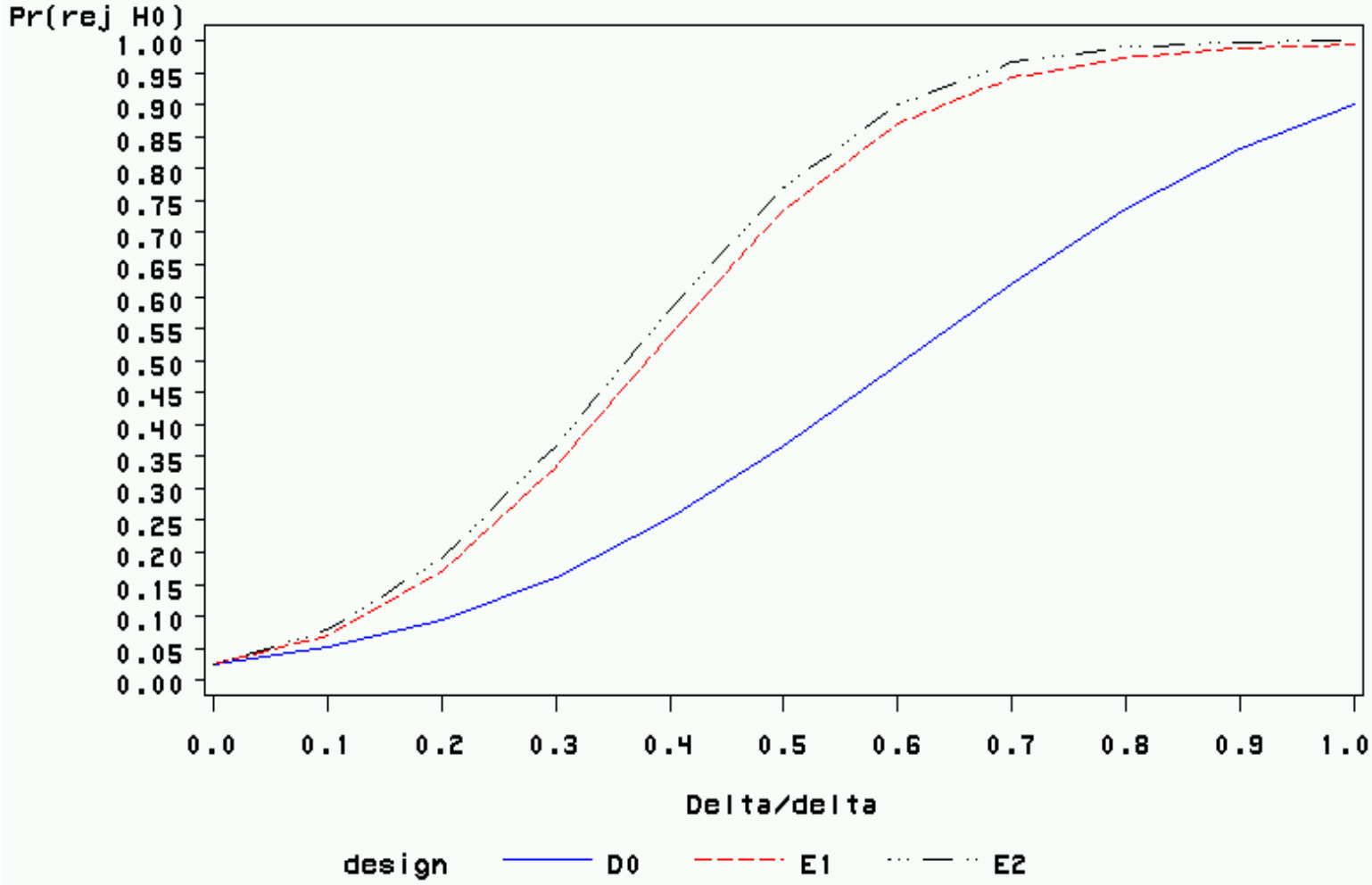
E2) Suppose that $\delta^* = 0.6\delta$ is known from the start (note: this is extremely difficult).

So plan n^* to detect $\Delta=\delta^*$ at $\alpha=0.025$, $\beta=0.10$.

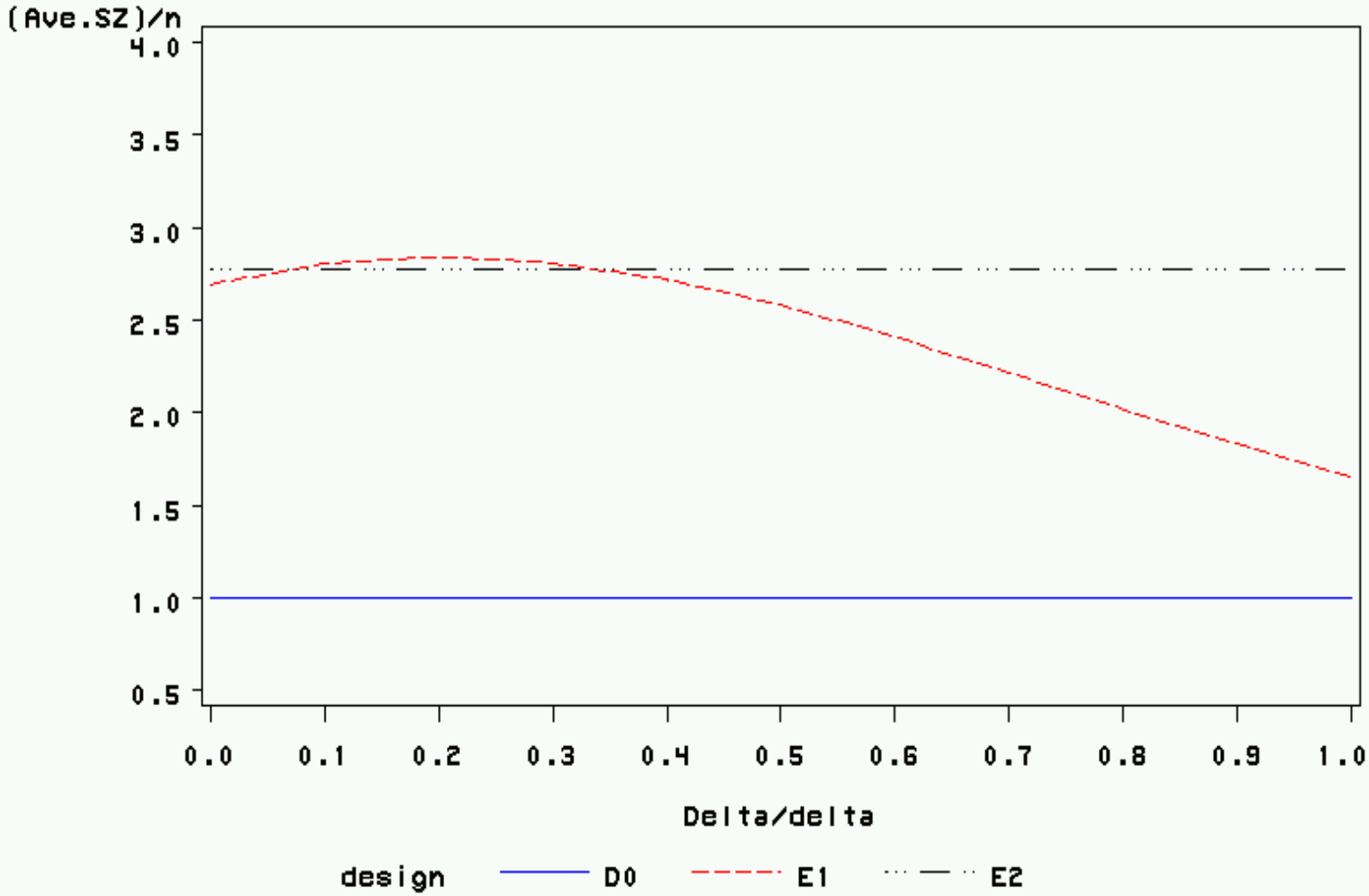
Do not change n^* throughout the trial.

Use the conventional Z test.

Pr(rej H_0) vs Δ/δ



(Ave SZ)/n vs Δ/δ


(Ave.SZ)/n
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Delta/delta
design
D0
E1
E2

Another valid adaptation method works on the original Z statistic for the entire data; equivalently, it works on the Z statistic for the data after time s with the weights s and $(1-s)$ also changing as a function of the past data.

Can replace Z_{1-s} with any test statistic W such that $W \mid Z_s$ follows $N(0, 1)$ under H_0 .

The α-level critical value C will change.

$$Z^* = \sqrt{s^*}Z_s + \sqrt{1-s^*}W$$

$$\Pr_{H_0}(Z^* > C^*)$$

$$= \Pr_{H_0}(\sqrt{1-s^*}W > C^* - \sqrt{s^*}Z_s)$$

$$= E_{H_0}\{\Phi(\frac{-C^* + \sqrt{s^*}Z_s}{\sqrt{1-s^*}})\}$$

The weight s changed to s* (depending on Z_s).

Notes on Adaptive Test Z*:

- Z* seems more natural because the original structure of the Z statistic is preserved for the entire data (one patient one vote)
- Z* can be much more efficient than U for large sample size increase
- the weights are functions of the interim data and so Z* is no longer normal
- extension to group sequential design is complicated (Brownian motion properties are no longer applicable to generate rejection boundary)

- Z^* requires the rule for changing n be specified, i.e.,
  - any different rule for changing n may result in different C^*;
  - once the rule is set, it cannot be changed during the course of the trial.
- if sample size does not change, then Z^* reduced to the original Z but the critical value C^* will have to change to C

- usual estimator (e.g., sample mean, MLE) compatible with Z^* has substantial bias
- mean unbiased estimator has very large variance and is not normally distributed

Examples

Proschan & Hunsberger (1995)
Lan & Trost (1997)
Li, Shih, Xie & Lu (2002)

All these methods incorporate futility stopping rule and the rejection regions also depend on the futility stopping rules. Thus, the futility stopping rules need to strictly follow. Overruling futility stopping may make the method invalid.

In fact, the adaptive test Z^* does not need to incorporate the futility stopping rule.
It seems more desirable that calculation of the critical value C^* of Z^* does not involve any consideration of futility stopping.

Logistics

Without any interim look at the trial data, statistical power is a guide for trial planning.

By monitoring interim data path, conditional power for future trend given the past data path may provide more relevant guidance for planning the future course of the trial.

Observing interim data path may generate anxiety which can be an issue with classical group-sequential designs and certainly with adaptive designs.

The greater unnecessary anxiety, the greater potential for bias might arise.

Points to consider in handling logistics

1) Establish standard operation procedure (SOP) in the protocol. Trial conduct complies with the SOP.
2) During interim adaptation, only unblind data that are necessary to be unblinded.
3) Have adaptation performed by an independent third party with no conflict of interest issue.

4) Logistics issues pertaining to traditional group-sequential designs also pertain to adaptive designs.
5) Adaptation entails careful planning at the protocol design stage (Bauer, Brannath, Posch, 2002).

Summary

Fixed information design is attractive because

1) logistics can be simplified
2) opportunity for operational bias can be reduced
3) statistical properties of commonly used statistics can be well assessed

However, design adjustments have taken place in most of Phase III clinical trials for drug development.

- so far recommend to avoid any adjustment affected by examination of internal sample path

Interim adaptation of fixed information design may be considered to potentially save an individual trial when it is strongly suspected that design specifications depart greatly from reality.

- need to consider practicality and flexibility
- need to ensure statistical validity
- need to assess potential loss in statistical efficiency with the adaptation if the suspicion leading to the adaptation is false
- efficiency needs to be evaluated on the basis of the entire development program, not just a single trial

Fully adaptive design with proper planning is attractive for development of medical products.

- change sample size or randomization allocation
- change study hypothesis (e.g., superiority vs. non-inferiority or equivalence)
- change test method
- change the primary endpoint from one pre-specified endpoint to another pre-specified one
- drop futile or unsafe treatment arms

and …….

Partial List of References

Bauer, Köhne (1994, Biometrics)
Bauer, Röhmel (1995, Stat. In Med.)
Lan, Trost (1997, ASA Proceedings)
Proschan, Hunsberger (1995, Biometrics)
Fisher (1998, Stat. In Med.)
Shen, Fisher (1999, Biometrics)
Cui, Hung, Wang (1999, Biometrics)
Bauer, Kieser (1999, Statistics in Medicine)
Posch, Bauer (1999, Biometrical J.)
Kieser, Bauer, Lehmacher (1999, Biometrical J.)
Lehmacher, Wassmer (1999, Biometrics)
Müller & Schäfer (2001, Biometrics)
Hellmich (2001, Biometrics)
Liu & Chi (2001, Biometrics)

Bauer, Brannath, Posch (2002, Method Inform Med)
Brannath, Posch, Bauer (2002, JASA) Cui (2002, Encyclopedia of Biopharmaceutical Statistics)
Li, Shih, Xie, Lu (2002, Biostatistics)
Lawrence (2002, J. of Biopharmaceutical Statistics)
Lawrence & Hung (2002, Biometrical J.)
Hung, Cui, Wang, Lawrence (2002, ASA Proceed.)
Jennison & Turnbull (2002, ASA talk)
Tsiatis & Mehta (2003, Biometrika ?)
Liu, Proschan, Pledger (2003, JASA)
Proschan, Liu, Hunsberger (2003, Statistics in Medicine)