Adaptive Designs from a DSMB Perspective: 
Some Controversies and Some Case Studies

Bruce Turnbull

Cornell University, Ithaca NY

http://legacy.orie.cornell.edu/ bruce/
Why now?

- High failure rate of Phase III trials
- Costs to pharma companies — $$$, patent life.
- Patients denied potential new effective treatments
What is Adaptive Design?

A. Big picture: A philosophy, a new paradigm;

Adapt to emerging information even in late stage trials.

versus

Old paradigm of “Learn first, then confirm”

B. Details: A collection of new statistical tricks and tools.
Adaptive Design of Clinical Trials: Early vs. Late Phase

*Early Phase trials:* Adaptive/ Bayesian approach is natural.

*Late Phase/Registration trials:* More controversial

   Control of Type I error is emphasized.
## The Current Controversy

Currently proposed adaptive methods . . .

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>are more efficient</td>
<td>are less efficient</td>
</tr>
<tr>
<td>are more flexible</td>
<td>are less flexible</td>
</tr>
<tr>
<td>have practical advantages</td>
<td>have logistical disadvantages</td>
</tr>
<tr>
<td>follow “scientific method”</td>
<td>results less scientifically credible</td>
</tr>
</tbody>
</table>
What Can Adaptive Designs Do? — (1)

1. Sample size re-estimation for a nuisance parameter,

2. Sample size modification to increase power in response to internal info,

3. Sample size modification to increase power in response to external info,

4. Testing for both superiority and non-inferiority,

5. Switching to a patient sub-population (enrichment),

……
What Can Adaptive Designs Do? — (2)

6. Multiple objectives with hierarchical structure,

7. Adaptive dose finding designs,

8. Reducing “white space” between Phases II and III,

9. Combining data from Phases II and III,

10. Joint planning of Phases II and III.

......
What Can Adaptive Designs Do? — (3)

11. Switching endpoints, e.g. oncology: OS — PFS — Tumor response,

12. Switching test statistics, (1yr vs 2yr survival, weighted logrank statistic)

13. Dropping or adding treatment arms,

14. Allocation to treatment to achieve balance in prognostic factors,

15. Reducing number of patients assigned to inferior treatments.

16. . . . . . .
Some Cautionary Examples

1. Keeping it simple (the “KISS” principle)

2. Perils of not recognizing early results are subject to high variability.

3. How adaptive designs may not work as expected!

4. Where adaptive designs are not needed.

5. Where adaptive designs are quite natural.
1. Keeping it simple (the “KISS” principle)

- Transparency, Credibility
  - Will the results impact practice?
  - Will the payers (e.g. CMMS) be convinced?

- Reproducibility: Audit or “independent re-analysis” (old JAMA (2005) policy).
1. Keeping it simple (the “KISS” principle)

Example 1A: Nutritional Prevention of Cancer (NPC) Trial

- Concern: Initial lag in treatment effect
  - Effect on power of logrank test
  - and sample size.
Example 1B: National Lung Cancer Screening Trial (NLST)


CT vs. X-ray screening. Sponsors NCI & ACRIN

- Concern: Crossing of hazard functions due to more early surgery in CT screened arm. Consequent insensitivity of logrank test and inappropriate early stopping boundary for harm/futility

- Team’s Solution — an adaptive weighted logrank test and nontraditional stopping boundaries.

- DSMB skeptical?

- Compromise ...
Example 1C. Bayesian adaptive design in Neurology Phase IIb/III trials


Recent reference: S. Chevet (SiM 2012).

“Bayesian adaptive clinical trials: A dream for statisticians only?”

A proposed neurology trial:

- Many (e.g. 16) dose levels. ED_{95} dose to be selected for confirmatory stage.
- Dose response modeled by spline function with priors on parameters.
- Adaptive randomized allocation
- Posteriors updated using MCMC
- Bayesian criteria used to time switchover from Ph IIb and the stopping of Ph III
Example 1C. Bayesian adaptive design in Neurology Phase IIb/III trials

Reaction

- Clinical PI: Extremely enthusiastic

- Statistical data center: Some pushback.
  - Concerns about credibility,
    transparency (proprietary ‘black box’ data processing),
  - Concerns about frequentist error rates.
    Is it possible to do enough simulations for an exhaustive investigation?
  - Concerns about any required independent audit analysis — (old JAMA (2005) policy).

- DSMB

- Regulators
Fact: Early results are subject to high variability

Decisions taken early at an interim look can be problematic:

Two examples:

A. Capricorn beta-blocker trial

B. Dropping treatment arms.

C. Premature stopping
A. Cautionary Example – The CAPRICORN Beta-blocker Trial

- 984 subjects with MI randomized to treatment or placebo.
  Endpoint: all-cause mortality (OS)

- During trial, Steering Committee decide to change primary endpoint to
  A. OS or cardiovascular hospitalization ($\alpha = .045$), and
  B. OS ($\alpha = .005$)

- In the end, neither A nor B was achieved.
  A because of a large number of non-specific hospitalizations;
  B because of low target $\alpha = .005$ was not quite met.

- Trial officially negative despite “convincing clinical benefit” (Editorial)

B. Dropping treatment arms.
Role of NIH official on DSMB

- Important stakeholder
- Unblinded
- Adaptations usually involve increased cost/funding.
  Uncomfortable for everyone!
- NIH can make interim decisions independent of DSMB
  Example: ADAPT
  (Alzheimers Disease Anti-Inflammatory Prevention Trial) — Pfizer
2C. Premature stopping ?? ADAPT

- ADAPT: Naproxen (Aleve—Bayer) vs. Celebrex vs. Placebo for Alzheimers prevention

- 9/2004, Merck withdrew Vioxx from the market due to increased CV events in a trial of colon cancer prevention. Trial termination recommended by DSMB at a regular meeting.

- 12/2004, For similar reasons, Pfizer terminated a similar trial of Celebrex. Trial termination recommended by DSMB.

- 3 days later NIH officials (not DSMB) stopped ADAPT trial

- In hindsight: Premature? Debate lingers to now (PRECISION)
  - adverse CV events: 23 (Na) vs 17 (Ce) vs 22 (Pl)
  - FDA Advisory Panel in Feb 2005 : “naproxen was the most appropriate comparator with which to evaluate the relative risks of new anti-inflammatory agents.”
**PRECISION**: Prospective Randomized Evaluation of Celecoxib Integrated Safety vs Ibuprofen or Naproxen
A non-inferiority trial.

**USA Today** (11 Feb 2014): “FDA Arthritis Advisory Committee (AAC) and Drug Safety and Risk Management (DSARM) Advisory Committee: Panel discussed whether PRECISION should continue, as some felt the trial was biased toward a null effect and that the results would not be useful. However, the majority of panelists felt that despite the difficulties with PRECISION, which has been delayed because of slow enrollment, the study should continue, as it could help inform clinical practice. ”

So much for confidentiality !!!

3. How adaptive designs may not work as expected!

A. Schizophrenia RCT — Design issues.

B. Seamless Ph II/III oncology RCT — Operational issues.

C. Bayesian adaptive Phase II enrichment oncology trial design.
3. How adaptive designs may not work as expected!

3A. Proposed Design of a Schizophrenia RCT


- Trt vs. Cntrl
- Endpoint: New Symptoms Assessment (NSA) score \( \theta = \mu_T - \mu_C \).
  \[ \alpha = .025; \]
  Power 0.8 at \( \theta = \Delta = 2; \)
  But \( \theta = 1.6 \) would also be clin. sig.

Propose: Adaptive sample size re-estimation plan — “promising zone approach”
Drawbacks of adaptive sample size re-estimation


“Start small, then ask for more” sounds seductive, but

- Usually inefficient — time, $$$, numbers of patients; Why?
  - Need for conservatism or non-standard test statistic to preserve alpha;
  - Ad hoc nature of sample size modification rule;
  - Subject to curse of high variability of interim results.

- Can be dishonest — “sucking in” investors (e.g. Wittes 2009)

- May cause logistical problems — enrolling new centers at short notice;

- May abrogate expensive decision to DSMB who may be reluctant — not stakeholders (e.g. Wittes SiM, 2002);

- Extent of sample size increase at interim gives away efficacy information to both insiders and outsiders who should otherwise remain blinded.
Business pressures can lead to misuse:

“A company that can’t afford to run a full trial might plan an adaptive trial that stipulates an ‘outrageously large’ effect for their treatment and contains a built-in plan to increase subject numbers if this noncredible effect size isn’t met. Then at some planned interim they look at the data and they say, ‘Oh, the observed effect size is smaller than anticipated.’ Still, that limited data can be used to lure investors to fund the larger trial.”

Janet Wittes (2009),
Business pressures can lead to misuse:

“A company that can’t afford to run a full trial might plan an adaptive trial that stipulates an ‘outrageously large’ effect for their treatment and contains a built-in plan to increase subject numbers if this noncredible effect size isn’t met. Then at some planned interim they look at the data and they say, ‘Oh, the observed effect size is smaller than anticipated.’ Still, that limited data can be used to lure investors to fund the larger trial.”

Janet Wittes (2009),

Example of an oncology trial where chair and individual members of DSMB received letters from SEC (Securities and Exchange Commission) after suspicious movements in stock price.
Multiply vert axis by 225 to get sample size. Plan (A) is constant at 1.96, Plan (B) at 2.2.
Red curve is distribution of interim estimate $\hat{\theta}_1$
Efficiency: Comparing designs
Sources of inefficiency in flexible, adaptive designs

Statistical “contortions” needed to preserve false positive rate.

Ad hoc nature of sample size modification rule

Typical sample size function for an optimal adaptive test

Typical sample size function for a conditional power adaptive design

“Conditional power” sample size modification rules differ qualitatively from optimal.
3. How adaptive designs may not work as expected!

3B. Experience with execution of a seamless Ph IIb/III oncology trial.

- Drug trial: Primary endpt: 28 day mortality.
  Required sample size  3000 subjects

- Three dose groups plus placebo. Randomized, double blind.

- After 200 patients, all but one of the active doses to be dropped.

- As per FDA instructions, this choice to be made by DSMB.
  Sponsor to remain blinded.

- Sponsor (and DSMB) very nervous about this arrangement.
  Sponsor draws up an algorithm involving a number of clinical and PK/PD variables by which DSMB will make their decision.
  A list of eleven “if then else” clauses are included in the DSMB charter, to which all members must swear to uphold.

- Study started and “ramped up”. Over 60 sites worldwide were “qualified”.


• At 1st interim (200 subjects), results on primary, co-primary and other clinical endpoints were promising. However, because of an unanticipated marginal finding on PK/PD variables, the DSMB was required to recommend stopping the trial for futility.

• Study was stopped, unblinded and findings reported to FDA.

• In the mean time, results from an additional 150 subjects who had been in the pipeline accrued. Again promising clinical results were observed.

• Sponsor submits plan to FDA for a new trial with conventional 2-arm placebo controlled group sequential design. Results of this trial will “stand on their own”.

• Sponsor appoints new DSMB, re-qualifies sites, gets new approvals from IRBs etc.

• New trial starts after a delay of 15 months.
3C. Bayesian adaptive Phase II enrichment oncology trial design.

- Drug trial: Primary endpt: PFS; Active drug vs control (blinded).
- Four histologic subtypes.
- Hope that drug is effective for all; but, if not, find subgroups that benefit.
- Hierarchical Bayes approach is natural – pooling of strength of evidence
- During enrollment, any subgroup could be stopped early for efficacy or futility based on a Bayesian updating algorithm (MCMC)
- DSMB to direct decisions based on updates at bimonthly telecon meetings.
3C. Bayesian adaptive Phase II enrichment oncology trial design.

What happened?

- Faster accrual and slower communications than anticipated meant that accrual was completed before DSMB could meet!
- Expense and effort in planning Bayesian adaptive design was wasted.
- During followup period (longer now, charter did not permit DSMB to make unblinding decision for futility or efficacy (only for safety).
- Should final analysis be Bayesian?
  If frequentist, should the analysis ignore the design?
4. Examples where adaptive methods may NOT be needed

A. VA Rehabilitation Trial — Adaptive Sample Size Modification

B. VA COURAGE Trial — Adaptive choice of goal: superiority or non-inferiority
Robot-Assisted Therapy for Long-Term Upper-Limb Impairment after Stroke


BACKGROUND
Effective rehabilitative therapies are needed for patients with long-term deficits after stroke.

METHODS
In this multicenter, randomized, controlled trial involving 127 patients with moderate-to-severe upper-limb impairment 6 months or more after a stroke, we randomly assigned 49 patients to receive intensive robot-assisted therapy, 50 to receive intensive comparison therapy, and 28 to receive usual care. Therapy consisted of 36 1-hour sessions over a period of 12 weeks. The primary outcome was a change in motor function, as measured on the Fugl-Meyer Assessment of Sensorimotor Recovery after Stroke, at 12 weeks. Secondary outcomes were scores on the Wolf Motor Function Test and the Stroke Impact Scale. Secondary analyses assessed the treatment effect at 36 weeks.

RESULTS
At 12 weeks, the mean Fugl-Meyer score for patients receiving robot-assisted therapy was better than that for patients receiving usual care (difference, 2.17 points; 95% confidence interval [CI], −0.23 to 4.58) and worse than that for patients receiving intensive comparison therapy (difference, −0.14 points; 95% CI, −2.94 to 2.65), but the differences were not significant. The results on the Stroke Impact Scale were significantly better for patients receiving robot-assisted therapy than for those receiving usual care (difference, 7.64 points; 95% CI, 2.03 to 13.24). No other treatment comparisons were significant at 12 weeks. Secondary analyses showed that at 36 weeks, robot-assisted therapy significantly improved the Fugl-Meyer score (difference, 2.88 points; 95% CI, 0.57 to 5.18) and the time on the Wolf Motor Function Test (difference, −8.30 seconds; 95% CI, −13.61 to −2.60) as compared with usual care but not with intensive therapy. No serious adverse events were reported.

CONCLUSIONS
In patients with long-term upper-limb deficits after stroke, robot-assisted therapy did not significantly improve motor function at 12 weeks, as compared with usual care or intensive therapy. In secondary analyses, robot-assisted therapy improved outcomes over 36 weeks as compared with usual care but not with intensive therapy. (ClinicalTrials.gov number, NCT00372411)
4A. Government Agency Sponsor (VA):
Sample size modification — “flexible conventional”.

- Study of rehabilitation treatments of patients with limb mobility impairment.
- Three treatments: Standard therapy versus intensive therapy versus intensive therapy plus medical device.
- Endpoint: Ability (Fugl-Meyer) score after 6 months. Range 0–226, High scores indicate better mobility.
• Initially an adaptive design (Lan & Trost) was proposed because “did not know what to expect”.

• Could assume scores were approx. normally distributed.

• It was agreed that an increase of 5 points would be of clinical importance.

• Really what was not known were the variances of the scores under each treatment.

• A conventional information monitoring design was implemented instead. (Mehta & Tsiatis, DIJ 2001).
  
  A goal in terms of information (not sample size) was set for each treatment arm.

• Trial went to conclusion. Published NEJM April 2010.
4B. Government (VA): COURAGE Trial

Superiority or non-inferiority design?

- Population: patients with stable coronary disease
- Study: to compare strategy of immediate PCI versus initial med therapy
- Primary endpoint: Non fatal MI or death (all cause)
- Sponsor’s goal was to show non-inferiority of med therapy
- A superiority trial was proposed to gain participation/enthusiasm of surgeon investigators
- Because of nesting of the two hyps, no alpha adjustment is required.
Optimal Medical Therapy with or without PCI for Stable Coronary Disease

William E. Boden, M.D., Robert A. O'Rourke, M.D., Koon K. Teo, M.B., B.Ch., Ph.D., Pamela M. Hartigan, Ph.D., David J. Maron, M.D., William J. Kostuk, M.D., Merrill Knudtson, M.D., Marcin Dada, M.D., Paul Caspersson, Ph.D., Crystal L. Harris, Pharm.D., Bernard R. Chaitman, M.D., Leslee Shaw, Ph.D., Gilbert Gosselin, M.D., Shah Nawaz, M.D., Lawrence M. Title, M.D., Gerald Gau, M.D., Alvin S. Blaustein, M.D., David C. Booth, M.D., Eric R. Bates, M.D., John A. Spertus, M.D., M.P.H., Daniel S. Berman, M.D., G.B. John Mancini, M.D., and William S. Weintraub, M.D., for the COURAGE Trial Research Group

ABSTRACT

BACKGROUND
In patients with stable coronary artery disease, it remains unclear whether an initial management strategy of percutaneous coronary intervention (PCI) with intensive pharmacologic therapy and lifestyle intervention (optimal medical therapy) is superior to optimal medical therapy alone in reducing the risk of cardiovascular events.

METHODS
We conducted a randomized trial involving 2287 patients who had objective evidence of myocardial ischemia and significant coronary artery disease at 50 U.S. and Canadian centers. Between 1999 and 2004, we assigned 1149 patients to undergo PCI with optimal medical therapy (PCI group) and 1138 to receive optimal medical therapy alone (medical-therapy group). The primary outcome was death from any cause and non-fatal myocardial infarction during a follow-up period of 2.5 to 7.0 years (median, 4.6).

RESULTS
There were 211 primary events in the PCI group and 202 events in the medical-therapy group. The 4.6-year cumulative primary-event rates were 19.0% in the PCI group and 18.5% in the medical-therapy group (hazard ratio for the PCI group, 1.05; 95% confidence interval [CI], 0.87 to 1.27; P=0.62). There were no significant differences between the PCI group and the medical-therapy group in the composite of death, myocardial infarction, and stroke (20.0% vs. 19.5%; hazard ratio, 1.05; 95% CI, 0.87 to 1.27; P=0.62); hospitalization for acute coronary syndrome (12.4% vs. 11.8%; hazard ratio, 1.07; 95% CI, 0.84 to 1.37; P=0.56); or myocardial infarction (13.2% vs. 12.3%; hazard ratio, 1.13; 95% CI, 0.89 to 1.43; P=0.33).

CONCLUSIONS
As an initial management strategy in patients with stable coronary artery disease, PCI did not reduce the risk of death, myocardial infarction, or other major cardiovascular events when added to optimal medical therapy. (ClinicalTrials.gov number, NCT00007657.)
• Trial went to conclusion with no significant differences between groups.

• Interpretation:
  – Was this a negative study failing to show PCI is better?, or
  – a positive study showing that the less expensive, less invasive meds trt is equivalent?

• Results of COURAGE published in NEJM April 2007.

• Impact. For VA practice? More generally?

• Post-mortem? Could/should this trial have been designed differently?
Testing for both superiority and non-inferiority

In retrospect, perhaps an adaptive design should have been used:

1. *Adaptive approach involving involve a change in the hypothesis being tested.*

Initially, design to prove superiority of PCI over meds, testing

\[
\theta \leq 0 \quad \text{vs} \quad \theta > 0.
\]

If this outcome starts to look unlikely, attention shifts to testing that the meds trt is non-inferior:

\[
\theta \leq -\delta \quad \text{vs} \quad \theta > -\delta.
\]

Power for the test of superiority is often set assuming a fairly large treatment effect, while the distance between hypotheses in the test of non-inferiority is smaller.

Since the test of non-inferiority requires a larger sample size, a change in objective is accompanied by a change in sample size. To maintain \(\alpha\) rates, these later outcomes must be downweighted, as in Cui et al. (1999).

2. Instead, a non-adaptive 3-decision group sequential design?

A more efficient option may have been to employ group sequential design with three possible decisions on termination. Boundaries control error rates $\alpha_{N}, \alpha_{S}, \beta_{N}, \beta_{S}$.

Such designs are able to achieve a low expected sample size for each possible value of the treatment effect (Öhrn & Jennison, *Stat. Med.*, 2010).
5. Sometimes the Adaptive Approach is the Natural One

- University based trial.

- Menopause symptoms. Compare treatments — medical, nutritional, behavioral.

- Limited funding stream necessitates a succession of small trials ("waves") of control versus one treatment at a time.
  Each study to be done in one center only.

- How should treatments to be used in the successive studies be chosen — "exploration vs. exploitation"?

- Funding agency (NIH) wants to see organized strategy for choosing the studies and a plan for statistical analysis (Meta-analysis?)
Key Takeaways:

1. Get DSMB involved before the design is etched in stone. “Take it or leave it” is not a good way to start a DSMB kickoff meeting.

   The DSMB needs to “buy in” to the design.

2. Make sure DSMB realizes early its responsibilities if it is to take charge of any trial modifications in an adaptive design.

3. Decision to use an adaptive design cannot be a substitute for "up-front" careful planning nor used to enable postponement of such planning.