Adaptive Designs for Clinical Trials

Insightfully Innovative

or

Irrelevantly Impractical

Stuart Pocock

London School of Hygiene and Tropical Medicine
Adaptive Designs for Clinical Trials

Insightfully Innovative

or

Irrelevantly Impractical
Adaptive Designs for Clinical Trials

unblinded interim results of ongoing trial

change some aspect of the trial design

according to a pre-planned Adaptive Charter

Focus here on pivotal Phase III trials
Types of Adaptive Design

After unblinded interim analysis:

**Increase sample size**
- sample size re-estimation

**Drop treatment arms/doses**
- seamless phase II/III

**Change entry criteria**
- enrichment design

**Change randomization ratio**
- play the winner

**Change primary endpoint(s)**
- challenging!
Are Adaptive Designs Useful?

flexibility appeals to trial sponsors, especially when trial is in “unexplored territory”

new methodology is fun

but they break the rules: interim results highly confidential

statistical penalties: need to preserve type I error practical issues: preserving trial’s integrity
EMEA Reflection Paper on Adaptive Designs
18 Oct ’07

“to modify design of ongoing Phase III trial is a contradiction to its confirmatory nature”

“rarely acceptable without further justification”

“adaptive designs should not alleviate rigorous planning”

“best for difficult experimental situations”

“ensure different stages can be justifiably combined”
THE BAD OLD DAYS

does bisoprolol reduce mortality in heart failure?

CIBIS trial began 1989 [Circulation ’94; 90 p1765]

bisoprolol vs placebo in 621 patients

**optimistic design:** underpowered trial

600 patients

**results:** 53 vs 67 deaths in mean 1.9 years

hazard ratio .80 (95% CI .56 to 1.15)  P = .22
CIBIS II trial  [Lancet ’99; 353 p9]

bisoprolol vs placebo in 2647 patients

**realistic design:** well powered trial

**results:** trial stopped early

156 vs 228 deaths in mean 1.3 yrs

hazard ratio .66 (95% CI .54 to .81)  P <.0001

the whole process took 10 years
unduly small first trial

↓

disappointment

↓

larger second trial

↓

success **eventually** (if treatment works)

**solutions:** be realistic in the first place

or

go for adaptive design
REALISM plus efficacy stopping rules for optimists

futility pessimists

PERFORM trial

terutroban vs aspirin for stroke/TIA patients

**realistic design:** 18,000 patients followed until 2340 have a primary event (stroke, MI, CV death)

90% power to detect 13% risk reduction, $\alpha = .05$

assuming 5% annual incidence

anticipate mean 3 years duration, trial is ongoing
PERFORM stopping guidelines

two efficacy looks:

at 40% point stop if $P < .0001$

at 70% point stop if $P < .001$

one futility look:

at 70% point stop if 95% CI excludes 7% risk reduction
“Astra Zeneca stroke drug fails in a clinical trial”  
[New York Times  27 Oct ’06]

The SAINT experience:
Two trials planned to run simultaneously

NXY-059 vs placebo in acute ischaemic stroke

primary outcome:
disability at 90 days, using modified Rankin scale (0 to 5)

90% power to detect common odds ratio 1.3
 require N=1700 patients in each trial
SAINT I (mostly European) recruited quickly
[New Eng J Med 2006;354 p588]

“NXY-059 significantly improved primary outcome”

1722 patients, common odds ratio 1.20 (95% CI 1.01 to 1.42)
P=.038

SAINT II (intended US) recruited slowly

July’05: increase sample size to 3200 patients based on SAINT I findings: 80% power to detect same result

Oct ’06:
3195 patients, common odds ratio 0.94 (95% CI 0.83 to 1.07)

NEJM Aug ’07 “NXY-059 is ineffective”
Lessons learnt from SAINT

in a tough field, any significant result is liable to be an exaggeration of the truth

SAINT I secondary outcomes were negative

SAINT II was not adaptive
ie its interim data (~ 1200 patients) were not used
SAINT II interim data negative, only seen by DMC, not involved in decision
SAINT II had no futility boundary

if truly adaptive, what would have happened?
Sample Size Re-estimation (Non-Adaptive)

DMC should not be involved

best done by Trial Executive/Sponsor

based on overall (blinded) event data

no statistical penalty

DMC Charter/monitoring plans affected
Sample Size Adjustment  (Non-Adaptive)

MIRACL trial in acute coronary syndromes  
(JAMA 4 April 2001)

atorvastatin vs. placebo

14% v 20% event rate,  $\alpha = .05$, 95% power

↓

2100 patients

13% overall event rate after 1260 patients

↓

DMC asked to advise, and declined!

↓

increase to 3000 patients by Steering Committee

to preserve power for same relative reduction

Result with 3086 patients: 14.8% vs. 17.4%  $P = .048$
Adaptive Sample Size Re-estimation

planned unblinded interim analysis

if observed treatment difference somewhat less than expected, increase trial size

conditional power calculation:
eg. choose new size to achieve (say) 90% chance of eventual $P<.05$, assuming observed interim difference is the truth

calculate $CP(N) = \text{power to achieve } p<.05 \text{ with sample size } N$, conditional on estimated difference

choose new $N^*$ so that $CP(N^*) = 90\%$ (say)

$N^*$ has an upper limit $N_{\text{max}}$
Conceptual Outline of the Adaptive Options at a Planned Interim Analysis

Interim Difference                Decision
→ Amazing                        Stop now for efficacy
      Favorable                    Continue to original N, CP(N) > 90%
⇒ Hopeful                        Extend to new N* so that CP(N*)=90%(say)
⇒ Not brilliant                  Extend to Nmax so that CP(Nmax)>30%(say)
Disappointing                    Continue to original N
→ Hopeless                       Stop now for futility or harm
→ conventional data monitoring boundaries (optional)
⇒ adaptive options
δ = specified difference in protocol power calculation
\hat{\delta} = estimated treatment difference at interim analysis

boundaries (and existence) of each option according to a pre-defined Adaptive Charter
CHAMPION trial uses adaptive design
cangrelor vs clopidogrel in 8750 ACS patients
primary endpoint: death, MI, revasc. at 48 hrs
interim analysis after 70% evaluated
5 options: stop early for efficacy
proceed to N = 8750 if $\geq 80\%$ conditional power
expand up to N = 15000 to have 80% CP*
proceed to N = 8750 if less promising
stop early for futility if $< 20\%$ CP
*also enrich to high-risk (or also plavix naive only) if interaction
detailed adaptive charter approved by FDA
CHAMPION trial in Acute Coronary Syndrome

[Mehta et al Circulation 2009]

primary endpoint: death, MI or revascularization within 48 hours

planned N = 8750 : 7.0% vs 8.7% \(\alpha = .05\) power 82%

interim analyses at \(N_I = 4375\) and 6125

group sequential boundaries for efficacy, futility and harm

adaptive interim look at \(N_I = 6125\)

if \(CP(N) < 80\%\) extend size to \(N^*\) such that \(CP(N^*) = 80\%\)
only if \(N^* < 15,000\), otherwise go to original \(N = 8750\)
CHAMPION Results [NEJM Nov 2009]

trial stopped early for futility

<table>
<thead>
<tr>
<th></th>
<th>cangrelor</th>
<th>clopidogrel</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3889</td>
<td>3865</td>
</tr>
<tr>
<td>primary endpoint at 48 hours</td>
<td>290</td>
<td>276</td>
</tr>
<tr>
<td>MI</td>
<td>278</td>
<td>256</td>
</tr>
<tr>
<td>ischemia-driven revascularisation</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>death</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

P = .59
Active Drug vs Placebo in Schizophrenia

[Cyrus Mehta, EAST]

primary endpoint: change in symptom score at 26 weeks

planned size $N = 442$ patients

$\delta = 2, \quad \sigma = 7.5 \quad 2\alpha = .05 \quad 80\%$ power

interim analysis: $N_I = 208$ patients with 26 week outcome
timed so that planned recruitment not yet complete

sponsor sets $N_{max} = 884$

no early stopping at $N = 208$
The Adaptive Design Specifications

planned $N = 442$, $\delta = 2$  \( \hat{\delta} \) = estimate of $\delta$ at $N_l = 208$

calculate conditional $CP(N)$ assuming true $\delta = \hat{\delta}$

1) if $CP(N) > 80\%$, continue to $N = 442$

2) if $CP(N)$ lies between 36\% and 80\% interim outcome is deemed **promising**:

   extend to $N^*$ so that $CP(N^*) = 80\%$
   if $N^*$ exceeds 882, extend to $N_{max} = 882$ instead

3) if $CP(N) < 36\%$, continue to $N = 442$
### Operating Characteristics of Adaptive Design (by simulation)

#### Overall

<table>
<thead>
<tr>
<th>True $\delta$</th>
<th>Fixed Power</th>
<th>Adaptive Power</th>
<th>Expected Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>61%</td>
<td>67%</td>
<td>509</td>
</tr>
<tr>
<td>1.8</td>
<td>71%</td>
<td>76%</td>
<td>506</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
<td>84%</td>
<td>499</td>
</tr>
</tbody>
</table>

#### When interim Outcome is Promising

<table>
<thead>
<tr>
<th>True $\delta$</th>
<th>Probability of “being Promising”</th>
<th>Fixed Power</th>
<th>Adaptive Power</th>
<th>Expected Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>.26</td>
<td>61%</td>
<td>83%</td>
<td>696</td>
</tr>
<tr>
<td>1.8</td>
<td>.26</td>
<td>69%</td>
<td>89%</td>
<td>690</td>
</tr>
<tr>
<td>2</td>
<td>.24</td>
<td>76%</td>
<td>93%</td>
<td>685</td>
</tr>
</tbody>
</table>
Statistical Penalty for Sample Size Re-estimation?

Cui et al [Biometrics ’99; 55 p853]

“Increasing sample size based on interim treatment difference can substantially inflate type I error in most practical situations”

give less weight to later data? illogical

Chen et al [Stats in Med ’04 ; 23 p 1023]

“Increasing sample size when unblinded interim result is promising will not inflate type I error. No statistical adjustment required”

Who is right?
Statistical Penalty for Sample Size Re-estimation

To preserve the Type I error

1) **Down-weight the later data? NO**
   
   [Cui et al. Biometrics ’99]

\[
z = \sqrt{\frac{n_1}{n_1 + n_2}} z_1 + \sqrt{\frac{n_2}{n_1 + n_2}} z_2^*
\]

illogical, need to weight equally
link to estimation

2) **Adjust final \( \alpha \) ? YES**
   
   [Gao, Ware & Mehta in press]

analogous to \( \alpha \) spending in data monitoring
adjustment is fairly small
No inflation of type I error if:

1) Only increase sample size when conditional power at interim analysis already exceeds around 30% to 40%
   and/or

2) One stops for futility at interim analysis if conditional power is less than 10%

Gao et al [J Biopharm Stats ‘08; 18 p1184]
Conceptual Outline of Adaptive Design:
\( n_{\text{max}}/n_2 = 2, \quad n_1/n_2 = 0.5, \quad 2\alpha = .05, \quad 1-\beta = 0.9 \)
Values of \( C_{p_{\text{min}}} \) that preserve \( \alpha \) for conventional final analysis

<table>
<thead>
<tr>
<th>Sample Size Ratios</th>
<th>CP_{min} Values for Targeted Conditional Powers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80%</td>
</tr>
<tr>
<td>Maximum Allowed (( n_{\text{max}}/n_2 ))</td>
<td></td>
</tr>
<tr>
<td>At Interim Look (( n_1/n_2 ))</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.42</td>
</tr>
<tr>
<td>1.5</td>
<td>0.41</td>
</tr>
<tr>
<td>1.5</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.32</td>
</tr>
<tr>
<td>3</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
</tr>
<tr>
<td>( \infty )</td>
<td>0.32</td>
</tr>
<tr>
<td>( \infty )</td>
<td>0.31</td>
</tr>
<tr>
<td>( \infty )</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Illustrating the slight conservatism of adaptive design when sample size increase confined to promising zone
Adaptive Design vs Group Sequential Design

**adaptive:** start smaller, expand if need to

**group sequential:** plan big, stop early if “good news”

group sequential is more statistically efficient?  
[Jennison and Turnbull Stats in Med 2006; 25 p917-]

group sequential requires greater up-front commitment

has overrun problems if outcome is long term
Practicalities of the Adaptive Approach

detailed planning: an Adaptive Charter

define initially, or while blinded to interim results

seek regulatory approval, show statistical rigor

who sees unblinded data and decides:
   Expert Panel (may be DMC or different)
   + independent data analyst

fixed rules or judgement allowed (eg safety issues)
Potential Problems with Adaptive Approach

sponsor stays blinded throughout?

If algorithm known, others can infer (guess) interim results

risk of wider unblinding in effect

consequences re: investigators
sponsor
investment analysts
others

could the trial itself be compromised?
Seamless Phase II/III Trials

indacaterol in COPD patients

Stage I (N = 115 per group, 7 groups)

75, 150, 300, 600 mg indacaterol

vs placebo vs formoterol vs tiotropium

2 week efficacy outcome, safety data

which two doses proceed to Stage 2?

[Barnes et al Pulm Pharm and Therap, in press]
Stage 2  (extra N = 285 per group)

dose A vs dose B vs placebo vs tiotropium

dose selection guideline:
  lowest dose meeting pre-defined efficacy criterion + next dose

  safety issues (& lack of efficacy) can affect selection

Data Monitoring Committee selected 150 and 300 mg

Stage I + Stage 2 (N = 400 per group)
re efficacy, safety, tolerability over 26 weeks
Statistical Adjustment

Bonferroni, ie $\alpha/4$ for each dose vs placebo
active control

merit of simplicity, but too conservative?

Advantages of Seamless Phase II/III

use all data on doses selected

gain in power, overall speed
Enrichment design

unblinded interim subgroup analyses

↓

restrict recruitment to “promising subgroups”

pre-define which subgroups & restriction criteria

but subgroup analyses (at interim) lack power

seriously avoid this one?

exceptions: distinct disease types, genetic markers
AMIHOT trials

intracoronary supersaturated oxygen therapy vs control after PCI in acute MI

primary outcome: infarct size %

**AMIHOT I overall** [N = 269 patients]
medians 11% vs 13%  \( P = 0.6 \)

**Subgroup: anterior MI, treated within 6 hours** [N = 105]
medians 9% vs 23%  \( P = 0.07 \)

**AMIHOT II, only in subgroup** [N = 301, 3 : 1 randomisation]
medians 20% vs 26.5%  \( P = 0.10 \)

“squeezing” the combined data  Bayesian  \( P = 0.023 \)
Changing the Primary Endpoint(s)

based on **blinded** interim data/external evidence

eg low event rate $\rightarrow$ broader composite

or

select endpoints which “worked” in related trial

no adjustment? credibility affected

DMC not involved
Changing Primary Endpoint(s)

based on **unblinded** interim data (**adaptive**)

much trickier

complete loss of credibility?

can one pre-define switch criteria
  statistical adjustments
CAPRICORN [Lancet 5 May 2001]

carvedilol vs placebo after MI
original primary endpoint: all cause mortality lower than expected in blinded interim analysis

downwards

cchange to co-primary endpoints:

1) all cause mortality .005
2) all cause mortality + CV hospital administration .045
## RESULTS

<table>
<thead>
<tr>
<th></th>
<th>carvedilol</th>
<th>placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N=975</strong></td>
<td></td>
<td><strong>N=984</strong></td>
</tr>
</tbody>
</table>
| all cause deaths     | 116        | 151     | **P=.031**
| all cause death      | 340        | 367     | **P=.3**
| or CV hosp\n         |            |         |

tricky to interpret
formally not sig at 5% level
FDA approval
Adaptive Designs are Intuitively Appealing, But….

Flexibility  
Efficiency  
Methodological stimulus

Lack of rigour  
Too much haste  
Useless theoretical fun

Some adaptive ideas workable, others not

for each application: attention to details
  statistical rigour
  practical implementation
  avoidance of bias

Learn from real experiences
do adaptive designs show **inconsistency in** approach?

Oscar Wilde

“**Consistency** is the last refuge of the unimaginative”

Aldous Huxley

“**Consistency** is contrary to nature, contrary to life. The only completely consistent people are dead.”