Adaptive Designs for Clinical Trials: Promises, myths, challenges and benchmarking

Ken Cheung

Columbia Biostatistics

MARKETPLACE

THE WALL STREET JOURNAL.

ed. * * * * MONDAY, JULY 10, 2006 B1

◀ **A record opening** for *'Pirates of the Caribbean: Dead Man's Chest'* Page B4

FDA Signals It's Open to Drug Trials That Shift Midcourse

By Anna Wilde Mathews

CLINICAL TRIALS of medicines are traditionally performed in a "blinded" fashion so that the findings will remain secret until the studies are completed. But regulators and the pharmaceutical industry are increasingly interested in starting to use a very different model that lets studies change as they go along, based on early results.

Drug companies have begun to perform such adaptive trials for their new medicines, hoping for more efficient tests that could save millions of dollars. The Food and Drug Administration, meanwhile, is sending increasingly encouraging signs that it is open to considering the results of such trials. In a move that could lay the groundwork for greater future use of such studies, Scott Gottlieb, an FDA deputy commissioner, is set to announce today plans to develop regulatory guidelines for adaptive trials. The FDA has also put together an internal team to work with its drug-review divisions on the adaptive designs, which are statistically complex.

"We think it's time to start exploring the appropriate use of these designs in the appropriate situations," says Robert T. O'Neill, director of the FDA drug center's office of biostatistics. Over the past year, all of the FDA drug-review divisions have seen at least one adaptive trial submitted by companies, he says.

The most ambitious adaptive designs would represent a big change from traditional clinical-trial practices, and the idea has sparked controversy among researchers. Now, once trials are set in motion, they are supposed to be left largely untouched until they are finished and the drug company finds out the results. One exception: The studies often have an independent data-monitoring board that has the power to shut a trial down for ethical or safety reasons.

Adaptive trials have aspects that are "fundamentally different from what we currently do," says Michael Krams, who joined **Wyeth** in April as assistant vice president for adaptive trials. The results of an ongoing study are watched closely, and changes to the design occur as it continues, guided by a complex plan developed in advance, typically through computer simulations. If one treatment looks more effective, a greater proportion of patients may be funneled to it. If one group of patients appears to be benefiting more, the trial might start adding a larger share of that type of person.

A Trial Basis

Some ways that adaptive designs may allow clinical trials to be adjusted based on early results:

- Route a larger share of patients to the treatment that seems to work the best
- Drop treatments that don't appear to be effective
- Add more of the type of patients who seem to be reacting best to a particular treatment
- Merge two different phases of drug development into one trial

Source: WSJ research

Pharmaceutical companies hope such approaches hold the potential for major savings, though so far they are largely focused on early-stage trials. Advocates of adaptive designs say they can involve a reduction of 30% or more in the number of patients needed in a trial, and can save time as well. They also say that adaptive trials carry major benefits for patients, who have reduced odds of getting a less-effective treatment.

"It helps us pick the winners and losers faster," says Steve Ruberg, director of global medical information sciences at **Eli Lilly** & Co. The company has three ongoing early-stage adaptive drug trials in areas including oncology and diabetes, he says.

Bristol-Myers Squibb Co. is planning a migraine drug trial that will use adaptive principles to help determine how much medicine to give. The study will start with 10 to 15 different doses, far more than

Please Turn to Page B8, Column 4

Breaking the Marwari Rules

Mittal's Bid for Arcelor Shows

In Brazil, Ford Has Discovered

Agenda

- Promises and myths:
  - Proof-of-concept study (Piclozotan)
  - Outcome adaptive allocation (ECMO)
  - Phase 1 dose finding (Lovastatin)
- Challenges (complexity) and benchmark
  - Phase 1/2 dose finding of a thrombolytic agent
- Discussion

FDA Guidance February 2010

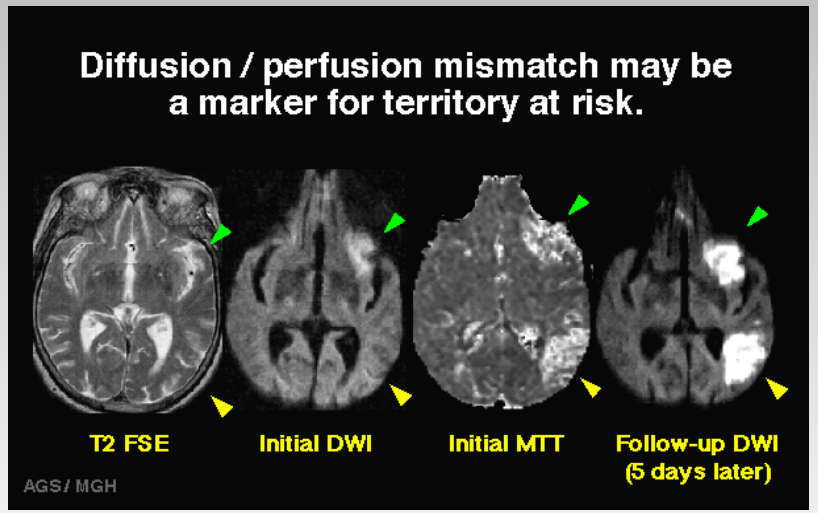
"An adaptive design clinical study is defined as a study that includes a *prospectively planned opportunity for modification … of the study design … based on analysis of data from subjects in the study*"

- More efficient (i.e., smaller N)
- Increase likelihood of success on study objective
- Yield improved understanding of the treatment effect (e.g., dose-response)

PROOF-OF-CONCEPT STUDY

Case Study: Piclozotan I

- Single-arm proof-of-concept (Phase 2)
- Enroll and treat *n* acute stroke patients with a new treatment
- *Primary endpoint:* MRI response = Indicator of no growth in infarct size by DWI.
- Research Questions:
  - Is the treatment good?
  - What is the response rate?
- Observe S_n = #MRI responses
- Design Question:
  - **What is n?**



Case Study: Piclozotan I

- Inputs for sample size calculation:
  1. A bad ("null") response rate, 25%
  2. Type I error rate $\alpha \leq 5\%$ under null
  3. A good ("alternative") response rate, 40%
  4. Power $\geq 80\%$ under alternative
- A fixed (non-adaptive) design
  - **n = 62**
  - Decision rule: conclude the treatment is good if $S_n \geq 22$

Case Study: Piclozotan I

- A two-stage adaptive design with the same error constraint: 5% type I error, 80% power
- *Stage 1:* Enroll $n_1 = 20$ subjects
- *Futility interim:*
  - Stop the trial and conclude *futility* if $S_{20} < 6$
- *Stage* 2 (if $S_{20} \geq 6$):
  - Enroll another $n_2 = 51$ subjects (i.e. **total n = 71**)
  - Conclude the treatment is good if $S_{71} \geq 24$; conclude futility otherwise

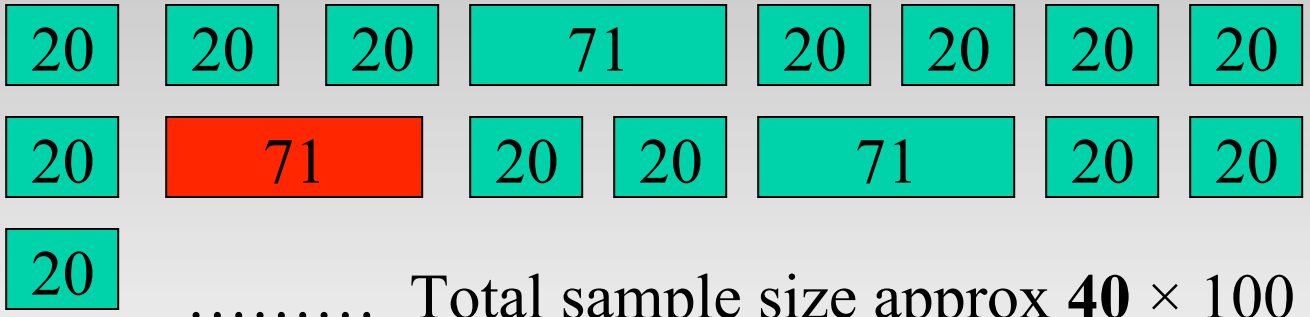
Adaptive design: Efficiency myth

- When the response rate is 25% ("null"), the *expected* sample size of the two-stage design is

$$20 \times 0.62 + 71 \times (1 - 0.62) \approx \mathbf{40}$$

where 0.62 = Pr (Stop at stage 1 | response rate = 25%)

- Thought exercise: Suppose 100 drug trials use this design



…… Total sample size approx **40** × 100 = 4,000

Adaptive design: Efficiency myth

- Using the fixed design in 100 similar drug trials will need 6,200 subjects.
- Comparison 1 (portfolio management):
  - **4,000 < 6,200** → Two-stage design is more efficient than the fixed design in terms of *expected* sample size under the null hypothesis.
- Comparison 2 (investigator's perspective)
  - *Maximum* sample size of two-stage design is **71 > 62**
  - Two-stage design makes sense if you believe the drug doesn't work

Adaptive design: Efficiency myth

- This numeric comparison demonstrates a potential tension between the perspectives of the individual investigator and the broader community…The individual investigator's interest resides in keeping the sample size of a single trial small. Given limited resources and finite numbers of stroke patients, the community's interest resides in keeping the average sample size small so that more trials can be performed.

-Cheung and Kaufmann, *Stroke,* 2011

Adaptive design: Efficiency myth

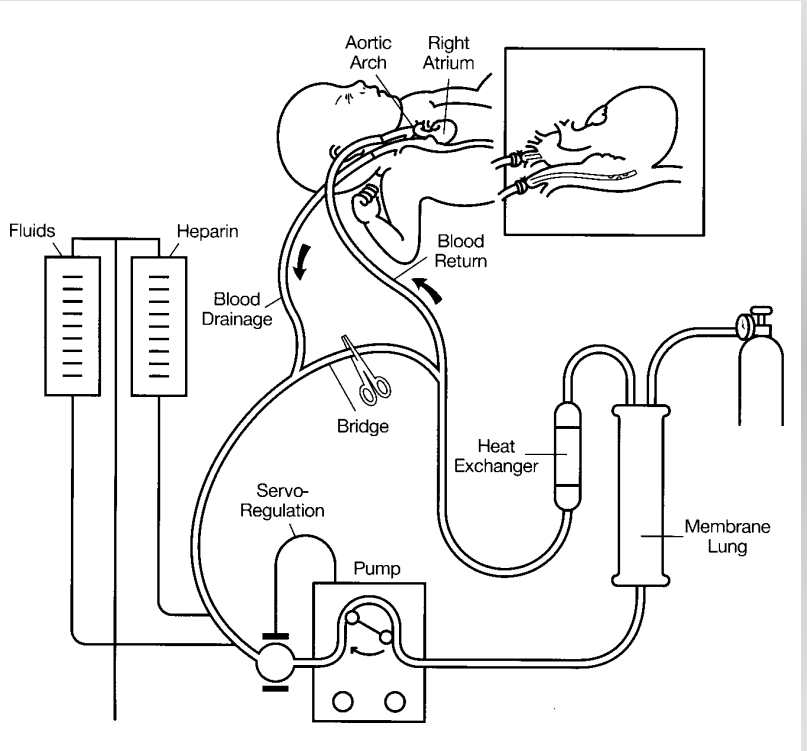
A Statistical theory (Neyman-Pearson) says

- For the same error constraints, the maximum sample size of *any* two-stage adaptive design is always at least as large as that of the single-stage *fixed* design

OUTCOME ADAPTIVE ALLOCATION

Case Study: Michigan ECMO

- Extracorporeal membrane oxygenation
- Indication: Persistent pulmonary hypertension of the newborn
- Bartlett et al. *Pediatrics* (1985)
  - U of Michigan
  - Historical survival rate
    - Infants with EMCO: 80%
    - Untreated: 20%
  - Play-the-winner design (Wei & Durham, *JASA,* 1978)



http://www.nichd.nih.gov/publications/pubs/images/Efig1p26.gif

Case Study: Michigan ECMO

Randomized play-the-winner (PTW)

- Adaptive randomization
- Use outcome data obtained during trial to influence randomization probabilities of treatments
- Goal: allocate as few patients as possible to a seemingly inferior treatments

Case Study: Michigan ECMO

Modified play-the-winner (Urn model)

A ball → ECMO

B ball → Standard control

If success on A, add another A ball ….

Randomized Consent Design

Results

- First infant: Randomization probability ECMO:Placebo = 1:1 → ECMO and survived
- Second infant: Randomization probability = 2:1 → Placebo and died
- Third infant: Randomization probability = 3:1 → ECMO and survived
- Fourth infant: Randomization probability = 4:1 → ECMO and survived …

	1	2*	3	4	5	6	7	8	9	10
ECMO	S		S	S	S	S	S	S	S	S
CONTROL		F								

***sickest patient**

Case Study: Michigan ECMO

- 2 additional infants treated with ECMO, both survived.
- Can superiority be proven with 1 placebo subject?
- P-Values, depending on method, ranged $.001 < .05 < .28$
- Consequence: Harvard ECMO, not without ethical difficulties due to equipoise; O'Rourke, et al. *Pediatrics (1989)*

Adaptive design: efficiency myth

- Motivation for adaptive randomization is ethics, not economics
- Tradeoff for better ethics:
  - Issues of proper analyses can be quite complicated
  - Require proper planning (e.g., PTW with bigger urn)
- Not all adaptive randomization are Bayesian

PHASE 1 DOSE FINDING

Phase 1 Dose Finding

- Primary objectives
  - Safety
  - Maximum tolerated dose (MTD)
- Design issues:
  - Trade-off between safety and efficacy
  - Sequential accrual
  - Oncology convention: 3+3 design
    - Exceed MTD if a dose has ≥2 toxicities in 6 patients
    - Dose is considered safe with 0/3 or 1/6

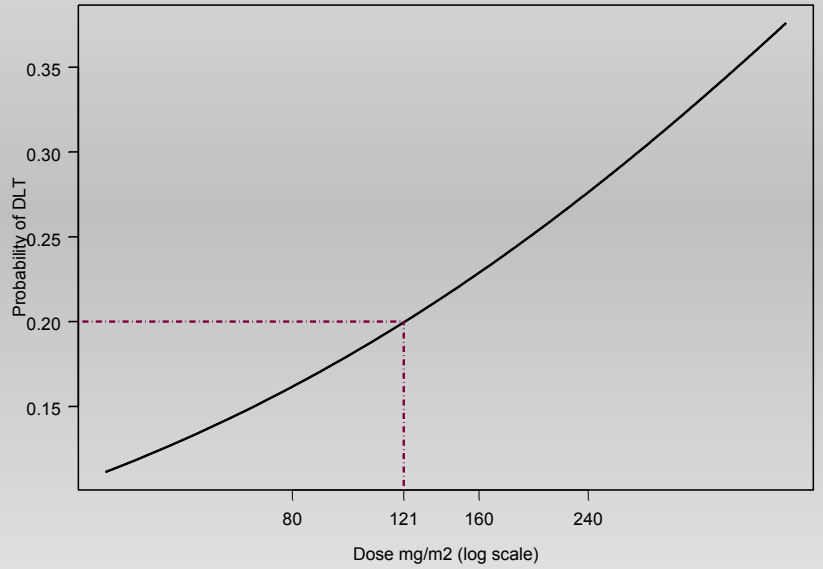
Phase 1 Dose Finding

- Some difficulties with 3+3
  - What if 2 toxicities in 7 patients?
  - Lack a quantitative definition of the MTD
  - Toxicity tolerance is lower in non-cancer population
- Continual reassessment method (CRM)
  - MTD = a dose associated with p percent toxicity
  - Model-based: use a dose-toxicity model to decide dose assignments of study subjects

Continual Reassessment Method (CRM)

Some operational details

1. Treat the first group of subjects at the prior MTD
2. Observe toxicity outcomes
3. Estimate the dose-toxicity curve
4. Treat a new group of subjects at the estimated MTD.
5. Go back to Step 2, until the sample size is reached.

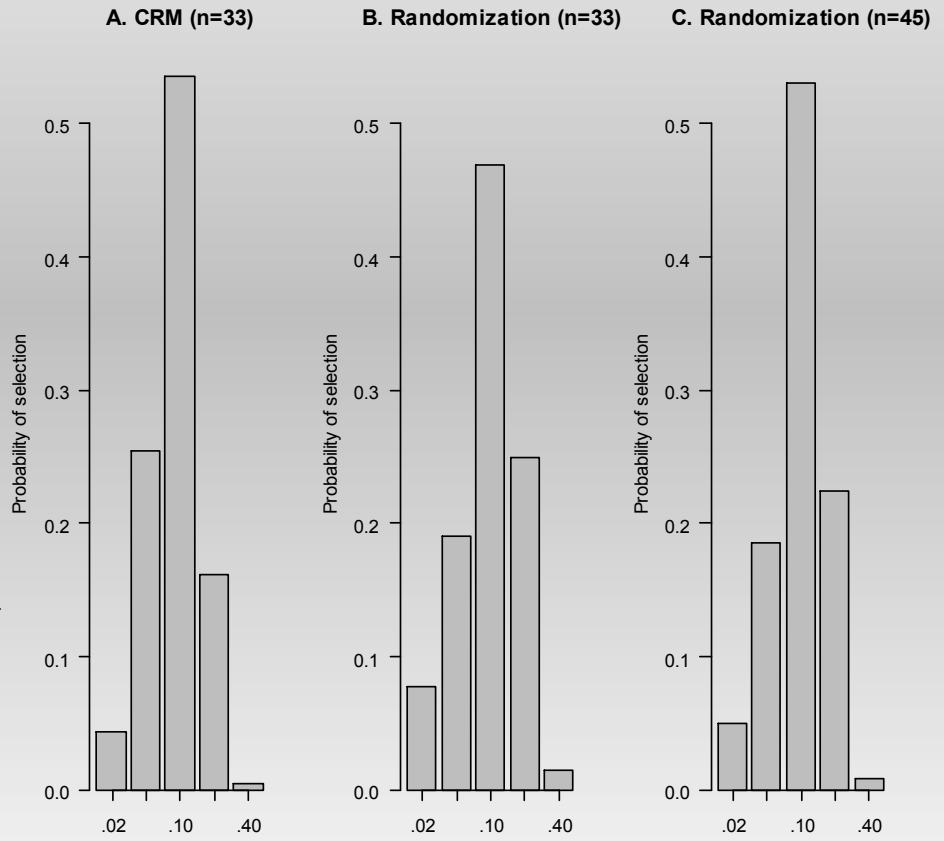


Case Study: NeuSTART

- 5 dose levels of lovastatin in acute stroke patients
- Allow p=10% toxicity: liver, muscle
- Use a CRM variant for dose escalation in N=33
- Results: MTD estimate = 8 mg/kg/day for 3 days
- The 3+3 method allows 1 toxicity out of 6

Case Study: NeuSTART

- *Why adaptive: Efficiency*
- Target rate: 10%
- Toxicity odds increases 2.5 times per dose level
- Logistic regression was used to estimate the MTD at trial's end



Cheung and Kaufmann, *Stroke,* 2011

Case Study: NeuSTART

Design characteristics	CRM	Randomization
(a) Probability of correctly selecting the MTD[a]	0.54	0.47
(b) Probability of selecting an overdose[a]	0.17	0.26
(c) Average number of subjects treated at	13	7
(d) Average number of subjects treated at an overdose	6	13
(e) Median of toxicity odds ratio estimate[a]	**5.2**	**2.6**

[a]The MTD and the odds ratio are estimated using logistic regression at the end of each simulated trial for both the CRM and the randomization design

Adaptive design: efficiency myth

- *Why adaptive:*
  - Ethics: Treat more patients on average at the right dose
  - Higher likelihood of finding the right dose
- Learn less on dose-response
  - Odds ratio: Can't answer how fast the toxicity or response increases beyond the MTD

CHALLENGE: COMPLEXITY
(IN CONTEXT OF DOSE FINDING)

Statistical world of dose finding (is long)

- Up-and-down designs (Storer, 1989)
- Continual reassessment method (O'Quigley et al, 1990)
- Biased coin design (Durham et al, 1997)
- EWOC (Babb et al, 1998)
- Curve-free method (Gasparini and Eisele, 2000)
- ... [apology for omission]
- A+B and stepwise designs (Lin and Shih, 2001; Cheung 2007)
- Stochastic approximation (Cheung, 2010)
- Stochastic optimization (Bartroff and Lai, 2010)
- …

- Classification
  - Model-based vs algorithm-based
  - Long memory vs short memory

- "Standard case"
  - MTD = p^{th} percentile of a tolerance distribution
  - Binary outcome (Y = 0 or 1)
  - Exchangeable patients; dose (X) is the only covariate Y ~ X

Statistical world of dose finding (is complex)

- "Nonstandard cases"
- Use non-binary endpoints ("Y")
  - Delayed toxicity; time to event (Cheung and Chappell, 2000)
  - Eff-tox trade-off (O'Quigley et al, 2001; Thall and Cook, 2004)
  - Ordinal outcome (TBS; Lee et al., 2010)
  - Continuous outcome (Cheung and Elkind, 2010; Hu and Cheung, 2012)
- Incorporate complex design ("X")
  - Drug combination (Thall; Ying and Yuan; etc.)
    - $Y \sim X_A + X_B$
  - Patient heterogeneity $Y \sim X + Z$

Clinical world of dose finding (is often much simpler)

3+3

Despite

- The abundance of statistical principles
- The willingness of clinical investigators

Why the gap?

- Adaptive (dose finding) designs are often too complex to be
  - Well specified
  - Accessible (N?)
  - Well understood

Guggenheim. Architect: Frank Lloyd Wright

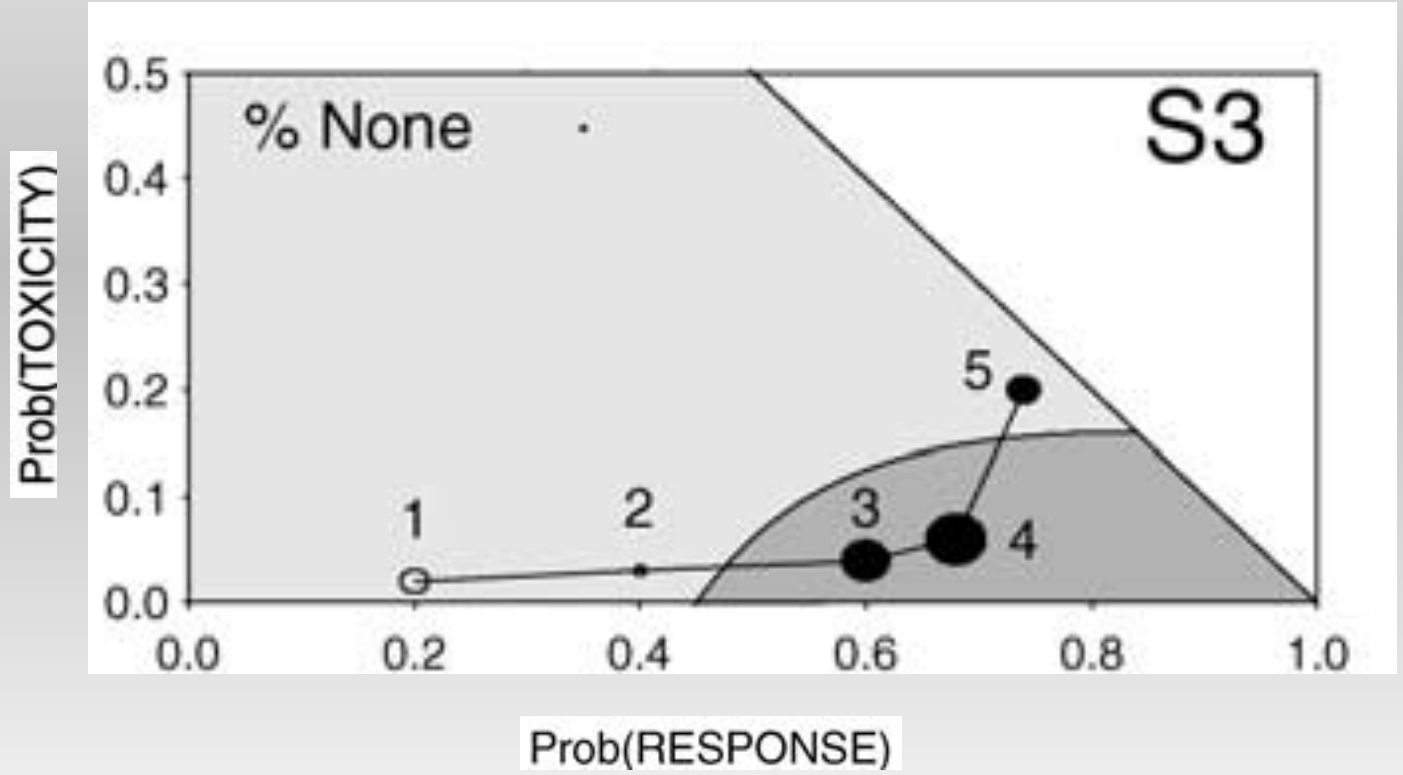
Filling the gap

- Automated algorithms to specify the CRM model
  - Lee and Cheung (2009, 2011, *Clinical Trials*)
  - Jia, Lee, and Cheung (2014, *Biometrika*)
- A sample size formula for the CRM
  - Cheung (2013, *Clinical Trials*)
- For general dose finding methods:
  - Cheung (2014, *Biometrics):* A Cramer-Rao type benchmark for diagnostics and gauging plausibility

Thrombolytic agent for stroke

- Dose finding of a thrombolytic agent for acute stroke
- Phase 1/2 study
- Trinary outcome:
  - Intracranial hemorrhage (Toxicity)
  - Reperfusion without hemorrhage (Response)
  - Neither
- Thall and Cook (2004):
  - Define desirability $\delta(p_E,p_T)$ as a function of response rate p_E and toxicity rate p_T
  - Aim to find a dose that maximizes $\delta(p_E,p_T)$

Thrombolytic agent for stroke



Thrombolytic agent for stroke

Thall and Cook (2004):

- Outcome-adaptive
- Bayesian, model-based dose finding method
  - N = 72
  - Assign patients at dose with maximum desirability based on interim data (CRM-like)
  - Consider two dose-response-toxicity models: Proportional odds (PO) and Continuation ratio (CR)
  - Number of model parameters: 6

Simulation results

Scenario 3

Model	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
Desirability	-0.48	-0.13	0.22	**0.32**	-0.26
PO✔	0	0	20	**72**	7
CR	0	2	32	**49**	16

Scenario 4

Model	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
Desirability	0.12	0.29	0.45	0.58	**0.69**
PO	0	2	10	34	**54**
CR✔	0	0	1	5	**94**

Dose Finding Benchmark

- A theoretical dose finding design that provides an upper limit of accuracy for any dose finding methods for a given design objective under a given scenario.

Dose Finding Benchmark

- Let $d(\pi)$ denote the design objective, e.g.,

  NeuSTART: $d(\pi) = \arg\min_k |\pi(k) - 0.10|$

  Thall and Cook: $d(\pi) = \arg\max_k \delta_k$
- π denotes the true dose-response curve
- Benchmark: $d(\pi^*)$ where π^* is a nonparametric optimal estimate of π based on complete outcome profile

Dose Finding Benchmark

- In an actual trial, we observe a partial outcome profile, e.g., a patient at dose 3 with toxicity

Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
?	?	Toxicity	Toxicity	Toxicity

- In a computer simulation, we can observe a complete profile by generating a uniform tolerance

Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
No toxicity	Toxicity	Toxicity	Toxicity	Toxicity

Dose Finding Benchmark

- *Ordinal outcome Y:* Takes values on $L+1$ possible values $\{w_0, w_1, \ldots, w_L\}$ with tail distribution $\pi(k)$ at dose k
- $Y_i(k)$ = Outcome for patient i at dose level k
- In simulation, randomly draw a tolerance profile: $U_{i1}, U_{i2}, \ldots U_{iL}$ iid Uniform(0,1)
- Generate complete outcome profile $Y_i(k)$ for patient i at dose level k as follows:
  - $Y_i(k) = w_l$ if $U_{i,l+1} > r_{l+1}(k)$ and $U_{ij} \le r_j$ for all $j=1,\ldots,l$
  - $r_j(k) = \pi_j(k) / \pi_{j-1}(k)$
- Nonparametric optimal $\pi^*(k)$ = average of $I\{Y_i(k) \ge w_l\}$

Simulation results

Scenario 3

Model	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
Desirability	-0.48	-0.13	0.22	**0.32**	-0.26
PO✔	0	0	20	**72**	7
CR	0	2	32	**49**	16
d(π*)	0	0	13	**85**	1

Scenario 4

Model	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
Desirability	0.12	0.29	0.45	0.58	**0.69**
PO	0	2	10	34	**54**
CR✔	0	0	1	5	**94**
d(π*)	0	0	0	5	**95**

Dose finding benchmark

- Features of a good benchmark:
  - Easy and quick to compute (not error prone)
  - Nonparametric (not favoring one parametric model over another)
  - Upper bound of accuracy for parametric methods
  - Sharp upper bound: warrant more work

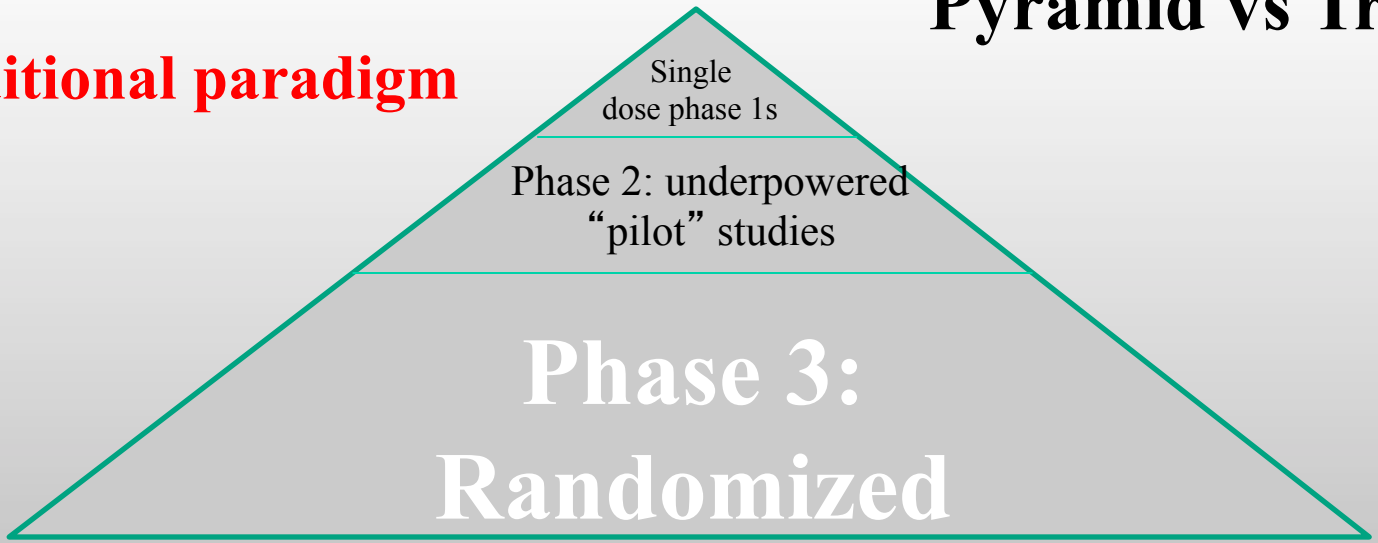
DISCUSSION

Summary

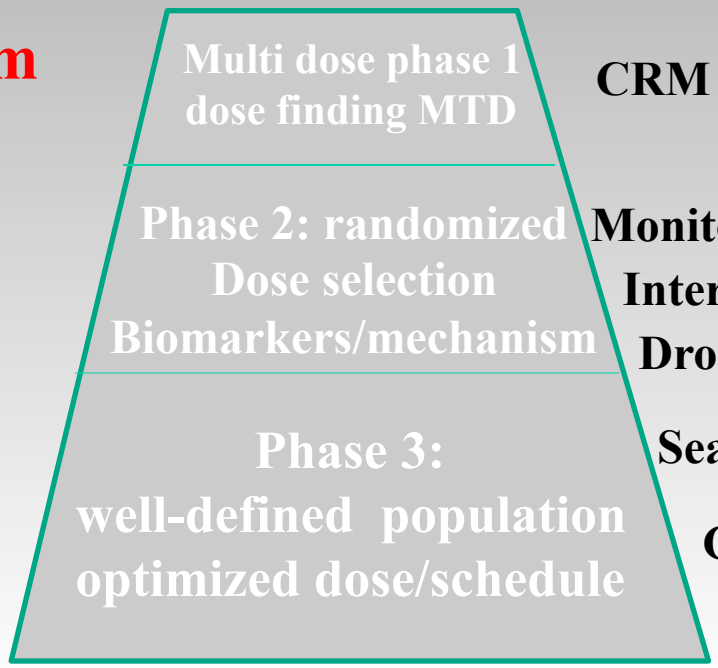
- Efficiency: AD is efficient from a portfolio management perspective
  - Funding agency buy-in
  - Incentive for individual investigators
- Success of study: AD often gains in terms of ethical costs, rather than economic costs
- Better understanding: AD gains by giving up something
  - Need to know exactly what is being given up
- More statistical work is needed on AD to make it more automated, accessible, and transparent

Pyramid vs Trapezoid

Traditional paradigm



"New" paradigm



CRM

Monitoring: SPRT/futility

Internal pilot

Drop-the-loser

Seamless phase 2/3

Group sequential

A D
D E
A S
P I
T G
I N
V S
E

REFERENCES

1. Cheung, Kaufmann: Efficiency perspectives on adaptive designs in stroke clinical trials. *Stroke* 2011;42,2990-4.
2. Cheung: Sample size formulae for the Bayesian CRM. *Clin Trials* 2013;10,852-61
3. Cheung: Simple benchmark for complex dose finding studies. *Biometrics* 2014 in press.
4. Elkind et al.: High dose lovastatin for acute ischemic stroke: results of the phase 1 dose escalation NeuSTART. *Cerebrovasc Dis* 2009;28,266-75.
5. FDA: Guidance for industry: adaptive design clinical trials for drugs and biologics. February 2010
6. Jia et al.: Characterisation of the likelihood CRM. *Biometrika* 2014 in press.
7. Lee, Cheung: Model calibration in the CRM. *Clin Trials* 2009;6,227-38.
8. Thall, Cook: Dose finding based on efficacy-toxicity tradeoffs. *Biometrics* 2004;60,684-93.