Bayesian Adaptive Clinical Trial Design

Jason Connor

ConfluenceStat

Jason@ConfluenceStat.com

412-860-3113

Day 1

Great Irony of Biostatistics

- Our job is to identify whether the newest, latest, greatest medical technologies are safe & efficacious and what works best for whom
  - Laser therapies, Whole genome diagnostics
  - Immunotherapies for cancer, etc

- Many statisticians believe our ‘technologies’ were as good as can be by 1933 or 1977 and nothing better can be invented

Great Irony of Biostatistics

- Donald Berry @ GBM AGILE kickoff: “Randomized clinical trials are 70 years old…what other technology doesn’t change in 70 years? Meanwhile, cancer biology is moving at light speed and potential treatments have to wait in the queue.”

- Take away: Realize the constraints (lack of) computing played on statistical methodology – and realize we are no longer constrained

Introductions

Plus reminder to self to confirm I’m recording

Decision Problem 1: Pandemic!

- A pandemic just hit the USA!!
- Patients are dying from a deadly disease
- 7-day survival rate is estimated to be less than 50% with standard care
- Patients who are alive at 7 days after initial symptoms typically have full recovery

Decision Problem 1: Pandemic!

- We need to determine best treatment of infected people
- Currently available therapies
  - Standard care with aforementioned ~50% mortality
  - 3 experimental anti-virals are ready to go
  - Each experimental arm is a novel anti-viral drug plus standard care
- Primary Endpoint:
  - Alive at 7 days after randomization (yes/no)

Allocation of Patients

- An effective treatment is any treatment that is better than standard care
- We will design the trial in stages, lets say we can enroll 80 patients per month
- You tell me where you want to assign patients
- I'll tell you how many on each drug survived

Interim Analyses

- At each interim analysis, you will receive efficacy data and will have to decide one of three things:
    1. Terminate the trial for futility, choose standard care as best option
    2. Stop the trial for success, choose optimal drug to treat all future patients
    3. Continue to collect data, allocating the next 80 patients to the four arms however you choose

Contest Points

- Team Competition
  - Each deceased patient costs 5 points
  - Every minute it takes to make a final decision costs 50 points (e.g., 20 minutes costs 1000 points)
  - If you claim a drug is superior to standard care (successful trial):
    1. If (in truth) the chosen drug is not superior to standard care, you lose 1,000 points
    2. If (in truth) the chosen drug is superior to standard care, you receive 2,000 points plus 200 for each % efficacy compared to control
  - If you claim standard care is best (futile trial):
    1. If (in truth) at least one of the drugs is superior to standard, you lose 1,000 points
    2. If (in truth) all drugs are not superior to standard, you receive 2,000 points

Instructions

- I'll create breakout rooms
- Talk among yourselves and decide how many patients (80 total) you would like to allocate to
  - Standard Care
  - Drug 1
  - Drug 2
  - Drug 3
- Aim for 3-4 minutes per iteration
- One member return to the main room and private message me with
  - Group Name, Patients to Placebo, Drug 1, Drug 2, Drug 3
  - For example "Group C: 20 20 20 20
- I'll write back your new total Deaths & N and % per group
- Repeat until you decide which is best or that none is better than standard care

Decisions

You private message to me, if you want 20 placebo & 20 on each drug:

Group C: 20 20 20 20

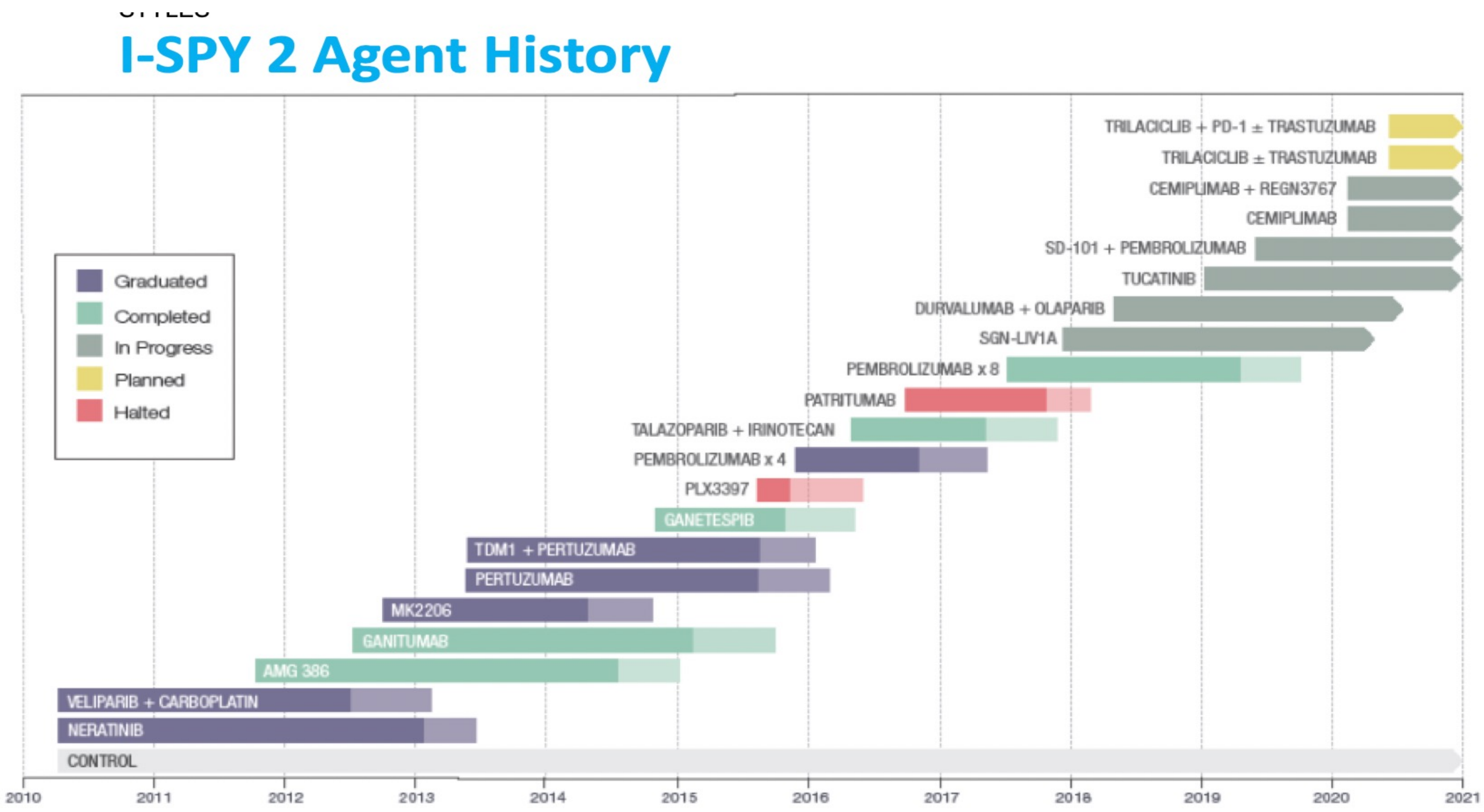
I private message back to you:

```
           N Alive PctAlive
Control 20    10      50%
Drug1   20     7      35%
Drug2   20    10      50%
Drug3   20    11      55%
Overall 80    38      48%
```

I will always give you THE SUM TOTALs so far

Go!

I-SPY2 in Breast Cancer



I-SPY 2 Agent History

Desirable Qualities of an RCT

- a

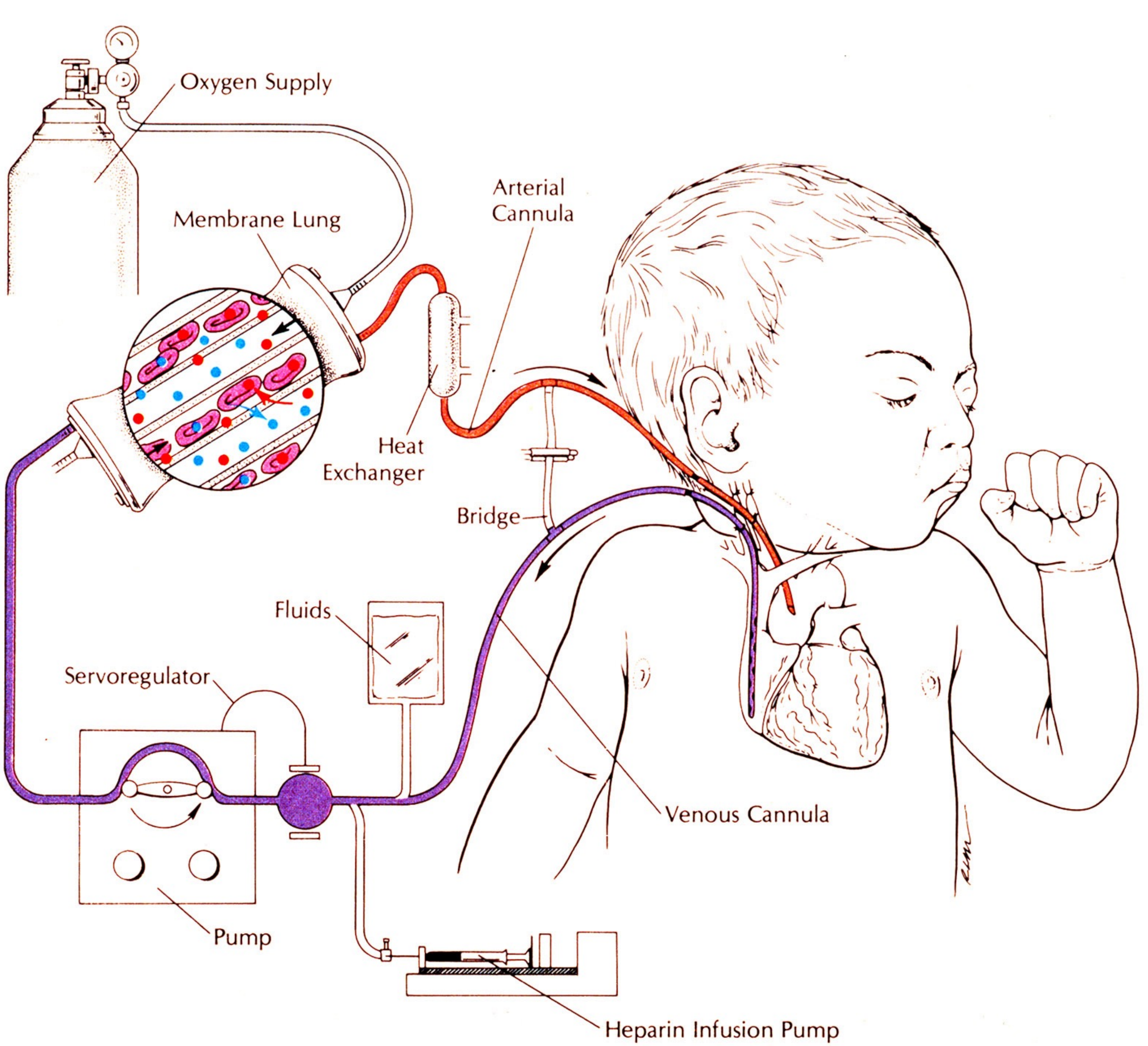
- b

Decision Problem #2

- New device to assist pre-mature infants
- Historical mortality rate >75%
- How to decide if new device is better than standard of care?

Decision Problem 2: ECMO

- Extracorporeal membrane oxygenation
- Oxygenates babies' blood & gives underdeveloped lungs & heart time to heal or grow
- Historical survival rates ≤ 25%
- Michigan trial: Randomized play the winner strategy
  - Bartlett, *Pediatrics,* 1985, 76: 479-487


Oxygen Supply
Membrane Lung
Arterial Cannula
Heat Exchanger
Bridge
Fluids
Servoregulator
Venous Cannula
Pump
Heparin Infusion Pump

Randomization Rules

- Randomize first patient 1:1 to treatment *t*
- If survives on treatment *t*, add 1 "*t*-colored" ball
- If dies on treatment *t*, add 1 other colored ball
- Treat 10 patients this way

- Expected number patients treated with better treatment > 5, "ethical"

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50				
2					
3					
4					
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO			
2					
3					
4					
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived		
2					
3					
4					
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived	1	2
2	0.67				
3					
4					
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived	1	2
2	0.67	CMT	Died	1	3
3	0.75				
4					
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived	1	2
2	0.67	CMT	Died	1	3
3	0.75	ECMO	Lived	1	4
4	0.80				
5					
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived	1	2
2	0.67	CMT	Died	1	3
3	0.75	ECMO	Lived	1	4
4	0.80	ECMO	Lived	1	5
5	0.83				
6					
7					
8					
9					
10					

ECMO Results

	Prob to			Balls in Urns	
	ECMO	TRT	Result	CMT	ECMO
Start				1	1
1	0.50	ECMO	Lived	1	2
2	0.67	CMT	Died	1	3
3	0.75	ECMO	Lived	1	4
4	0.80	ECMO	Lived	1	5
5	0.83	ECMO	Lived	1	6
6	0.86	ECMO	Lived	1	7
7	0.88	ECMO	Lived	1	8
8	0.89	ECMO	Lived	1	9
9	0.90	ECMO	Lived	1	10
10	0.91	ECMO	Lived	1	11

What Would You Decide?

- ECMO 9/9 CMT 0/1*

  * The 1 on CMT was the sickest of all patients

- As a statistician / clinical trialist do you have sufficient information to declare ECMO more efficacious than standard of care?

What Would You Decide?

- ECMO 9/9 CMT 0/1*

  * The 1 on CMT was the sickest of all patients

- As a statistician / clinical trialist do you have sufficient information to declare ECMO more efficacious than standard of care?

- As a parent would you dare *not* request ECMO for your premature baby?

Lessons of ECMO

- Questions the trials designers should have asked *before* the trial
  - How do we calculate a p-value?

Lessons of ECMO

- Questions the trials designers should have asked *before* the trial
  - How do we calculate a p-value?
  - Published p-values for this data (Stat Sci Nov 1989)

0.00049

0.001

0.003

0.009

0.038

0.045

undefined

0.051

0.083^{F}

0.280

0.500

0.617

1.000

Lessons of ECMO

- Questions the trials designers should have asked *before* the trial
  - How do we calculate a p-value?
  - Will the medical community believe our results?
    - Will we have enough data to sway opinions of people with a wide range of prior beliefs
  - What are trial results likely to look like?
  - What if everyone is randomized to ECMO?
    - If CMT success = 30% and ECMO success = 90%

      6% chance all 10 patients will be randomized to ECMO

Follow-Up Trials

- Harvard
  - Stage 1: randomize equally until 4 deaths in one arm
  - Stage 2: assign all to other arm until 4 deaths or stat sig.
  - 6/10 conventional therapy (60%)
  - 9/9 & 19/20 on ECMO (97%)
  - *Pediatrics,* 1989, 84: 957-963
- U.K
  - 63/93 on ECMO (68%)
  - 38/92 on conventional therapy (41%)

    Lancet, 1996, 348: 75-82
- Were these study designs ethical?
- Do we have an irrational commitment to blinded RCTs?
- Do we have an irrational commitment to $p < 0.05$?
- Does lack of $p<0.05$ mean equipoise until we see $p<0.05$?

Why are Study Designs (Usually) Fixed

- It’s easiest to calculate type I error rates if the design parameters of the trial are all constant
- Results obtained using “Standard approaches” are generally considered valid
- Logistically simpler to execute
- Fixed designs are less sensitive to drift in the characteristics of subjects over time
  - Fears worse than reality
- We could do the math 40 years ago
  - We still can but we can also do more sophisticated things now too

Why are Study Designs (Usually) Fixed

- It's easiest to calculate type I error rates if the design parameters of the trial are all constant
- Results obtained using "Standard approaches" are generally considered valid
- Logistically simpler to execute
- Fixed designs are less sensitive to drift in the characteristics of subjects over time
  - Fears worse than reality
- We could do the math 40 years ago
  - We still can but we can also do more sophisticated things now too

But now we can simulate trials &
do much more complicated calculations

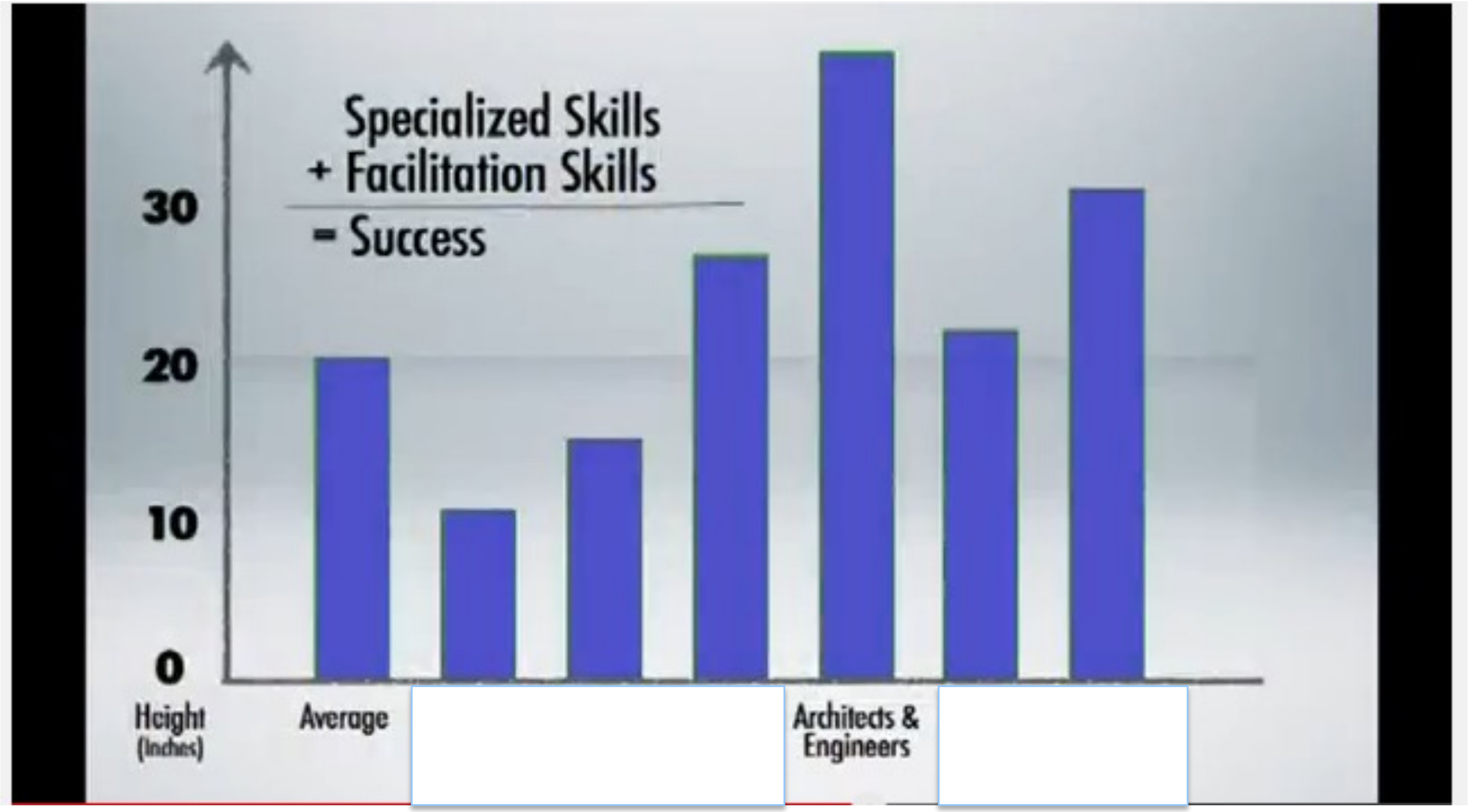
Digression:
The Marshmallow Design Challenge

The Marshmallow Design Challenge
Peter Skillman

- 4-person team
- 18 minutes
- 20 pieces of raw spaghetti
- 1 meter of tape
- 1 meter of string
- 1 marshmallow

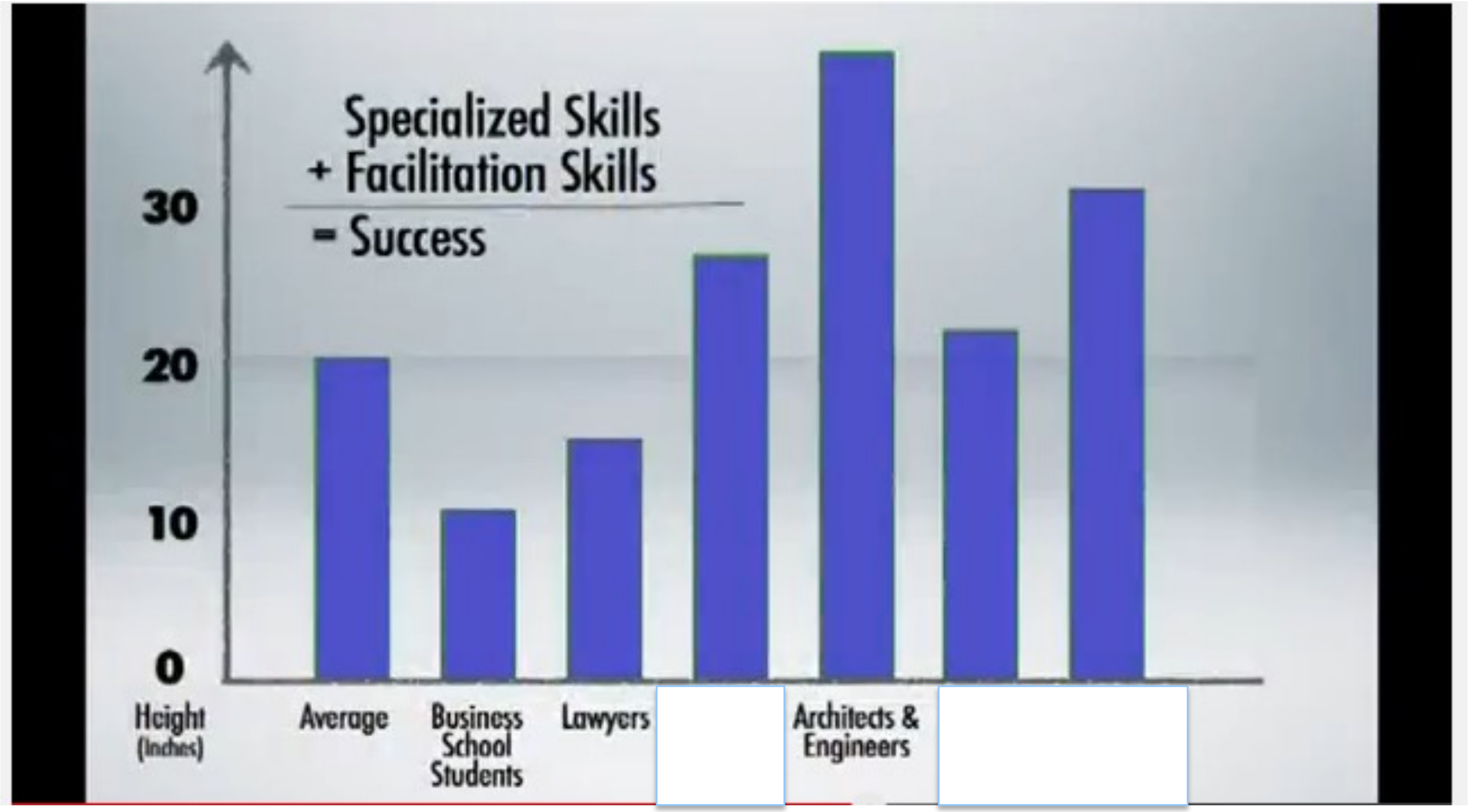
Peter Skillman Marshmallow Design Challenge
https://www.youtube.com/watch?v=1p5sBzMtB3Q

The Marshmallow Design Challenge



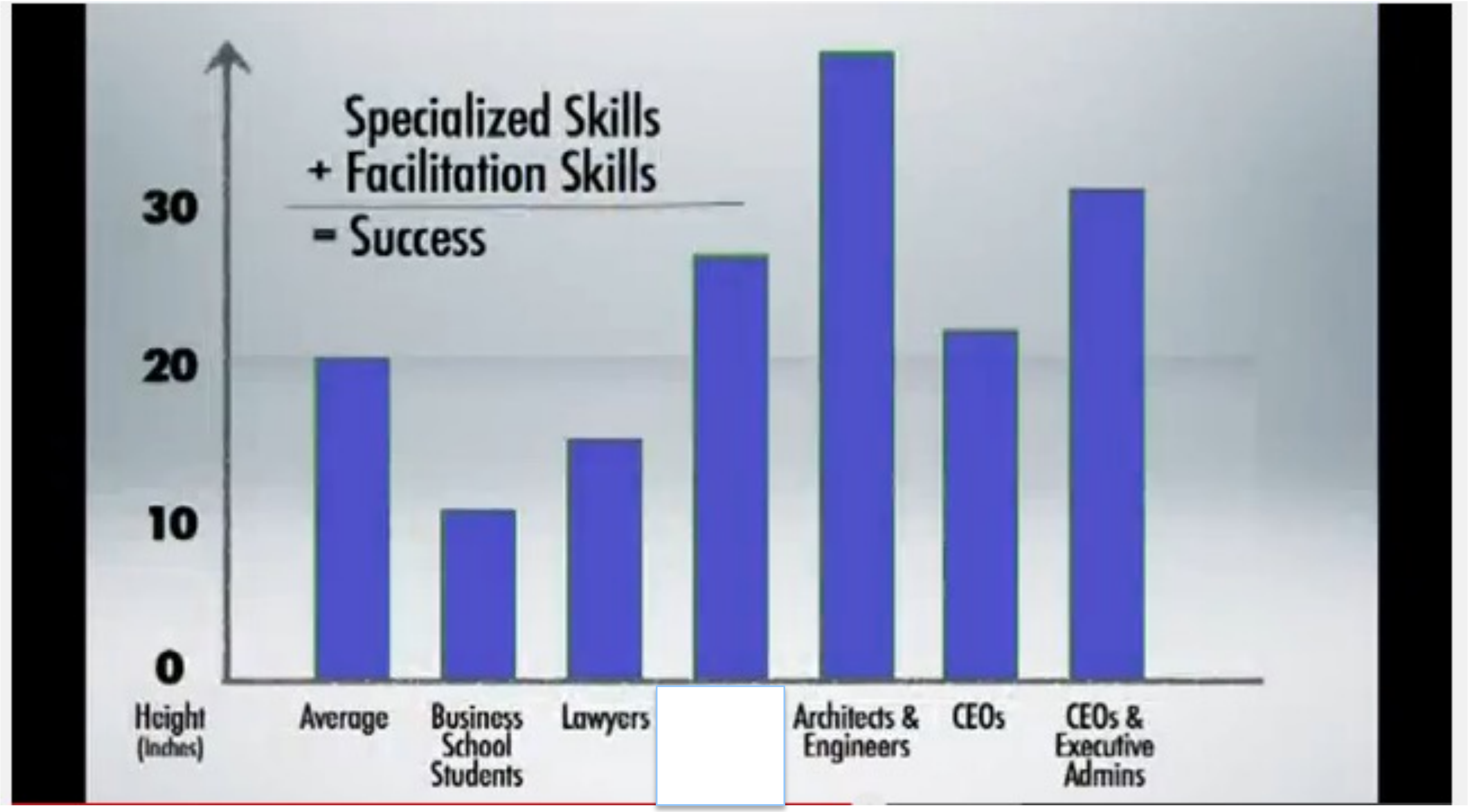
Tom Wujec: Build a tower, build a team.
https://www.youtube.com/watch?v=H0_yKBitO8M

The Marshmallow Design Challenge



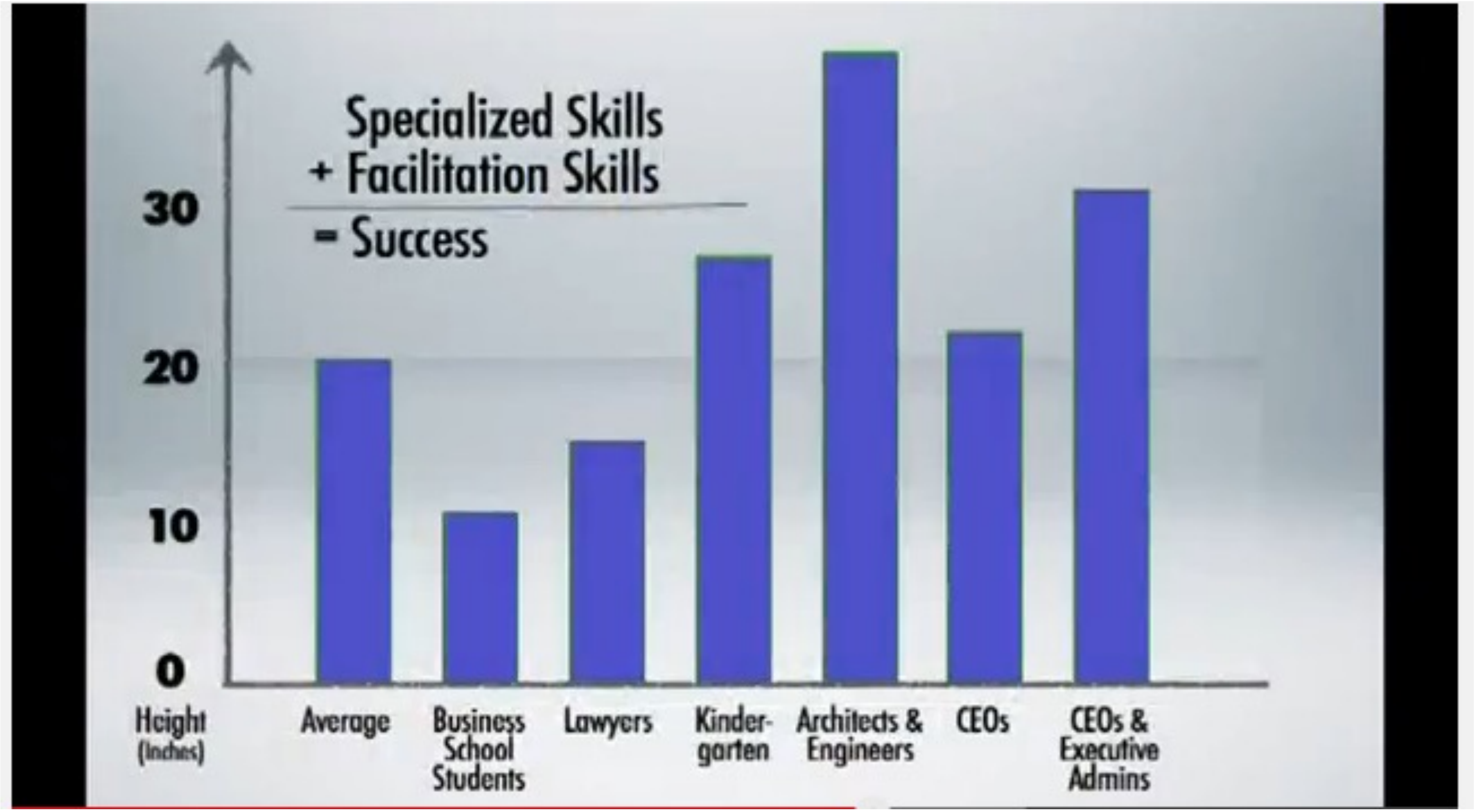
Tom Wujec: Build a tower, build a team.
https://www.youtube.com/watch?v=H0_yKBitO8M

The Marshmallow Design Challenge



Tom Wujec: Build a tower, build a team.
https://www.youtube.com/watch?v=H0_yKBitO8M

The Marshmallow Design Challenge



Tom Wujec: Build a tower, build a team.
https://www.youtube.com/watch?v=H0_yKBitO8M

The Marshmallow Design Challenge Peter Skillman

- Kindergartners
  - Don’t waste time seeking power
  - Don’t sit around talking about the problem
  - Try, fail, try, fail until time runs out
  - They all grab stuff and try things
  - Usually keep the marshmallow on top when trying

- MBA grads
  - Spend a lot of time talking
  - Trained to find single best plan
  - Trained never to fail
  - Last thing they do it put the marshmallow on top (and often watch the whole tower collapse)

The Marshmallow Design Challenge
Peter Skillman

- You learn by doing and failing & redoing
  - With Simulation we can do this cheap, fast, and ethically!
- Work in parallel
- Doing multiple iterations is good
- All projects have resource constraints

ECMO: Trial & Error Design by Simulation

```
p.ecmo <- 0.75;      p.cmt <- 0.25

group.vec <- NULL;    outcome.vec <- NULL
outcome <- matrix(nrow=100000, ncol=5)

for(s in 1:100000){
urn <- c(1,1)
for(pt in 1:10){
  group <- sample(c("C","E"), 1, prob=urn)
  result <- rbinom(1, 1, ifelse(group=="C",p.cmt, p.ecmo))
  if(group=="C"){
      if(result==1){
           urn[1] <- urn[1] + 1
      }else{
           urn[2] <- urn[2] + 1
      }
  }else{
      if(result==1){
           urn[2] <- urn[2] + 1
      }else{
           urn[1] <- urn[1] + 1
      }
}
group.vec[pt] <- group
outcome.vec[pt] <- result
}
tab <- table(factor(group.vec, levels=c("C","E")), factor(outcome.vec,
levels=0:1))
outcome[s,] <- c(c(tab), fisher.test(tab, alternative='greater')$p.value)
print(s)
}
```

```
### Pr no patients on control
mean((outcome[,1]+outcome[,3]) == 0)
### Pr no patients on ECMO
mean((outcome[,2]+outcome[,4]) == 0)
### Pr more on ECMO than control
mean((outcome[,1]+outcome[,3]) < (outcome[,2]+outcome[,4]))
### Pr more equal on each
mean((outcome[,1]+outcome[,3]) == (outcome[,2]+outcome[,4]))
### Pr more on control than ECMO
mean((outcome[,1]+outcome[,3]) > (outcome[,2]+outcome[,4]))

### More ECMO than control success
mean((outcome[,3]) < (outcome[,4]))
### 4 or more ECMO than control successes
mean((outcome[,3] + 4) <= (outcome[,4]))
```

ECMO: Prospective Simulation

Operating Characteristics	CMT 25% ECMO 75%	CMT 25% ECMO 25%
Pr(All patients randomized to ECMO)	2.5%	0.04%
Pr(All patients randomized to CMT)	0.04%	0.04%
Pr(Majority to ECMO)	72%	36%
Pr(5 ECMO & 5 CMT)	14%	27%
Pr(Majority to CMT)	14%	36%

Pr(Fisher P-value < 5%)
Pr(Chi-square P-value < 5%)
	12%
	32%
		0.1%
		1.9%

Pr(# ECMO Success > # CMT Successes)	89%	38%
Pr(# ECMO Success ≥ # CMT Success + 4)	59%	2.7%

ECMO: Prospective Simulation

Operating Characteristics	CMT 25% ECMO 75%	CMT 25% ECMO 25%
Pr(All patients randomized to ECMO)	2.5%	0.04%
Pr(All patients randomized to CMT)	0.04%	0.04%
Pr(Majority to ECMO)	72%	36%
Pr(5 ECMO & 5 CMT)	14%	27%
Pr(Majority to CMT)	14%	36%

Pr(Fisher P-value < 5%)
Pr(Chi-square P-value < 5%)
	12%
	32%
		0.1%
		1.9%

Pr(# ECMO Success > # CMT Successes)	89%	38%
Pr(# ECMO Success ≥ # CMT Success + 4)	59%	2.7%

Power

Type I error

ECMO: Prospective Simulation

Operating Characteristics	CMT 25% ECMO 75%	CMT 25% ECMO 25%
Pr(All patients randomized to ECMO)	2.5%	0.04%
Pr(All patients randomized to CMT)	0.04%	0.04%
Pr(Majority to ECMO)	72%	36%
Pr(5 ECMO & 5 CMT)	14%	27%
Pr(Majority to CMT)	14%	36%

Pr(Fisher P-value < 5%)
Pr(Chi-square P-value < 5%)
	12%
	32%
		0.1%
		1.9%

Pr(# ECMO Success > # CMT Successes)	89%	38%
Pr(# ECMO Success ≥ # CMT Success + 4)	59%	2.7%

Power

Type I error

ECMO Iterate Design

N	Decision Rule # ECMO Successes vs. # CMT Successes	Power when ECMO 75% CMT 25%	Type I error ECMO 25% CMT 25%
10	1 or more	89%	38%
10	4 or more	59%	2.7%
10	3 or more	72%	8.1%

ECMO Iterate Design

N	Decision Rule # ECMO Successes vs. # CMT Successes	Power when ECMO 75% CMT 25%	Type I error ECMO 25% CMT 25%
10	4 or more	59%	2.7%
10	3 or more	72%	8.1%
15	4 or more	79%	5.9%
15	5 or more	71%	2.3%

ECMO Iterate Design

N	Decision Rule # ECMO Successes vs. # CMT Successes	Power when ECMO 75% CMT 25%	Type I error ECMO 25% CMT 25%
10	4 or more	59%	2.7%
10	3 or more	72%	8.1%
15	4 or more	79%	5.9%
15	5 or more	71%	2.3%
16	4 or more	82%	6.7%
16	5 or more	74%	2.8%

ECMO Iterate Design

N	Decision Rule # ECMO Successes vs. # CMT Successes	Power when ECMO 75% CMT 25%	Type I error ECMO 25% CMT 25%
10	4 or more	59%	2.7%
10	3 or more	72%	8.1%
15	4 or more	79%	5.9%
15	5 or more	71%	2.3%
16	4 or more	82%	6.7%
16	5 or more	74%	2.8%
18	5 or more	80%	3.5%

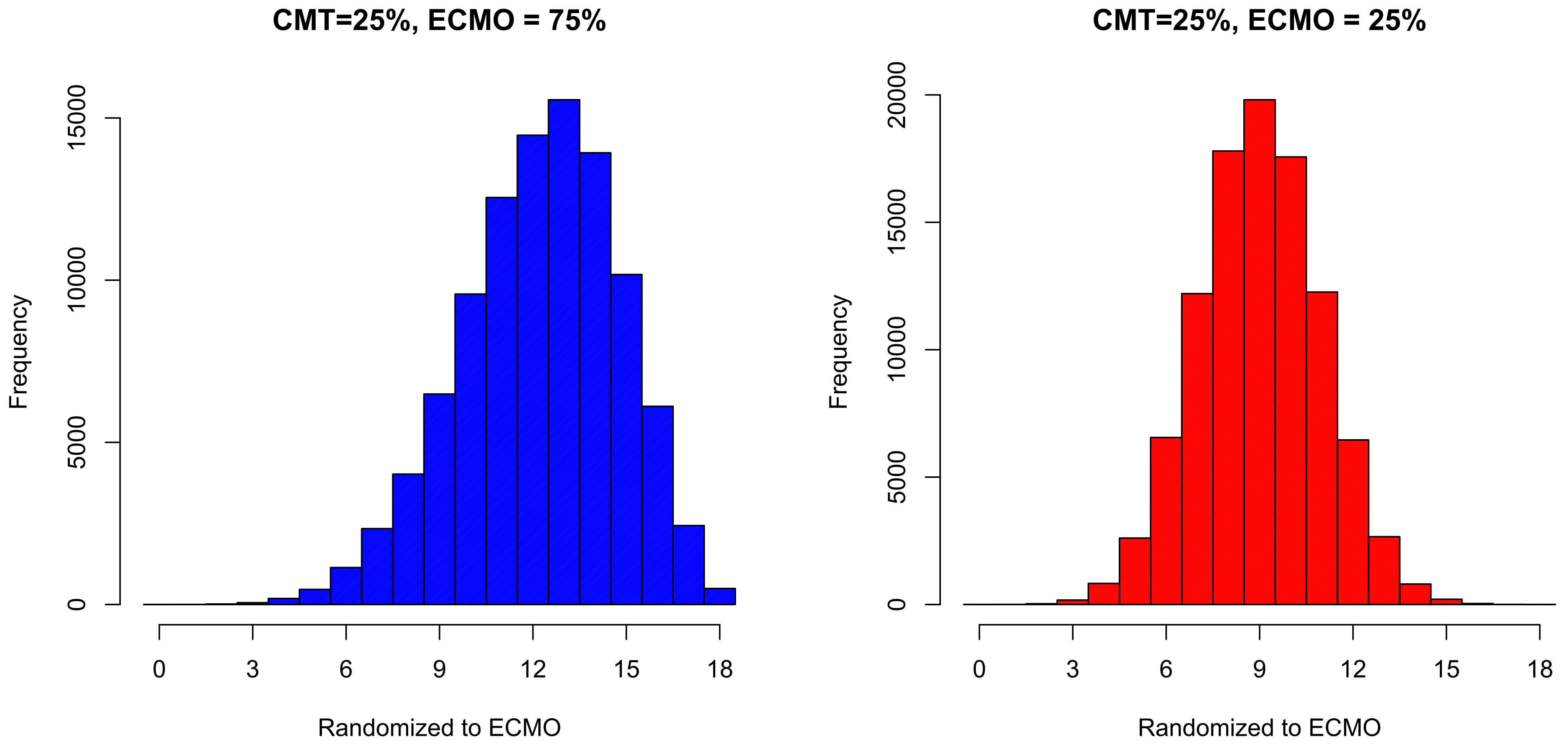
Fisher's exact test: 59% power @ 1-sided 5.0%.

ECMO Iterate Design

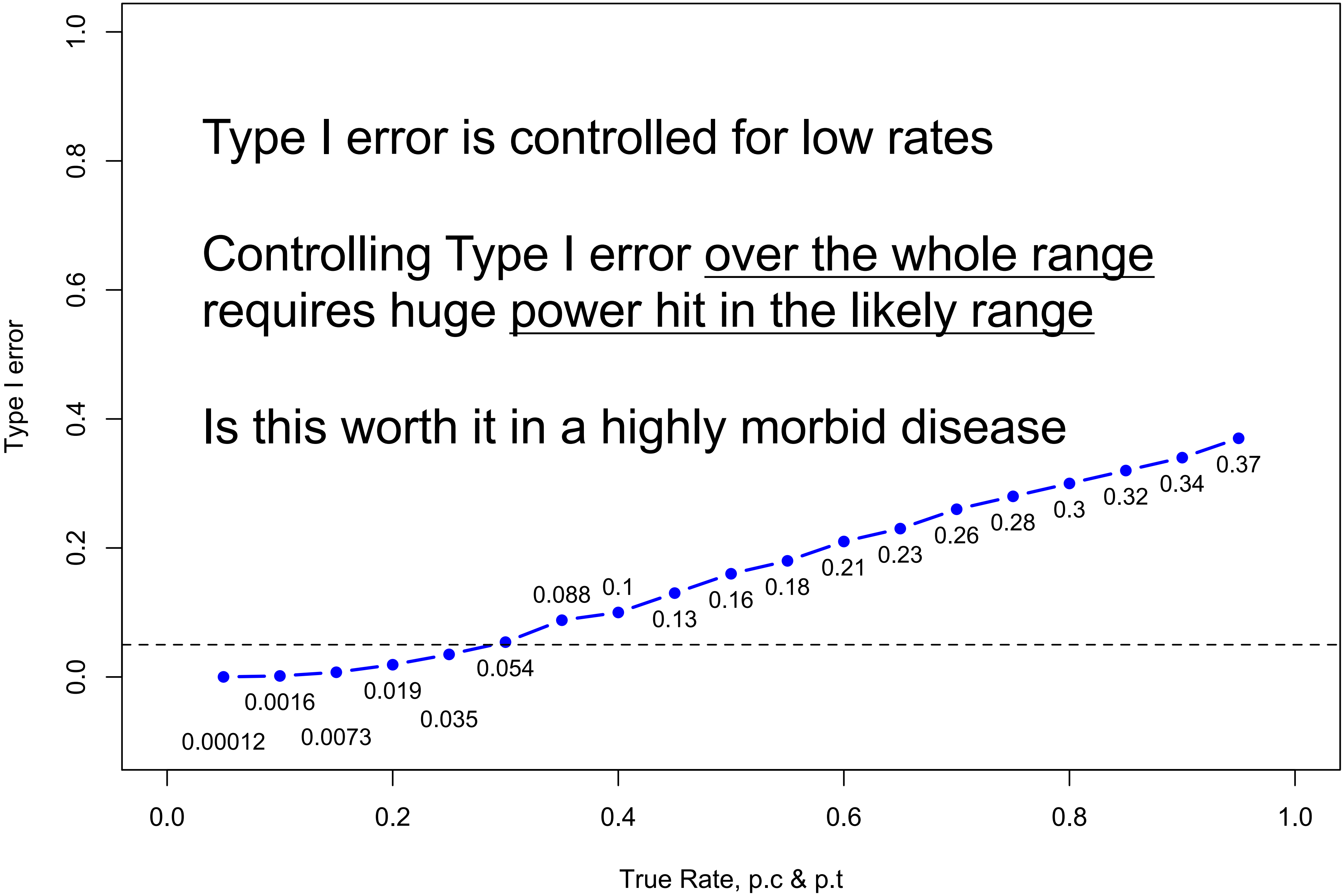
N	Decision Rule ECMO v CMT	Power 75v25	ECMO S/N	CMT S/N	T1error 25v25	ECMO S/N	CMT S/N
10	4 or more	59%	4.9 / 6.5	0.9 / 3.5	2.7%	1.25 / 5	1.25 / 5
10	3 or more	72%			8.1%		
	8 more patients		5.7 more	2.3 more		4 more	4 more
18	5 or more	80%	9.2 / 12.2	1.4 / 5.8	3.5%	2.25 / 9	2.25 / 9

Standard trial with 18 patients has 58% power with 5% Type I error
Always randomized half to CMT; E(survive) = 10.6 vs. 9

ECMO with 18 patients



No free lunch



Type I error is controlled for low rates

Controlling Type I error over the whole range requires huge power hit in the likely range

Is this worth it in a highly morbid disease

When designing trials I believe we should

- Remember that most ‘standard’ statistical methods were developed for agriculture
- Remember that current trialists were trained by people who were trained by people who had seeds as patients
- Remember most statistical methodology is based on asymptotic theory
  - Because we couldn’t do math then that we can do it now
- Forget much of what we know about clinical trials & hypothesis testing & asymptotic theory
- Hire smart people with their heart in the right place
- Balance treating the next patient well & producing valuable long-term evidence
- Think much harder about the ‘right’ Type I error rate
  - Nothing sacred about 0.05
- Design trials by trial & error by using simulation, iterate designs with doctors, patients, payers, regulators
- Not let within-trial patient benefit be a side effect of quality research

Part 2

What are Adaptive Trials?

Trials in which key design parameters change during trial execution based upon *a priori* predefined rules and accumulating data from the trial to achieve goals of validity, scientific efficiency, and safety

- Planned: All possible adaptations defined *a priori*
- Well-defined: Criteria for adapting clearly explained
- Key parameters: *Not* minor inclusion or exclusion criteria, routine amendments, etc.
- Validity: Reliable statistical inference

What are Adaptive Designs?

- Adaptive Design:
  - A design that “changes” depending on observed values in the trial
- Prospective Adaptive Design:
  - A design that has pre-specified dynamic aspects that are determined by the accruing information

Every time I say “Adaptive Design” I mean “Prospectively Adaptive Design”

What are Adaptive Trials?

Trials that change based on prospective rules & the accruing information

- Adaptive sample sizes based on predictive probabilities
  - Stop early for success
  - Terminate early for futility
- Adaptive randomization
  - For statistical efficiency
  - For improved patient treatment
  - Drop/Re-enter arms or dose groups
- Adaptive accrual rate
- Combination therapies
- Adapt to responding sub-populations
- Adaptive borrowing of information
- Seamlessly combine phases of development
  - Phase 2/3 designs: Operationally vs. Inferentially seamless

When is Adaptation Most Valuable

- Outcomes or biomarkers available rapidly relative to time required for entire trial
- Substantial morbidity, risks, costs
- Large uncertainty regarding relative efficacy, adverse event rates, variability, patient population in trial, etc.
- Logistically practical
- Able to secure buy-in of stakeholders

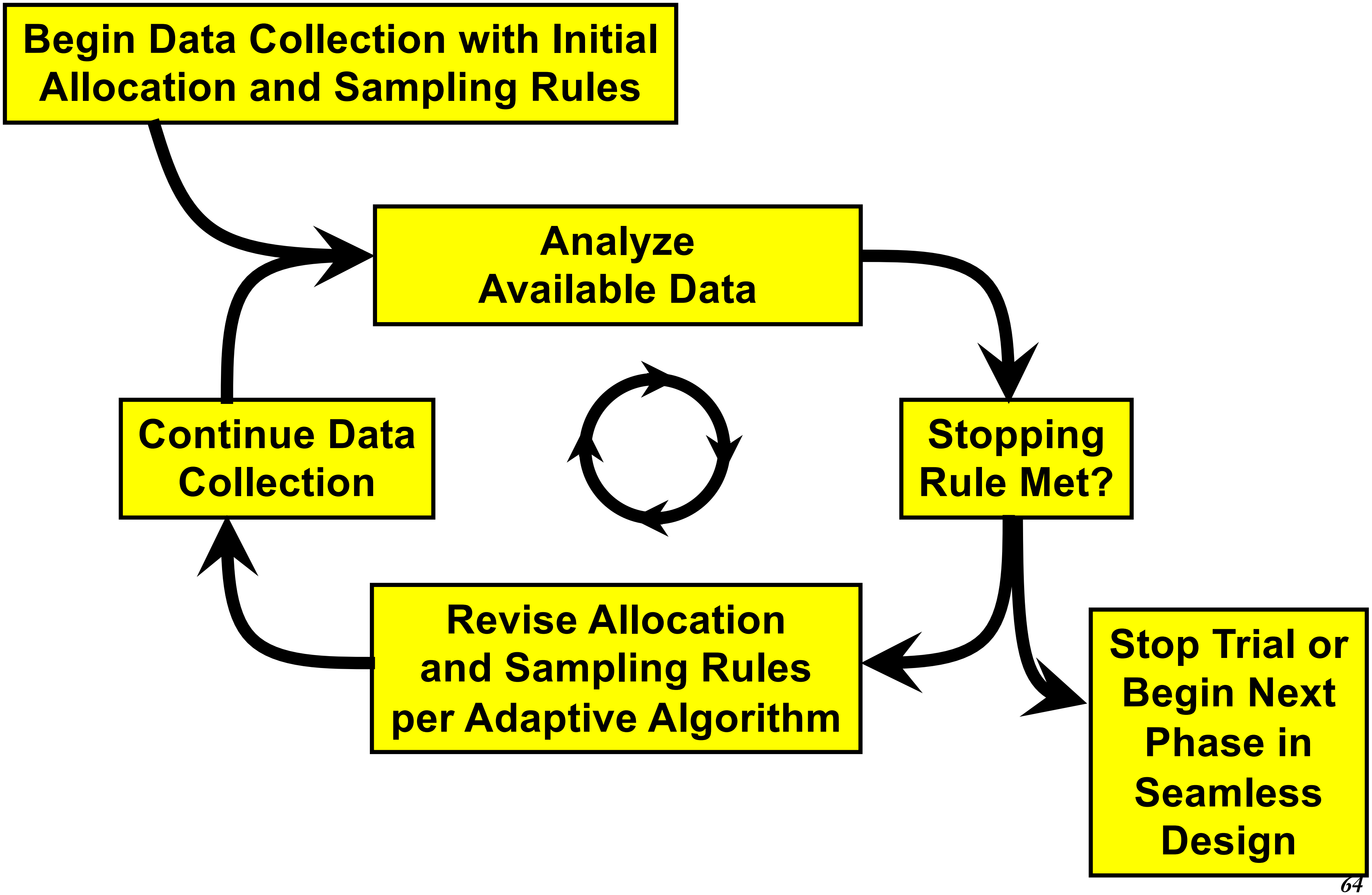
Drawbacks of Adaptation

- Infeasible if time from patient accrual to final outcomes long vs. total accrual time
- Adaptive design take much more forethought & buy-in from more stakeholders
- Determining traditional Type I and II error rates more difficult
  - Rely on simulation
- People fear new
  - Most statisticians have never designed or analyzed an adaptive trial
  - Some regulatory personnel unfamiliar with
  - Funders (e.g. venture capitalists and NIH) unfamiliar with
  - DMCs / IRBs may not understand
  - Clinicians may not understand

Drawbacks of Adaptation

- Logistical issues
  - Design stage is longer
  - Data needs to be entered & transmitted quickly
  - Data needs to be checked / validated quickly
  - Events need to be adjudicated quickly
  - Drug supply concerns for adaptive randomization
    - Fear of unblinding
  - Need centralized randomization
    - Use web or phone systems
  - Need to have lots of people / systems well & correctly connected

Typical Prospective Adaptive Design



Begin Data Collection with Initial Allocation and Sampling Rules

Analyze Available Data

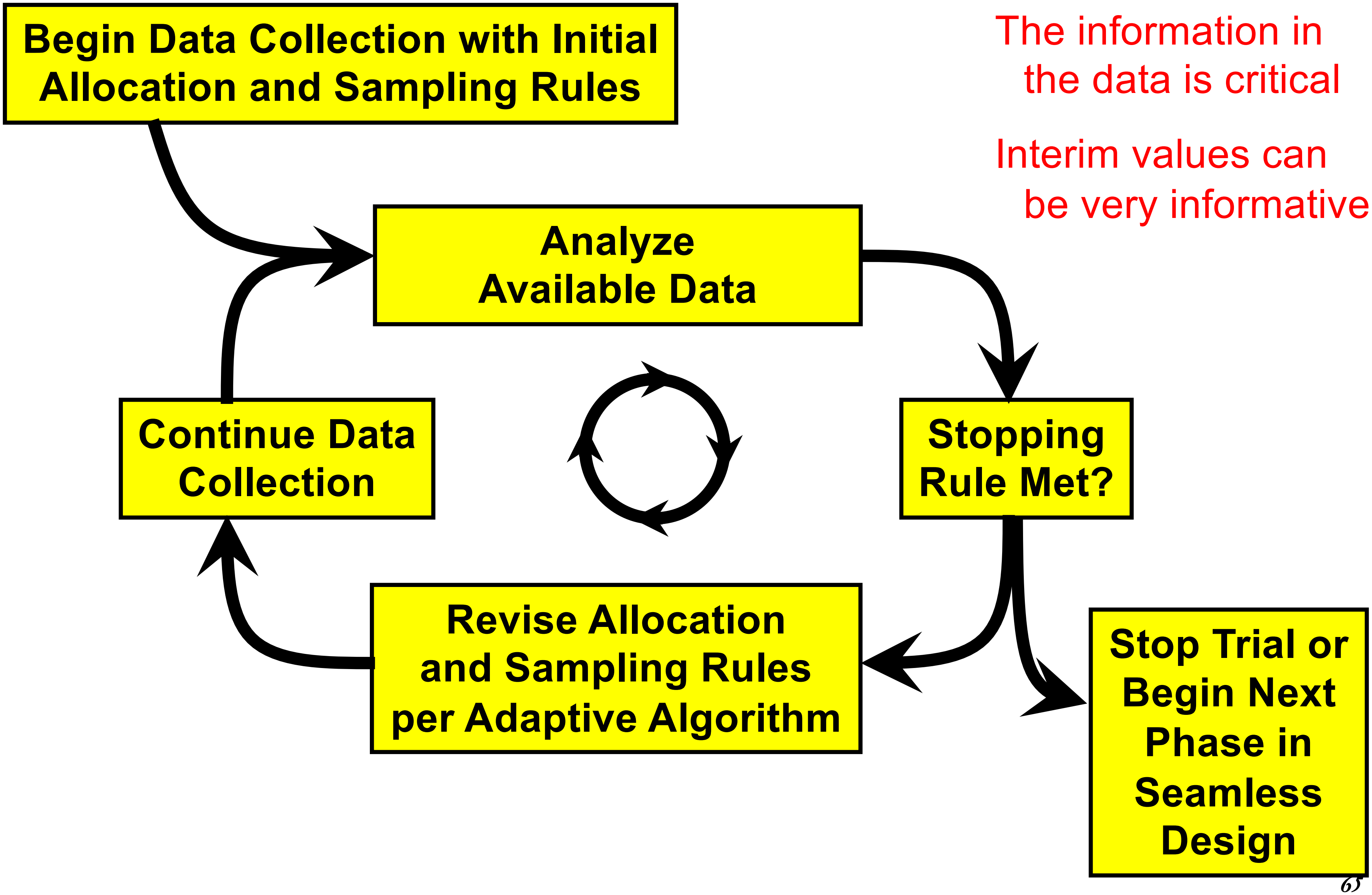
Stopping Rule Met?

Stop Trial or Begin Next Phase in Seamless Design

Revise Allocation and Sampling Rules per Adaptive Algorithm

Continue Data Collection

Typical Prospective Adaptive Design



Begin Data Collection with Initial Allocation and Sampling Rules

Analyze Available Data

Stopping Rule Met?

Revise Allocation and Sampling Rules per Adaptive Algorithm

Continue Data Collection

Stop Trial or Begin Next Phase in Seamless Design

The information in the data is critical

Interim values can be very informative

JAMA 2006;296:1955-1957.

Who To Involve

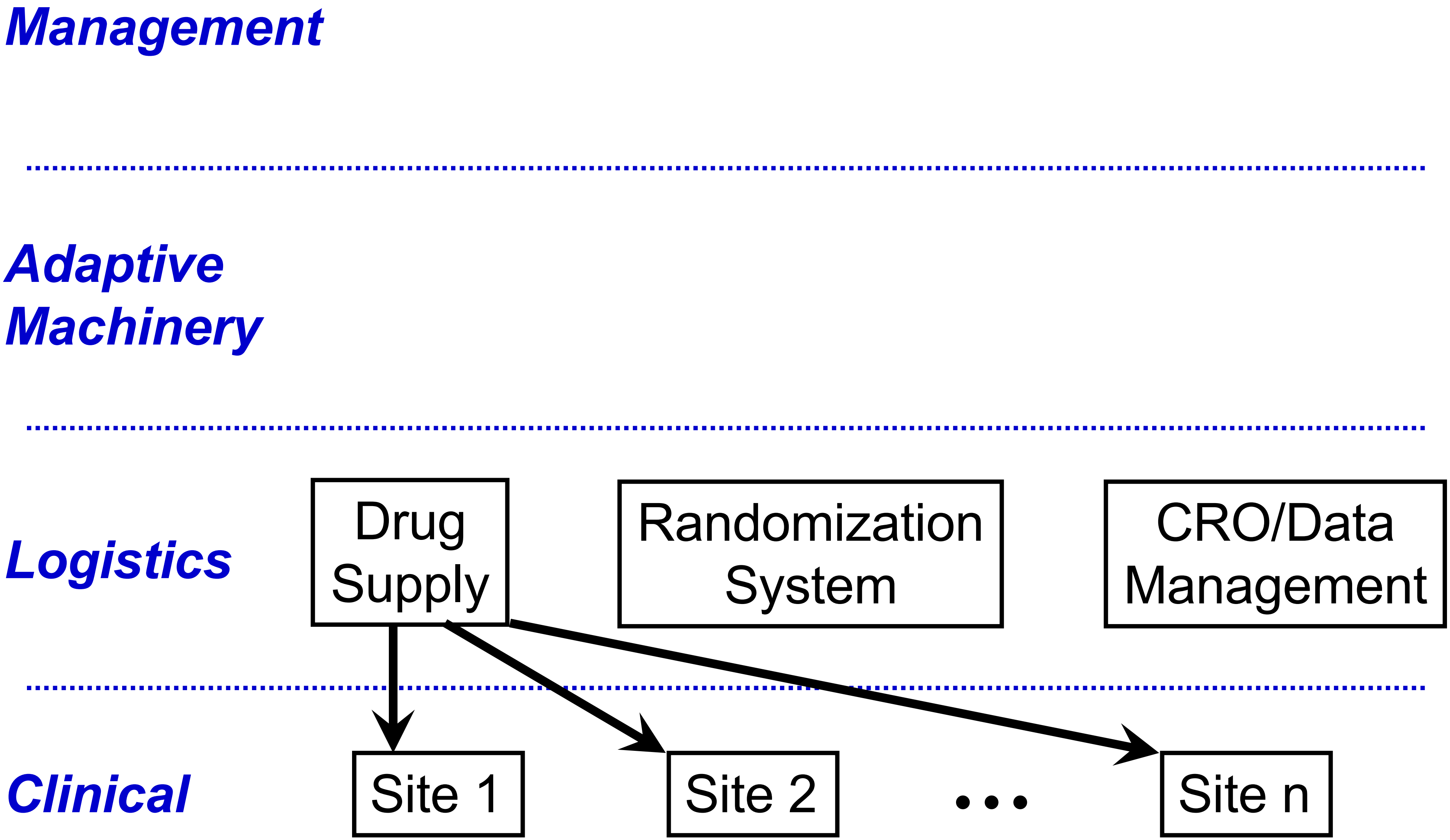
- Sponsor
  - Project leaders
  - Statisticians
  - PK/PD
  - Clinical experts
  - Business leaders
  - Patient advocates
- Clinical site IRBs
- Data Safety Monitoring Board
- IVRS/IWRS service
- CRO who will house data
- Regulatory agencies
- Patient advocacy groups?
  - Treat patients in trial best vs. get drug to market sooner?
- Payers

Adaptive Designs & Collaborators

- Requires buy-in and educating IRB, DSMB, decision-makers, study teams, investigators, and subjects
- Requires more time, resources, and upfront planning, especially at the protocol-design stage
- Show sponsor many many example trials
  - Also great for debugging
- Complex study designs typically require more statistical assumptions, rigorous calculations, and extensive simulations (operating characteristics)
- But also more robust to deviations from our assumptions
- Operationally challenging
  - Work with CROs as early as possible, fit statistical parts within infrastructure
- Make sure sponsors understands what adaptive designs are not

Components of an Adaptive Trial

Thanks Roger Lewis



Management

Adaptive Machinery

Logistics

Drug Supply

Randomization System

CRO/Data Management

Clinical

Site 1

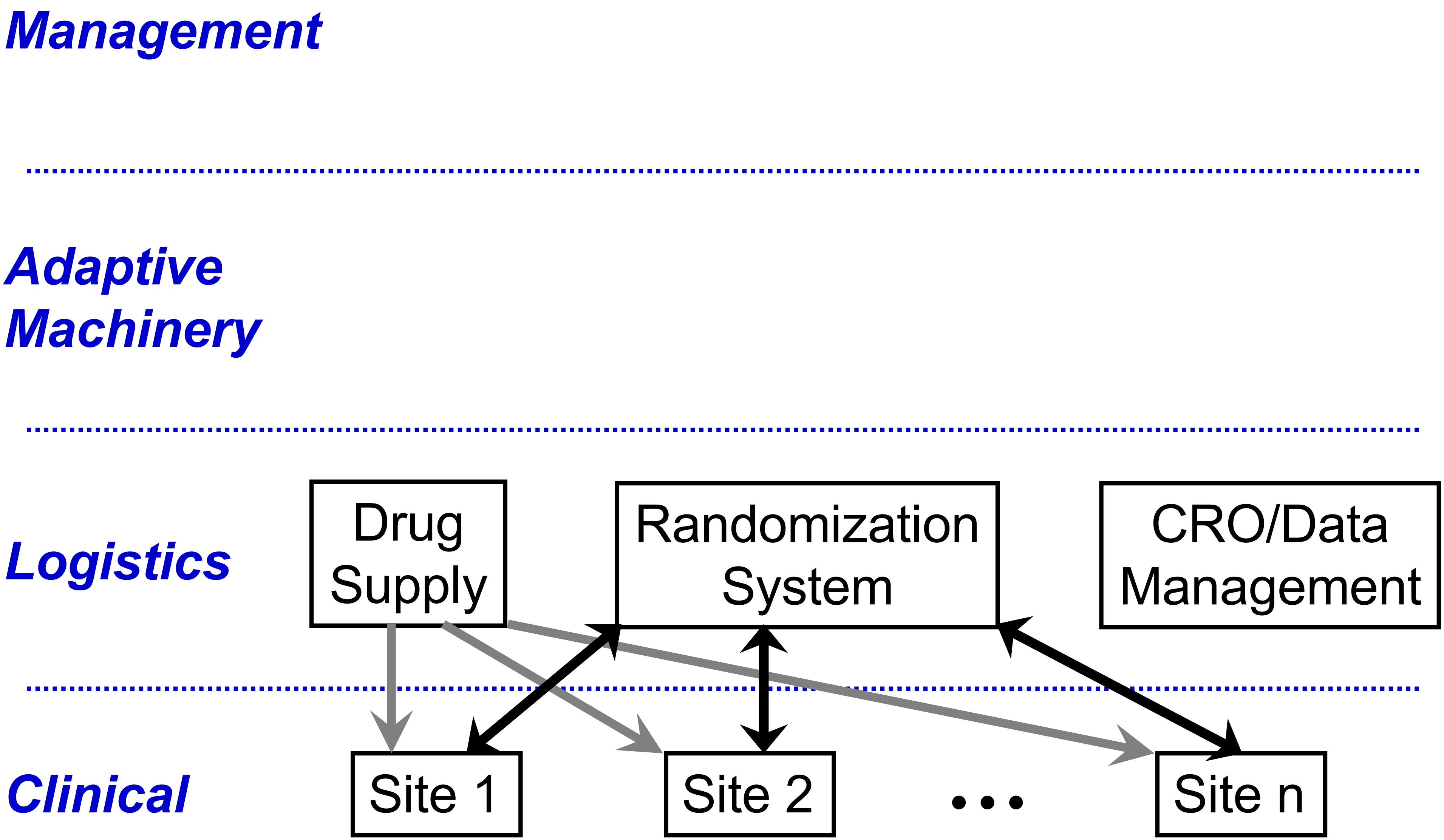
Site 2

• • •

Site n

Components of an Adaptive Trial

Thanks Roger Lewis



Management

Adaptive Machinery

Logistics

Drug Supply

Randomization System

CRO/Data Management

Clinical

Site 1

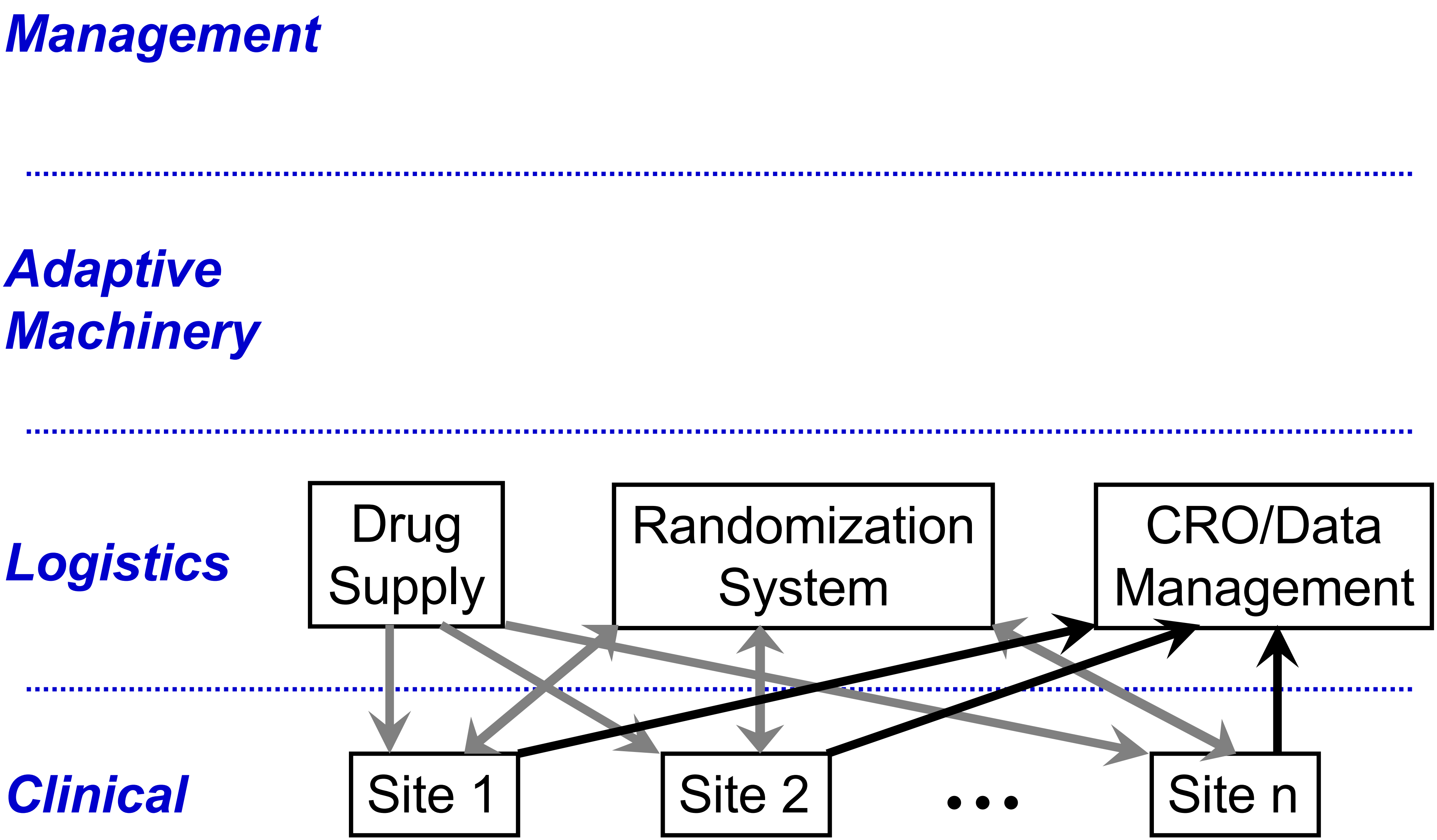
Site 2

• • •

Site n

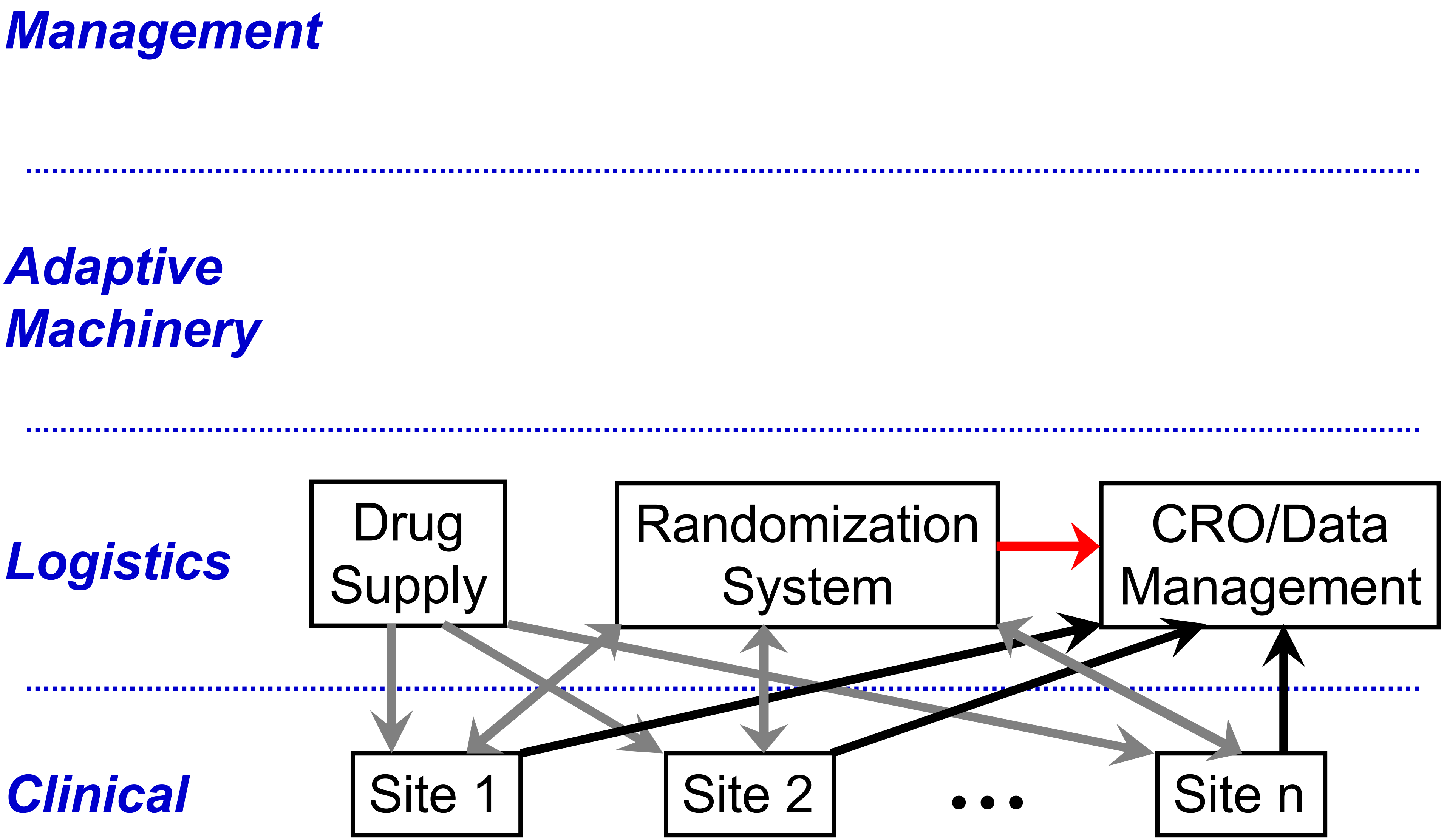
Components of an Adaptive Trial

Thanks Roger Lewis



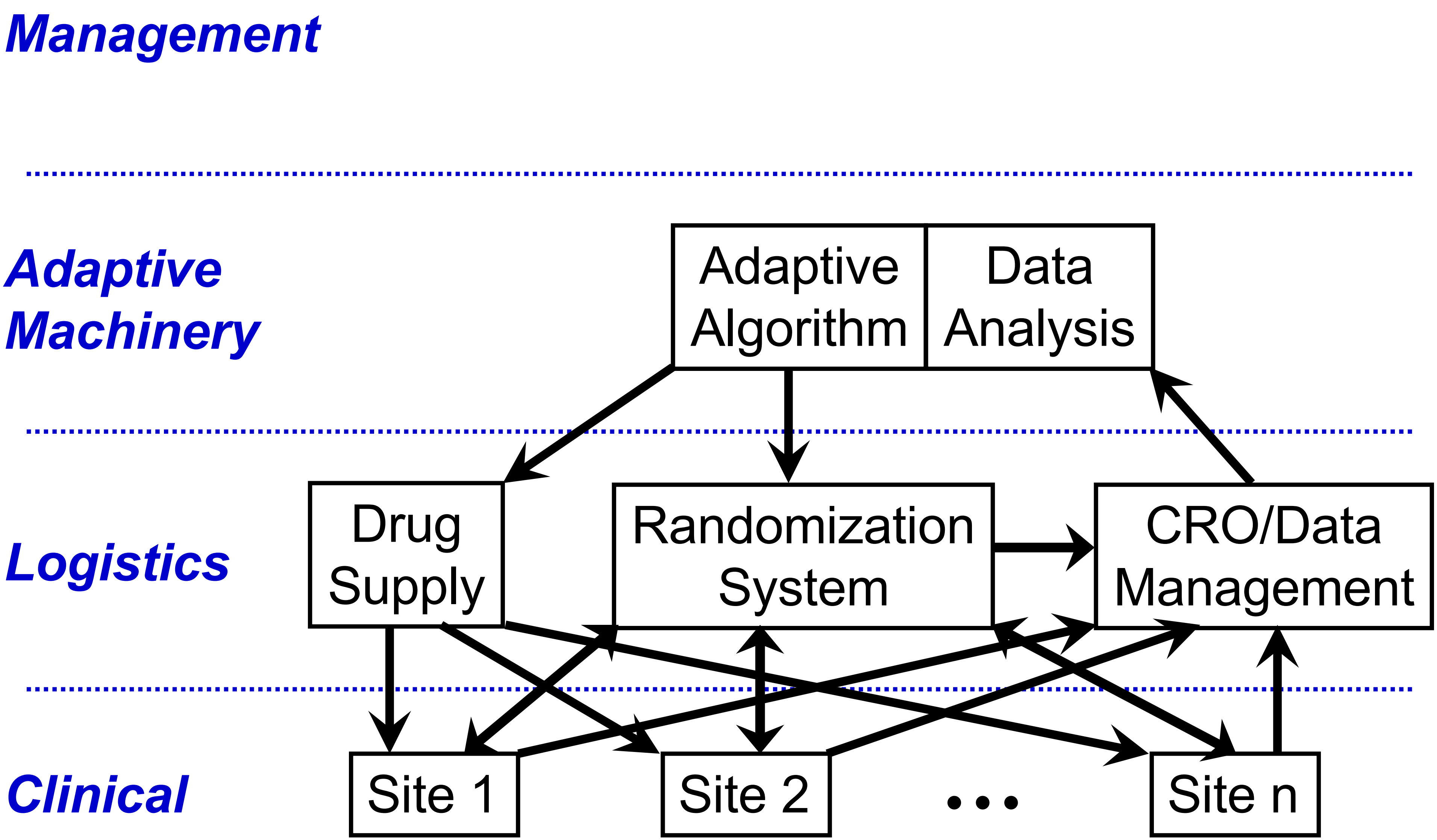
Components of an Adaptive Trial

Thanks Roger Lewis



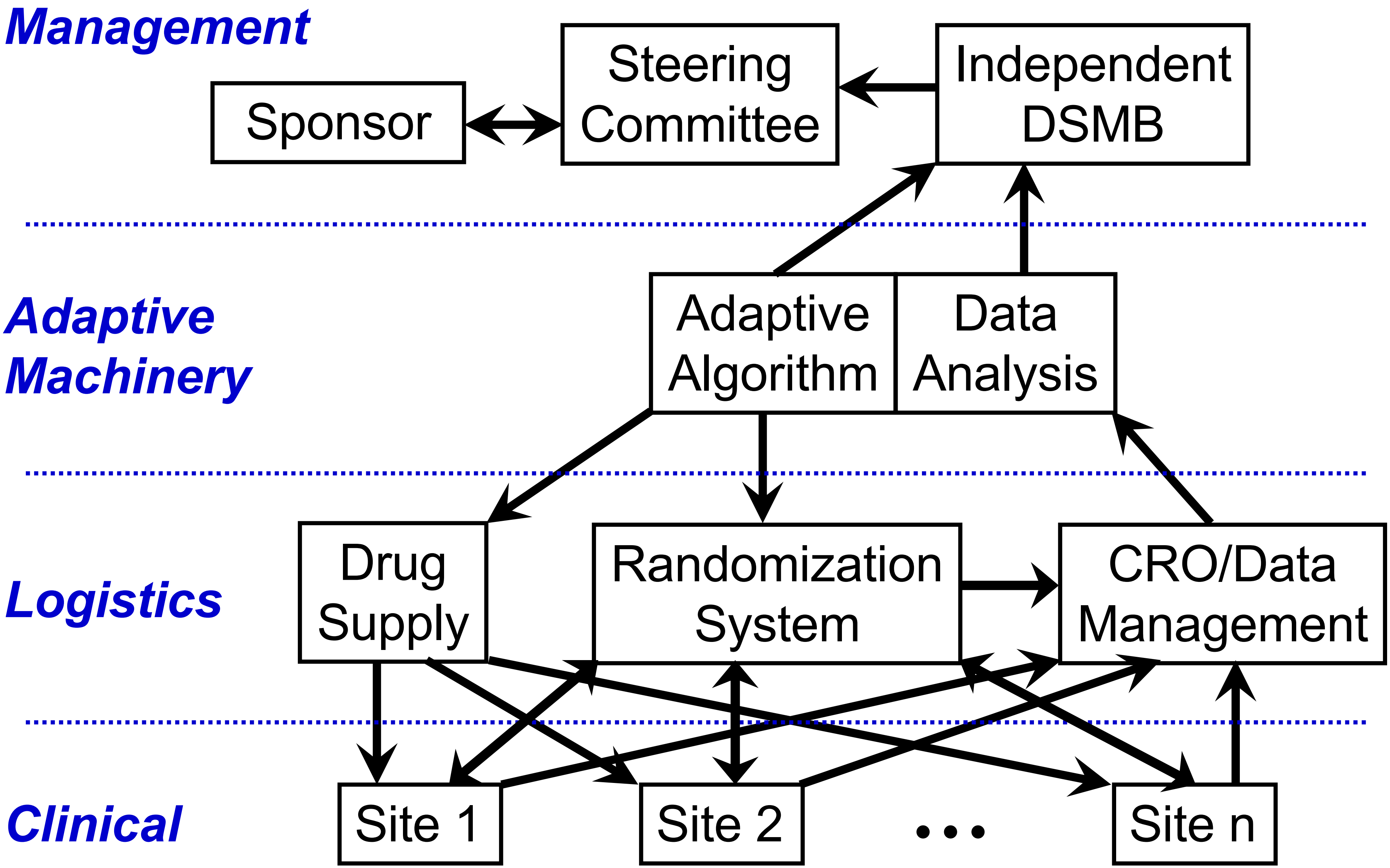
Components of an Adaptive Trial

Thanks Roger Lewis



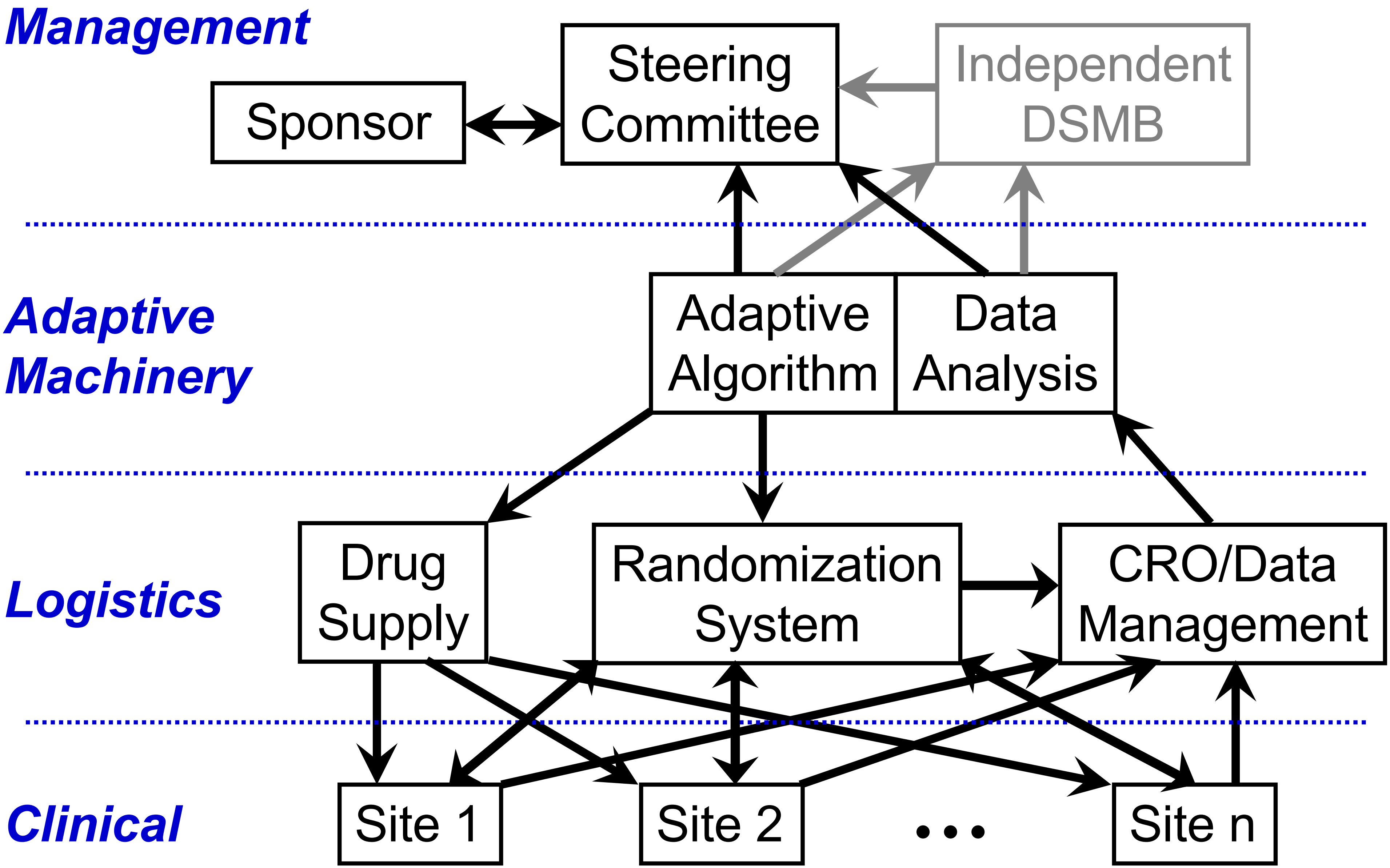
Components of an Adaptive Trial

Thanks Roger Lewis



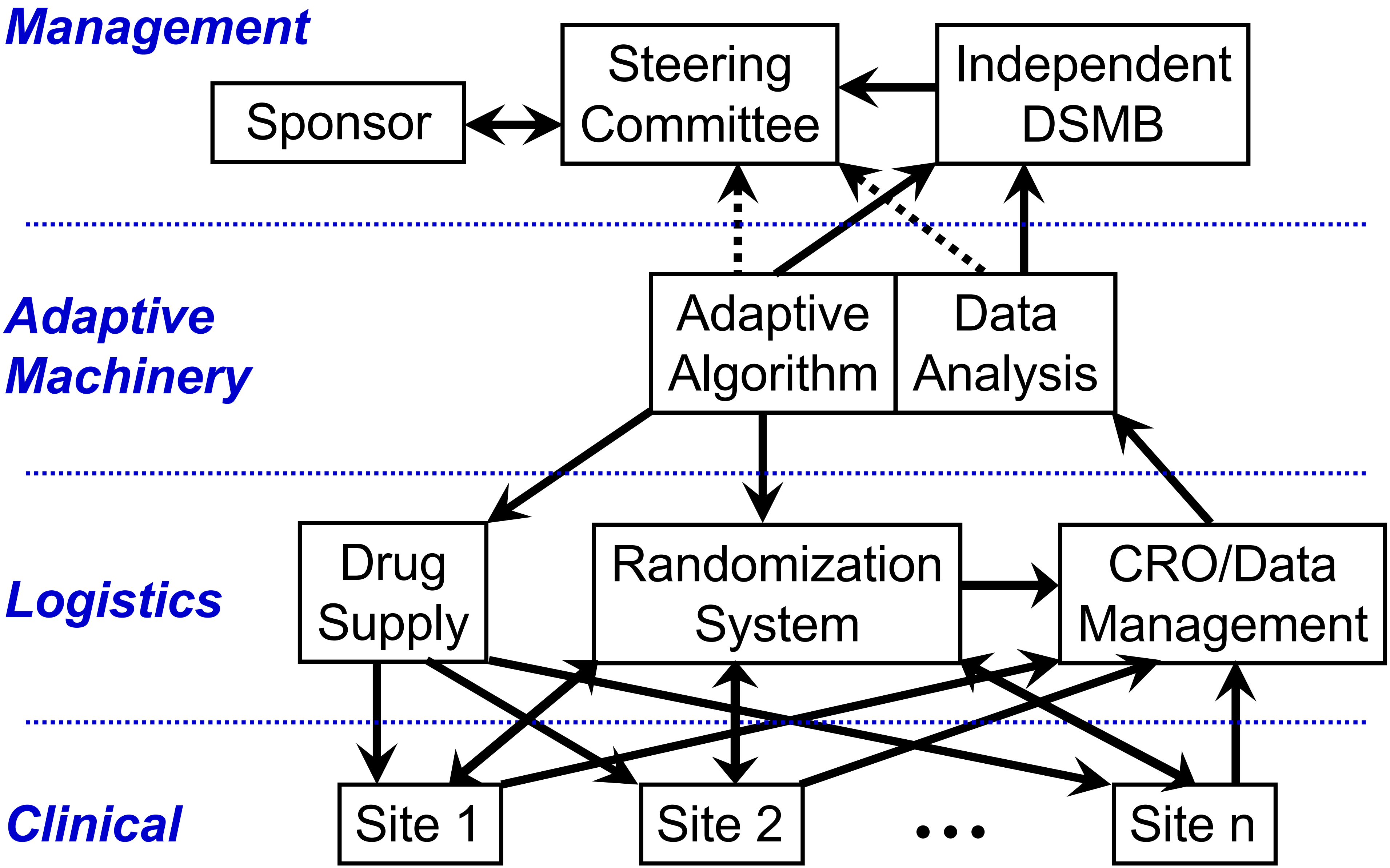
Components of an Adaptive Trial

Thanks Roger Lewis



Components of an Adaptive Trial

Thanks Roger Lewis



Data Safety Monitoring Boards

- Purpose
  - To ensure continued safety, validity, feasibility, and integrity of the clinical trial
  - To ensure the trial is conducted according to *a priori* plan, including adaptation
- Structure
  - Learn phase: usually includes internal personnel
  - Confirm phase: generally includes only independent, external members

Data Safety Monitoring Boards

- What’s different in an adaptive trial?
  - Requires expertise to assess whether the planned adaptations continue to be safe and appropriate
  - May increase need to include sponsor personnel
  - Ideally expertise to ensure everything is working
- What’s unchanged in an adaptive trial?
  - The DSMB ensures completion of the trial *as planned, including the adaptation*
  - It is the trial that’s adaptive, not the DSMB

IRB Review

- IRBs review/approve the full protocol, including the planned adaptations
- No new review when adaptations made
  - IRBs may request to be informed (e.g., new sample size, dropping of a surgical arm)
- Amendments are different
  - Not preplanned
- Irony
  - Little changes (amendments) may require IRB review
  - Big changes (adaptations) are defined by design and only reviewed/approved once

Acceptability to Key Stakeholders

- FDA
  - FDA Critical Path Initiative
  - 2010 Guidance for the Use of Bayesian Statistics in Medical Device Trials
  - 2019 Guidance for Adaptive Design Clinical Trials for Drugs and Biologics
  - Joint Regulatory Science initiative with NIH
- EMEA & PCORI Guidances
- Journals
  - Surprisingly clinical journals care little about design
    - Ever see a medical journal with smaller font for the methods?
  - We've had to argue to let journals give us <u>more</u> space for the design

Acceptability to Key Stakeholders

- NIH
  - ADAPT-IT sponsored by NIH Common Fund
    - Redesigning four neurologic emergency trials using adaptive designs
  - READAPT sponsored by NHLBI
  - ISPY-2 initiated by NIH Institute

  - Very good about seeking expertise to judge adaptive grants
    - Fear is innovate statistical methods will be reviewed by conventional (anti-adaptive) reviewers
    - Most institutes very good about seeking those with expertise to review methods

FDA Critical Path Initiative

From FDA website:

> Many of the tools used today to predict and evaluate product safety and efficacy are badly outdated from a scientific perspective. We have not made a concerted effort to apply new scientific knowledge -- in areas such as gene expression, analytic methods, and bioinformatics -- to medical product development. There exists tremendous opportunities to create more effective tests and tools, if we focus on the hard work necessary to turn these innovations into reliable applied sciences.

http://www.fda.gov/scienceresearch/specialtopics/criticalpathinitiative/ucm077015.htm

FDA Critical Path Initiative

From FDA website:

Inefficient clinical trial designs. Innovative clinical trial design may make it possible to develop accepted protocols for smaller but smarter trials. For example, new statistical techniques may make it possible to reduce the number of people who need to receive placebo or to adaptively change the trial based on ongoing results.

50% of Phase 3 trials failing

$800 million per successful NME (new chemical entity)

Ann. Rev. Medicine, Woodcock & Woosley, 2008

Critical Path Initiative

- Areas of improvement
  - Development & use of biomarkers (for prediction)
    toward personalized medicine
  - Modernizing clinical trial methodologies & processes
  - Aggressive use of bioinformatics
    including disease modeling & trial simulation
  - Improvement in manufacturing technologies
  - 76 discrete projects that could improve product development & product use

US FDA 2006, "Innovation or Stagnation: Critical Path Opportunities Report & List."

www.fda.gov/oc/initiatives/criticalpath/reports/opp_report.pdf

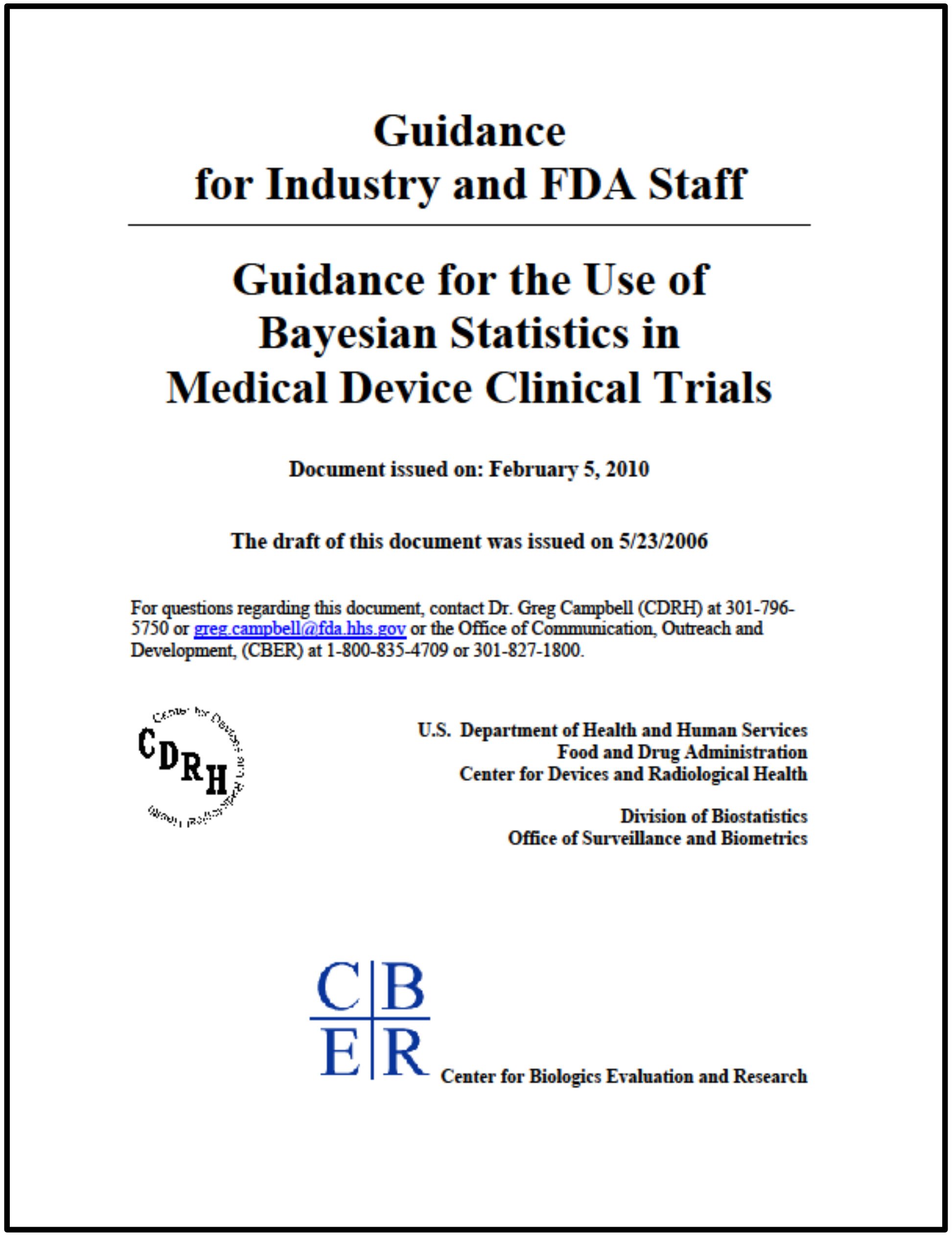
Is Now a Prime Time for Adaptive Designs in Clinical Trials?

- It's well past time
- Virtually every large pharmaceutical company, 100+ device companies, and dozens of biotech companies are investing in adaptive designs
  - Many device companies have completed adaptive designs
- What is the likelihood that these designs will lead to regulatory approval when such approval is warranted?
- Is there a gap between perceived risk to sponsors and the real risk?
  - Does industry overestimate FDAs conservatism?

Time has been Right for Adaptive Designs

- Janet Woodcock, FDA's CDER Director, 2006
  - Improved utilization of adaptive and Bayesian methods could help resolve low success rate of and expense of phase 3 clinical trials
- Margaret Hamburg, FDA Commissioner 2010
  - "The final guidance on the use of Bayesian statistics is consistent with the FDA's commitment to streamline clinical trials, when possible, in order to get safe and effective products to market faster."
- CDRH produced guidelines for Bayesian statistics Feb 5, 2010
  - "Agency says Bayesian statistical methods could trim costs, boost efficiency" from press release
  - "They beauty is you do not end up doing a trial that is too big or too small; you end up doing a trial that is just right." Greg Campbell
- CDER/CBER produced guidance for adaptive designs Nov 2019
  - Generally supportive of well-characterized adaptation by design
  - Appropriately cautious

FDA Guidance Documents



Guidance
for Industry and FDA Staff

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials

Document issued on: February 5, 2010

The draft of this document was issued on 5/23/2006

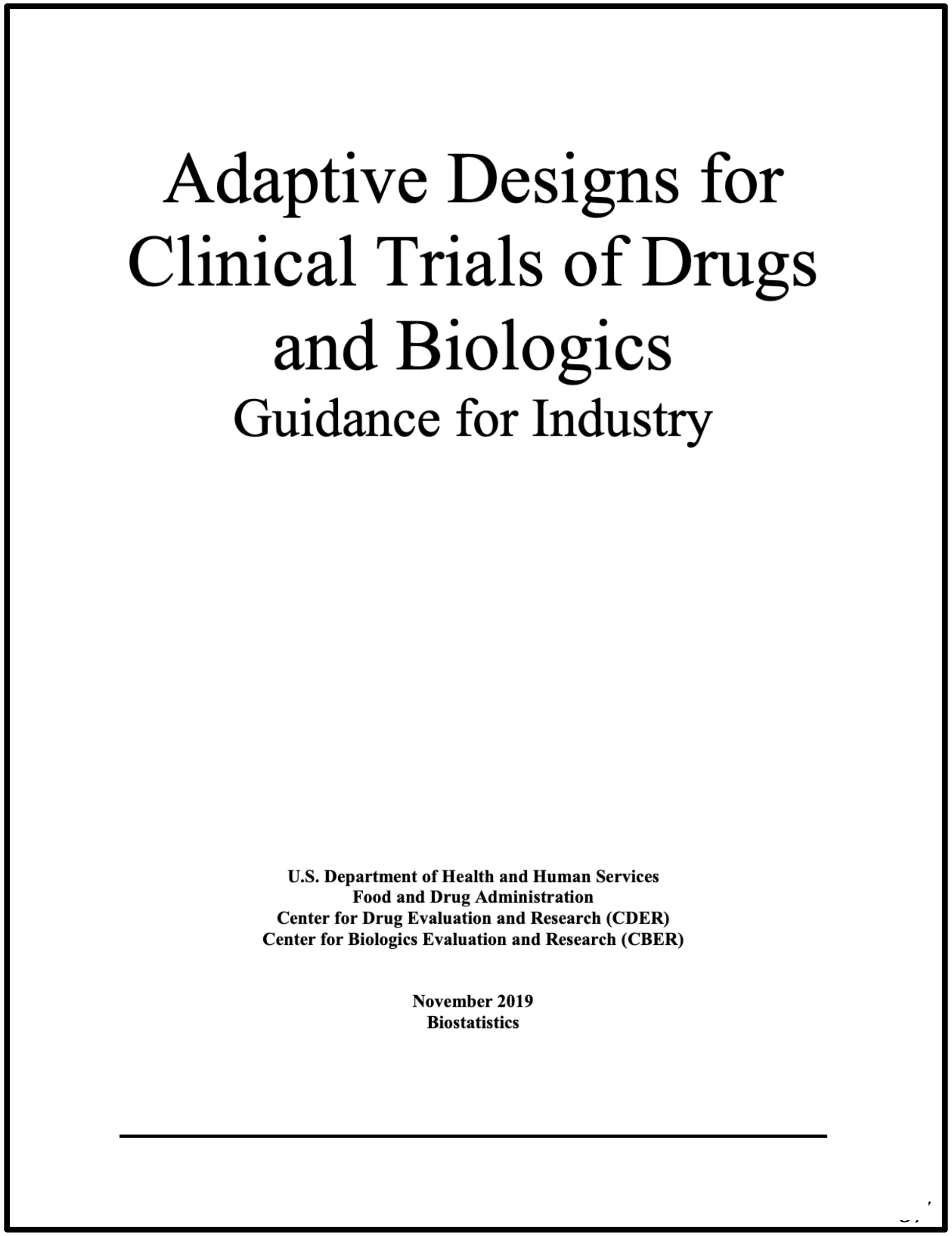
For questions regarding this document, contact Dr. Greg Campbell (CDRH) at 301-796-5750 or greg.campbell@fda.hhs.gov or the Office of Communication, Outreach and Development, (CBER) at 1-800-835-4709 or 301-827-1800.

CDRH

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health

Division of Biostatistics
Office of Surveillance and Biometrics

CBER **Center for Biologics Evaluation and Research**



Adaptive Designs for Clinical Trials of Drugs and Biologics
Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

November 2019
Biostatistics

Online Tools & Resources

- MD Anderson
  - http://biostatistics.mdanderson.org/SoftwareDownload/
  - Lots of good utilities, including "Adaptive Randomization" to help with response adaptive trials
  - Allows 10 arms; minimum number of patients before adapting randomization scheme; maximum number of patients or length of trial
  - Free
- Commercial resources increasingly available
- Usually I code my own

Some Current Areas of Application

- Alzheimer’ s Disease
- Aneurysm
- Asthma
- Atrial Fibrillation
- Cancer Diagnostics
- Cancer Screening
- Cancer Therapeutics
- Crohn’ s Disease
- Diabetes
- DVT
- Ebola
- Heart Valves
- Ebola
- Emphysema
- HIV
- Libido
- Lymphoma
- Lung Cancer
- Lupus
- Migraines
- Multiple Sclerosis
- Obesity
- Pain
- Parkinson’s
- Pandemic Flu
- Pre-term Labor
- Rheumatoid Arthritis
- Sepsis
- Smoking Cessation
- Spinal Cord Injury
- Spinal Implants
- Stroke
- Tinnitus
- Uterine Cancer
- Vaccines

Decision Problem 3: ESETT Trial

A multicenter, randomized, double-blind,

comparative effectiveness study of

fos-phenytoin, levetiracetam, and valproic acid

in subjects with benzodiazepine-refractory

Status Epilepticus:

The Established Status Epilepticus Treatment Trial

Acknowledgements

ADAPT-IT U01-NS073476

- William Barsan, MD
  University of Michigan
- Donald Berry, PhD
  MD Andersson, Berry Consultants
- Roger Lewis, MD, PhD
  UCLA Harbor
- Scott Berry, PhD
  Berry Consultants
- Valerie Durkalski, PhD, Yuko Palesch, PhD
  Medical University of South Carolina
- Michael Fetters, MD
  University of Michigan
- Shirley Fredrickson, RN, MS
  University of Michigan
- Will Meurer, MD
  University of Michigan
- Robin Conwit, MD, Scott Janis, MD
  NINDS

ESETT U01-NS088034

- Jaideep Kapur, MD
  University of Virginia
- Kristine Broglio, MS
  Berry Consultants
- Jordan Elm, PhD, Wenle Zhao, PhD
  Medical University of South Carolina
- James Chamberlain, MD
  Children's National Med Center
- Nathan Fountain, MD
  University of Virginia
- Daniel Lowenstein, MD
  UCSF
- Shlomo Shinnar, MD, PhD
  Albert Einstein COM
- Rob Silbergliet, MD
  University of Michigan
- David Treiman, MD
  Barrow Neurological Institute

Research Question

- How to treat seizing patients who've failed benzodiazapine?
  - fosphenytoin (fPHT)
  - levetiracetam (LVT)
  - valproic acid (VPA)

Comparative Effectiveness

- No control group
  - – Three drugs start out equal
  - – Want to know which is best
- What is Type I error in CER?
  - – Consequence of Type I error less in CER
- Really want to know
  - – Which drug is best … with measure of certainty
  - – Which drug is worst … with measure of certainty

Trial Overview

- Primary endpoint
  - cessation of seizure within 20 minutes
  - no further intervention within 1 hour
  - no significant adverse event
- Powered to identify 15% difference in response rate
  - Min 400, Max 795 Patients (to get 720)
- Stratify randomization by age

Bayesian Adaptive Design Features

- Adaptively allocate to favor better treatments
- Drop poor performing arms
  - Relative to one another
  - Relative to 25% goal
- Stop early if we know the answer

  or know we won't know
  - Efficacy stop if treatment clearly better
  - Futility stop if unlikely to ID a 'best' or 'worst'
    - Do not stop if 1 worse and other 2 equally good
  - Futility stopping if all arms bad

Randomization Options

- Let r_d = randomization probability to dose d
- Let p_d = probability arm d has highest (best) response rate
- Randomization weighting by C

$$r_d = \frac{p_d^C}{p_1^C + p_2^C + p_3^C + \ldots + p_D^C}$$

Randomization Options

$$r_d = \frac{p_d^C}{p_1^C + p_2^C + p_3^C + \ldots + p_D^C}$$

- $C = 0$, equal randomization ($r_d = 1/\text{Number of Groups}$)
- $C = 1$, proportional to probability best ($r_d = p_d$)
- $C \geq 1$
  - strongly favor 1 arm earlier in the trial, even when treatments are equal
  - more subjects likely assigned to the best treatment
  - C ➞ big means assign all to best treatment, play the leader
- $0 < C < 1$
  - weakly favor better
  - fewer subjects likely assigned to best treatment
  - more even distribution early in trials
  - randomization less affected by early events
- $C = n/N$, trial begins with c = 0 and ends with c = 1

Adaptive Allocation

- Randomize 300 patients equally
- At 300 & then every 100 adaptively allocate to

$$r_t \propto \sqrt{\frac{\Pr\left(p_t = \max(p)\right) Var(p_t)}{n_t}}$$

  - Favor better performing treatments
  - Favor treatments with greater uncertainty
  - Every 100 = About every 6 months | expected accrual
- If allocation probability < 5%, suspend accrual
- If Pr(Success > 0.25) < 0.05 drop arm

Early Stopping

- Analyses begin after 400 patients and repeat every additional 100 patients accrued
- Early Success Stopping:
  - – If arm has 97.5% probability of having highest success rate
    - i.e. $\Pr(p_t = max(p)) > 0.975$
- Early Futility Stopping
  - – If all doses have Pr(Success > 0.25) < 0.05
  - – If predicted probability of success (ID ‘winner’ or ‘loser’ at the max N=795) < 0.05

Example Trial: 300 pt analysis

	N Enrolled Observed Response Rate			Pr(Max Effective Trt)			Pr(Allocation)			Pred Prob
Look	LVT	fPHT	VPA	LVT	fPHT	VPA	LVT	fPHT	VPA	
300	51/100 51%	55/100 55%	64/100 64%	0.025	0.092	0.88	0.12	0.22	0.66	0.71

Example Trial: 400 pt analysis

	N Enrolled Observed Response Rate			Pr(Max Effective Trt)			Pr(Allocation)			Pred Prob
Look	LVT	fPHT	VPA	LVT	fPHT	VPA	LVT	fPHT	VPA	
300	51/100 51%	55/100 55%	64/100 64%	0.025	0.092	0.88	0.12	0.22	0.66	0.71
Next 100	6/11 55%	19/26 73%	39/63 62%							
400	57/111 51%	74/126 59%	105/163 64%	0.01	0.16	0.83	0.09	0.34	0.57	0.50

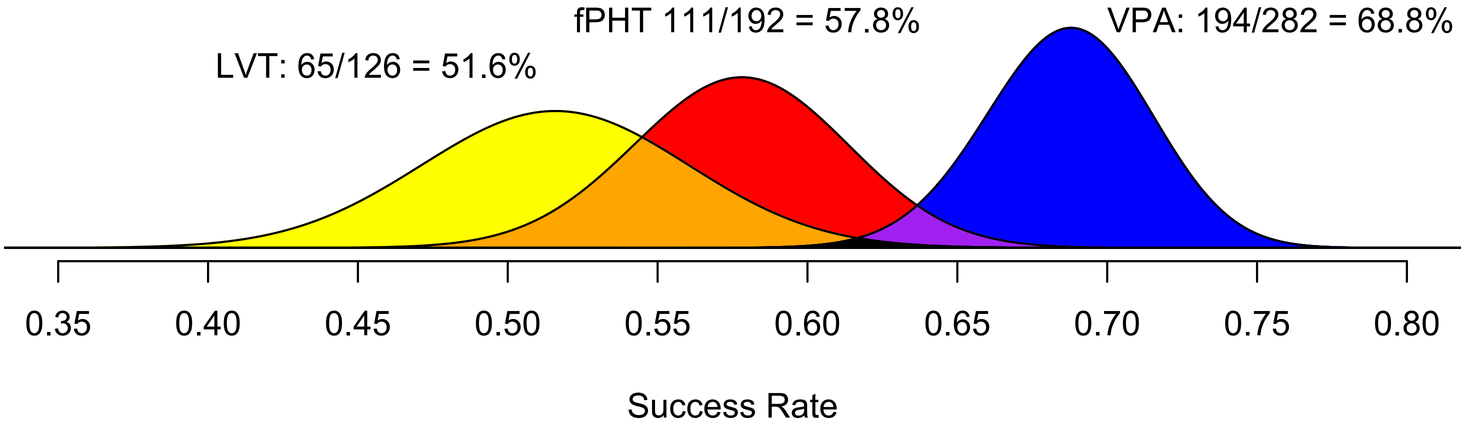
Example Trial: 500 pt analysis

	N Enrolled Observed Response Rate			Pr(Max Effective Trt)			Pr(Allocation)			Pred Prob
Look	LVT	fPHT	VPA	LVT	fPHT	VPA	LVT	fPHT	VPA	
300	51/100 51%	55/100 55%	64/100 64%	0.025	0.092	0.88	0.12	0.22	0.66	0.71
400	57/111 51%	74/126 59%	105/163 64%	0.01	0.16	0.83	0.09	0.34	0.57	0.50
Next 100	5/12 42%	20/38 53%	34/50 68%							
500	62/123 50%	94/164 57%	139/213 65%	0.004	0.056	0.94	0.08	0.23	0.69	0.59

Example Trial: 600 pt analysis

	N Enrolled Observed Response Rate			**Pr(Max Effective Trt)**			**Pr(Allocation)**			**Pred Prob**
Look	LVT	fPHT	VPA	LVT	fPHT	VPA	LVT	fPHT	VPA	
300	51/100 51%	55/100 55%	64/100 64%	0.025	0.092	0.88	0.12	0.22	0.66	0.71
400	57/111 51%	74/126 59%	105/163 64%	0.01	0.16	0.83	0.09	0.34	0.57	0.50
500	62/123 50%	94/164 57%	139/213 65%	0.004	0.056	0.94	0.08	0.23	0.69	0.59
Next 100	3/3 100%	17/28 61%	55/69 80%							
600	65/126 52%	111/192 58%	194/282 69%	0.000 0.87	0.008 0.13	0.992 0.00	Trial Stops Early for Identifying Best Treatment			

Example Trial: Final Evaluation



Treatment	Observed	%	95% CI	Pr(Best)	Pr(Worst)
VPA	194/282	68.8%	(.632, .739)	0.992	0.0005
fPHT	111/192	57.8%	(.507, .646)	0.007	0.138
LVT	65/126	51.6%	(.429, .601)	0.0005	0.862

Difference	Observed	95% CI	Pairwise Comparison
VPA – fPHT	0.110	(0.022, 0.197)	Pr(VPA>fPHT) = 0.993
VPA – LVT	0.172	(0.069, 0.272)	Pr(VPA>LVT) > 0.999
fPHT - LVT	0.062	(-0.049, 0.172)	Pr(fPHT>LVT) = 0.862

Comparison to without Adaptive Randomization

	Adaptive Randomization			Fixed Randomization		
Scenario **3 Efficacy Rates**	**Power** **Best/Wst**	**Mean** **N**	**% to** **Best**	**Power** **Best/Wst**	**Mean** **N**	**% to Best**
Null 0.5 – 0.5 – 0.5	0.013 0.018	507		0.023 0.007	499	
One Good 0.5 – 0.5 – 0.65	0.89 0.03	483	48	0.87 0.04	497	33
Two Good 0.5 – 0.65 – 0.65	0.11 0.67	679	84	0.10 0.79	687	67
One Middle One Good 0.5 – 0.575 – 0.65	0.50 0.25	586	47	0.44 0.31	599	33
All Bad 0.25– 0.25 – 0.25	0.011 0.020	524		0.023 0.008	509	
All Very Bad 0.10 – 0.10 – 0.10	0.006 0.01	400		0.008 0.02	400	

Comparison to without Adaptive Randomization

	Adaptive Randomization			Fixed Randomization		
Scenario **3 Efficacy Rates**	**Power** **Best/Wst**	**Mean** **N**	**% to** **Best**	**Power** **Best/Wst**	**Mean** **N**	**% to Best**
Null 0.5 – 0.5 – 0.5	0.013 0.018	507		0.023 0.007	499	
One Good 0.5 – 0.5 – 0.65	0.89 0.03	483	48	0.87 0.04	497	33
Two Good 0.5 – 0.65 – 0.65	0.11 0.67	679	84	0.10 0.79	687	67
One Middle One Good 0.5 – 0.575 – 0.65	0.50 0.25	586	47	0.44 0.31	599	33
All Bad 0.25– 0.25 – 0.25	0.011 0.020	524		0.023 0.008	509	
All Very Bad 0.10 – 0.10 – 0.10	0.006 0.01	400		0.008 0.02	400	

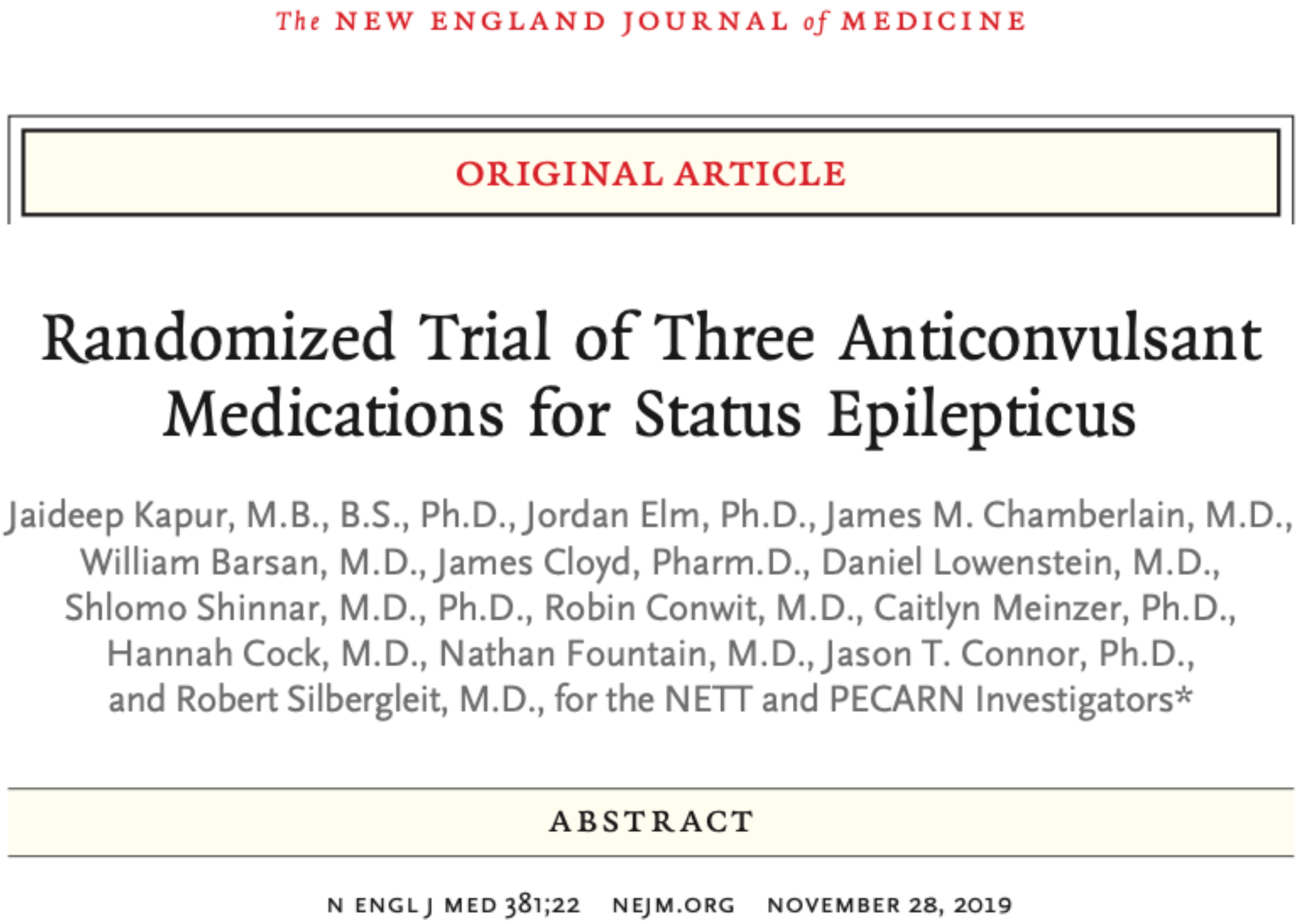
Comparison to without Adaptive Randomization

	Adaptive Randomization			Fixed Randomization		
Scenario **3 Efficacy Rates**	**Power Best/Wst**	**Mean N**	**% to Best**	**Power Best/Wst**	**Mean N**	**% to Best**
Null 0.5 – 0.5 – 0.5	0.013 0.018	507		0.023 0.007	499	
One Good 0.5 – 0.5 – 0.65	0.89 0.03	483	48	0.87 0.04	497	33
Two Good 0.5 – 0.65 – 0.65	0.11 0.67	679	84	0.10 0.79	687	67
One Middle One Good 0.5 – 0.575 – 0.65	0.50 0.25	586	47	0.44 0.31	599	33
All Bad 0.25– 0.25 – 0.25	0.011 0.020	524		0.023 0.008	509	
All Very Bad 0.10 – 0.10 – 0.10	0.006 0.01	400		0.008 0.02	400	

Comparison to without Adaptive Randomization

	Adaptive Randomization			Fixed Randomization		
Scenario 3 Efficacy Rates	**Power Best/Wst**	**Mean N**	**% to Best**	**Power Best/Wst**	**Mean N**	**% to Best**
Null 0.5 – 0.5 – 0.5	0.013 0.018	507		0.023 0.007	499	
One Good 0.5 – 0.5 – 0.65	0.89 0.03	483	48	0.87 0.04	497	33
Two Good 0.5 – 0.65 – 0.65	0.11 0.67	679	84	0.10 0.79	687	67
One Middle One Good 0.5 – 0.575 – 0.65	0.50 0.25	586	47	0.44 0.31	599	33
All Bad 0.25– 0.25 – 0.25	0.011 0.020	524		0.023 0.008	509	
All Very Bad 0.10 – 0.10 – 0.10	0.006 0.01	400		0.008 0.02	400	

Results



The NEW ENGLAND JOURNAL of MEDICINE

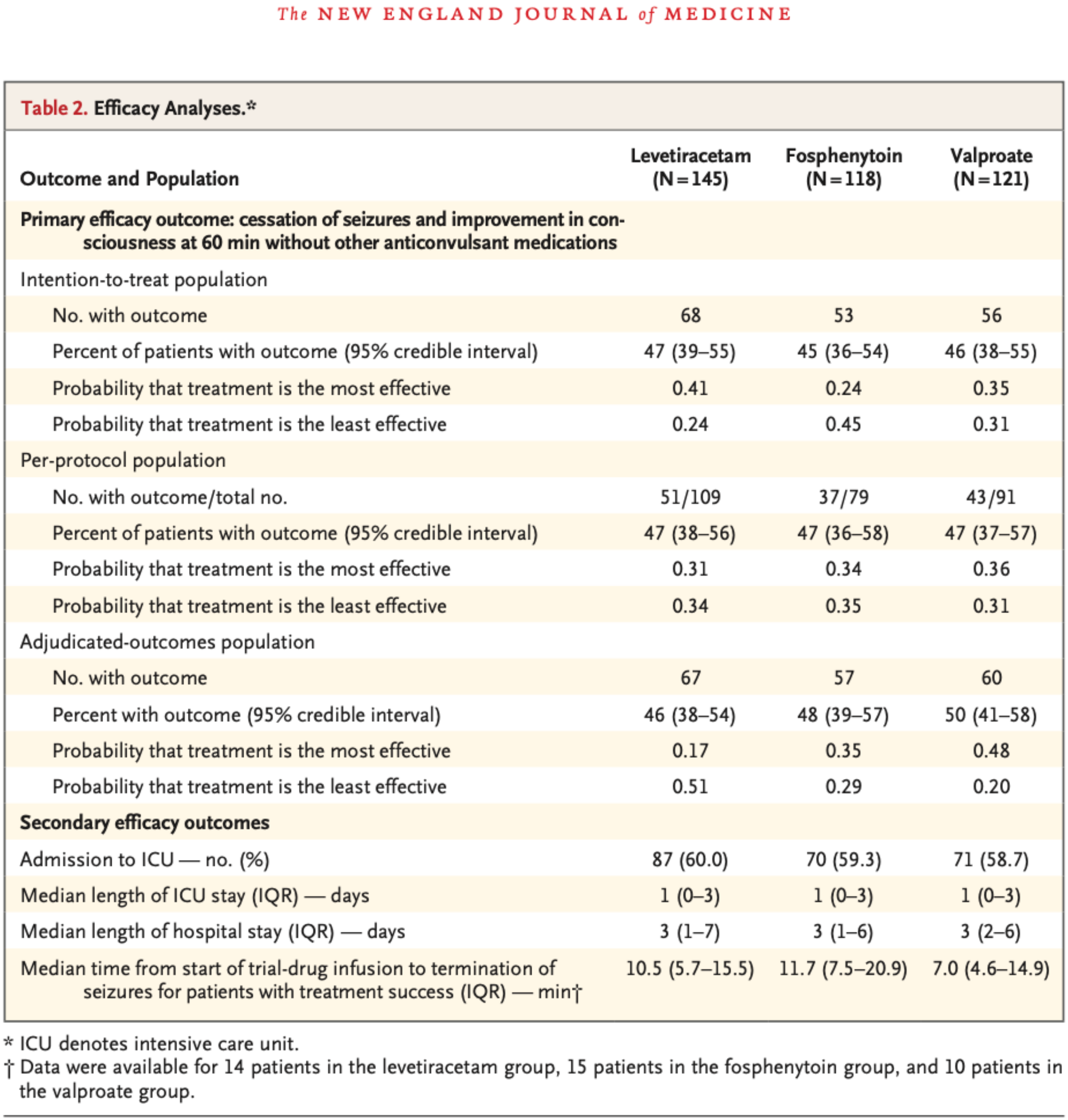
ORIGINAL ARTICLE

Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus

Jaideep Kapur, M.B., B.S., Ph.D., Jordan Elm, Ph.D., James M. Chamberlain, M.D.,
William Barsan, M.D., James Cloyd, Pharm.D., Daniel Lowenstein, M.D.,
Shlomo Shinnar, M.D., Ph.D., Robin Conwit, M.D., Caitlyn Meinzer, Ph.D.,
Hannah Cock, M.D., Nathan Fountain, M.D., Jason T. Connor, Ph.D.,
and Robert Silbergleit, M.D., for the NETT and PECARN Investigators*

ABSTRACT

N ENGL J MED 381;22 NEJM.ORG NOVEMBER 28, 2019



The NEW ENGLAND JOURNAL of MEDICINE

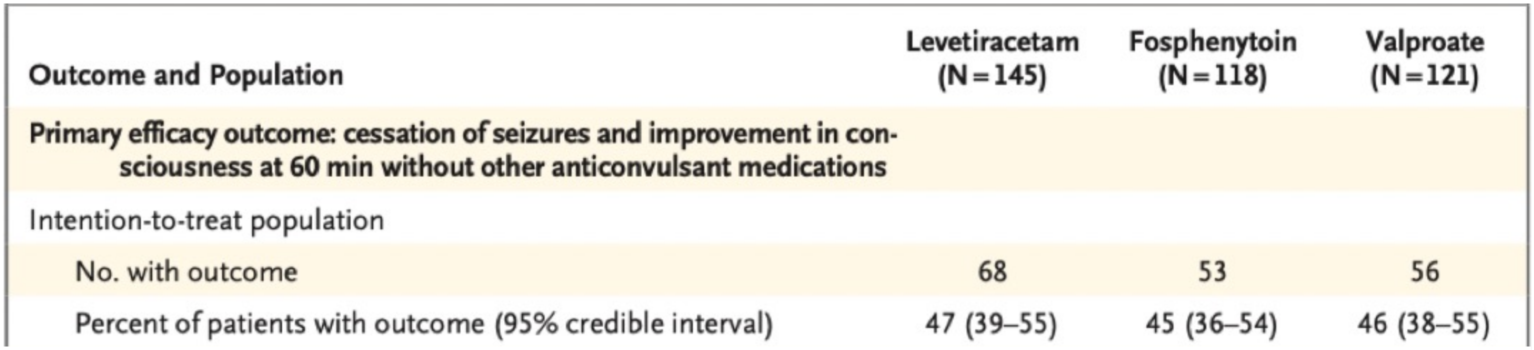
Table 2. Efficacy Analyses.*

Outcome and Population	Levetiracetam (N=145)	Fosphenytoin (N=118)	Valproate (N=121)
Primary efficacy outcome: cessation of seizures and improvement in consciousness at 60 min without other anticonvulsant medications			
Intention-to-treat population			
No. with outcome	68	53	56
Percent of patients with outcome (95% credible interval)	47 (39–55)	45 (36–54)	46 (38–55)
Probability that treatment is the most effective	0.41	0.24	0.35
Probability that treatment is the least effective	0.24	0.45	0.31
Per-protocol population			
No. with outcome/total no.	51/109	37/79	43/91
Percent of patients with outcome (95% credible interval)	47 (38–56)	47 (36–58)	47 (37–57)
Probability that treatment is the most effective	0.31	0.34	0.36
Probability that treatment is the least effective	0.34	0.35	0.31
Adjudicated-outcomes population			
No. with outcome	67	57	60
Percent with outcome (95% credible interval)	46 (38–54)	48 (39–57)	50 (41–58)
Probability that treatment is the most effective	0.17	0.35	0.48
Probability that treatment is the least effective	0.51	0.29	0.20
Secondary efficacy outcomes			
Admission to ICU — no. (%)	87 (60.0)	70 (59.3)	71 (58.7)
Median length of ICU stay (IQR) — days	1 (0–3)	1 (0–3)	1 (0–3)
Median length of hospital stay (IQR) — days	3 (1–7)	3 (1–6)	3 (2–6)
Median time from start of trial-drug infusion to termination of seizures for patients with treatment success (IQR) — min†	10.5 (5.7–15.5)	11.7 (7.5–20.9)	7.0 (4.6–14.9)

* ICU denotes intensive care unit.

† Data were available for 14 patients in the levetiracetam group, 15 patients in the fosphenytoin group, and 10 patients in the valproate group.

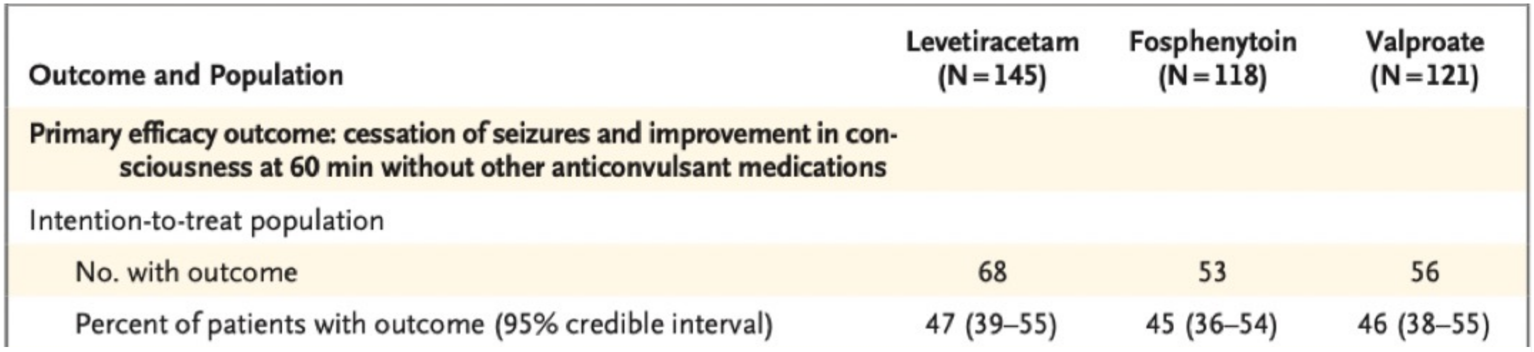
Imagine frequentist test



Outcome and Population	Levetiracetam (N=145)	Fosphenytoin (N=118)	Valproate (N=121)
Primary efficacy outcome: cessation of seizures and improvement in consciousness at 60 min without other anticonvulsant medications			
Intention-to-treat population			
No. with outcome	68	53	56
Percent of patients with outcome (95% credible interval)	47 (39–55)	45 (36–54)	46 (38–55)

```
> x <- c(68,53,56); n <- c(145,118,121)
> mat <- cbind(x, n-x)
> chisq.test(mat)
Pearson's Chi-squared test
data:  mat
X-squared = 0.10527, df = 2, p-value = 0.9487
```

Imagine frequentist test



Outcome and Population	Levetiracetam (N=145)	Fosphenytoin (N=118)	Valproate (N=121)
Primary efficacy outcome: cessation of seizures and improvement in consciousness at 60 min without other anticonvulsant medications			
Intention-to-treat population			
No. with outcome	68	53	56
Percent of patients with outcome (95% credible interval)	47 (39–55)	45 (36–54)	46 (38–55)

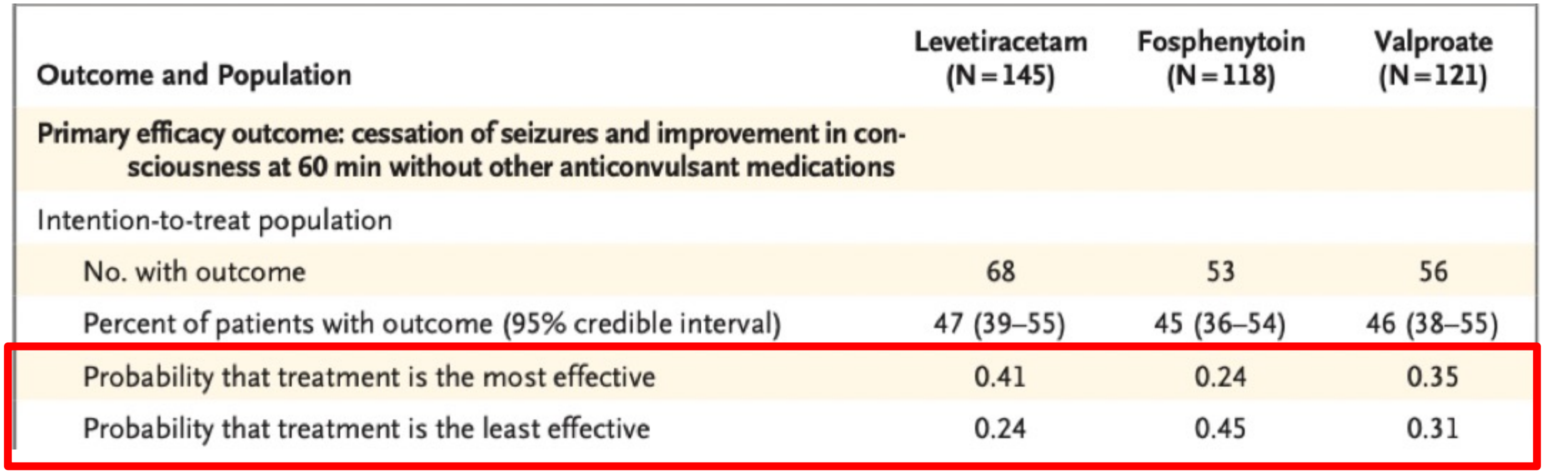
```
> x <- c(68,53,56); n <- c(145,118,121)
> mat <- cbind(x, n-x)
> chisq.test(mat)
Pearson's Chi-squared test
data:  mat
X-squared = 0.10527, df = 2, p-value = 0.9487
```

Can’t reject Ho: $p_{lev} = p_{fos} = p_{VPA}$ with p-value = 0.95

But you have to choose a treatment

How sure are you that you’ve chosen the best one?

Imagine frequentist test



Outcome and Population	Levetiracetam (N=145)	Fosphenytoin (N=118)	Valproate (N=121)
Primary efficacy outcome: cessation of seizures and improvement in consciousness at 60 min without other anticonvulsant medications			
Intention-to-treat population			
No. with outcome	68	53	56
Percent of patients with outcome (95% credible interval)	47 (39–55)	45 (36–54)	46 (38–55)
Probability that treatment is the most effective	0.41	0.24	0.35
Probability that treatment is the least effective	0.24	0.45	0.31

```
> x <- c(68,53,56); n <- c(145,118,121)
> mat <- cbind(x, n-x)
> chisq.test(mat)
Pearson's Chi-squared test
data:  mat
X-squared = 0.10527, df = 2, p-value = 0.9487
```

Can't reject Ho: $p_{lev} = p_{fos} = p_{VPA}$ with p-value = 0.95
But you have to choose a treatment
How sure are you that you've chosen the best one?

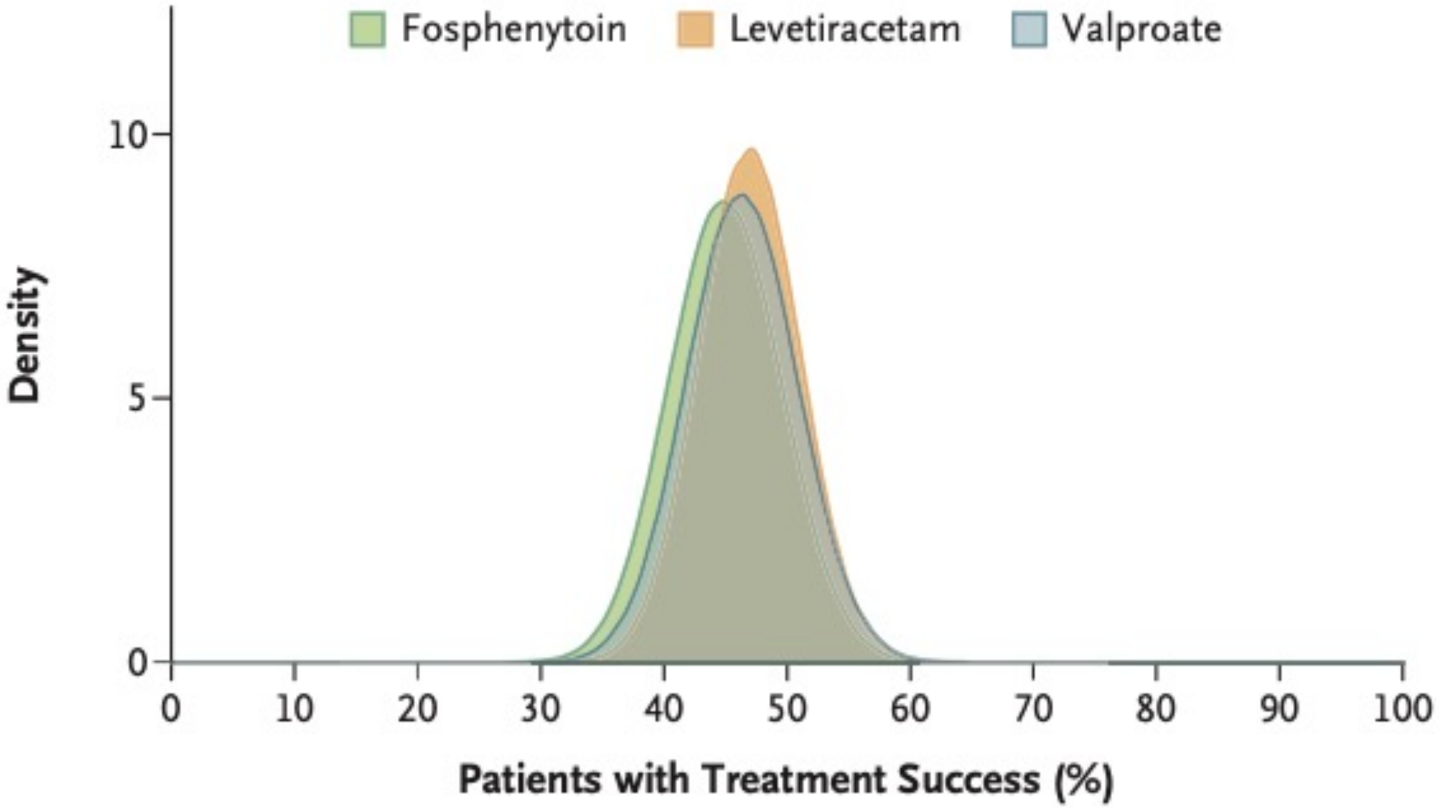
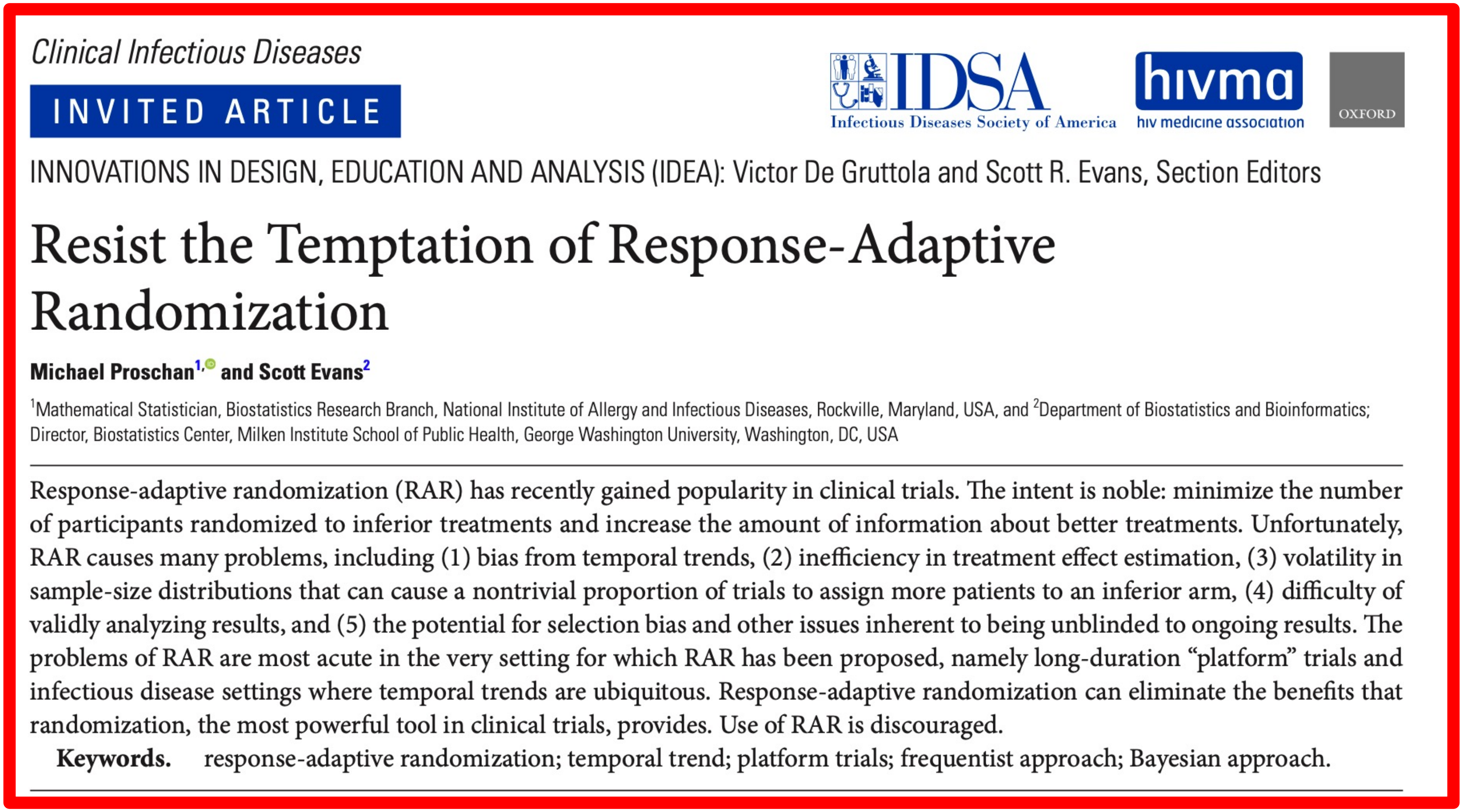


Figure 2. Posterior Probabilities of Success According to Treatment Group for the Primary Outcome of Cessation of Status Epilepticus at 60 Minutes.

The relative posterior probabilities of treatment success with regard to the primary outcome for each drug are shown. The percentage of patients with treatment success was 47% (95% credible interval, 39 to 55) in the levetiracetam group, 45% (95% credible interval, 36 to 54) in the fosphenytoin group, and 46% (95% credible interval, 38 to 55) in the valproate group.

Thoughts on Adaptive Randomization

Usual Criticisms



Clinical Infectious Diseases

INVITED ARTICLE

INNOVATIONS IN DESIGN, EDUCATION AND ANALYSIS (IDEA): Victor De Gruttola and Scott R. Evans, Section Editors

Resist the Temptation of Response-Adaptive Randomization

Michael Proschan[1] and Scott Evans[2]

[1]Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, USA, and [2]Department of Biostatistics and Bioinformatics; Director, Biostatistics Center, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

Response-adaptive randomization (RAR) has recently gained popularity in clinical trials. The intent is noble: minimize the number of participants randomized to inferior treatments and increase the amount of information about better treatments. Unfortunately, RAR causes many problems, including (1) bias from temporal trends, (2) inefficiency in treatment effect estimation, (3) volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm, (4) difficulty of validly analyzing results, and (5) the potential for selection bias and other issues inherent to being unblinded to ongoing results. The problems of RAR are most acute in the very setting for which RAR has been proposed, namely long-duration "platform" trials and infectious disease settings where temporal trends are ubiquitous. Response-adaptive randomization can eliminate the benefits that randomization, the most powerful tool in clinical trials, provides. Use of RAR is discouraged.

Keywords. response-adaptive randomization; temporal trend; platform trials; frequentist approach; Bayesian approach.

Usual Criticisms

- Too few on control
  - Controls randomization rate is usually fixed
  - Use adaptive randomization for different doses
  - At least put a minimum on controls %

- Early, wrong adaptation leads to bias
  - Require burn-in prior to adaptation
  - ESSET didn't start until N=300

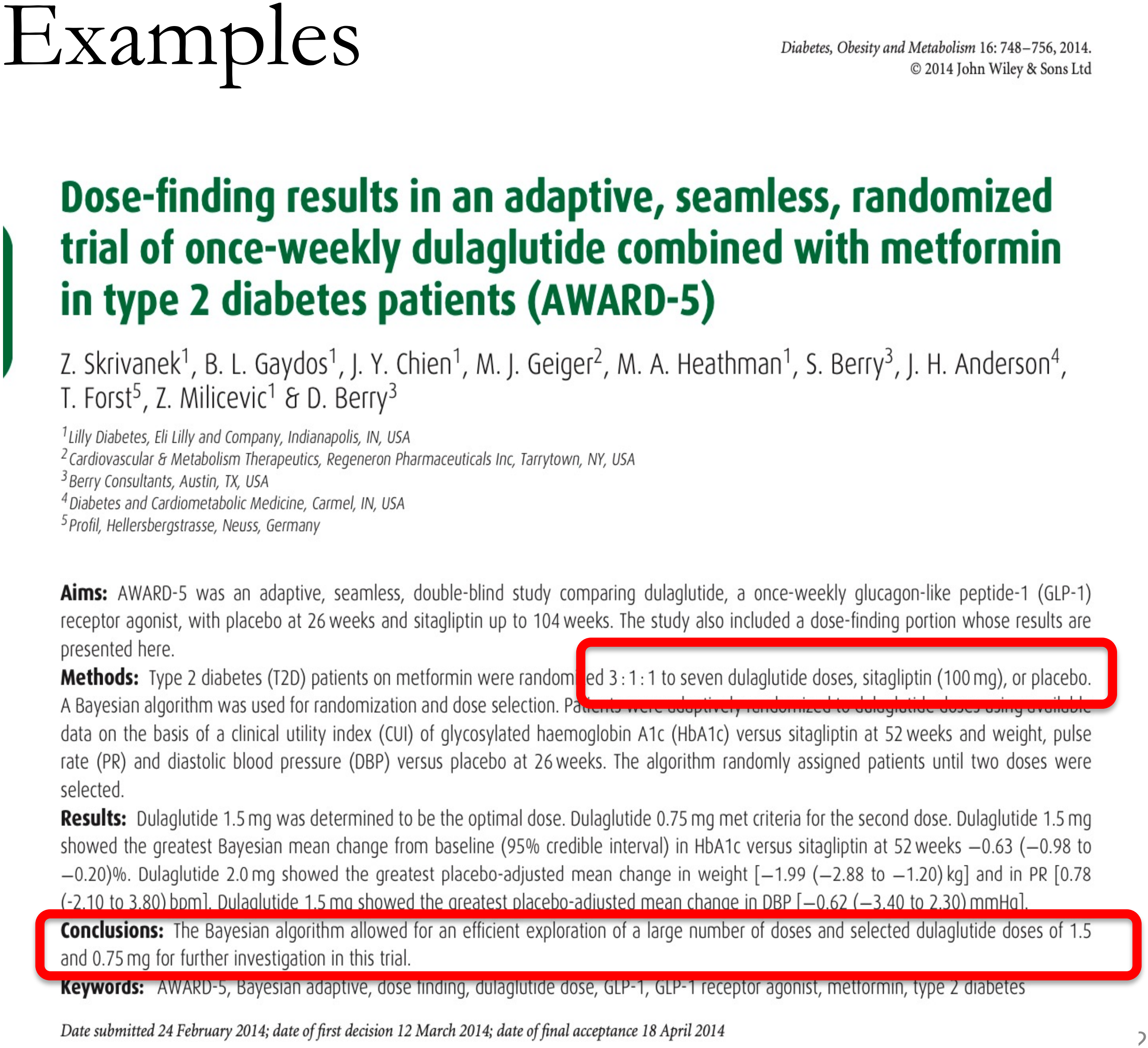
Usual Criticisms

- Drift makes uninterpretable
  - Legit concern
  - Very rare to observe
  - Standard methods also don’t work well here either
  - Nearly all stats methods require i.i.d.
    - If there is drift, data isn’t i.i.d.
    - Even standard methods require assumption that treatment effect is the same even though population is drifting
    - If drift is high, results may be uninterpretable

Usual Criticisms

- RAR can be unblinding
  - I agree with this
  - Ideally the treatment are masked so even if randomization probabilities change, investigators can not tell (e.g. ESETT)
  - If blinding not possible, perhaps have somewhat large minimum bounds or do arm dropping

RAR Examples



Diabetes, Obesity and Metabolism 16: 748–756, 2014.
© 2014 John Wiley & Sons Ltd

Dose-finding results in an adaptive, seamless, randomized trial of once-weekly dulaglutide combined with metformin in type 2 diabetes patients (AWARD-5)

Z. Skrivanek[1], B. L. Gaydos[1], J. Y. Chien[1], M. J. Geiger[2], M. A. Heathman[1], S. Berry[3], J. H. Anderson[4], T. Forst[5], Z. Milicevic[1] & D. Berry[3]

[1] *Lilly Diabetes, Eli Lilly and Company, Indianapolis, IN, USA*
[2] *Cardiovascular & Metabolism Therapeutics, Regeneron Pharmaceuticals Inc, Tarrytown, NY, USA*
[3] *Berry Consultants, Austin, TX, USA*
[4] *Diabetes and Cardiometabolic Medicine, Carmel, IN, USA*
[5] *Profil, Hellersbergstrasse, Neuss, Germany*

Aims: AWARD-5 was an adaptive, seamless, double-blind study comparing dulaglutide, a once-weekly glucagon-like peptide-1 (GLP-1) receptor agonist, with placebo at 26 weeks and sitagliptin up to 104 weeks. The study also included a dose-finding portion whose results are presented here.

Methods: Type 2 diabetes (T2D) patients on metformin were randomized 3 : 1 : 1 to seven dulaglutide doses, sitagliptin (100 mg), or placebo. A Bayesian algorithm was used for randomization and dose selection. Patients were adaptively randomized to dulaglutide doses using available data on the basis of a clinical utility index (CUI) of glycosylated haemoglobin A1c (HbA1c) versus sitagliptin at 52 weeks and weight, pulse rate (PR) and diastolic blood pressure (DBP) versus placebo at 26 weeks. The algorithm randomly assigned patients until two doses were selected.

Results: Dulaglutide 1.5 mg was determined to be the optimal dose. Dulaglutide 0.75 mg met criteria for the second dose. Dulaglutide 1.5 mg showed the greatest Bayesian mean change from baseline (95% credible interval) in HbA1c versus sitagliptin at 52 weeks −0.63 (−0.98 to −0.20)%. Dulaglutide 2.0 mg showed the greatest placebo-adjusted mean change in weight [−1.99 (−2.88 to −1.20) kg] and in PR [0.78 (-2.10 to 3.80) bpm]. Dulaglutide 1.5 mg showed the greatest placebo-adjusted mean change in DBP [−0.62 (−3.40 to 2.30) mmHg].

Conclusions: The Bayesian algorithm allowed for an efficient exploration of a large number of doses and selected dulaglutide doses of 1.5 and 0.75 mg for further investigation in this trial.

Keywords: AWARD-5, Bayesian adaptive, dose finding, dulaglutide dose, GLP-1, GLP-1 receptor agonist, metformin, type 2 diabetes

Date submitted 24 February 2014; date of first decision 12 March 2014; date of final acceptance 18 April 2014

Viele, Broglio, McGlothlin, Saville

Article

CLINICAL TRIALS

Clinical Trials
2020, Vol. 17(1) 52–60
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774519877836
journals.sagepub.com/home/ctj
SAGE

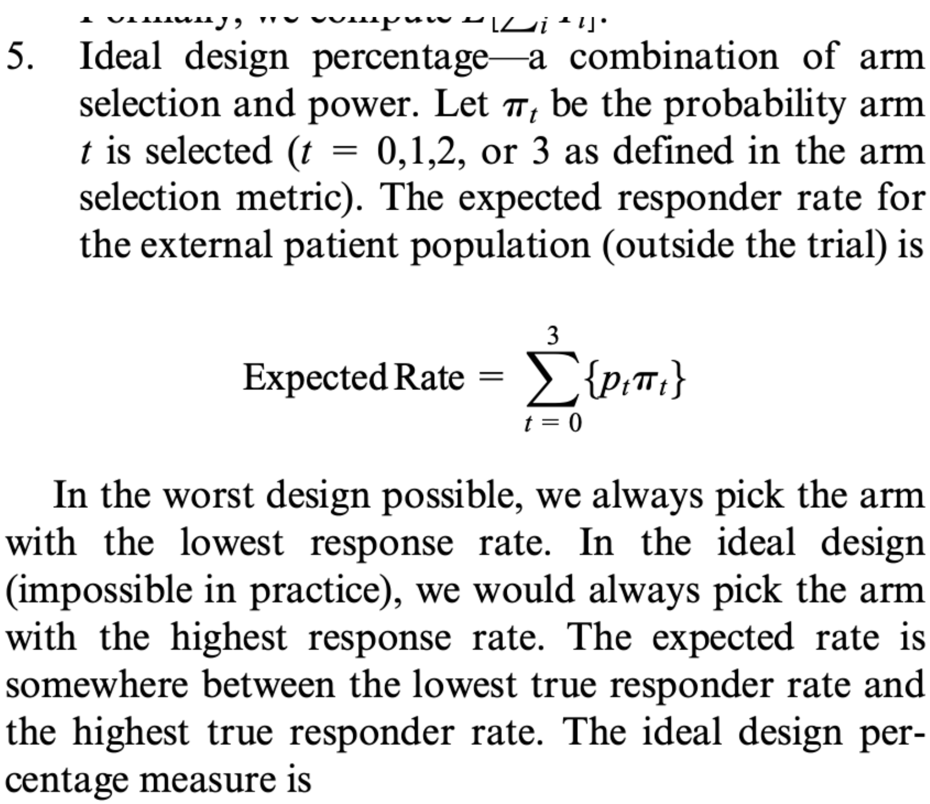
Comparison of methods for control allocation in multiple arm studies using response adaptive randomization

Kert Viele[1], Kristine Broglio[1], Anna McGlothlin[1] and Benjamin R Saville[1,2]

Viele, Broglio, McGlothlin, Saville

- Evaluate via
  - Power
  - Prob choosing the best arm
  - Mean square error (bias)
  - Expected number of responders
  - Ideal design percentage

Ideal Design Percentage

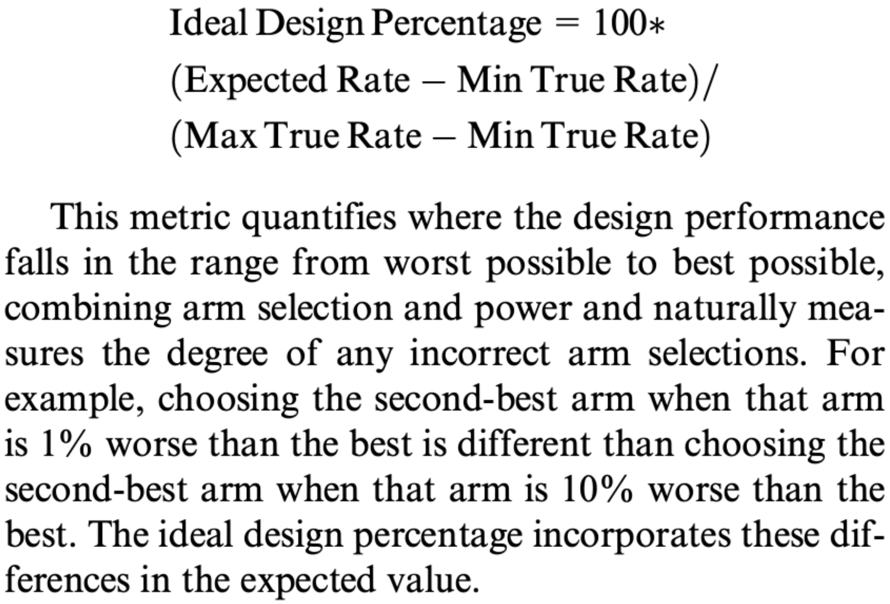


Formally, we compute $E[\sum_i Y_i]$.

5. Ideal design percentage—a combination of arm selection and power. Let π_t be the probability arm t is selected (t = 0,1,2, or 3 as defined in the arm selection metric). The expected responder rate for the external patient population (outside the trial) is

$$\text{Expected Rate} = \sum_{t=0}^{3} \{p_t \pi_t\}$$

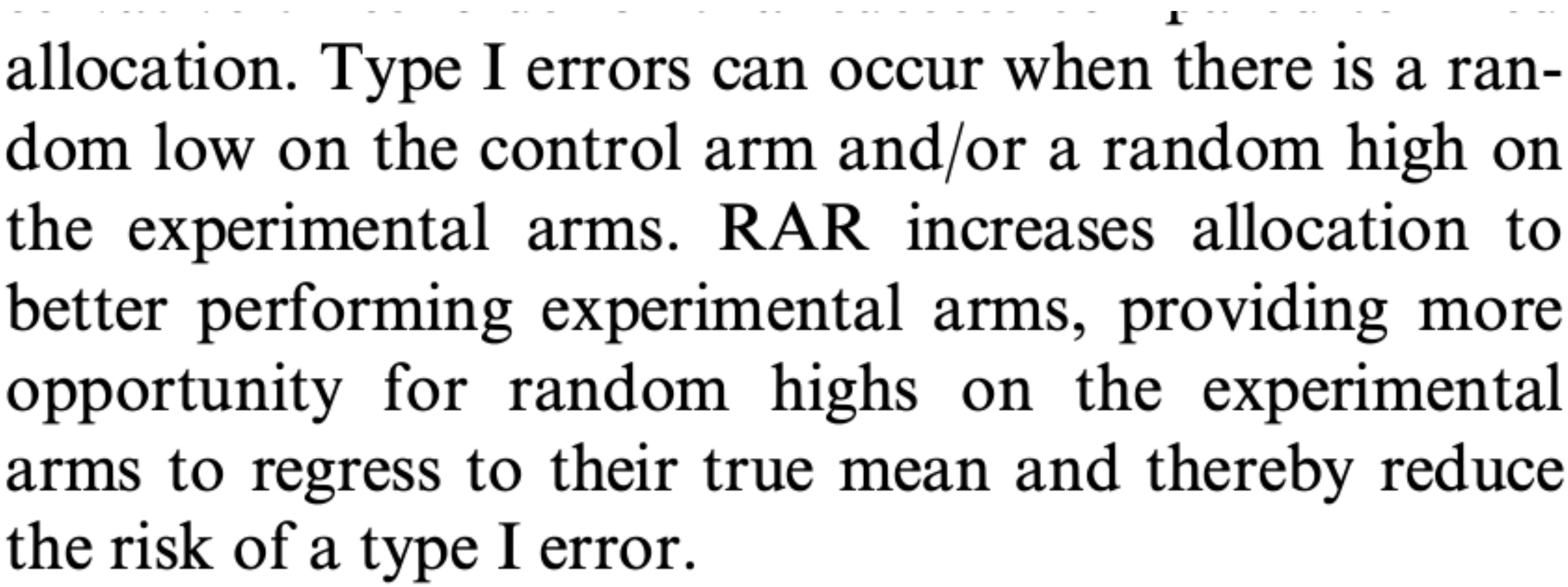
In the worst design possible, we always pick the arm with the lowest response rate. In the ideal design (impossible in practice), we would always pick the arm with the highest response rate. The expected rate is somewhere between the lowest true responder rate and the highest true responder rate. The ideal design percentage measure is



$$\text{Ideal Design Percentage} = 100* \\ (\text{Expected Rate} - \text{Min True Rate})/ \\ (\text{Max True Rate} - \text{Min True Rate})$$

This metric quantifies where the design performance falls in the range from worst possible to best possible, combining arm selection and power and naturally measures the degree of any incorrect arm selections. For example, choosing the second-best arm when that arm is 1% worse than the best is different than choosing the second-best arm when that arm is 10% worse than the best. The ideal design percentage incorporates these differences in the expected value.

Type I error



allocation. Type I errors can occur when there is a random low on the control arm and/or a random high on the experimental arms. RAR increases allocation to better performing experimental arms, providing more opportunity for random highs on the experimental arms to regress to their true mean and thereby reduce the risk of a type I error.

Choosing the best arm

Rmatch essentially puts the control % = max best arm %
- control never goes to 0 and goes to 50% as 1 arm appears best

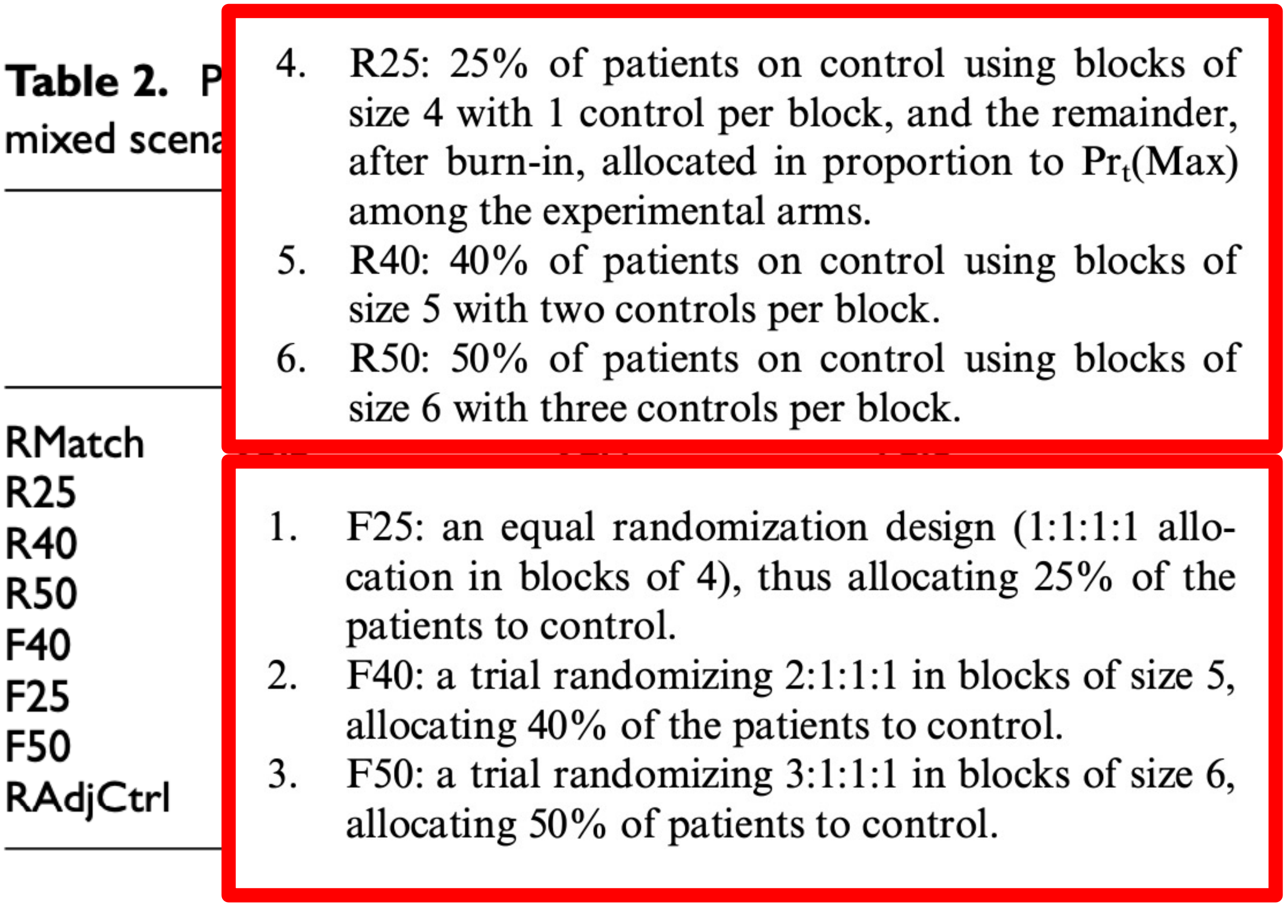


Table 2. P
mixed scena

RMatch
R25
R40
R50
F40
F25
F50
RAdjCtrl

4. R25: 25% of patients on control using blocks of size 4 with 1 control per block, and the remainder, after burn-in, allocated in proportion to $\Pr_t(\mathrm{Max})$ among the experimental arms.
5. R40: 40% of patients on control using blocks of size 5 with two controls per block.
6. R50: 50% of patients on control using blocks of size 6 with three controls per block.

1. F25: an equal randomization design (1:1:1:1 allocation in blocks of 4), thus allocating 25% of the patients to control.
2. F40: a trial randomizing 2:1:1:1 in blocks of size 5, allocating 40% of the patients to control.
3. F50: a trial randomizing 3:1:1:1 in blocks of size 6, allocating 50% of patients to control.

RAdjCtrl means control % can go to 0 or 1
- cases used exclusively by anti-RAR advocates

Power with RAR

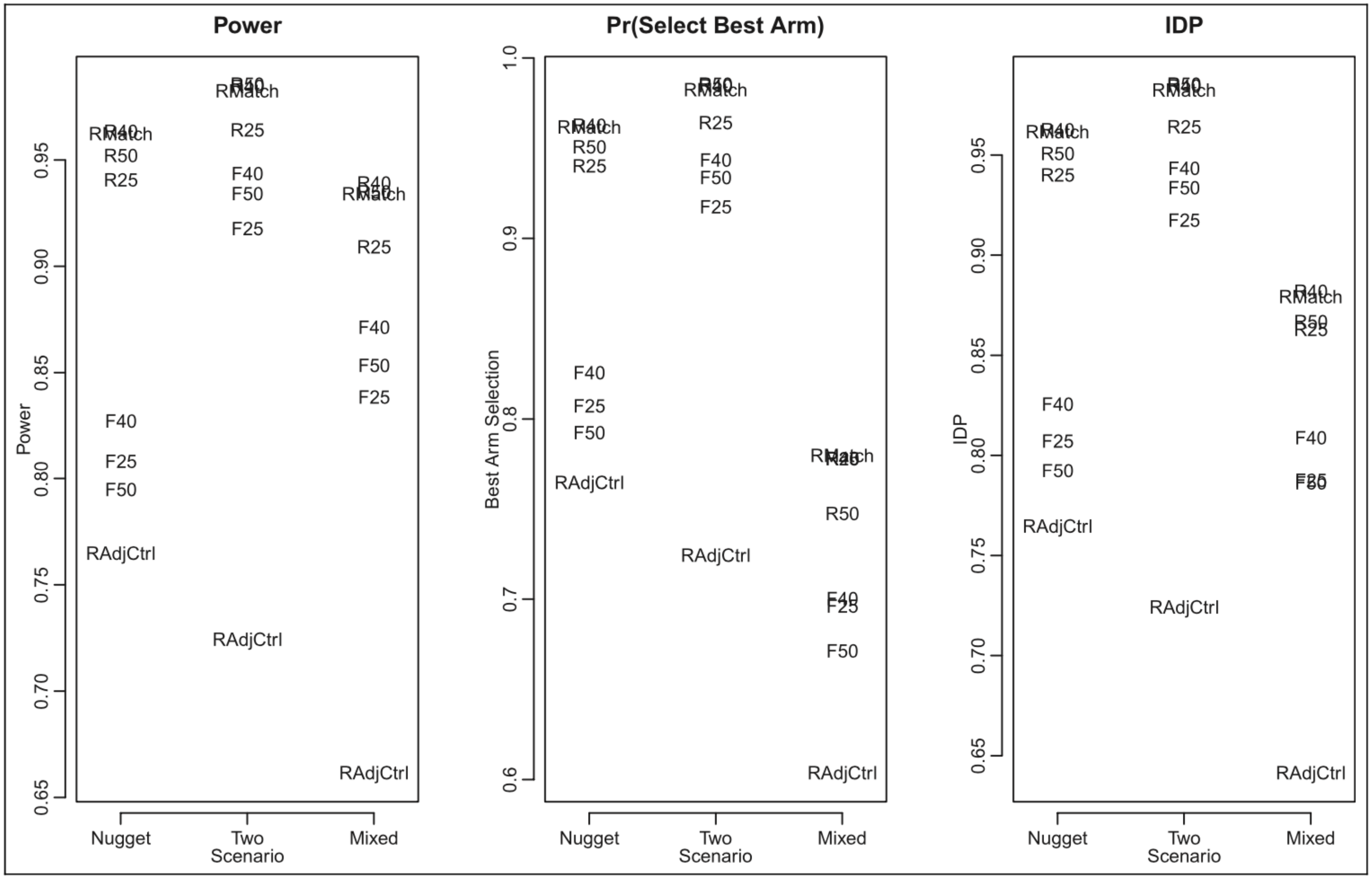


Figure 1. Results for external patient considerations. Left panel shows power, middle panel the probability of selecting the best arm, and the right panel ideal design percentage. The text string within each column indicates the design, either "F25,""F40," or "F50" for the fixed designs, "R25,""R40," and "R50" for the RAR designs with fixed control, and "RAdjCtrl" and "RMatch" for the RAR designs that alter control throughout the trial.

Accuracy of Estimate Effect

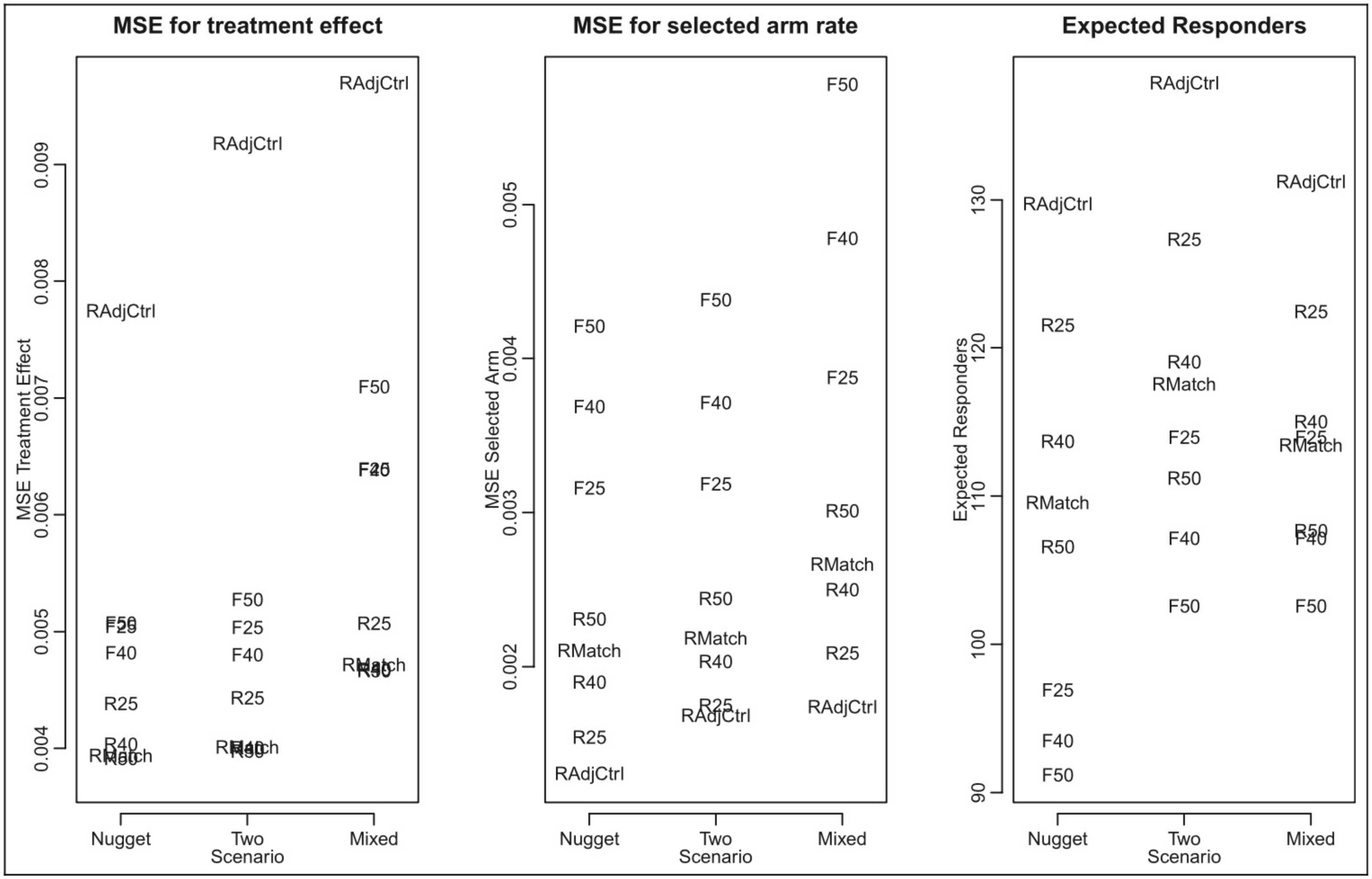


Figure 2. The left and middle panel show mean square error for estimated treatment effects (left) and the response rate on the selected arm (middle), while the right panel shows the expected number of responders. The text string indicates the design as shown in Figure 1.

Choosing the best arm

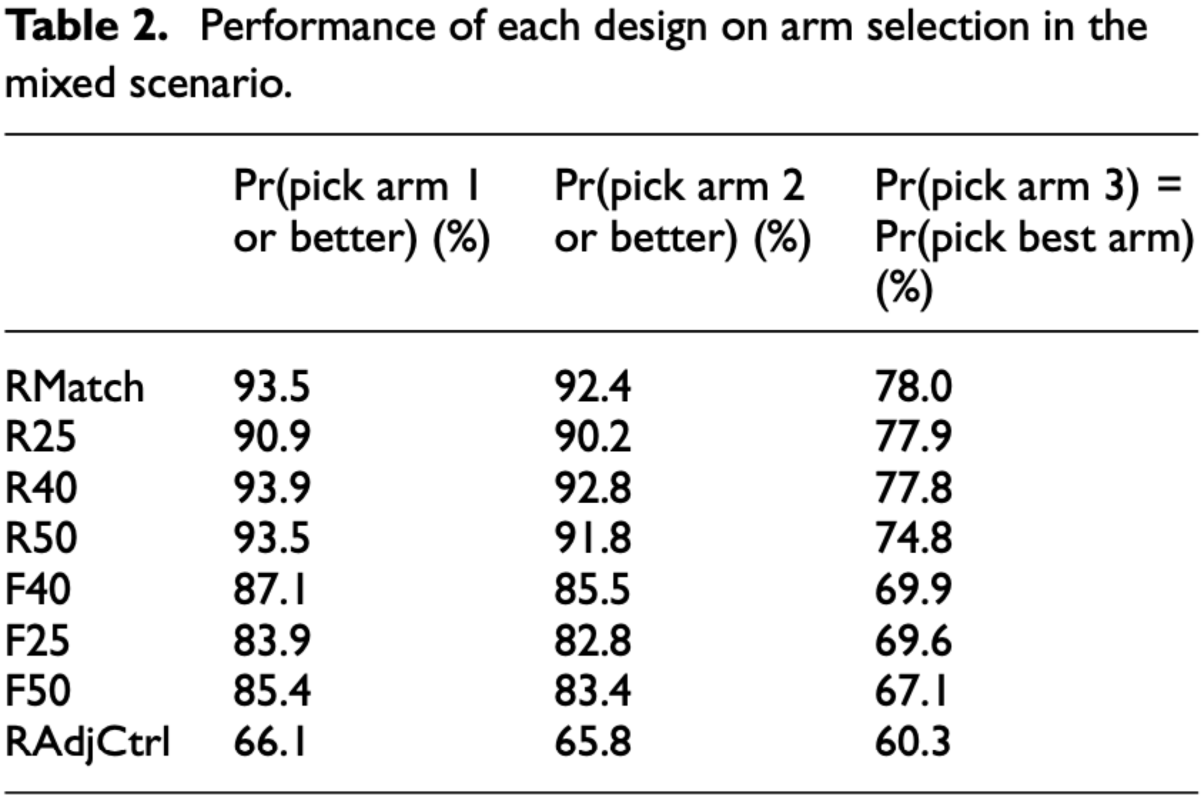


Table 2. Performance of each design on arm selection in the mixed scenario.

	Pr(pick arm 1 or better) (%)	Pr(pick arm 2 or better) (%)	Pr(pick arm 3) = Pr(pick best arm) (%)
RMatch	93.5	92.4	78.0
R25	90.9	90.2	77.9
R40	93.9	92.8	77.8
R50	93.5	91.8	74.8
F40	87.1	85.5	69.9
F25	83.9	82.8	69.6
F50	85.4	83.4	67.1
RAdjCtrl	66.1	65.8	60.3

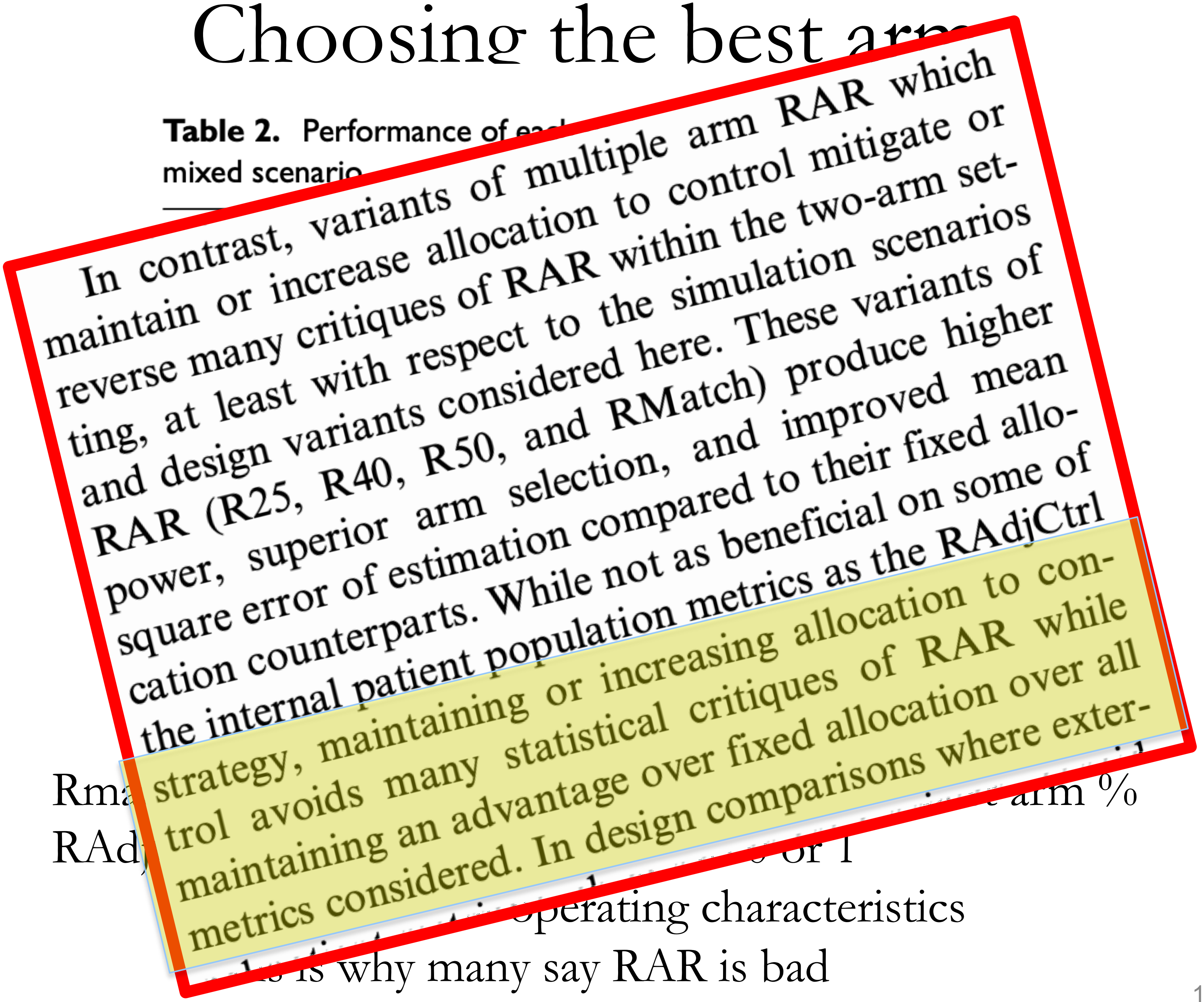
Mixed case with

Control = 35%
Arm 1 = 45%
Arm 2 = 55%
Arm 3 = 65%

Choosing the best arm

Table 2. Performance of ea

mixed scenario



In contrast, variants of multiple arm RAR which maintain or increase allocation to control mitigate or reverse many critiques of RAR within the two-arm setting, at least with respect to the simulation scenarios and design variants considered here. These variants of RAR (R25, R40, R50, and RMatch) produce higher power, superior arm selection, and improved mean square error of estimation compared to their fixed allocation counterparts. While not as beneficial on some of the internal patient population metrics as the RAdjCtrl strategy, maintaining or increasing allocation to control avoids many statistical critiques of RAR while maintaining an advantage over fixed allocation over all metrics considered. In design comparisons where exter-

Rma arm %

RAdj 0 or 1

operating characteristics

is why many say RAR is bad

Adaptive Randomization[1]

- Pros
  - Resolve conflict of healer vs. investigator
  - Maximize number of patients assigned more effective therapy
  - Consistent with current theories of continuous quality improvement
- Cons
  - Must be one (or few) outcome(s) of interest
  - Outcomes must be apparent in a short timeframe relative to accrual time
  - May be statistically less efficient
  - Estimates affected by population drift during accrual

[1] Used with permission, Robert Truog, http://www.bioethics.nih.gov/slides04/truog.ppt

Conclusions for Adaptive Designs in Comparative Effectiveness Research

- Adaptive trials / adaptive CER processes more closely mimic real-life human learning & decision making
- Ongoing projects: Learn & Adapt
  - randomize patients to best products
  - drop treatments/strategies that prove less effective
  - include new treatments as they come to market
  - provide constant sharing of information
  - encourage better patient management

Why Adapt?

The Prospective Postmortem

- Consider whether any adaptations might be added to *prospectively* address *potential* regrets

Why Adapt?
The Prospective Postmortem

- Consider whether any adaptations might be added to *prospectively* address *potential* regrets

- Be honest with yourself in design Phase
  - We overestimate treatment effects
  - We underestimate variability
  - Because we need to justify a doable trial
  - Because we can't be honest in grant proposals

Equipoise

- Would you rather be the last patient enrolled in a clinical trial or the first person treated after its results are published?

- Declaration of Helsinki:
  - "considerations related to the well-being of the human subject should take precedence over the interests of science and society"

ESSET Code

Definitions, Trial Parameters

```
rm(list=ls())
## All times in months
library(VGAM)
v = list(
  ### Event, success probabilities for IV, IV+2nd therapy, Oral, Oral + 2nd therapy
  S3 = c(## There are success rates for the three groups
     0.50,      # fPHT
     0.50,      # LVT
     0.50       # VPA
  ),

  # Maximum sample size & max sample size for Stage 1
  MaxN = 795,
  # Priors
  a = rep(1, 3),
  b = rep(1, 3),
  # First look and look every
  firstlook = 300,
  firststop = 400
  lookevery = 100,
  # Min to randomized
  minpr = 0.05,
  # simulations
  nsims = 1000,
  badlim = 0.25,
  # critv to (a) for 'best'
  #          (b) for 'worst'
  #          (c) to stop for futility (i.e Pred prob a winner or loser id'd)
  #          (d) for worse than 25%
 critv = c(.975, .975, 0.05, 0.05)
)
```

Response Rates

MaxN

Priors

Sample Size &
Timing of Looks

Critical values for stopping

```
simtrials <- function(v){

  co <- ppcutoffs(v$critv[3])

  #out.mat
  # (1) N
  # (2-4) N per group
  # (5-7) Rank as 1, 2, 3 (according to prob best)
  # (8)   Sig best (1 2 or 3 or 0 if none)
  # (9)   Sig worst (1 2 or 3 or 0 if none)
  # (10) Final conclusion
  #           1 = overall futility stop,
  #           2 = stop early for winner
  #           3 = stop early for winner & loser
  #           4 = stop early for loser and futility (not possible in ours)
  #           5 = max overall futility
  #           6 = max and loser
  #           7 = max and winner
  #           8 = max & winner & loser
  #  (11-13) Final Pr(best)
  #  (14-16) Final Pr(2nd)
  #  (17-19) Final Pr(worst)
  #  (20-22) Successes per group
  #  (23-25) Ever drop arm? (rand goes to 0 at any pt)
```

Creates a big matrix to store simulation results

```
out.mat <- matrix(NA, nrow=v$nsims, ncol=25)
  for(s in 1:v$nsims){
    ad <- c(1,1,1)
    ## Rand assignment for first FirstLook pts & generate outcome
    group <- rep(NA, v$MaxN)
    group[1:v$firstlook] <- rand.new(v$firstlook, c(1,1,1))
    y <- rep(NA, v$MaxN)
    y[1:v$firstlook] <- sim.endpoint(group[1:v$firstlook], v$S3)
    look1 <- interim(v$firstlook, y, group, v, co)
#    print(round(look1,3))
    # Track if arm every dropped
    ad <- ad * as.numeric(look1[12:14]>0)
    n.now <- v$firstlook
    print(c(s,n.now))
 ## Now loop through Stage 1
    while(look1[1]==1){
      new <- min(v$MaxN-n.now, v$lookevery)
      group[(n.now+1):(n.now+new)] <- rand.new(new, look1[12:14])
      y[(n.now+1):(n.now+new)] <- sim.endpoint(group[(n.now+1):(n.now+new)], v$S3)
      look1 <- interim(n.now+new, y, group, v, co)
#      print(round(look1,3))
      ad <- ad * as.numeric(look1[12:14]>0)
      n.now <- n.now+new
      print(c(s,n.now))
    }
```

Simulate group assignment & response to tx

First interim look

Simulate group assignment & response to tx

Do interim looks

```
    mx <- look1[3:5];    mn <- look1[6:8]
    winner <- ifelse(max(mx) > v$critv[1], (1:3)[mx==max(mx)], 0)
    loser <-  ifelse(max(mn) > v$critv[2], (1:3)[mn==max(mn)], 0)
```

See if best or worst identified

```
   if(look1[2]==1){
       whystop <- 1    ## futility
     }else if(look1[2]==3){
       if(loser>0){
         whystop <- 3
       }else{
         whystop <- 2
       }
     }else if(look1[2]==2){
       if(winner==0 & loser==0)   { whystop <- 5}
       else if(winner>0 & loser>0){ whystop <- 8}
       else if(winner>0)          { whystop <- 7}
       else if(loser>0)           { whystop <- 6}
       else{print("error why stop at max?")}
       else{print("error, why did trial stop?")}
```

See if stopping rules met

```
 out.mat[s,1:25] <- c(n.now, look1[18:20], order(mx), winner, loser,
                 whystop,look1[c(3,4,5,9,10,11,6,7,8,15,16,17)],1-ad)
}
    out.mat <- data.frame(out.mat)
    names(out.mat) <- c("N","N1","N2","N3",…
     return(out.mat)
```

Print out simulation results

```
sumtrial <- function(outmat){
  mat <- matrix(nrow=4, ncol=9)
  out <- table(factor(outmat[,10], levels=1:8))
#              Ntotal SDN phat Rank1  Rank2  Rank3  SigBest SigWorst Drop
#      fPHT
#      LVT
#      VPA             --
#      Total
  mat[1:3,1] <- apply(outmat[,2:4], 2, mean)
  mat[1:3,2] <- apply(outmat[,2:4], 2, sd)
  mat[1:3,3] <- c(mean(outmat[,20]/outmat[,2]), mean(outmat[,21]/outmat[,3]), 
mean(outmat[,22]/outmat[,4]))
  mat[1,4:6] <- table(factor(outmat[,5], levels=3:1))/dim(outmat)[1]
  mat[2,4:6] <- table(factor(outmat[,6], levels=3:1))/dim(outmat)[1]
  mat[3,4:6] <- table(factor(outmat[,7], levels=3:1))/dim(outmat)[1]
  mat[1:3,7] <- table(factor(outmat[,8], levels=1:3))/dim(outmat)[1]
  mat[1:3,8] <- table(factor(outmat[,9], levels=1:3))/dim(outmat)[1]
  mat[1:3,9] <- apply(outmat[,23:25], 2, mean)
  mat[4,1] <- mean(outmat[,1])
  mat[4,2] <- sd(outmat[2])
  mat[4,3] <- mean(rowSums(outmat[,20:22]) / rowSums(outmat[2:4]))
  mat[4,4:6] <- NA
  mat[4,7] <- sum(mat[1:3,7])
  mat[4,8] <- sum(mat[1:3,8])
  mat[4,9] <- NA
  mat <- data.frame(mat)
  names(mat) <- c("N","SD","Phat","Best","Mid","Worst","SigBest","SigWorst","Drop")
  dimnames(mat)[[1]] <- c("fPHT","LVT","VPA","Total")
  return(list(out, mat))
}
```

Takes the results of ‘simtrials’ and
Produces prettier output

```
interim <- function(N, y, group, v, co){
   ## Runs trial returns:
   #  (1) go (0=stop, 1=keep going)
   #  (2) why stop (1=3-way fut, 2=max n, 3=1 winner)
   #  (3-5) Pr each is best
   #  (6-8) Pr each is worst
   #  (9-11) x/N for each group
   #  (12-14) rand probs
   ns <- table(factor(group[1:N], levels=1:3))
   tab <- table(factor(group[1:N],levels=1:3), factor(y[1:N], levels=0:1))
```

Does interim analysis
Calc posteriors, new
rand probs,
Pred prob of success
at max

```
   post1 <- rbeta(10000, v$a[1]+tab[1,2], v$b[1]+tab[1,1])
   post2 <- rbeta(10000, v$a[2]+tab[2,2], v$b[2]+tab[2,1])
   post3 <- rbeta(10000, v$a[3]+tab[3,2], v$b[3]+tab[3,1])
```

Calc posteriors

```
   vr <- as.numeric(( (v$a+tab[,2])*(v$b+tab[,1])) / ((v$a+v$b+ns)^2 * (v$a+v$b+ns+1)))
   top <- apply(cbind(post1,post2,post3), 1, max)
   bot <- apply(cbind(post1,post2,post3), 1, min)

   best <- c(mean(post1==top), mean(post2==top), mean(post3==top))
   worst <- c(mean(post1==bot), mean(post2==bot), mean(post3==bot))
   middle <- 1-best-worst
```

Calc prob each is
best & worst

```
   toobad <- 1-c(pbeta(v$badlim, v$a[1]+tab[1,2], v$b[1]+tab[1,1]),
               pbeta(v$badlim, v$a[2]+tab[2,2], v$b[2]+tab[2,1]),
               pbeta(v$badlim, v$a[3]+tab[3,2], v$b[3]+tab[3,1]))
```

Calc Pr(p<0.25)

```
   wt <- sqrt(best * vr / as.numeric(ns));     wt <- wt/sum(wt)
   wt[wt < v$minpr] <- 0;    wt[toobad < v$critv[4]] <- 0
   if(sum(wt) > 0){
   wt <- wt/sum(wt)
   }
```

Calc new rand prob

```
 #####PRED PROBS; only do if all 3 arms left
   if((N >= v$firststop) & (N < v$MaxN) & (prod(wt>0)> 0)){
     drop <- 0
     left <- v$MaxN - N
     left <- ceiling(rep(left/3, 3))
     ns.total <- ns+left
     winlose <- 0
     counter <- 1

     while((winlose < co[counter,1]) & (winlose >= co[counter,2]) & (counter < 1000)){
       y.end <- tab[,2] + rbetabin.ab(3, left, v$a+tab[,2], v$b+tab[,1])
       post1f <- rbeta(10000, v$a[1]+y.end[1], v$b[1]+ns.total[1]-y.end[1])
       post2f <- rbeta(10000, v$a[2]+y.end[2], v$b[2]+ns.total[2]-y.end[2])
       post3f <- rbeta(10000, v$a[3]+y.end[3], v$b[3]+ns.total[3]-y.end[3])
       topf <- apply(cbind(post1f,post2f,post3f), 1, max)
       botf <- apply(cbind(post1f,post2f,post3f), 1, min)
       bestf <- c(mean(post1f==topf), mean(post2f==topf), mean(post3f==topf))
       worstf <- c(mean(post1f==botf), mean(post2f==botf), mean(post3f==botf))
       winlose <- winlose + ifelse((max(bestf)>v$critv[1]) | (max(worstf)>v$critv[2]),
1, 0)
       counter <- counter + 1
#       print(c(winlose/counter, counter))
     }
     ppwin <- winlose/counter
   }else{
     drop <- 1
     ppwin <- v$critv[3]+1 #  If missing just make bigger than the crit value.
   }
```

Calc pred prob of success
At Max N

```
   ## Stopping:
   if(N < v$firststop){
     go <- 1
     whystop <- NA
   }else if(N >= v$MaxN){
     go <- 0
     whystop <- 2
   }else if(max(best) > v$critv[1]){
     go <- 0
     whystop <- 3
   }else if(ppwin < v$critv[3]){
     go <- 0
     whystop <- 1
   }else if(wt[1]==0 & wt[2]==0 & wt[3]==0){
     go <- 0
     whystop <- 1
   }else{
     go <- 1
     whystop <- NA
   }

  return(as.numeric(c(go, whystop, best, worst, middle, wt, tab[,2], ns, ppwin, drop)))
}
```

Track IF stop
And WHY stop

Thanks for a great class

What did you like?
What worked?
What did not?

Survey for Tomorrow

- Question 1
  - More examples, less detail per example
  - Fewer examples, more detail per example

- What to cover
  - Platform trials
  - Phase 1, Borrowing
  - Device trials
  - In depth example of Phase 2, Dose finding trials
  - In depth example of Phase 3, Goldilocks trial