Adaptive, Bayesian, and Complex Clinical Trials: What, Why, and How

Scott M. Berry
scott@berryconsultants.com
@statberry
September 23, 2020

Berry Consultants
Statistical Innovation

1

Outline

- What are Adaptive Designs?
- Why Bayes?
- What are 'Complex' trials?
- What are Simulations?
- Examples
  - Phase 1 Pediatric & Adults
  - Trulicity
  - ICECAP
  - DAWN

Berry Consultants

2

2

Austin Bradford Hill

- Credited with designing the first randomized clinical trial in humans
- Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948; 2:769-782.

Born	8 July 1897
Died	18 April 1991 (aged 93)
Nationality	United Kingdom
Occupation	Epidemiologist statistician
Known for	"Bradford Hill" criteria
Awards	Guy Medal (Gold, 1953)

3

3

Table I.–*Condition on Admission*

General Condition	S Group	C Group	Max. Evening Temp. in First Week*	S Group	C Group	Sedimentation Rate	S Group	C Group
Good .	8	8	98-98·9° F. (36·7-37·15° C.)	3	4	0-10	0	0
Fair . .	17	20	99-99·9° F. (37·2-37·75° C.)	13	12	11-20	3	2
Poor . .	30	24	100-100·9° F. (37·8-38·25° C.)	15	17	21-50	16	20
			101° F (38·3° C.) +	24	19	51+	36	29
Total	55	52	Total	55	52	Total	55	51†

* Temperature by mouth in all but six cases. † Examination not done in one case.

Table II.–*Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission*

Radiological Assessment	Streptomycin Group		Control Group	
Considerable improvement . .	28	*51%*	4	*8%*
Moderate or slight improvement	10	*18%*	13	*25%*
No material change	2	*4%*	3	*6%*
Moderate or slight deterioration	5	*9%*	12	*23%*
Considerable deterioration . .	6	*11%*	6	*11%*
Deaths	4	*27%*	14	*27%*
Total	55	*100%*	52	*100%*

4

4

Randomized Clinical Trials

- Incredible innovation in health care and science
- Pre-1948 relied on anecdote and observational studies
- For 50 years the 'science of the clinical trial' barely changed
- Trials are "long boxes" designed to answer a single question
  - "Not sustainable" —Janet Woodcock, FDA
- ***Trial design science is being innovated***

5

5

Adaptive Designs

- What is an adaptive design?
  – A design that has pre-specified dynamic aspects that are determined by the accruing information
  – Adaptive … "By Design"

JAMA 2006; 296:1955-1957

6

6

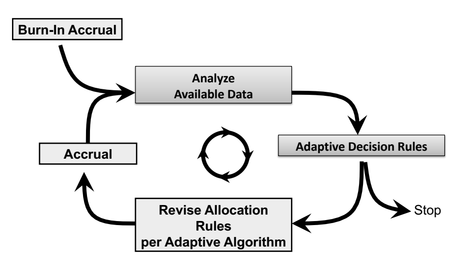
Adaptive Promise

- During the course of the trial things are learned that – *had you known before the trial started* – you would have adapt the design.
  – Learned: It is important that the trial learn about the important aspects, and efficiently.
    - Dose-response models, Longitudinal models, prediction, imputation, biomarkers,…
  – Adapt: The dynamically moving aspects of the trial: prospective changes

7

7

Typical Adaptive Design



8

8

What Phase/Stage of CT?

- Phase I:
  - Sample size
  - Dose escalation
  - Combination of arms
  - Seamless phase I-II
- Phase II/Pilot:
  - Sample size
  - Dose allocation
  - Introduce/Drop arms
  - Enrichment
  - Prediction of Phase III
  - Seamless phase II-III
  - Platform Trials
- Phase III/Confirmatory:
  - Sample size
  - Multiple Arms
  - Accrual Interim Analyses
  - Futility Analyses
  - Timing of Conclusions
  - Enrichment
  - Platform Trials
- Phase IV:
  - Sample size
  - Timing of Conclusions
  - Indications
  - Platform Trails

Berry Consultants

9

9

Therapeutic Areas/Diseases

- Oncology
- Migraine
- Lupus
- Sepsis
- Diabetes
- Obesity
- Stroke
- Tinnitus
- MS
- CHD
- Smoking Cessation
- Gastroparesis
- Alzheimers
- Atrial Fibrillation
- Cancer diagnostic
- Disc Disease
- Contraceptives
- Valves/stents
- Asthma
- Emphysema
- PFO
- RA
- Sleep Apnea
- Chronic Cough
- Osteoparesis
- Parkinsons
- Pain
- Hydrocephalus
- HIV
- Schizophrenia
- Crohns
- Spinal Cord Injury
- Hep C
- Preterm Labor
- Constipation
- Micturition
- Drooling
- PO Ileus
- DVT
- Sexual health
- Emesis
- Statins
- Infections
- OAB
- TB
- Head Trauma
- Cardiac Arrest
- ALS
- Alcohol Abuse
- Drug Abuse
- CHF
- Influenza
- Epilepsy
- BPS
- Crohns
- Drug Resistant Path.
- Many Diagnostics
- Hypertension
- Insomnia
- CMV
- Amyloidosis
- Sickle Cell Disease
- COPD
- GNE Myopathy
- SMA
- RSV
- Prater-Willi
- EBOLA

Berry Consultants

10

10

Statistical Limits?

- What statistical aspects of a problem may provide limitations for adaptation?
  - –Time to Information

Berry Consultants

11

11

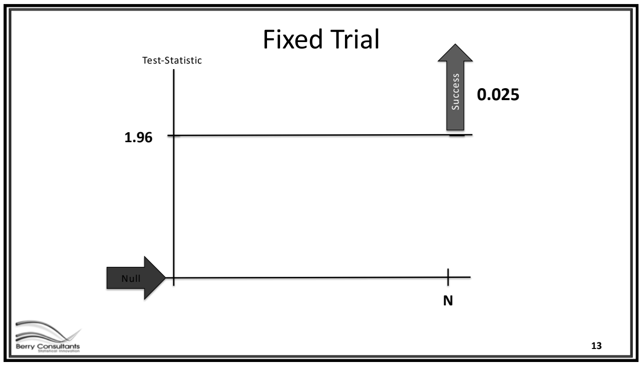
Simple Group Sequential

- Think of a trial with a single analysis after a sample size of N
- We can use a critical value of the test-statistic, such that the type I error is the needed level (say one-sided 0.025): 1.96

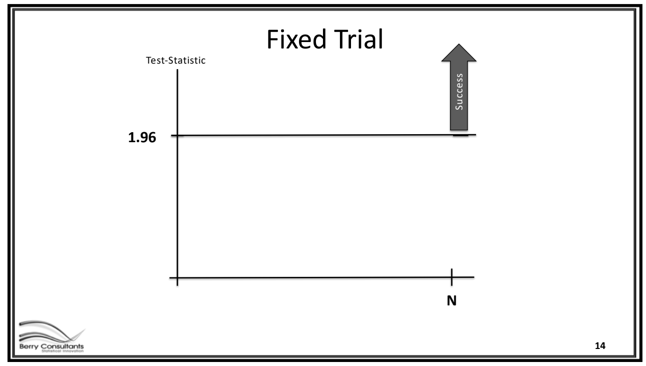
Berry Consultants

12

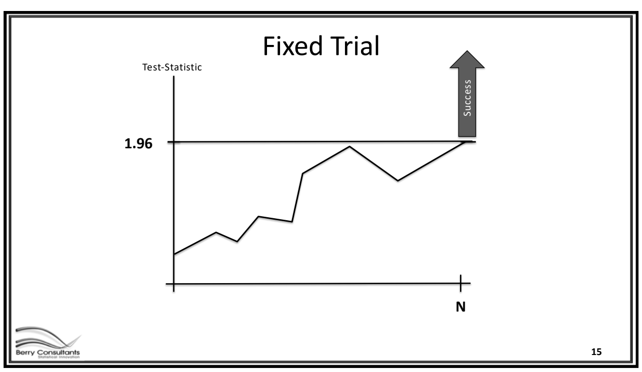
12


Fixed Trial
Test-Statistic
Success
0.025
1.96
Null
N
Berry Consultants
13

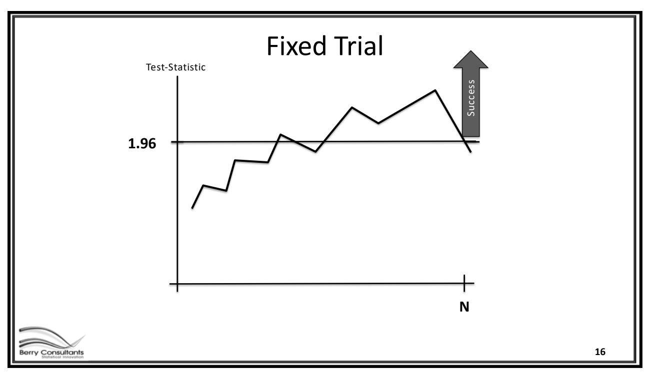
13


Fixed Trial
Test-Statistic
Success
1.96
N
Berry Consultants
14

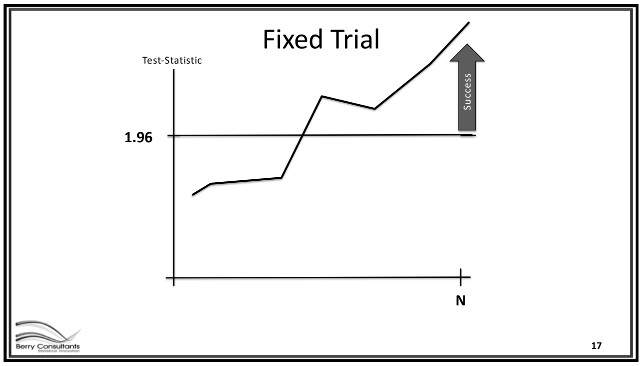
14


Fixed Trial
Test-Statistic
Success
1.96
N
Berry Consultants
15

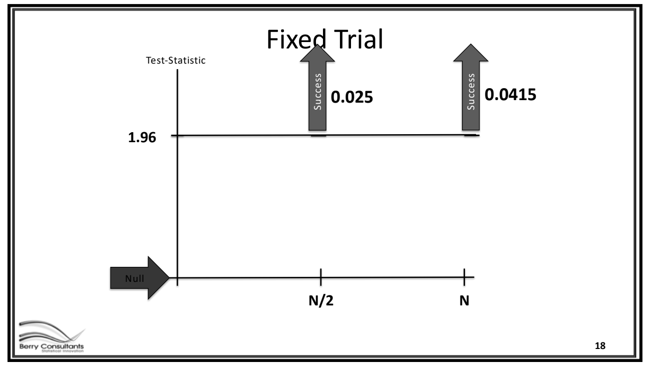
15


Fixed Trial
Test-Statistic
Success
1.96
N
Berry Consultants
16

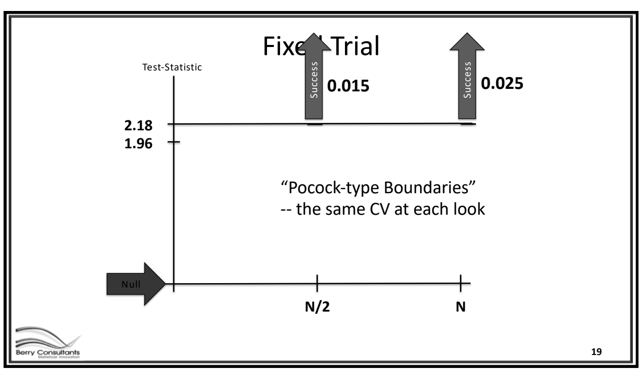
16


Fixed Trial
Test-Statistic
1.96
Success
N
Berry Consultants
17

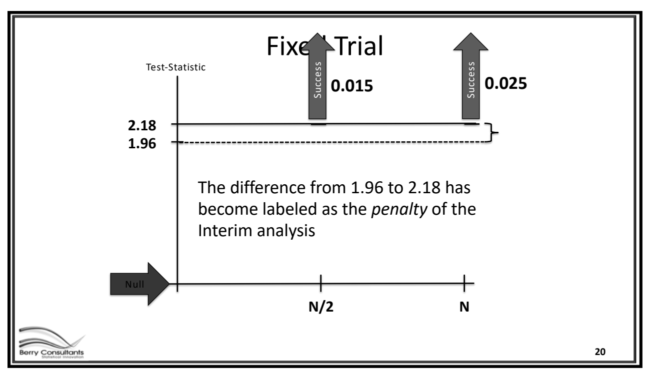
17


Fixed Trial
Test-Statistic
1.96
Success
0.025
Success
0.0415
Null
N/2
N
Berry Consultants
18

18

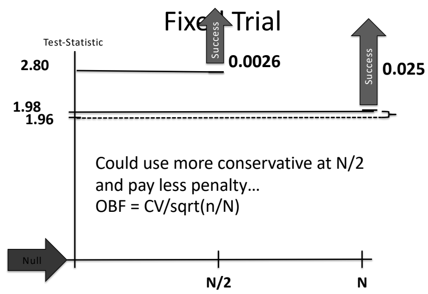

Fixed Trial
Test-Statistic
2.18
1.96
Success
0.015
Success
0.025
“Pocock-type Boundaries”
-- the same CV at each look
Null
N/2
N
Berry Consultants
19

19


Fixed Trial
Test-Statistic
2.18
1.96
Success
0.015
Success
0.025
The difference from 1.96 to 2.18 has become labeled as the *penalty* of the Interim analysis
Null
N/2
N
Berry Consultants
20

20

Fixed Trial



Could use more conservative at N/2
and pay less penalty…
OBF = CV/sqrt(n/N)

Berry Consultants

21

21

Group Sequential

- You can be very aggressive (1.96 at N/2) … to very conservative… but you need to adjust the CV to win the trial depending on these looks
  - We can do the math to find these values
- Cautionary Note:
  - You do NOT pay a penalty for *looking at the data* you pay a penalty for an ACTION that could result in an increase in the probability of success
    - Futility, safety, adjust randomization, bigger N,…

Berry Consultants

22

22

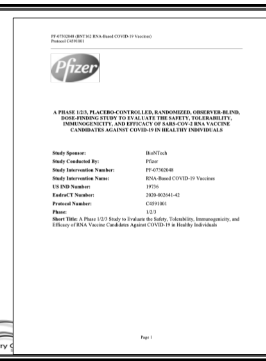
Actions/Data

	Action	
At Interim	Decrease N	Increase N
Data is Positive	Increase T1E Success Stopping	Decrease T1E RAR Promising Zone
Data is Negative	Decrease T1E Futility Stopping	Increase T1E Promising Zone Goldilocks

Berry Consultants

23

23


PF-07302048 (BNT162 RNA-Based COVID-19 Vaccines)
Protocol C4591001

Pfizer

A PHASE 1/2/3, PLACEBO-CONTROLLED, RANDOMIZED, OBSERVER-BLIND, DOSE-FINDING STUDY TO EVALUATE THE SAFETY, TOLERABILITY, IMMUNOGENICITY, AND EFFICACY OF SARS-COV-2 RNA VACCINE CANDIDATES AGAINST COVID-19 IN HEALTHY INDIVIDUALS

Study Sponsor:	BioNTech
Study Conducted By:	Pfizer
Study Intervention Number:	PF-07302048
Study Intervention Name:	RNA-Based COVID-19 Vaccines
US IND Number:	19736
EudraCT Number:	2020-002641-42
Protocol Number:	C4591001
Phase:	1/2/3

Short Title: A Phase 1/2/3 Study to Evaluate the Safety, Tolerability, Immunogenicity, and Efficacy of RNA Vaccine Candidates Against COVID-19 in Healthy Individuals

Page 1

COVID Vaccine Trial

VE for the first primary objective will be evaluated . Overwhelming efficacy will be declared if the first primary study objective is met. The criteria for success at an interim analysis are based on the posterior probability (ie, P[VE >30%|data]) at the current number of cases. Overwhelming efficacy will be declared if the posterior probability is higher than the success threshold.

The success threshold for each interim analysis will be calibrated to protect overall type I error at 2.5%. Additional details about the success threshold or boundary calculation at each interim analysis will be provided in the SAP.

24

24

COVID Vaccine Trial

PF-07302048 (BNT162 RNA-Based COVID-19 Vaccines)
Protocol C4591001

A PHASE 1/2/3, PLACEBO-CONTROLLED, RANDOMIZED, OBSERVER-BLIND, DOSE-FINDING STUDY TO EVALUATE THE SAFETY, TOLERABILITY, IMMUNOGENICITY, AND EFFICACY OF SARS-COV-2 RNA VACCINE CANDIDATES AGAINST COVID-19 IN HEALTHY INDIVIDUALS

Study Sponsor:	BioNTech
Study Conducted By:	Pfizer
Study Intervention Number:	PF-07302048
Study Intervention Name:	RNA-Based COVID-19 Vaccines
US IND Number:	19736
EudraCT Number:	2020-002641-42
Protocol Number:	C4591001
Phase:	1/2/3

Short Title: A Phase 1/2/3 Study to Evaluate the Safety, Tolerability, Immunogenicity, and Efficacy of RNA Vaccine Candidates Against COVID-19 in Healthy Individuals

Page 1

The study will stop for lack of benefit (futility) if the predicted probability of success at the final analysis or study success is <5%. The posterior predictive POS will be calculated using a beta-binomial model. The futility assessment will be performed for the first primary endpoint and the futility boundary may be subject to change to reflect subsequent program-related decisions by the sponsor.

Berry Consultants

25

25

Adaptive Design

Table 5. Interim Analysis Plan and Boundaries for Efficacy and Futility

Analysis	Number of Cases	Success Criteria[a]	Futility Boundary
		VE Point Estimate (Case Split)	**VE Point Estimate (Case Split)**
IA1	32	76.9% (6:26)	11.8% (15:17)
IA2	62	68.1% (15:47)	27.8% (26:36)
IA3	92	62.7% (25:67)	38.6% (35:57)
IA4	120	58.8% (35:85)	N/A
Final	164	52.3% (53:111)	

Abbreviations: IA = interim analysis; N/A = not applicable; VE = vaccine efficacy.
Note: Case split = vaccine : placebo.
a. Interim efficacy claim: P(VE >30%|data) > 0.995; success at the final analysis: P(VE >30%|data) > 0.986.

Berry Consultants

26

26

COVID Vaccine Trial

PF-07302048 (BNT162 RNA-Based COVID-19 Vaccines)
Protocol C4591001

A PHASE 1/2/3, PLACEBO-CONTROLLED, RANDOMIZED, OBSERVER-BLIND, DOSE-FINDING STUDY TO EVALUATE THE SAFETY, TOLERABILITY, IMMUNOGENICITY, AND EFFICACY OF SARS-COV-2 RNA VACCINE CANDIDATES AGAINST COVID-19 IN HEALTHY INDIVIDUALS

Study Sponsor: BioNTech
Study Conducted By: Pfizer
Study Intervention Number: PF-07302048
Study Intervention Name: RNA-Based COVID-19 Vaccines
US IND Number: 19736
EudraCT Number: 2020-002641-42
Protocol Number: C4591001
Phase: 1/2/3

Short Title: A Phase 1/2/3 Study to Evaluate the Safety, Tolerability, Immunogenicity, and Efficacy of RNA Vaccine Candidates Against COVID-19 in Healthy Individuals

Page 1

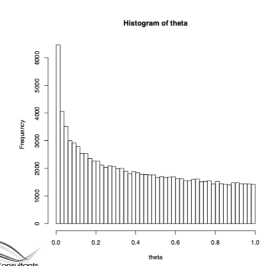
A minimally informative beta prior, beta (0.700102, 1), is proposed for θ = (1-VE)/(2-VE). The prior is centered at θ = 0.4118 (VE=30%) which can be considered pessimistic. The prior allows considerable uncertainty; the 95% interval for θ is (0.005, 0.964) and the corresponding 95% interval for VE is (-26.2, 0.995)

Berry Consultants

27

27

Prior for Theta




```
> ve = (2*theta - 1)/(theta - 1)
```

```
> round(quantile(ve,prob=(1:19)/20),3)
     5%     10%     15%     20%     25%     30%     35% 
-12.415  -5.203  -2.824  -1.654  -0.971  -0.505  -0.180 
    40%     45%     50%     55%     60%     65%     70% 
  0.066   0.256   0.407   0.529   0.629   0.713   0.783 
    75%     80%     85%     90%     95% 
  0.840   0.888   0.929   0.961   0.986 
```

> length(ve[ve>.30])/length(ve)
[1] 0.53689

Berry Consultants

28

28

Operating Characteristics

Table 6. Statistical Design Operating Characteristics: Probability of Success or Failure for Interim Analyses

Vaccine Efficacy (%)	Interim Analysis 1 (Total Cases = 32)		Interim Analysis 2 (Total Cases = 62)		Interim Analysis 3 (Total Cases = 92)		Interim Analysis 4 (Total Cases = 120)
	Probability of Success (Cases in Vaccine Group ≤6)	Probability of Failure (Cases in Vaccine Group ≥15)	Probability of Success (Cases in Vaccine Group ≤15)	Probability of Failure (Cases in Vaccine Group ≥26)	Probability of Success (Cases in Vaccine Group ≤25)	Probability of Failure (Cases in Vaccine Group ≥35)	Probability of Success (Cases Vaccine Group ≤35)
30	0.006	0.315	0.003	0.231	0.002	0.239	0.002
50	0.054	0.078	0.051	0.056	0.063	0.103	0.075
60	0.150	0.021	0.160	0.010	0.175	0.019	0.160
70	0.368	0.003	0.310	<0.001	0.195	0.001	0.085
80	0.722	<0.001	0.238	<0.001	0.037	<0.001	0.003

Table 7. Statistical Design Operating Characteristics: Probability of Success for Final Analysis and Overall

Vaccine Efficacy (%)	Final Analysis (Total Cases = 164)	Overall Probability of Success
	Probability of Success (Cases in Vaccine Group ≤53)	
30	0.007	0.021
50	0.196	0.439
60	0.220	0.866
70	0.036	>0.999
80	<0.001	>0.999

Berry Consultants

29

29

Kert Viele Tweetorial…

Kert Viele
@KertViele

(1/n) The Pfizer SARS-Cov-2 vaccine trial has 4 interim analyses. Each interim has a set number of events and certain splits result in declaring efficacy or futility. If they were to announce "the trial is continuing after interim X", what might we learn?
pfe-pfizercom-d8-prod.s3.amazonaws.com/2020-09/C45910…

6:34 PM · Sep 19, 2020 · Twitter Web App

Berry Consultants

30

30

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19
(2/n) First off, I wouldn't make these announcements. The trial is continuing per protocol. We need the trial to finish per protocol to conclude anything. I think people misinterpret these announcements all the time (hence this tweetorial….). Other opinions differ….

1 14

Berry Consultants

31

31

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19
(3/n) The interims require specific splits. The first interim, for example, happens with 32 events (instances of COVID). If 6 or less of these are in the vaccine arm, the trial declares efficacy. If 15 or more are in the vaccine arm, the trial declares futility.

Table 5. Interim Analysis Plan and Boundaries for Efficacy and Futility

Analysis	Number of Cases	Success Criteria[a]	Futility Boundary
		VE Point Estimate (Case Split)	VE Point Estimate (Case Split)
IA1	32	76.9% (6:26)	11.8% (15:17)
IA2	62	68.1% (15:47)	27.8% (26:36)
IA3	92	62.7% (25:67)	38.6% (35:57)
IA4	120	58.8% (35:85)	N/A
Final	164	52.3% (53:111)	

Abbreviations: IA = interim analysis; N/A = not applicable; VE = vaccine efficacy.
Note: Case split = vaccine : placebo.
a. Interim efficacy claim: P(VE >30%|data) > 0.995; success at the final analysis: P(VE >30%|data) > 0.986.

Berry Consultants

32

32

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(4/n) Thus, knowing the required splits and knowing "the trial is continuing", we can determine the data must lie in a range. If the announcement says "we are continuing after the first interim" we know those 32 events had between 7 and 14 events on the vaccine arm.

1 9

Berry Consultants

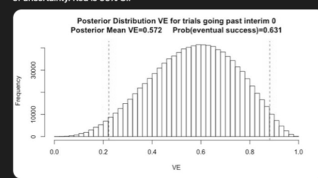
33

33

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(5/n) If I had prior beliefs about the vaccine, I can update them when I hear "the trial continued after interim X. Suppose my original beliefs looks like the graph below (Beta(4,3) for stat folks). I'm generally optimistic, but have a lot of uncertainty. Red is 95% CI.



1 11

Berry Consultants

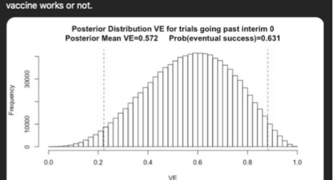
34

34

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(6/n) My prior mean vaccine efficacy is 57%. My prior guess is that the trial has a 63% chance of declaring efficacy. This isn't the power, which assumes the vaccine works. This 63% incorporates my uncertainty on whether the vaccine works or not.



1 7

Berry Consultants

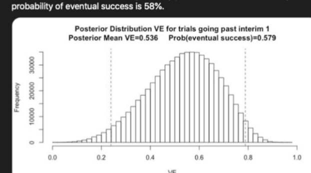
35

35

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(7/n) After a "continue" at interim 1, here are my revised belief. Importantly, they haven't changed much. I'm eliminated some of the values near 100% VE (would have stopped for efficacy) but my posterior mean 0.536 and my probability of eventual success is 58%.



1 6

Berry Consultants

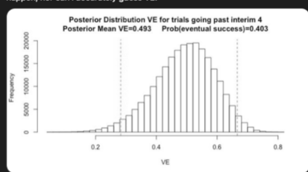
36

36

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(8/n) While continuing interims do make me a little more pessimistic, even continuing after interim 4 my posterior mean is 0.493, probability of eventual success 40%, and CI roughly between 30-70%. I'm never "sure" of what will happen, nor can I accurately guess VE.



1 6

Berry Consultants

37

37

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(9/n) If you walked into the trial incredibly optimistic or pessimistic (but with at least some uncertainty!), then your beliefs would change. The incredible optimist would be disheartened the trial didn't stop early for efficacy, the pessimist would grow more hopeful.

1 8

Berry Consultants

38

38

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(10/n) But if your prior was focused on the 30-70% range, most of that relative probability is unchanged. If the trial is continuing, you tend to stick with that range and have to just wait and see what happens (I'm happy to do these plots for anyone's preferred prior).

1 6

Berry Consultants

39

39

Kert Viele Tweetorial…

Kert Viele @KertViele · Sep 19

(11/n) Quick note, people are WAY too quick to analyze timing of these announcements, etc. If you want to see overreactions to interims…google Oncothyreon, too optimistic early and then disappointment later.

Oncothyreon shares routed as Stimuvax PhIII trial soldi…
Merck KGaA told analysts today that after assessing its late-stage study for the lung cancer vaccine Stimuvax …
fiercebiotech.com

1 8

Kert Viele @KertViele · Sep 19

(12/12) Again, I wouldn't make these announcements on principle, but the information conveyed publicly is fairly small. If the trial is continuing, the vaccine likely is somewhere between a dud and good, and it might end up winning or losing the trial.

3 2 15

Berry Consultants

40

40

Criticism

Eric Topol Retweeted

Reuters @Reuters · Sep 17

Pfizer is betting that early stages of its coronavirus vaccine trial will show that its candidate works. But some experts say such an interim analysis could overstate its effectiveness reut.rs/3iHsplf

4 43 67

Berry Consultants

41

41

Criticism

Eric Topol @EricTopol · Sep 17

The release of @moderna_tx puts the squeeze on @Pfizer to do so. 👍
pfe-pfizercom-d8-prod.s3.amazonaws.com/2020-09/C45910...
The big problem: 4—four—interim analyses at 32, 62, 92, 120, 164 events (infections), essentially engineering a trial to stop early with so many looks. Not good. (@moderna has 2)

Interim Analysis Plan and Boundaries for Efficacy and Futility

Analysis	Number of Cases	Success Criteria[a]	Futility Boundary
		VE Point Estimate (Case Split)	VE Point Estimate (Case Split)
IA1	32	76.9% (6:26)	11.8% (15:17)
IA2	62	68.1% (15:47)	27.8% (26:36)
IA3	92	62.7% (25:67)	38.6% (35:57)
IA4	120	58.8% (35:85)	N/A
Final	164	52.3% (53:111)	

Abbreviations: IA = interim analysis; N/A = not applicable; VE = vaccine efficacy.
Note: Case split = vaccine : placebo.
a. Interim efficacy claim: P(VE >30%|data) > 0.995; success at the final analysis: P(VE >30%|data) > 0.986.

21 103 207

Berry Consultants

42

42

Criticism

Eric Topol @EricTopol · Sep 17

—Rule #1: Don't stop a clinical trial early unless it is for a safety reason or futility
—Otherwise, doing so means "overwhelming efficacy" and that it is unethical to proceed with placebo
—These trials are on soft grounds for that claim
—They need to be completed

2 32 136

Berry Consultants

43

43

Bayesian Statistics

- Reverend Thomas Bayes (1702-1761)
- *Essay towards solving a problem in the doctrine of chances* (1764)

- This paper, on inverse probability, led to the name *Bayesian Statistics*

1

1

Bayes Theorem

$$\Pr\left(A_i \mid B\right) = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\sum_{j=1}^{k}\Pr(B \mid A_j)\Pr(A_j)}$$

$$f\left(\theta \mid X\right) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)\,d\theta}$$

2

2

Compare P-Values/Posteriors

- Inspired by Steve Ruberg Example
- You have a bag of coins, mixed fair coins and a single 2-headed coin
  - Assume a null (H_0) of "fair coin"
  - Alternative (H_1) of "2-headed coin"

- Flip the coin independently n times…

3

3

Data/P-Values

DATA	P-Value
1/1	0.50
2/2	0.25
3/3	0.125
4/4	0.0625
5/5	0.0312
6/6	0.0156
7/7	0.00781
8/8	0.00391
9/9	0.00195
10/10	0.000977
11/11	0.000488
12/12	0.000244

4

4

Bayesian Analysis

- What about a Bayesian analysis?
- Can’t do a Bayesian analysis unless there is a prior probability the coin is fair/2-headed
  – What if there are 50% of the coins in the bag as fair and 2-headed
  – What if there is 1 in 1000 coins being 2-headed

5

5

Data/P-Values

DATA	P-Value	Pr(Fair Coin)	
		50% each	0.001 2-headed
1/1	0.50	0.333	0.998
2/2	0.25	0.200	0.996
3/3	0.125	0.111	0.992
4/4	0.0625	0.0588	0.984
5/5	0.0312	0.0303	0.968
6/6	0.0156	0.0154	0.940
7/7	0.00781	0.00775	0.886
8/8	0.00391	0.00389	0.796
9/9	0.00195	0.00194	0.661
10/10	0.000977	0.000976	0.493
11/11	0.000488	0.000488	0.327
12/12	0.000244	0.000244	0.196
16/16	0.000015	0.000015	0.015

6

6

Bayesian Calculations

- Data: 13 S's and 4 F's
- Parameter = π = P(S)
- For ANY design with these results, the likelihood function is

$$\Pr(\text{data} \mid p) \propto p^{13}(1-p)^4$$

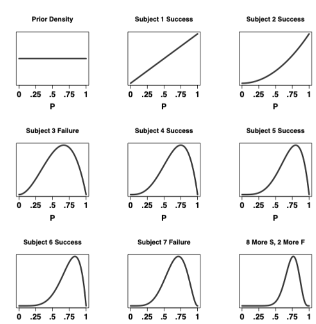
- Posterior probabilities…
  – Lets assume a Beta(1,1)….

7

7

Bayesian Analyses of All…

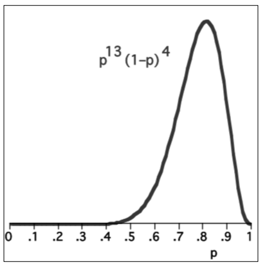
Or Updated sequentially



8

8

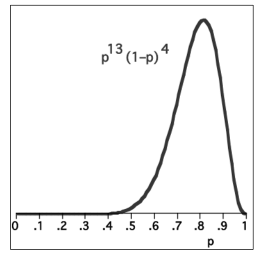
Likelihood function of π



9

9

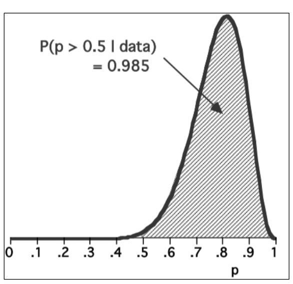
Posterior density of π for uniform prior: Beta(14,5)



10

10

Pr[π > 0.5]



11

11

PREDICTIVE PROBABILITIES

- Distribution of future data?
- P(next is an A) = ?
- Critical component of experimental design
- In monitoring trials

12

12

Predictive Distribution

- The posterior distribution of a future observation of X_i…

$$\left[x_{n+1} \mid x_1, \ldots, x_n\right] = \int \left[x_{n+1} \mid \theta\right]\left[\theta \mid x_1, \ldots, x_n\right] d\theta$$

- The distributions support is on the values of X, not the parameters space
- Convolution of X with respect to the variability in the parameter space

13

13

Suppose 17 more observations

*P(*A wins *x* of next 17 | data*)*

*= EP(*A wins *x* | data*, π)*

$$= E\left[\binom{17}{x} p^x (1-p)^{17-x} \mid data, p\right]$$

Beta-Binomial Distribution

14

14

Possible Calculation

$$\int \binom{17}{x} p^x (1-p)^{17-x} \frac{\Gamma(14)\Gamma(5)}{\Gamma(19)} p^{13} (1-p)^4 \, dp$$

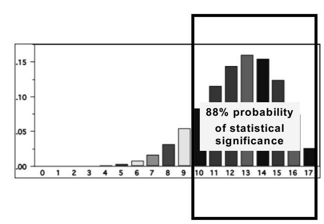
- Simulate a π from the beta(14,5)
- Simulate an x from binomial(17, π)
- Distribution of x's is beta-binomial--the predictive distribution

15

15

Predictive distribution

Predictive distribution of # of successes in next 17 tries:

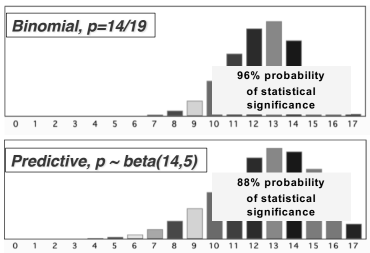


Has more variability than any binomial ⇒

16

16

Best fitting binomial vs. predictive probabilities



17

17

Posterior and Predictive…same?

- Clinical Trial, 100 subjects. H_A: π > 0.25? FDA will approve if # success ≥ 33 [post > 0.95, beta(1,1)]
- See 99 subjects, 32 successes
- Pr[π > 0.25 | data] = 0.955
- Predictive prob trial success = 0.327

18

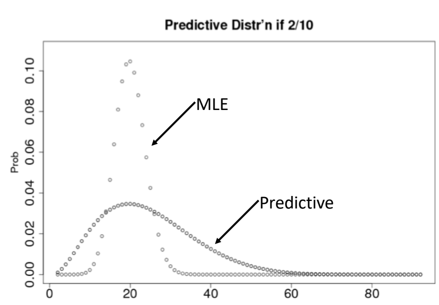
18

Example of Predictive Prob

- Same Trial, 33+ out of 100 is a SUCCESS
- Look at data at *n*=10
- Predict remainder of 90 subjects
- Predictive Prob accounts for uncertainty and “only” 10% of data observed

19

19



20

20

Predictive, Posterior, MLE Project

S@10	Post Prob >0.25	Pred Prob 33+	MLE Proj Prob 33+
0	.042	.0096	0
1	.197	.070	$6.6\text{x}10^{-11}$
2	.455	.234	.00097
3	.713	.487	.279
4	.885	.737	.948
5	.966	.900	.99991
6	.992	.973	1
7	.9988	.995	1

21

21

Interpretation

- Predictive is VERY different than posterior probability
- If you were using frequentist MLE to project you need to have constraints on # subjects before method "kinda works"
- If there is a constraint, it should be on # for MLE not on % of the subjects
- Predictive distribution handles both of these and does not need "constraints"

22

22

The Likelihood Principle

The likelihood function

$L_X(\pi) = f(X \mid \pi)$

contains all the information in an experiment relevant for inferences about π

$$\frac{L_X(\theta)\pi(\theta)}{\int L_X(\theta)\pi(\theta)d\theta}$$

23

23

Consequence of Bayes rule: The Likelihood Principle

The likelihood function

$L_X(\pi) = f(X \mid \pi)$

contains all the information in an experiment relevant for inferences about π

Assume: $L(\theta \mid x) = aL(\theta \mid y)$

$$f(\theta \mid x) = \frac{L(\theta \mid x)\pi(\theta)}{\int L(\theta \mid x)\pi(\theta)d\theta} = \frac{aL(\theta \mid y)\pi(\theta)}{\int aL(\theta \mid y)\pi(\theta)d\theta} = f(\theta \mid y)$$

24

24

Example

- Data: 13 A's and 4 B's
- Parameter = π = P(A wins)
- Likelihood $\propto p^{13}(1-p)^4$
- Frequentist conclusion? Depends on design

25

25

Frequentist hypothesis testing

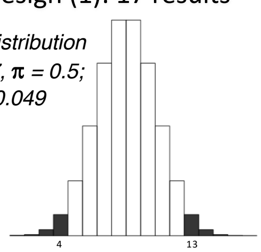
- P-value = Probability of observing data as or more extreme than results, assuming H_0.
- P-V = P(tail of dist. / H_0)
- Four designs:
  (1) Observe 17 results
  (2) Stop trial once both 4 A's and 4 B's
  (3) Interim analysis at 17, stop if 0 - 4 or 13 - 17 A's, else continue to n = 44
  (4) Stop when "enough information"

26

26

Design (1): 17 results

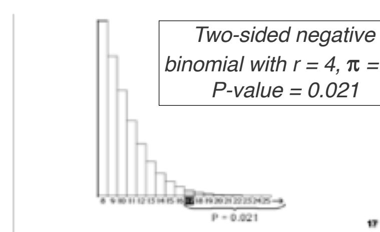
Binomial distribution with n = 17, π = 0.5; P-value = 0.049



27

27

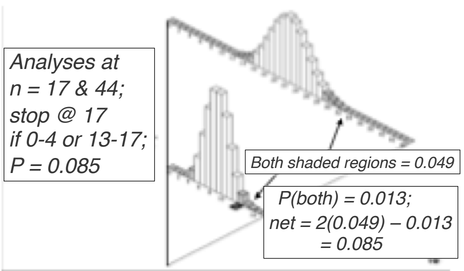
Design (2): Stop when both 4 A's and 4 B's



28

28

Design (3): Interim analysis at n=17, possible total is 44



29

29

Design (4): Scientist's stopping rule: Stop when you know the answer

- Cannot calculate P-value
- Strictly speaking, frequentist inferences are impossible

30

30

Bayesian Stopping Rule

- The Bayesian answer is the **same** in all these trials (assuming independent, identically distributed observations)
- The design – what didn't happen – affects the frequentist based approaches (and bias, and type I error, etc)
  - Violation of the likelihood principle

Critically important for adaptive designs

$f(x|\theta)$ is incredibly restrictive in the space of x!

31

31

Bayes and COVID-19

- Everyone is Bayesian… Why?
- The trial design is unknown! Our REMAP-CAP trial might be 200 if might be 10,000
- Interims are being done monthly, not based on sample size
- Alpha-spending very challenging (impossible)
- Multiple trials have multiple arms, disconnected in time – need modeling
- May need historical controls
- Uncertain trial design…

32

32

JAMA | Original Investigation | CARING FOR THE CRITICALLY ILL PATIENT

Effect of Hydrocortisone on Mortality and Organ Support in Patients With Severe COVID-19

The REMAP-CAP COVID-19 Corticosteroid Domain Randomized Clinical Trial

The Writing Committee for the REMAP-CAP Investigators

Table 2. Primary Outcome

Outcome/analysis[a]	Fixed-dose hydrocortisone (n = 137)	Shock-dependent hydrocortisone (n = 141)	No hydrocortisone (n = 101)
Primary outcome, organ support-free days			
Median (IQR)	0 (-1 to 15)	0 (-1 to 13)	0 (-1 to 11)
Subcomponents of organ support-free days			
In-hospital deaths, No. (%)	41 (30)	37 (26)	33 (33)
Organ support-free days among survivors, median (IQR)	11.5 (0 to 17)	9.5 (0 to 16)	6 (0 to 12)
Primary analysis of the primary outcome, using covariate data from all severe-state participants with COVID-19 (n = 576)[b]			
Adjusted odds ratio			
Mean (SD)	1.47 (0.35)	1.26 (0.31)	1 [Reference]
Median (95% CrI)	1.43 (0.91 to 2.27)	1.22 (0.76 to 1.94)	1 [Reference]
Probability of superiority to no hydrocortisone, %	93	80	
Secondary analysis of the primary outcome, restricted to corticosteroid domain participants (n = 379) with no adjustment for intervention assignment in other domains[c]			
Adjusted odds ratio			
Mean (SD)	1.49 (0.35)	1.28 (0.30)	1 [Reference]
Median (95% CrI)	1.45 (0.93 to 2.30)	1.24 (0.80 to 1.95)	1 [Reference]
Probability of superiority to no hydrocortisone, %	95	83	

33

33

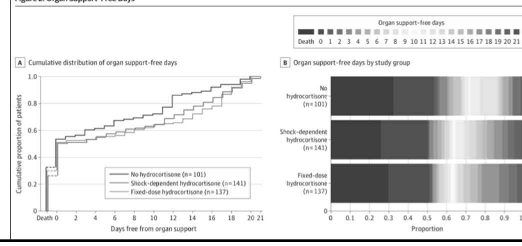
JAMA | Original Investigation | CARING FOR THE CRITICALLY ILL PATIENT

Effect of Hydrocortisone on Mortality and Organ Support in Patients With Severe COVID-19

The REMAP-CAP COVID-19 Corticosteroid Domain Randomized Clinical Trial

The Writing Committee for the REMAP-CAP Investigators

Figure 2. Organ Support-Free Days

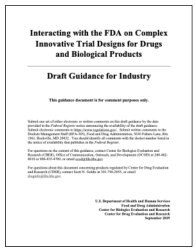


34

34

Complex Innovative Designs

"For the purposes of this guidance, CID includes trial designs that have rarely or never been used to date to provide substantial evidence of effectiveness in new drug applications or biologics license applications. A common feature of many CIDs is the need for simulations rather than mathematical formulae to estimate trial operating characteristics (Section III of this guidance)."


Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products

Draft Guidance for Industry

This guidance document is for comment purposes only.

1

Complex Innovative Designs

"For the purposes of this guidance, CID includes trial designs that have rarely or never been used to date to provide substantial evidence of effectiveness in new drug applications or biologics license applications. A common feature of many CIDs is the need for simulations rather than mathematical formulae to estimate trial operating characteristics (Section III of this guidance)."


Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products

Draft Guidance for Industry

This guidance document is for comment purposes only.

2

- Lots of example of complex designs to come…

3

DIAN-TU

- Randy Bateman, PI
- Dominantly Inherited Alzheimer Network (DIAN) is an international research partnership of leading scientists determined to understand a rare form of Alzheimer's disease (ADAD) that is caused by a gene mutation.
- Autosomal Dominant Alzheimer's Disease (ADAD) is caused by rare inherited gene mutations in the APP, PSEN1, or PSEN2 genes which lead to early-onset AD (<60 years old)
  - 40-80% of 41.2/100,000 (AD < 60 y.o)

4

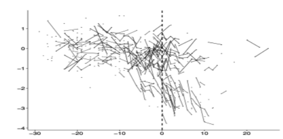
DIAN-TU: Design

- Two drugs 3:1; approx. 60 vs 20
- A single analysis takes place when the last enrolled reaches 4 years; fixed sample size; simple design
- Each arm is compared to placebo (well combined ~ 60 vs 40)
- Analysis is posterior probability of superiority…

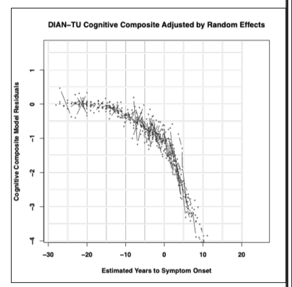
Page 5

5

DIAN-TU: Analysis



Hassenstab J, Lim YY, Fagan AM, Ma S, Xiong C, Bateman RJ, Morris JC, and the Dominantly Inherited Alzheimer Network. (2015). CSF Biomarkers and Cognitive Decline in Autosomal Dominant Alzheimer Disease. Presentation at the 2015 Alzheimer's Association International Conference. Washington, DC, USA.



Page 6

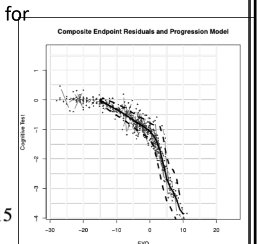
6

DIAN-TU: Bayesian Analysis

- Let Y_{ij} be the jth cognitive observation for subject i=1,…,k, with n_i observations
- Let E_{ij} be the EYO value for (i,j)

$$Y_{ij} = \gamma_i + f\left(E_{ij} - \delta_i \mid \alpha\right) + \varepsilon_{ij}$$

$$f(t) = \begin{cases} 0 & t \le -15 \\ (1 + \lfloor t \rfloor - t)\alpha_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\alpha_{\lfloor t \rfloor + 1} & -15 < t \le 15 \\ \alpha_{15} & t > 15 \end{cases}$$



Page 7

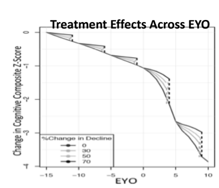
7

DIAN-TU: Bayesian Analysis

Treatment Effect

$$Y_{ij} = \gamma_i + g\left(E_{ij} \mid \delta_i, \theta\right) + \varepsilon_{ij}$$

$$g(E \mid \delta_i, \theta) = \begin{cases} f(E - \delta_i \mid \alpha) & E \le T \\ f(T - \delta_i \mid \alpha) + \exp(\theta)\left[f(E - \delta_i \mid \alpha) - f(T - \delta_i \mid \alpha)\right] & E > T \end{cases}$$



- T is time of intervention (on EYO scale)

Page 8

8

DIAN-TU: Bayesian Analysis

- Test

$$H_0 : \exp(\theta) \geq 1$$

$$H_A : \exp(\theta) < 1$$

- If the posterior probability of $\exp(\theta) < 1$ is greater than 0.985* then claim superiority

Page 9

9

DIAN-TU: Bayesian Analysis


ALZFORUM
NETWORKING FOR A CURE

LOGIN

HOME | ABOUT US | NEWSLETTER

NEWS | WEBINARS | DATABASES | PAPERS | PROFESSIONAL RESOURCES | VIRTUAL EXHIBIT HALL | ABOUT AD | MY ALZFORUM

ADD TO MY LIBRARY | FOLLOW NEWS COMMENTS | EMAIL | SHARE | PRINT

Topline Result for First DIAN-TU Clinical Trial: Negative on Primary

ARTICLE | COMMENTS | REFERENCES | FURTHER READING

10 Feb 2020

DIAN participants and investigators today are grappling with difficult news. A topline analysis of the first Phase 2/3 clinical trial that the DIAN-TU trials platform mounted for carriers of dominantly inherited Alzheimer's disease mutations showed that both of the trial's investigational drugs missed the primary endpoint. That endpoint would have been a statistically significant difference between drug and placebo on the DIAN Multivariate Cognitive Endpoint (DIAN-MCE), a composite of four cognitive tests developed by DIAN for this stage and type of Alzheimer's disease. The trial evaluated two investigational anti-amyloid drugs, Roche's gantenerumab and Lilly's solanezumab, against placebo and additional progression data drawn from the DIAN observational study.

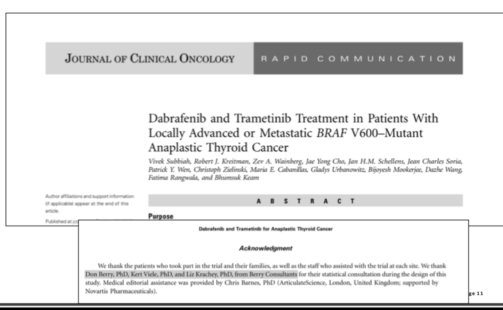
ANNOTATE

To make an annotation you must Login or Register.

Page 10

10

ROAR Trial


JOURNAL OF CLINICAL ONCOLOGY | RAPID COMMUNICATION

Dabrafenib and Trametinib Treatment in Patients With Locally Advanced or Metastatic *BRAF* V600–Mutant Anaplastic Thyroid Cancer

Vivek Subbiah, Robert J. Kreitman, Zev A. Wainberg, Jae Yong Cho, Jan H.M. Schellens, Jean Charles Soria, Patrick Y. Wen, Christoph Zielinski, Maria E. Cabanillas, Gladys Urbanowitz, Bijoyesh Mookerjee, Dazhe Wang, Fatima Rangwala, and Bhumsuk Keam

Author affiliations and support information (if applicable) appear at the end of this article.

Published at jo

ABSTRACT

Purpose

Dabrafenib and Trametinib for Anaplastic Thyroid Cancer

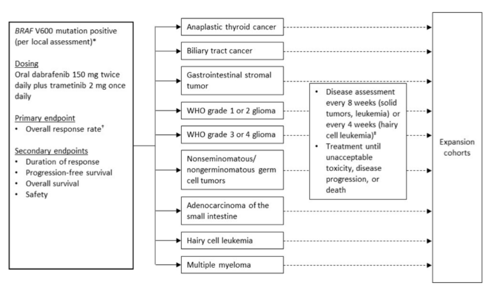
Acknowledgment

We thank the patients who took part in the trial and their families, as well as the staff who assisted with the trial at each site. We thank Don Berry, PhD, Kert Viele, PhD, and Liz Krachey, PhD, from Berry Consultants for their statistical consultation during the design of this study. Medical editorial assistance was provided by Chris Barnes, PhD (ArticulateScience, London, United Kingdom; supported by Novartis Pharmaceuticals).

ge 11

11

ROAR Trial: Basket Trial



Page 12

12

ROAR Trial: Analysis Methods

Statistical Analysis

To address the small sample size per histologic cohort, we used an adaptive design with a Bayesian hierarchical model (Data Supplement) that increases the power to detect clinically meaningful differences in overall response rate by borrowing information across histologic cohorts while controlling the type 1 error rate. This design allowed for multiple interim evaluations of the accumulating data to determine if at least one histologic cohort should discontinue enrollment early because of either success or futility.

13

ROAR Trial: Bayesian Model

$$\theta_g = log\left(\frac{\pi_g}{1-\pi_g}\right) - log\left(\frac{R_g}{1-R_g}\right).$$

Typical BHM

$\theta_g \sim N(\mu, \tau^2)$

$\mu, \tau^2 \sim hyper-prior$

Cluster BHM

$p_G \sim DPM(\alpha)$

"Dirichlet Process Mixture"

Within Each group: **Typical BHM**

14

ROAR Trial: ATC Results

"For the 15 patients with ATC in the primary analysis cohort, the Bayesian estimate of the primary end point—confirmed overall response rate on the basis of investigator assessment—was 69% (95% credible interval, 46.9% to 86.9%). "

11/15 = 0.733

"The posterior probability was 100% that the overall response rate of 69% exceeded the historical control response rate of 15% (Data Supplement), thereby meeting the protocol-specified rules for early stopping for efficacy. "

15

What are Simulations?

We are inundated with "simulations" being used as *predictions*

Berry Consultants

16

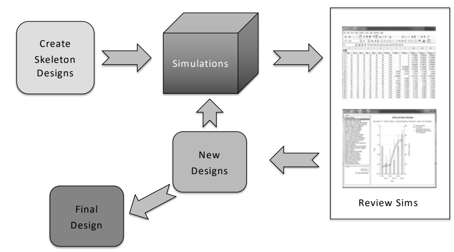
Role of Simulations

- This is common for PK/PD scientists – *predict* what will happen in humans
- This is not how simulations are used in creating *in silico designs*
- The "simulation evaluation" is nothing more than *numerical integration*
- Calculating operating characteristics *exactly*

Berry Consultants 17

17

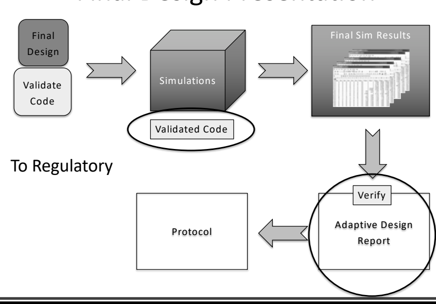
Design Process



Berry Consultants 18

18

Final Design Presentation



Berry Consultants 19

19

Adaptive Design Report

- A report that presents the full details of the design; adaptations, modeling, and simulations
  - –Allow completely reproducible results
- We have focused on the design and not why the design

Berry Consultants 20

20

Bayesian Adaptive Trial Design for the Influence of Cooling duration on Efficacy in Cardiac Arrest Patients (ICECAP)

Berry Consultants, LLC
January 16, 2016

1.0 Background

This is a multicenter, randomized adaptive clinical trial in comatose adult survivors of out of hospital cardiac arrest that have already been rapidly cooled. A maximum of 1800 patients will be enrolled. The primary endpoint is a weighted modified Rankin Score (mRS) measured at 90 days after the return of spontaneous circulation. This trial will enroll patients with and without initial shockable rhythms. Within each rhythm type, patients will be adaptively randomized to a cooling duration. The primary objective of this trial is to characterize, for each rhythm type, the duration-response curve for cooling and to identify the duration that provides the maximum treatment effect, referred to as the *target* duration. This document describes the complete details of the adaptive trial design, including statistical model, design parameters, and the operating characteristics.

1.1 Treatment Arms

All subjects will have already been rapidly cooled (33 °C +/− 0.5°C) and will then be randomized to one of ten treatment arms for the duration of cooling. The ten treatment arms are 6, 12, 18, 24, 30, 36, 42, 48, 60, or 72-hours of cooling. For the statistical modeling we label these arms as h=1,2,...,10.

1.2 Primary Endpoint

The primary endpoint is the 90-day modified Rankin score (mRS). The mRS is an integer-valued outcome from 0 to 6, with smaller values indicating a better neurological status. A summary of each of the outcomes is:

0 = No symptoms
1 = No significant disability; able to carry out all usual activities
2 = Slight disability; can look after own affairs, but unable to carry out all previous activities
3 = Moderate disability; require some help
4 = Moderately severe disability; unable to attend to own bodily needs without assistance
5 = Severe disability; bedridden and requiring constant nursing care and attention
6 = Dead

ICECAP Study Design and Simulation Report 2

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint

21

21

1.4 Analysis Population

All analyses will be based on the intent-to-treat ITT population. The ITT patient population will include all patients randomized, where patients will be included in the treatment arm to which they were randomized, regardless of the duration of cooling applied.

It is anticipated that subjects may initially be perceived to be of one rhythm type and later found to be of the other rhythm type. In the ITT analyses, these subjects will be included with their corrected rhythm type, but their treatment arm will remain as randomized.

Any subjects that are missing or withdraw from the study and have an unknown 90-day mRS will be included in the analyses of the primary endpoint with multiple imputation according to the longitudinal model described in Section 2.2.

1.5 Randomization

The first 200 subjects enrolled overall between the two rhythm types will be randomized equally (1:1:1) between three treatment arms: 12 hours, 24 hours, and 48 hours. After these first 200 patients have been enrolled, patients will be allocated to treatment arms based on response adaptive randomization (RAR). RAR will be applied separately within each rhythm type. During RAR, interim analyses will be conducted monthly in order to update the response adaptive randomization probabilities. The response adaptive randomization probabilities determined at each interim analysis will be applied for the newly enrolling subjects until the next interim analysis and update, one month later. Section 2 describes the calculation of the response adaptive randomization probabilities.

The 6, 60, and 72-hour duration treatment arms are initially closed to enrollment. The 6-hour duration treatment arm will be opened for a rhythm type r if there are at least 300 subjects enrolled across all arms in that rhythm type and there is at least a 0.33 probability that 6 hours is the target duration for that rhythm type. The 60 and 72-hour duration treatment arms will be opened incrementally. These arms open to both rhythm types simultaneously if there are at least 10 subjects on the next shorter duration across both rhythm types and, for at least one of the rhythm types, there is at least a 0.33 probability that the target duration for that rhythm type is at or above that next shorter duration.

1.6 Stopping Rules

Interim analyses begin after 200 patients have been enrolled and will occur monthly. At each interim analysis, the trial may stop for futility if no cooling duration greater than 6 hours is found to be more effective than the 6-hour duration. Futility will be assessed separately for each rhythm type. Therefore, the trial could be declared futile for one rhythm type, and yet continue to enroll subjects

ICECAP Study Design and Simulation Report 4

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules

22

22

of the opposite rhythm type. If both rhythm types are not stopped for futility, the trial will continue to enroll to the maximum sample size of 1800 patients. Specifically, a rhythm type will stop for futility if

1. At least 300 patients have been randomized to the 6-hour duration arm
2. There is at least a 50% probability that the 6-hour duration treatment arm is the target duration

2.0 Statistical Modeling

This section describes the statistical models used in the adaptive design and the primary analysis. The modeling is Bayesian in nature. All conclusions about each treatment arm will be based on a duration-response model. We use an inverted U-shaped duration-response model for modeling the mean weighted 90-day mRS for each treatment arm. Identical *instances* of the model are used for each rhythm type. During the trial, at the interim analyses, patients with unknown 90-day mRS values are modeled using a longitudinal model based on earlier, 30-day mRS, values. This same longitudinal model is used to handle missing data for the primary analysis at the conclusion of the trial.

The inverted U-shaped duration-response model is described in Section 2.1, the longitudinal model is described in Section 2.2, the Bayesian quantities calculated at each interim analysis are described in Section 2.3, and the calculation of the response adaptive randomization probabilities is described in Section 2.4.

Let the observed weight for a patient i be W_i. The weighted 90-day mRS values, W_i, are modeled as normally distributed observations:

$$[W_i] \sim N(\theta_{h,r}, \sigma_r^2),$$

where the treatment arm is h and the rhythm type is r. The variance components are modeled separately for each rhythm type, with weak prior distributions:

$$[\sigma_r^2] \sim Inverse-Gamma(0.001, 0.001) \text{ for } r=1,2.$$

2.1 Duration Response Model for Weighted mRS

We model the mean weighted 90-day mRS across the ten treatment arms with a duration-response model. The duration response model is a critical aspect to the adaptive design and the primary analysis. The duration response model restricts the shape of the duration response curve to have 3 phases – an increasing phase, a plateau phase, and a decreasing phase. We create a parametric family for this inverted-U duration response model. For each rhythm type a separate and identical instance of the model is used, therefore we present the details for a single instance.

ICECAP Study Design and Simulation Report 5

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model

23

23

than γ_1, if γ_1 is less than 72-hours, or the longest duration if γ_1 is greater than 72-hours.

Because the modeling is done separately for each rhythm type, we add a subscript r on the parameters to notate the specific rhythm type. The prior distributions for the "change-points" are defined on the space where $\gamma_1^r < \gamma_2^r$ as follows:

$$[\gamma_1^r] \sim N(4, 10^2) \text{ for } r=1,2$$

and

$$[\gamma_2^r] \sim N(8, 10^2) \text{ for } r=1,2.$$

The following priors for each instance are utilized:

$$[\beta_0^r] \sim N(4, 4^2) \text{ for } r=1,2$$

$$[\beta_j^r] \sim E(1) \text{ for } j=1,\dots,4;\ r=1,2,$$

where $E(\lambda)$ is an exponential distribution with parameter λ.

For the final analysis the selection of the most likely optimal duration within each rhythm type is selected using this model described above. For the determination of a positive treatment effect the posterior distribution using a hierarchical model is used. The following hierarchical prior structure is used:

$$[\gamma_1^r] \sim N(\mu_1, \tau^2) \text{ for } r=1,2$$

and

$$[\gamma_2^r] \sim N(\mu_2, \tau^2) \text{ for } r=1,2.$$

With hyperparameters

$$[\mu] \sim N(4, 10^2)$$

and

$$[\mu] \sim N(8, 10^2)$$

and

$$[\tau^2] \sim G^{-1}(0.001, 0.001).$$

ICECAP Study Design and Simulation Report 7

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model

24

24

2.2 Longitudinal Modeling

At each interim analysis, and at the final analysis, there will be subjects who have an unknown 90-day mRS value. We use the 30-day mRS value as possibly predictive of the 90-day mRS, allowing subjects with this earlier measurement to be included in the analyses of the 90-day measurement. This modeling is referred to as the longitudinal model. The longitudinal model allows for learning the relationship between the 30-day and 90-day mRS values as the accruing empirical data is used to determine the strength of the association between the two values for each treatment arm and rhythm type. Analyses of the 90-day mRS values are performed with multiple imputation from the longitudinal model for patients with an unknown 90-day mRS value.

The longitudinal model maps the 7 possible 30-day mRS values to the 7 possible 90-day mRS values. We present a Markovian structure for the "transitions" from the 30-day mRS state (k) to the 90-day mRS state (j). The transition matrix, P, represents the matrix of probabilities for a subject transitioning from the 30-day mRS state (rows) to the 90-day mRS state (columns). For example, p_{32} is the probability a subject that is a 30-day mRS of 2 becomes a 3 at 90-days.

$$P = \begin{bmatrix} p_{00} & p_{01} & & \cdots & p_{05} & p_{06} \\ p_{10} & p_{11} & & & p_{15} & p_{16} \\ & & \ddots & & & \\ \vdots & \vdots & & p_{kj} & \vdots & \vdots \\ & & & \ddots & & \\ p_{60} & p_{61} & & & p_{65} & p_{66} \end{bmatrix}$$

We place a Dirichelet prior distribution on each vector of probabilities from each state at 30 days to each state at 90 days:

$$\begin{pmatrix} p_{j0} & \cdots & p_{j6} \end{pmatrix}^T \sim \text{Dirichlet}\left(\begin{pmatrix} \alpha_{j0} & \cdots & \alpha_{j6} \end{pmatrix}^T\right) \text{ for } j = 0,\ldots,6.$$

For all treatment arms and both rhythm types we assume the following prior weights for the transitions (included is the assumption that all transitions from a 30-day mRS of 6 are to a 90-day mRs of 6):

ICECAP Study Design and Simulation Report 8

1. Background (Design)
    1. Treatment Arms
    2. Primary Endpoint
    3. Primary Analysis
    4. Analysis Population
    5. Randomization
    6. Stopping Rules
2. Modeling
    1. Duration Model
    2. Longitudinal Model

25

25

2.3 Bayesian Quantities

The following Bayesian quantities are calculated for each analysis. Each of these quantities are calculated using the data from all subjects in the trial—those with complete 90-day mRS values and those with only 30-day mRS values.

2.3.1 Most Likely Target Duration

The target duration is defined as the shortest duration that achieves the maximum treatment effect. As described in Section 2.1, the target duration is defined based on the duration-response model as the shortest duration greater than γ_2, if γ_2 is less than 72-hours, or the longest duration if γ_2 is greater than 72-hours.

At each analysis the posterior probability that each treatment arm is the target duration is calculated for each rhythm type. The treatment arm with the highest posterior probability of being the target duration for a particular rhythm type is the most likely target duration for that rhythm type. We label the probability a treatment arm h is the target duration for rhythm type r as

$$\Pr_{h,r}^{h^*}$$

2.3.2 Posterior Variance

For each treatment arm h and rhythm type r we calculate $\theta_{h,r}$, the posterior mean weighted 90-day mRS and $V(\theta_{h,r})$, the posterior variance of $\theta_{h,r}$.

2.3.3 Posterior probability superior to smaller duration

Within each rhythm type r, and each duration h, we calculate the posterior probability that it has a superior mean weighted 90-day mRS as compared to each treatment arm with a shorter cooling duration. Because of the duration-response model, it is sufficient to calculate only the posterior probability that a treatment arm is superior to the 6-hour duration arm. We refer to the posterior probability that each treatment arm is superior to a shorter duration treatment arm as

$$\Pr\left(\theta_{h,r} > \theta_{1,r} \mid \text{data}\right).$$

This quantity is calculated using both independent priors and using the hierarchical priors.

ICECAP Study Design and Simulation Report 10

1. Background (Design)
    1. Treatment Arms
    2. Primary Endpoint
    3. Primary Analysis
    4. Analysis Population
    5. Randomization
    6. Stopping Rules
2. Modeling
    1. Duration Model
    2. Longitudinal Model
    3. Bayesian Quantities

26

26

2.4 Adaptive Randomization

During the response adaptive randomization stage, separate allocation schemes are created for each rhythm type. The specification of the vector of probabilities for randomization is defined in this section. Randomization probabilities to each treatment arm are weighted according to the posterior probability that each treatment arm is the target duration. The goal of the adaptive randomization is to allocate subjects to the arms most likely to be the target duration, but also to learn effectively about the duration-response curve.

Let the number of subjects enrolled on duration arm h for rhythm type r be $n_{h,r}$. The posterior variance of the mean weighted 90-day mRS for each treatment arm and rhythm type is $V(\theta_{h,r})$. The probability a treatment arm is the most likely target duration for a rhythm type is $\Pr_{h,r}^{h^*}$. A *variance component*, $V_{h,r}$, is constructed for each duration arm within each rhythm type. The variance component is

$$V_{h,r} = \Pr_{h,r}^{h^*}\sqrt{\frac{V(\theta_{h,r})}{n_{h,r}+1}} \text{ for } h=1,2,\ldots,10 \text{ and } r = 1,2$$

The randomization probabilities for each treatment arm within each rhythm type is proportional to $V_{h,r}$.

$$q_{h,r} = \frac{V_{h,r}\delta_{h,r}}{\sum_{h=1}^{10} V_{h,r}\delta_{h,r}} \text{ for } h=1,2,\ldots,10.$$

The treatment arms for 6-hours (h=1), 60 hours (h=9), and 72 hours (h=10) will initially be closed, but may be later opened according to the rules described in Section 1.5. If a treatment arm is open we set $\delta_h = 1$, and if a treatment arm is closed we set $\delta_h = 0$. Therefore, the adaptive allocation scheme favors the arms most likely to be the target duration, but also favors arms with a greater variability (uncertainty) around the primary endpoint or a smaller sample size.

ICECAP Study Design and Simulation Report 11

1. Background (Design)
    1. Treatment Arms
    2. Primary Endpoint
    3. Primary Analysis
    4. Analysis Population
    5. Randomization
    6. Stopping Rules
2. Modeling
    1. Duration Model
    2. Longitudinal Model
    3. Bayesian Quantities
    4. Adaptive Randomization

27

27

3.0 Section intentionally left blank

Reserved for future use.

ICECAP Study Design and Simulation Report 12

1. Background (Design)
    1. Treatment Arms
    2. Primary Endpoint
    3. Primary Analysis
    4. Analysis Population
    5. Randomization
    6. Stopping Rules
2. Modeling
    1. Duration Model
    2. Longitudinal Model
    3. Bayesian Quantities
    4. Adaptive Randomization
3. Example Trials

28

28

4.0 Operating Characteristics

The operating characteristics of the design are calculated using simulation. We consider a variety of scenarios for the true duration-response curve. There are many possible performance measures for this design including the probability of selecting each duration, the probability cooling is considered effective, the mean number of patients allocated to each duration, and the probability of opening the 6, 60, or 72-hour durations. In order to calculate such properties of the design, we create a methodology for simulating virtual patients. This is described in Section 4.6.

Section 4.1 addresses the type I error of the design. We define Type I error as the probability, across all doses, that cooling is considered effective when in truth, the duration response curve is flat.

Sections 4.2, 4.3, and 4.4 address the probability of success, or power of the design when then there is a positive duration-response relationship. These sections show the properties of the design for different assumed effect sizes and shapes of the duration-response curves. There are two objectives of this trial and we define the probability of success for each. The first objective is to select a target duration. For each scenario, we define clinically acceptable durations. First there is the optimal duration, which is the shortest duration that achieves the maximum treatment effect. We additionally identify durations that are considered clinically acceptable. These durations must be within 12 hours of the optimal duration and must achieve at least 70% of the maximum treatment effect. We define power for this objective as the probability any one of these clinically acceptable durations is selected as the target duration. The second objective is to determine if there is a positive duration-response curve (i.e. that cooling is effective). We define power for this objective in scenarios where there is positive effect of cooling, as the probability the duration-response curve is found to be positive, regardless of the selected target duration.

4.1 Final Analysis and Type I Error

This section discusses the final primary analysis and the threshold for success. The final primary analysis within each rhythm type is to test whether the duration-response is positive for a range of durations. The parametric structure of the duration-response model is one that forces the effect to be one of the following:

1. Decreasing effect over all durations
2. A plateau effect (same response) for all durations
3. A plateau effect for a range of durations then a negative effect for longer durations.
4. Increasing effect over all durations
5. Increasing effect for a range of durations then becoming a plateau
6. Increasing effect for a range of durations, then a plateau effect for a range of durations, then becoming a negative effect.

ICECAP Study Design and Simulation Report 13

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model
   2. Longitudinal Model
   3. Bayesian Quantities
   4. Adaptive Randomization
3. Example Trials
4. Operating Characteristics
   1. Null Scenarios

29

29

extended cooling). The simulation frequency of meeting the primary efficacy outcome, of a posterior probability of at least 97.5% for a positive duration-response model for each rhythm type, regardless of the target duration selected, are reported in the last two columns. The two rhythm types are the same, and each with equal effects across all durations, so the average probability of a positive duration response is reported. In Scenarios 7 through 13 the results per rhythm type are important and are listed (shock/non-shock).

Null Case	Accrual	Long	Prob Shock	Probability of Each mRS Value							Probability Positive Duration Response
				0	1	2	3	4	5	6	
1	Slow	1	0.50	.20	.15	.15	.05	.05	.10	.30	0.0354
2	Slow	2	0.50	.20	.15	.15	.05	.05	.10	.30	0.0392
3	Med	1	0.50	.20	.15	.15	.05	.05	.10	.30	0.0454
4	Fast	1	0.50	.20	.15	.15	.05	.05	.10	.30	0.0374
5	Slow	1	0.50	.10	.10	.15	.10	.05	.10	.40	0.0318
6	Slow	1	0.50	.25	.20	.20	.10	.05	.10	.10	0.0444
7	Slow	1	0.55	.20	.15	.15	.05	.05	.10	.30	0.0352/0.0308
8	Slow	1	0.60	.20	.15	.15	.05	.05	.10	.30	0.0464/0.0376
9	Slow	1	0.65	.20	.15	.15	.05	.05	.10	.30	0.0436/0.0408
10	Slow	1	0.70	.20	.15	.15	.05	.05	.10	.30	0.0384/0.0256
11	Slow	1	0.75	.20	.15	.15	.05	.05	.10	.30	0.0392/0.0264
12	Slow	1	0.80	.20	.15	.15	.05	.05	.10	.30	0.0452/0.0248
13	Slow	1	0.85	.20	.15	.15	.05	.05	.10	.30	0.0424/0.0304

Each of the simulations is based upon 2000 simulated trials. The simulated type I error is typically smaller than 0.05 (a one-sided type I error). The different underlying distribution of mRS can have subtle effects on the behavior of the dose-response model. Cases 1 and 5 use the same assumptions, except for the distribution of mRS and have simulated type I errors of 0.0318 or 0.0454. The accrual rate and longitudinal pattern have little effect on the simulated type I error. Scenario 6 has a higher likelihood of good outcomes and has a higher type I error rate, 0.0444. The proportion of patients with shockable rhythms versus non-shockable rhythms does not have a large effect on the simulated type I error, unless the rate of subjects within one rhythm type is very small (< 20%), then the type I error is deflated for the smaller group.

In addition to these simulated null hypotheses, we create a post-simulation plan to determine the simulated type I error for the trial based on the parameters observed in the trial. At the conclusion of the trial, the design will be re-simulated using the actual subject data. Within each rhythm type all subjects will be placed in a list with their arm assignment removed. Subjects will be randomly sampled or "bootstrapped" from this list and assigned a duration. This creates a setting where, theoretically, all durations have the same treatment effect as all subjects are drawn from the same pool. The post-simulation type I error will be reported as the type I error for the trial. The 97.5% threshold will be used for primary success, regardless.

ICECAP Study Design and Simulation Report 15

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model
   2. Longitudinal Model
   3. Bayesian Quantities
   4. Adaptive Randomization
3. Example Trials
4. Operating Characteristics
   1. Null Scenarios

30

30

4.2.1 Small Treatment Effect at 18 Hours (Scenario 21.21)

In this scenario, increasing cooling duration provides a small treatment effect and the duration response curve plateaus at 18 hours. The 6-hour arm is enrolled to with 53% probability, the 60-hour arm is enrolled to with 14% probability, and the 72-hour arm is enrolled to with only 1% probability. The trial continues to completion with 98% probability (i.e. will stop early for futility only 2% of the time).

The 18 hours duration, and the two surrounding arms, 12- and 24-hours have the largest sample sizes. The adaptive randomization tends to place patients on these doses in order to identify where the treatment effect plateau begins.

The durations considered clinically acceptable are the 18-, 24-, and 30-hour durations. They are each within 12 hours of the 18-hour duration and offer the same treatment effect. One of these arms is selected with 64% probability in each rhythm type. In total, across all treatment arms, the positive duration-response curve is identified with 55% probability.

Table 4.2.1A: Summary Scenario Operating Characteristics

	Shockable	Non-Shockable
Pr(Select Acceptable Target Duration)	0.643	0.637
Pr(Positive Duration Response)	0.556	0.547
Pr(Futility)	0.022	0.019
Pr(Open 6 Hours)	0.540	0.523
Pr(Open 60 Hours)	0.138	0.147
Pr(Open 72 Hours)	0.014	0.012

Table 4.2.2B: Detailed Scenario Operating Characteristics

	Shockable Rhythm				Non-Shockable Rhythm			
	Mean Weighted mRS	N	Pr (Target)	Pr (Target & Eff)	Mean Weighted mRS	N	Pr (Target)	Pr (Target & Eff)
6 Hour	4.85	29.6	0.076	0.000	4.85	28.0	0.088	0.000
12 Hour	5.39	156.6	0.168	0.056	5.39	155.6	0.160	0.062
18 Hour	*5.93*	*150.1*	*0.333*	*0.210*	*5.93*	*152.1*	*0.322*	*0.190*
24 Hour	*5.93*	*168.7*	*0.228*	*0.155*	*5.93*	*169.5*	*0.240*	*0.169*
30 Hour	*5.93*	*110.3*	*0.082*	*0.058*	*5.93*	*110.4*	*0.075*	*0.053*
36 Hour	5.93	94.2	0.046	0.032	5.93	94.4	0.047	0.032
42 Hour	5.93	82.0	0.038	0.025	5.93	82.1	0.033	0.021
48 Hour	5.93	100.2	0.024	0.016	5.93	100.5	0.028	0.017
60 Hour	5.93	7.3	0.004	0.004	5.93	7.6	0.006	0.002
72 Hour	5.93	0.5	0.000	0.000	5.93	0.4	0.000	0.000

ICECAP Study Design and Simulation Report 18

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model
   2. Longitudinal Model
   3. Bayesian Quantities
   4. Adaptive Randomization
3. Example Trials
4. Operating Characteristics
   1. Null Scenarios
   2. Alternative Scenarios

31

31

curve. However, this is a challenging scenario for power regardless of the maximum sample size.

4.8 Simulating Virtual Subjects

In order to simulate the design we need to create a methodology for simulating individual subjects, including their longitudinal results. These assumptions do not affect the analysis described above and are only made to simulate virtual subjects.

The approach to simulate a subject is to simulate their 90-day mRS outcome, then conditional on this value simulate their 30-day mRS outcome. This reverse order approach allows the probabilities of the 90-day mRS outcomes to be set explicitly, and the truth of the durations to be controlled easier.

In order to simulate subjects a probability vector for each of the 90-day outcomes is assumed. Let the probability of the 7 outcomes be $p_0, \ldots, p_6$. The true assumed mean weight for the treatment arm is

$$\sum_{m=0}^{6} p_m W(m).$$

The design uses the 30-day mRS outcome in the longitudinal model. The following probabilities are used to simulate the 30-day mRS outcome (columns) conditional on the 90-day mRS outcome (row).

$$\begin{array}{c c} & \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} .3 & .3 & .2 & .15 & .04 & .01 & 0 \\ .1 & .5 & .2 & .15 & .04 & .01 & 0 \\ .05 & .05 & .5 & .25 & .14 & .01 & 0 \\ .01 & .05 & .05 & .5 & .35 & .04 & 0 \\ .01 & .01 & .01 & .07 & .7 & .2 & 0 \\ .01 & .01 & .01 & .01 & .26 & .7 & 0 \\ .01 & .01 & .01 & .01 & .06 & .4 & .5 \end{bmatrix} \end{array}.$$

Therefore, by specifying the 7 mRS outcomes creates a true mean response for each duration. For example, for the null hypothesis, the probabilities for the 7 outcomes are assumed to be:

mRS	0	1	2	3	4	5	6
Probability	0.20	0.15	0.5	0.05	0.05	0.10	0.30

The mean weighted truth for this scenario is 4.85. The truth for each of the 10 durations can be varied to create many different relative effects as well as different

ICECAP Study Design and Simulation Report 50

1. Background (Design)
   1. Treatment Arms
   2. Primary Endpoint
   3. Primary Analysis
   4. Analysis Population
   5. Randomization
   6. Stopping Rules
2. Modeling
   1. Duration Model
   2. Longitudinal Model
   3. Bayesian Quantities
   4. Adaptive Randomization
3. Example Trials
4. Operating Characteristics
   1. Null Scenarios
   2. Alternative Scenarios
5. Simulating Virtual Subjects

32

32

Questions?

Simulations

- Would be very cool – emergency simulation for clinical trials!
- Surely would have FACTS on board

1

Leadership Talk

Examples

- We started simulating a trial design – and compared it to a fixed trial design.
- The straw fixed trial design had 80% power for the nice effect size of "delta"… that was a great trial – it had 80% power
- We simulated the same trial designs and showed them single simulated trials that lost when the truth was delta – and described that 20% of the trials we simulated failed when the truth of the drug was delta
  - They were shocked and disappointed…

15

2

Leadership Talk

Examples

- There are many that don't understand power is risk – they assume power is just a restriction statisticians place on trials…
  - The best statisticians can get smaller N, yet still 80% powered
- They are thrilled if they get a smaller n… not understanding power is really a risk thing…
- Now in part this is our fault!

16

3

Leadership Talk

What's the Issue?

- The perception is that 'we' isolate ourselves within the work stream to provide routine contributions to the project
  - Power calculations
  - Protocol verbiage
  - SAP
  - Programming
- And we speak a different language: 'statistics'

17

4

Leadership Talk

Language of Statistics

- We love it, we can talk for hours about the difference between [X|θ] and [θ|X]…
  - Almost nobody cares
  - It's our science
- All too often these discussions happen in our language, we make them learn it (and they don't know it)
  - This is not leadership!
- *We have to speak their language*; the disease, the science, the drug, the team, the company

27

5

Leadership Talk

Consultant/Stats

- Very common that the question you are asked is not their real question!
  - "What sample size do I need for this phase II trial?"
  - "How many doses should we have in this trial to understand the dose-response?"
  - "What is the penalty for taking an early look?"
- All these questions have huge "it depends" on them and part of it is they are trying to speak our language
  - Answer to all is "well, lets back up a bit… why are we doing this trial?"
  - "What do we know?"
  - "What are we trying to learn?"
  - "What happens after this trial?"

28

6

Leadership Talk

Consultant/Stats

- We need to be the ones bridging the science gap – not your teammates
- We chuckle at their not understanding us
  - Powering at 80% for delta doesn't mean you need to see delta to win…
- This is our fault for making them bridge that gap – we should be putting everything in their language, their units, … we do the translating!

29

7

Leadership Talk

Tools

- Simulations
- Modeling (dose-response, longitudinal, hierarchical,…)
- Borrowing data?
- Adaptations: Sample size, futility, enrichment, baskets, platforms, add arms, subtract arms, combine trials (seamless), combine goals
- Graphics!

33

8

Leadership

- What is my point?

Simulation allows us to speak the language of the clinician, the trialist, the sponsor, etc, not *just* the statistical language

Without simulation our tools have been limited, and hence our role has been limited. With simulation our answers are better and our role is expanded

9

Examples

- Dose-Finding Trials; Select the right dose?
- Does RAR improve the chance we pick the right dose?
- What is the risk that our Bayesian borrowing for the control arm gets the wrong answer?
- What go/no-go decision optimizes our drug development?
- What are the average number of subjects we treat above the MTD using this CRM?
- In a basket trial does borrowing help or hurt our estimation?
- Does this design affect the speed to market of an effective drug?

10

Warning #1

Clinical Trial Simulation means different things to different people and everyone will be skeptical of it.

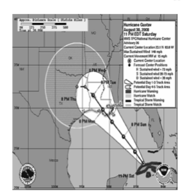

Tweet

Stephen John Senn
@stephensenn

Simulation is a means of verifying the wrong conclusions your incorrect model led you to by carrying out integration with the help of random numbers.

6:23 AM · Sep 18, 2020 from Stafford, England · Twitter for iPhone

11



We are inundated with “simulations” being used as *predictions*

12

12

Role of Simulations

- This is common for PK/PD scientists – *predict* what will happen in humans
- This is not how simulations are used in creating *in silico designs*
- The "simulation evaluation" is nothing more than *numerical integration*
- Calculating operating characteristics *exactly*

13

13

Warning #2

Less useful as a final task

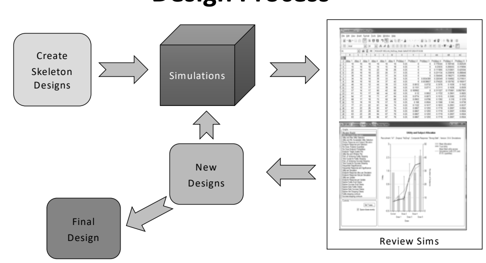
14

Warning #3

This is not meta-analysis: preplanning what you are going to simulate is limiting and defeats the purpose

15

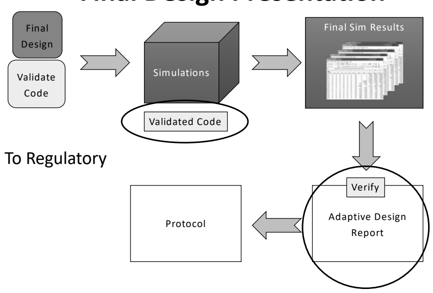
Design Process



16

16

Final Design Presentation



17

17

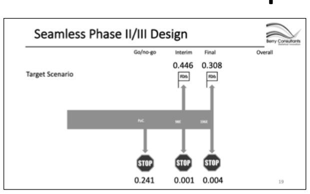
FACTS

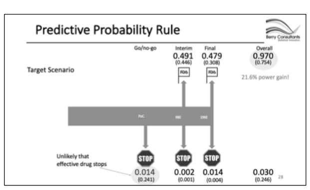
- Interesting time point in "history" of FACTS…
- ~2006 we were all in a room deciding…

Is the best software tool a collection of named designs (aircraft carrier) or a collection of choices to be crossed and explored?

18

Example This Month





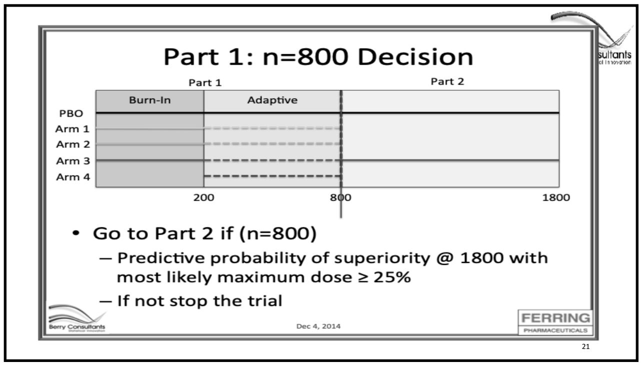
Progressed to 5-look design, effect of 2:1 on timing, Change randomization during?,

19

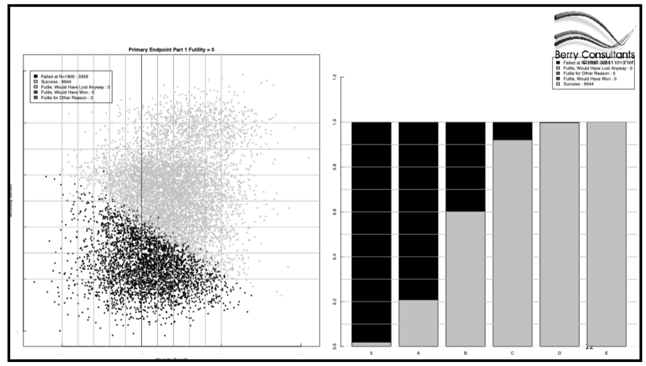
Example This Month

- A phase III trail, Two active arms vs. PBO; 1:1:1
- Slow enrollment ~50 per arm
- 4-week endpoint
- Should we explore arm-dropping? Futility Stopping? Flexible sample size?
  - "No, the trial is 80% powered so we cant make good decisions before that time point."
  - These types of decisions are being made non quantitatively by non-quantitative people…
  - Simulations can provide invaluable uses, very quickly…

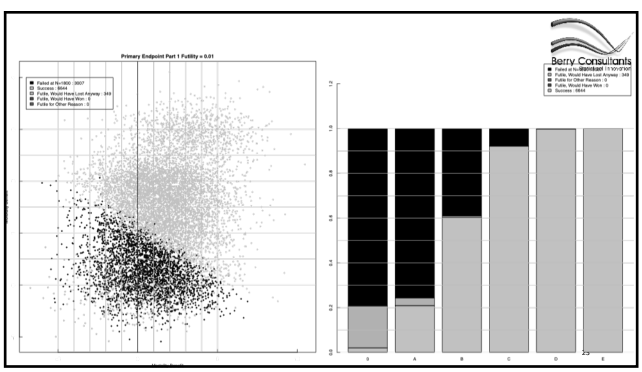
20


Part 1: n=800 Decision
Part 1
Part 2
Burn-In
Adaptive
PBO
Arm 1
Arm 2
Arm 3
Arm 4
200
800
1800
• Go to Part 2 if (n=800)
– Predictive probability of superiority @ 1800 with most likely maximum dose ≥ 25%
– If not stop the trial
Berry Consultants
Dec 4, 2014
FERRING PHARMACEUTICALS
21

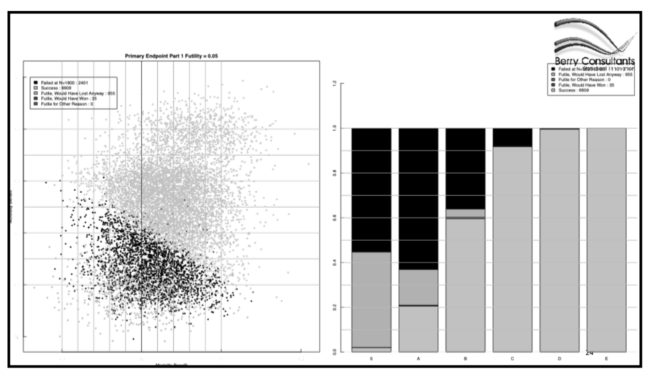
21


Primary Endpoint Part 1 Futility = 0
Berry Consultants
22

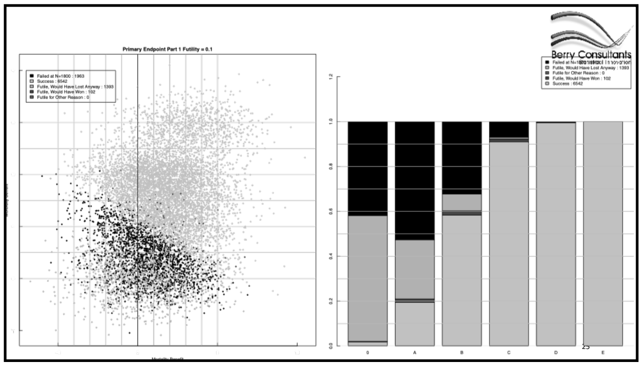
22


Primary Endpoint Part 1 Futility = 0.01
Berry Consultants

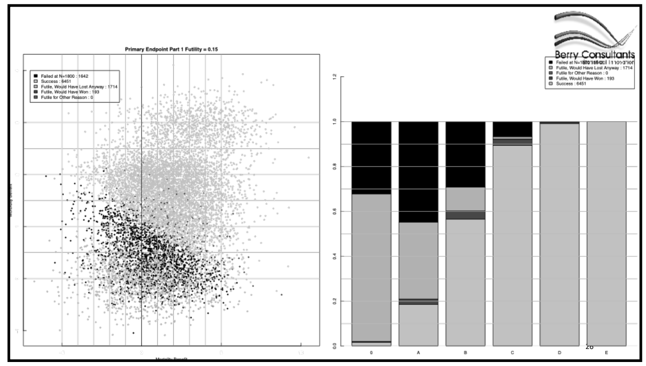
23


Primary Endpoint Part 1 Futility = 0.05
Berry Consultants
24

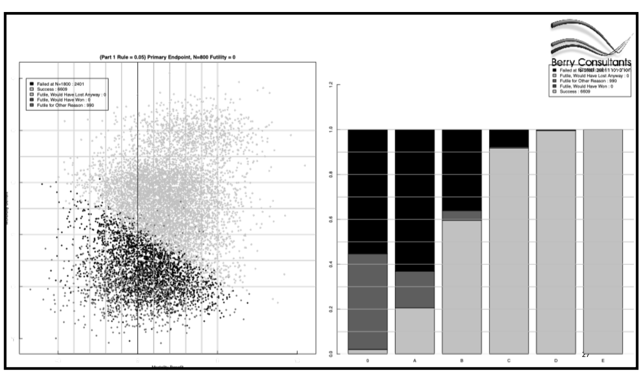
24


Primary Endpoint Part 1 Futility = 0.1
Failed at N=1800 : 1963
Success : 6542
Futile, Would Have Lost Anyway : 1393
Futile, Would Have Won : 102
Futile for Other Reason : 0
Berry Consultants
S
A
B
C
D
E

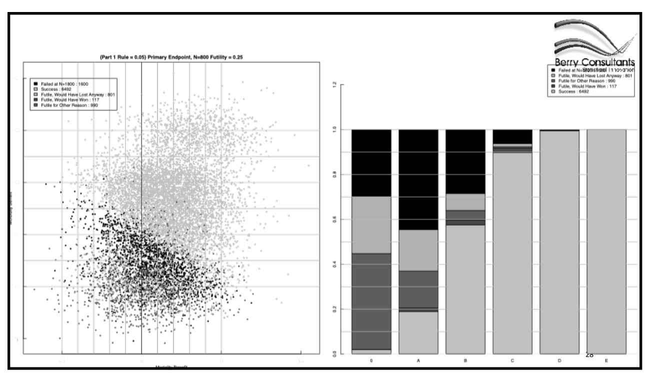
25


Primary Endpoint Part 1 Futility = 0.15
Failed at N=1800 : 1642
Success : 6451
Futile, Would Have Lost Anyway : 1714
Futile, Would Have Won : 193
Futile for Other Reason : 0
Berry Consultants
S
A
B
C
D
E

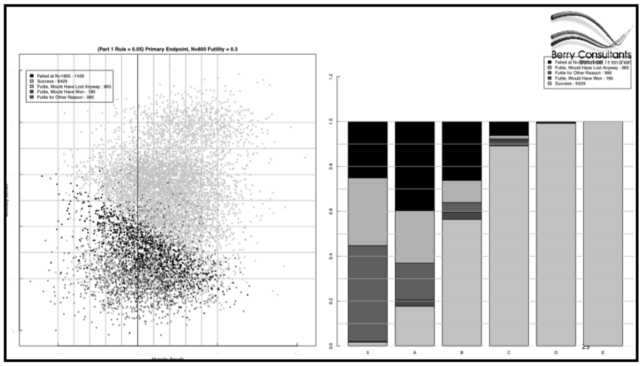
26


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0
Failed at N=1800 : 2401
Success : 6609
Futile, Would Have Lost Anyway : 0
Futile, Would Have Won : 0
Futile for Other Reason : 990
Berry Consultants
S
A
B
C
D
E

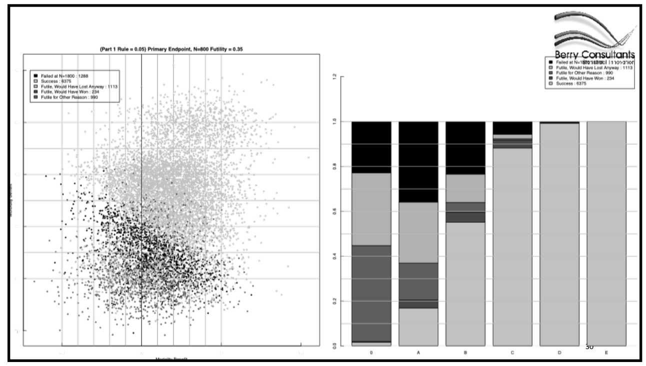
27


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0.25
Failed at N=1800 : 1600
Success : 6492
Futile, Would Have Lost Anyway : 801
Futile, Would Have Won : 117
Futile for Other Reason : 990
Berry Consultants
S
A
B
C
D
E

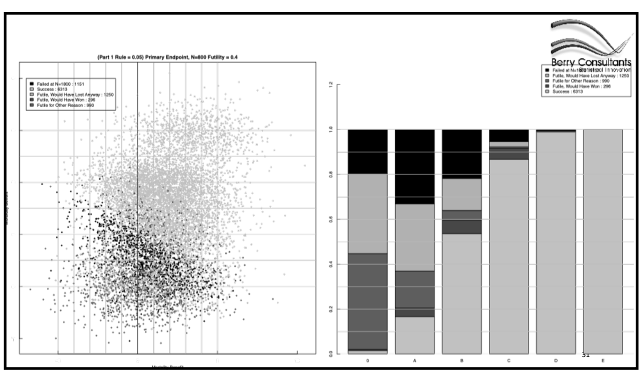
28


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0.3
Failed at N=1800 : 1436
Success : 6429
Futile, Would Have Lost Anyway : 965
Futile, Would Have Won : 180
Futile for Other Reason : 990
Berry Consultants

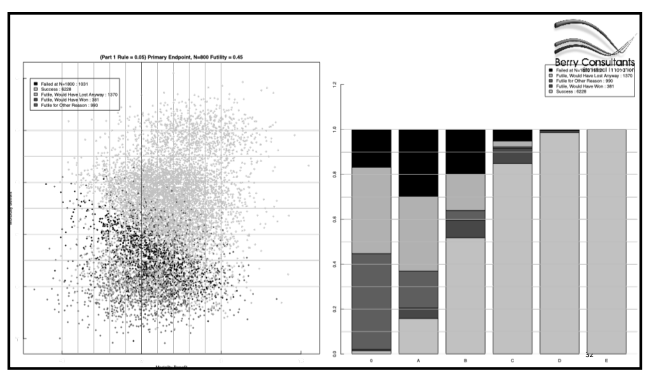
29


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0.35
Failed at N=1800 : 1288
Success : 6375
Futile, Would Have Lost Anyway : 1113
Futile, Would Have Won : 234
Futile for Other Reason : 990
Berry Consultants

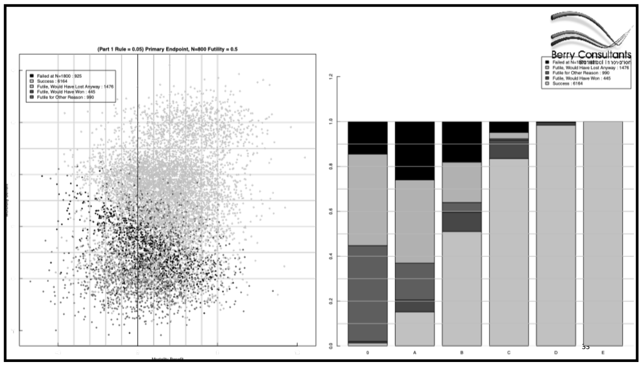
30


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0.4
Failed at N=1800 : 1151
Success : 6313
Futile, Would Have Lost Anyway : 1250
Futile, Would Have Won : 296
Futile for Other Reason : 990
Berry Consultants

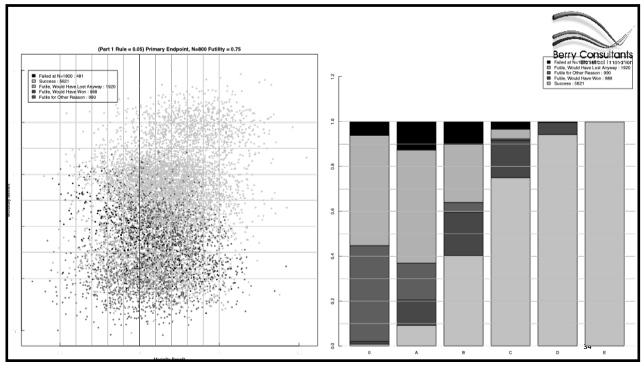
31


(Part 1 Rule = 0.05) Primary Endpoint, N=800 Futility = 0.45
Failed at N=1800 : 1031
Success : 6228
Futile, Would Have Lost Anyway : 1370
Futile, Would Have Won : 381
Futile for Other Reason : 990
Berry Consultants

32



33



34

Warnings of Simulations in Consulting

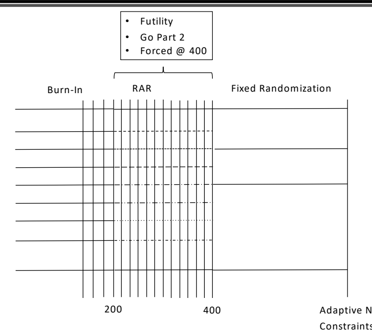
- Simulations are **distrusted** until the team sees how you use them – and then they're **loved**
- The presentation of the results are very important …
- Example trials are critically important
- Algorithms, predictive probabilities, etc are black boxes…
  - Show real data -> Conclusions;

35

Example: Diabetes II/III seamless

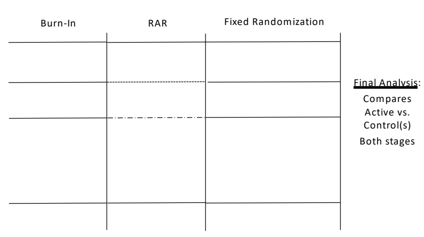
- 7 dose + PBO + Active Control
  - Interims every 2 weeks
  - RAR based on 4 endpoints
    - HbA1c, Weight Loss, DBP, HR with utility function
  - 200-400 make decision:
    - Go to Phase III (pick 1 or 2 doses); open more phase III
    - Stop futility
  - Phase III part powered by phase II
  - Entirely prospectively planned
    - Algorithms, Rules, Decisions, Analyses

1



2

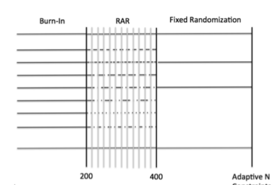
2



3

3

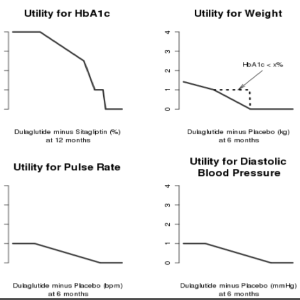
- Bayesian repeated measures & dose-response models for four endpoints
- Single utility function connecting 4 endpoints on one scale
- Predictive probability of statistical success



4

4

Utility of Drug?



Berry Consultants

5

5

Development

- Built "exact" trial in software (in silico)
  - Accrual Rate
  - Missing Data (function of outcome)
  - Same primary analysis, models, utility functions, dose selection, cut-offs, data delay,…
  - Wide range of "truth scenarios"
- Maximized design through simulations
  - Over 300 scenarios in the null
  - Interesting was that the LOCF ANCOVA had inflated type I errors – as large as 6-8% (aim 2.5%)

Berry Consultants

6

6

Diabetes II/III seamless

- Trial ran (for 3,467,321st time!)
- Shifted at 200 -- very successful!
  - Ran *exactly* as planned, spawned other phase III

Journal of Diabetes Science and Technology
Volume 6, Issue 6, November 2012
© Diabetes Technology Society

ORIGINAL ARTICLE

Application of Adaptive Design Methodology in Development of a Long acting Glucagon-like Peptide-1 Analog (Dulaglutide): Statistical Design and Simulations

Zachary Skrivanek, Ph.D.,[1] Scott Berry, Ph.D,[2] Don Berry, Ph.D.,[2,3] Jenny Chien, Ph.D.,[1] Mary Jane Geiger, M.D., Ph.D.,[1] James H. Anderson, Jr., M.D.,[4] and Brenda Gaydos, Ph.D.[5]

Berry Consultants

7

FierceBiotech
THE BIOTECH INDUSTRY'S DAILY MONITOR

NEWS TOPICS ANALYSIS FEATUR

UPDATED: FDA hands Eli Lilly a big win, OKs dulaglutide for diabetes

September 18, 2014 | By John Carroll

SHARE
Email
64
Tweet

An embattled Eli Lilly ($LLY) won a major battle today, gaining the FDA's approval to market dulaglutide for Type 2 diabetes. It will be sold as Trulicity.

With Novo Nordisk ($NVO) already digging in to defend its position around Victoza, the once-weekly treatment has been widely billed as a likely blockbuster. The Phase III program has long represented Eli Lilly's best shot at

Lilly Diabetes President

Peak sales projections for dulaglutide are all over the map. Cowen has pegged the potential at $700 million, with Bernstein's Tim Anderson now projecting $1.3 billion in 2020. That's not enough to make up for the patent losses, but it would go a long way to providing some credibility for an R&D group that is drawing an increasing amount of critical scrutiny.

Berry Consultants

8


FierceBiotech
THE BIOTECH INDUSTRY'S DAILY MONITOR
NEWS TOPICS ANALYSIS FEATUR
UPDATED: FDA hands Eli Lilly a big win, OKs dulaglutide for diabetes
September 18, 2014 | By John Carroll
SHARE
Email
64
Tweet
An embattled Eli Lilly ($LLY) won a major battle today,
With Novo Nordisk ($NVO) already digging in to defend its position around Victoza, the once-weekly treatment has been widely billed as a likely blockbuster. The Phase III program has long represented Eli Lilly's best shot at
Lilly Diabetes President
“Projecting $1.3 billion in 2020”
Peak sales projections for dulaglutide are all over the map. Cowen has pegged the potential at $700 million, with Bernstein's Tim Anderson now projecting $1.3 billion in 2020. That's not enough to make up for the patent losses, but it would go a long way to providing some credibility for an R&D group that is drawing an increasing amount of critical scrutiny.
Berry Consultants
Statistical Innovation

9

ICECAP Example

- Part of the ADAPT-IT (U01-NS073476) grant
- Funded by NIH & FDA
  - Get interaction with FDA on designs
- Bring adaptive exploration to 5 trials (NETT Trials)
- Study the barriers to adoption
  - Mixed methods assessment of the process and barriers
  - We are being studied

1

1

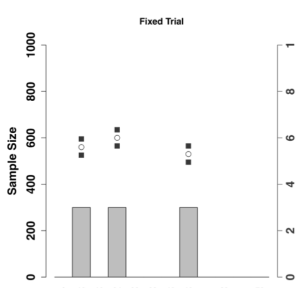
ICECAP

- ICECAP – Hypothermia after post cardiac arrest coma
  - Background
    - Two small surface cooling trials demonstrated efficacy (different durations and endovascular cooling more frequently used)
    - Medically accepted that this works
    - No FDA approval
  - Goals
    - To identify optimum cooling duration
    - Gain additional insight into efficacy (functional form of duration response model)
    - What types of strokes vs. duration
  - Fixed Design:
    - 300? On 12, 24, 48 hours cooling

2

2

Example Outcome of Fixed



- Idealized Outcome?
- Answer All your questions?
- Do anything differently?

3

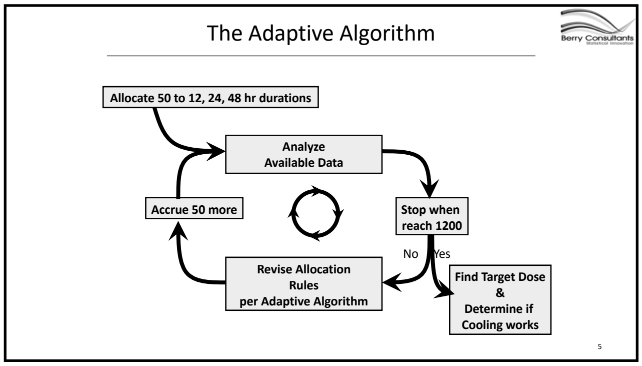
3

Initial skeleton

- Start with 12, 24, 48-hour durations (say 50/arm)
- Then analyze data and randomize to the best duration
  - Allow randomization to a much wider grid:
  - 6, 12, 18, 24, 30, 36, 42, 48, 60, 72, 84, 96
- Continue updating, say every 50 patients
- Continue to end of trial
  - Early stopping?
  - Endpoint 0,1,2 on mRs

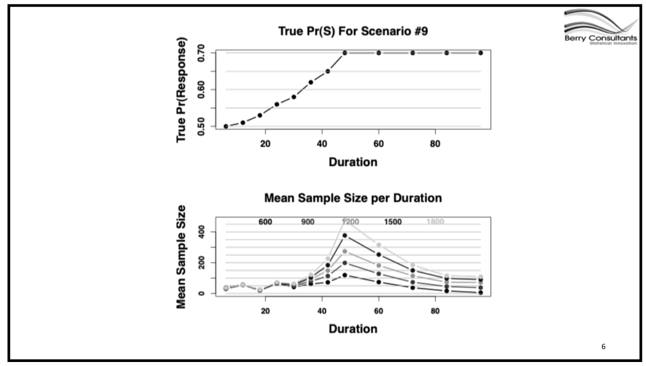
4

4



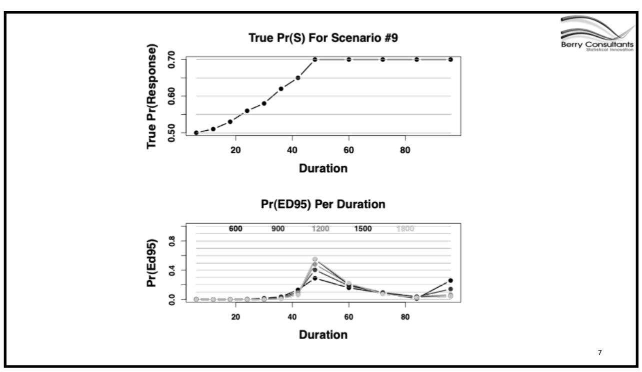
The Adaptive Algorithm
Berry Consultants
Allocate 50 to 12, 24, 48 hr durations
Analyze Available Data
Accrue 50 more
Stop when reach 1200
No
Yes
Revise Allocation Rules per Adaptive Algorithm
Find Target Dose & Determine if Cooling works
5

5



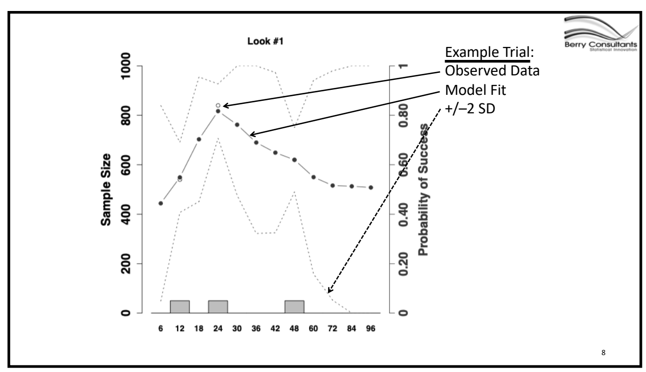
True Pr(S) For Scenario #9
Berry Consultants
True Pr(Response)
Duration
Mean Sample Size per Duration
600
900
1200
1500
1800
Mean Sample Size
Duration
6

6



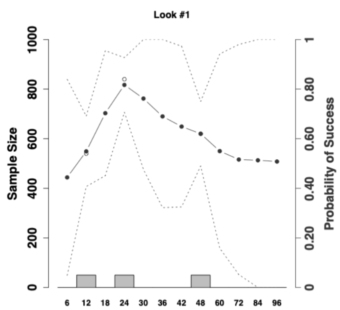
True Pr(S) For Scenario #9
Berry Consultants
True Pr(Response)
Duration
Pr(ED95) Per Duration
600
900
1200
1500
1800
Pr(Ed95)
Duration
7

7



Look #1
Berry Consultants
Example Trial:
Observed Data
Model Fit
+/–2 SD
Sample Size
Probability of Success
8

8

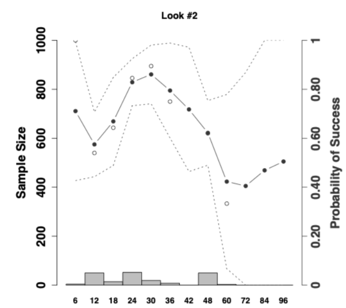


Next Blocked Randomization:

Arm	R:Block
6	2
12	0
18	7
24	1
30	9
36	3
42	0
48	0
60	1
72	--
84	--
96	--

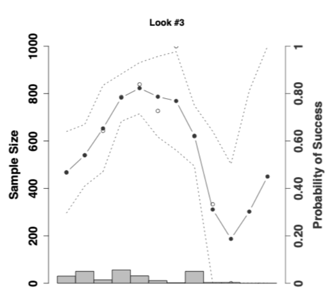
9

9



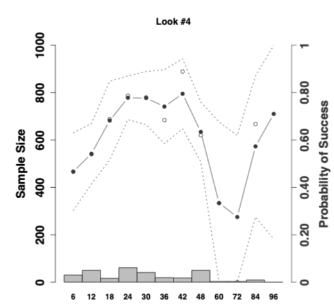
10

10



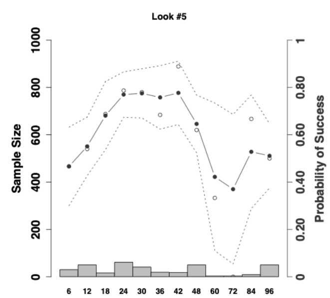
11

11



12

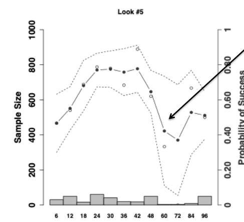
12



13

13

Issues?

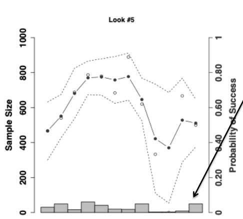


Unrealistic DR-curve

14

14

Issues?



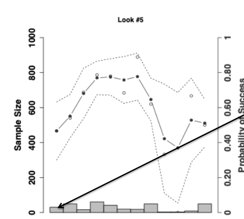
Unrealistic DR-curve

Escalates too quickly

15

15

Issues?



Unrealistic DR-curve

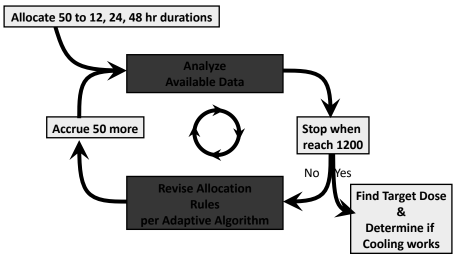
Escalates too quickly

De-escalates too easily

16

16

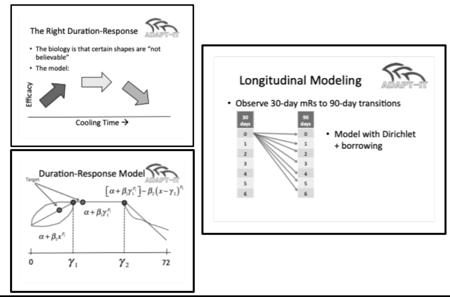
The Adaptive Design



17

17

Statistical Modeling



18

18

Weighted Analysis

- Assign a weight to each outcome

$$w(0),w(1),(2),\ldots,w(6)$$

- Evaluate the "Average Weight" as the quantity of interest for a treatment

$$\theta_d=\sum_{j=0}^{6}p_d(j)*w(j)$$

19

19

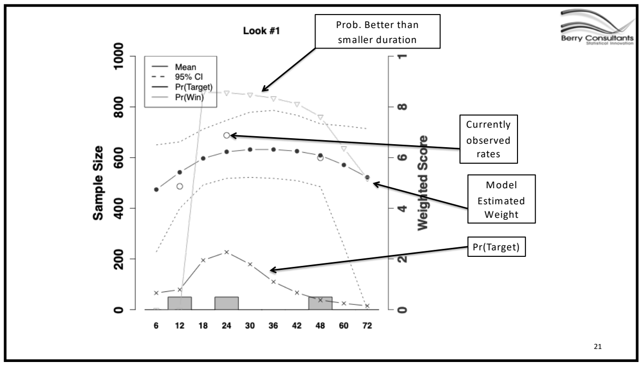
Weight Selection

Approach	0	1	2	3	4	5	6
Dichotomous	1	1	1	0	0	0	0
Equal	6	5	4	3	2	1	0
ICECAP	10	9	8	6	0	0	0

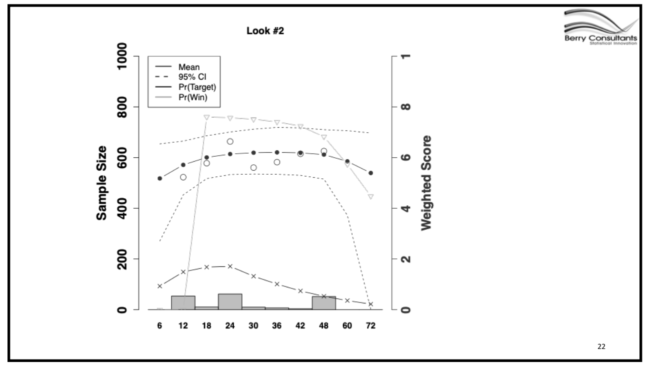
1 1 2 6

20

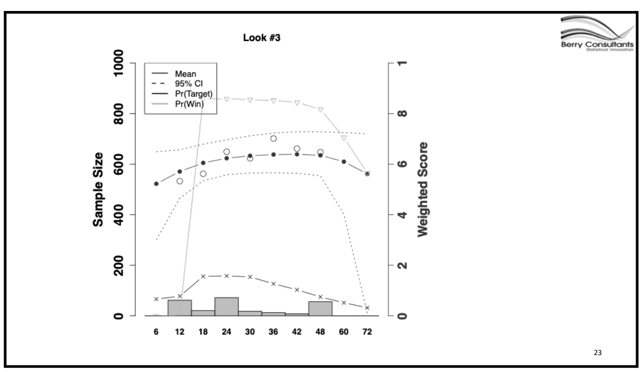
20


Look #1
Prob. Better than smaller duration
Currently observed rates
Model Estimated Weight
Pr(Target)
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
Berry Consultants
21

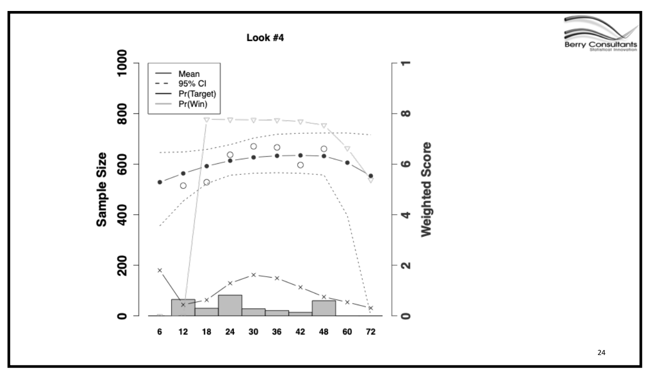
21


Look #2
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
Berry Consultants
22

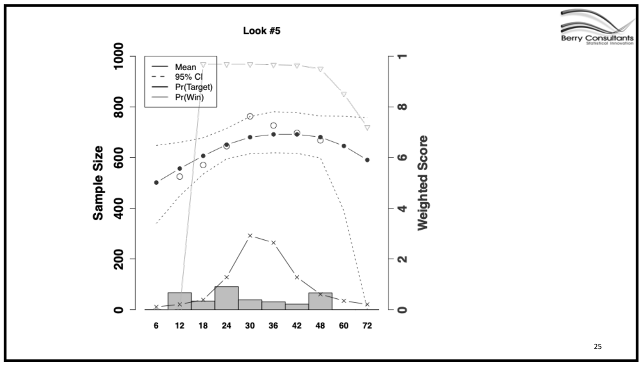
22


Look #3
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
Berry Consultants
23

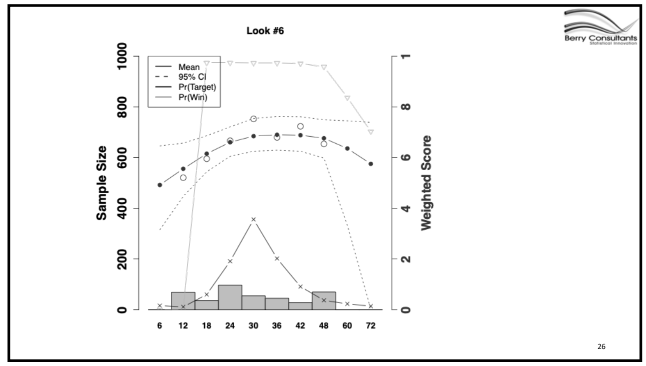
23


Look #4
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
Berry Consultants
24

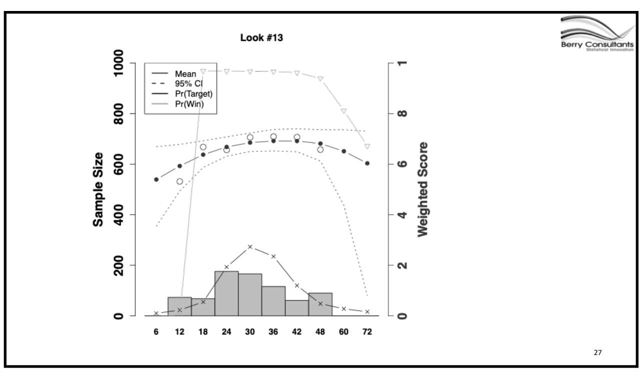
24


Look #5
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
25

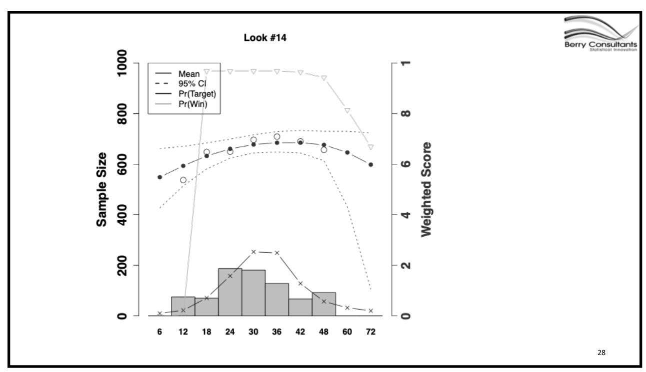
25


Look #6
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
26

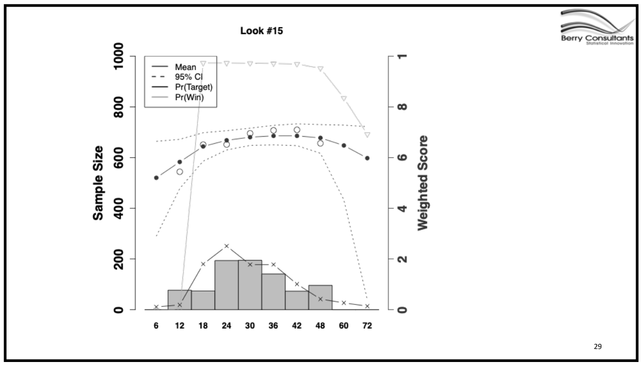
26


Look #13
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
27

27

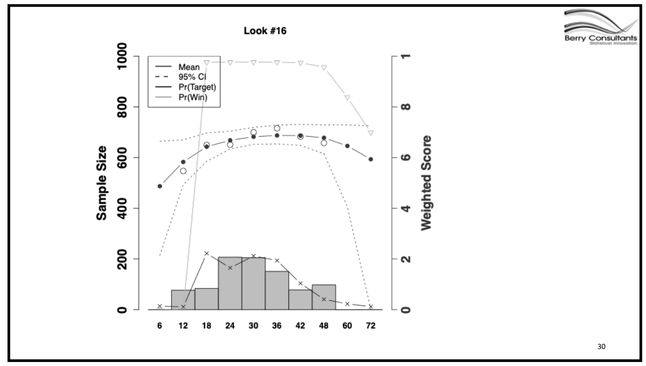

Look #14
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0 200 400 600 800 1000
0 2 4 6 8 1
6 12 18 24 30 36 42 48 60 72
28

28



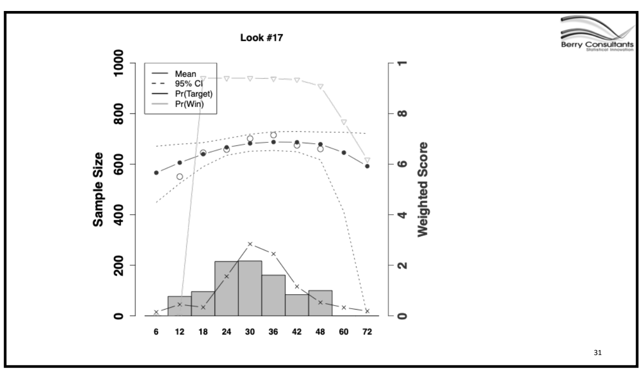
Look #15
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0
200
400
600
800
1000
0
2
4
6
8
1
6
12
18
24
30
36
42
48
60
72
29

29



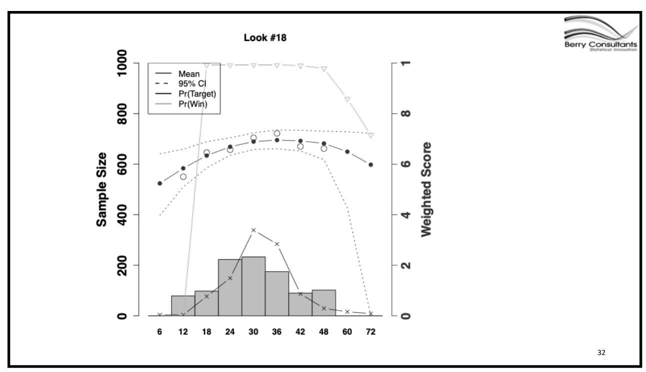
Look #16
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0
200
400
600
800
1000
0
2
4
6
8
1
6
12
18
24
30
36
42
48
60
72
30

30



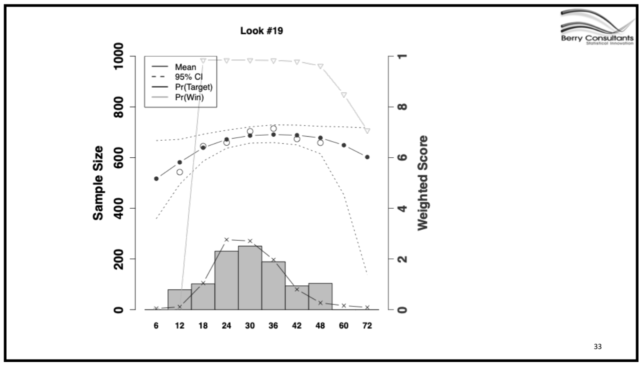
Look #17
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0
200
400
600
800
1000
0
2
4
6
8
1
6
12
18
24
30
36
42
48
60
72
31

31

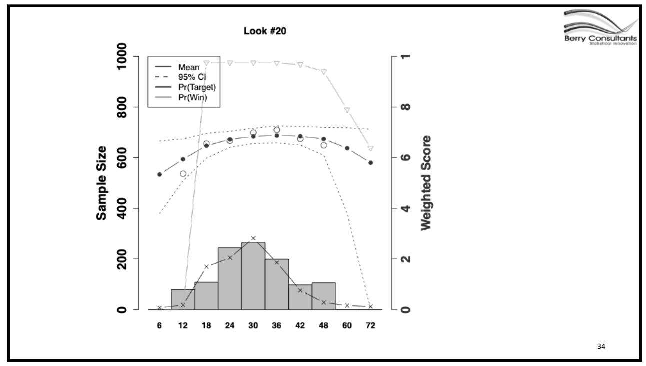


Look #18
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
0
200
400
600
800
1000
0
2
4
6
8
1
6
12
18
24
30
36
42
48
60
72
32

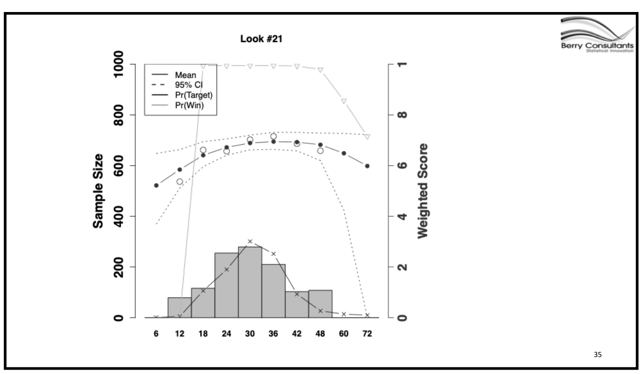
32


Look #19
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
33

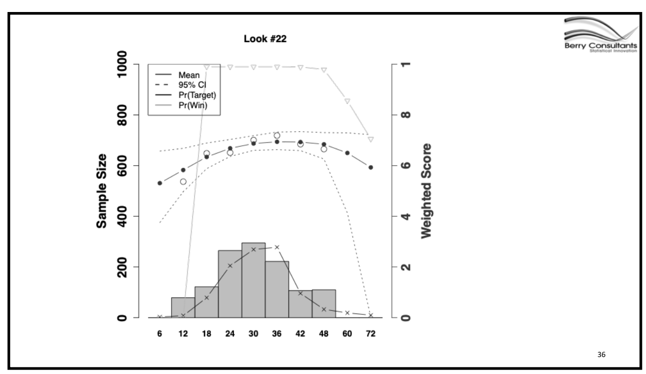
33


Look #20
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
34

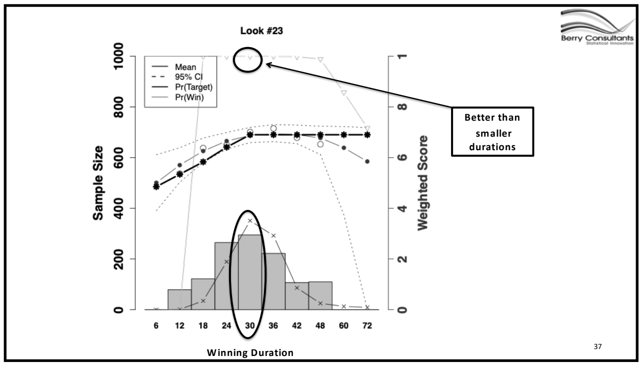
34


Look #21
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
35

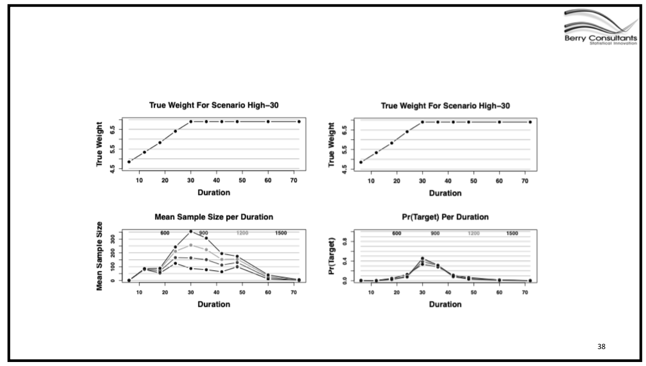
35


Look #22
Berry Consultants
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
36

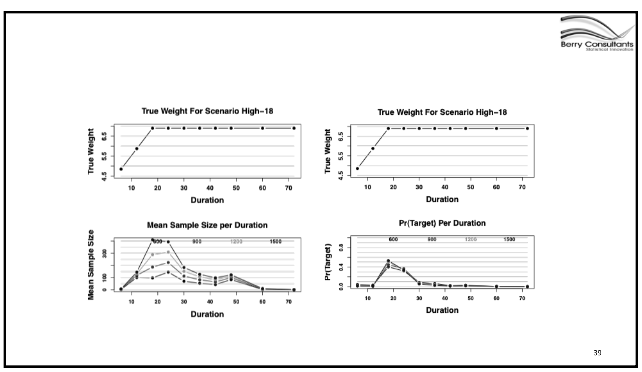
36


Berry Consultants
Look #23
Mean
95% CI
Pr(Target)
Pr(Win)
Sample Size
Weighted Score
Better than smaller durations
Winning Duration
37

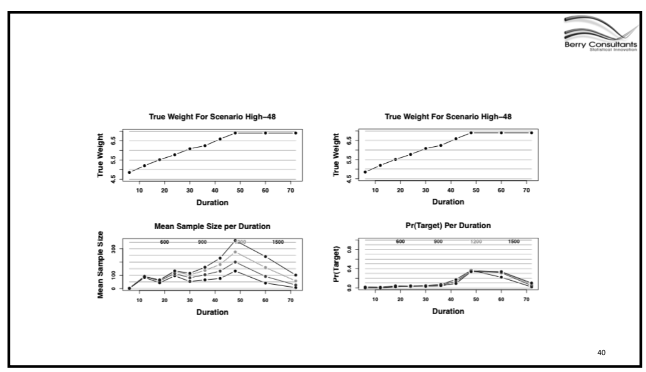
37


Berry Consultants
True Weight For Scenario High−30
True Weight For Scenario High−30
True Weight
Duration
Mean Sample Size per Duration
Mean Sample Size
Pr(Target) Per Duration
Pr(Target)
600
900
1200
1500
38

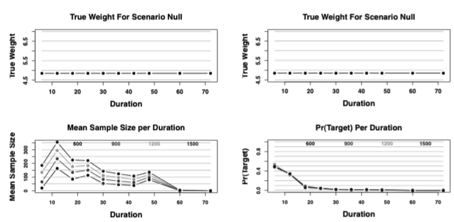
38


Berry Consultants
True Weight For Scenario High−18
True Weight For Scenario High−18
True Weight
Duration
Mean Sample Size per Duration
Mean Sample Size
Pr(Target) Per Duration
Pr(Target)
600
900
1200
1500
39

39


Berry Consultants
True Weight For Scenario High−48
True Weight For Scenario High−48
True Weight
Duration
Mean Sample Size per Duration
Mean Sample Size
Pr(Target) Per Duration
Pr(Target)
600
900
1200
1500
40

40



41

41

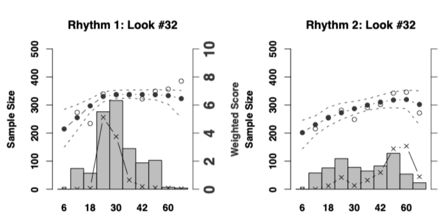
Role Simulations

- Incredible Learning Tool
  - Team, Regulators, Funders, DSMB, Operations
- Changed Models
- Changed measures of success
- Endpoint (dichotomous) wasn't correct
  - Weighted one
- Needed both rhythm types (shockable and non-shockable)
  - Possibly different duration, relative efficacy
- All recognized through flight simulator
  - Single example trials critical

42

42

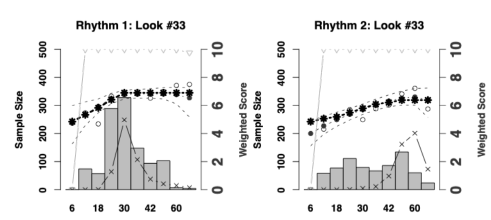
Updated



43

43

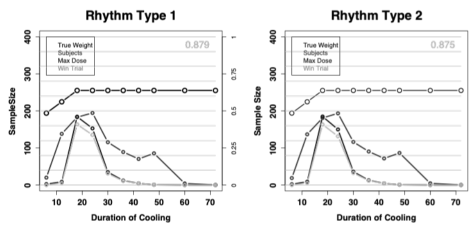
Updated



44

44

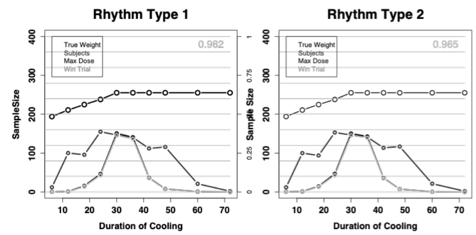
Operating Characteristics



45

45

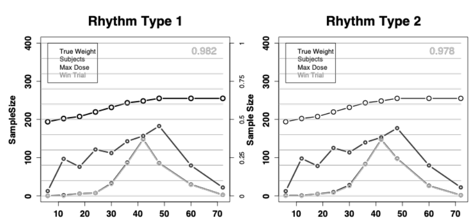
Operating Characteristics



46

46

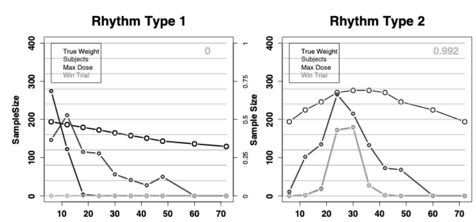
Operating Characteristics



47

47

Operating Characteristics



48

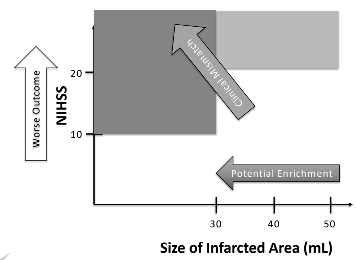
48

DAWN

- Endovascular Thrombectomy for ischemic stroke (approved ≤ 8 hours)
- New trial enrolling 6-24 hours since last seen well
- "Clinical Mismatch"

1

DAWN



2

Endpoint

- 90-day mRs
- Primary Analysis:
  - Weighted utility score:

mRs	0	1	2	3	4	5	6
Weight	10	9.1	7.6	6.5	3.3	0	0

- Trial a success if Pr(W_D > W_C) > 0.986*
  - My be adjusted if enrichment occurs*

3

Design

- Interims at 150, 200, 250, 300, 350, 400, … max of 500
- At 150, …, 400 can "enrich" to smaller entry criterion
  - Infarct size of 0-30; 0-35; 0-40; 0-45
- Could Stop for Expected Success (at 200+ interims)
- Could Stop for Futility

4

Enrich?

- If predictive probability of success by enriching increases by 10%+ then we enrich
  - Can be multiple steps
- If the posterior probability of benefit in ‘last 5 tail’ is less then 40% then drop the last 5 (enrich)
- If enrich we restrict the population for the final analysis as well

Berry Consultants

5

5

Adaptive Design Model

$$[Y_X] \sim N(\alpha_X + \theta_X, \sigma^2)$$

$$[\alpha_X] \sim NDLM^2(0, \lambda^2)$$

Experimental Arm

$$[\theta_X] \sim NDLM^2(0, \gamma^2)$$

Berry Consultants

JSM 2014

6

6

Final Analysis Model

$$[Y_X] \sim N(\alpha_X + \theta, \sigma^2)$$

$$[\alpha_X] \sim NDLM^2(0, \lambda^2)$$

Experimental Arm

$$[\theta] \sim N(0, 100^2)$$

Berry Consultants

JSM 2014

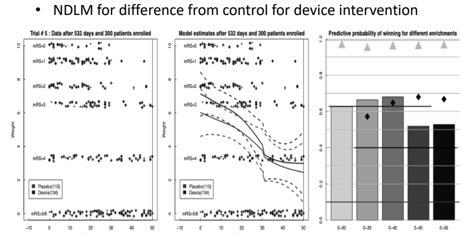
7

7

Modeling: Example Trial; n=300

Bayesian Models for outcomes as a function of infarct size,

- NDLM Model for mean Control results
- NDLM for difference from control for device intervention



Berry Consultants

8

8

"Stopping"

- Futility: Stop the trial for futility if the predictive probability of success by the cap is < 0.10 (including for any enrichment)
- Expected Success: If the predictive probability for the currently enrolled patients is > 0.99 then stop enrollment and follow all through primary endpoint
  - Must enroll at least +100 beyond enrichment

Berry Consultants

9

9

Critical Value Adjustment

- The critical value of 0.986 is used unless there is enrichment
- If we enrich, restrict the primary on only remaining group (discard some randomized)
- Boost CV:

$$\Phi\left(\Phi^{-1}(0.986)\sqrt{1+\frac{N_{drop}}{N_{keep}+N_{new}}}\right)$$

- E.g. N_{drop}=50; N_{keep}=300; N_{new}=100; cv= 0.9906

Berry Consultants

10

10

Simulation Constructed

- Trial fully and extensively simulated
- Modeling decisions, robustness, and cut-off optimization
- Control of type I error by simulation
  - Early stopping
  - Enrichment adjustment

Berry Consultants

11

11

DAWN Actual Result

- At the 150-interim there was *no enrichment*
  - no futility
  - No expected success possible
- At 200-interim PP > 0.9999; no enrichment; stop for expected success!
- Followed for 90 days; success at full data primary analysis

Berry Consultants

12

12

The NEW ENGLAND JOURNAL *of* MEDICINE

ORIGINAL ARTICLE

Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct

R.G. Nogueira, A.P. Jadhav, D.C. Haussen, A. Bonafe, R.F. Budzik, P. Bhuva, D.R. Yavagal, M. Ribo, C. Cognard, R.A. Hanel, C.A. Sila, A.E. Hassan, M. Millan, E.I. Levy, P. Mitchell, M. Chen, J.D. English, Q.A. Shah, F.L. Silver, V.M. Pereira, B.P. Mehta, B.W. Baxter, M.G. Abraham, P. Cardona, E. Veznedaroglu, F.R. Hellinger, L. Feng, J.F. Kirmani, D.K. Lopes, B.T. Jankowitz, M.R. Frankel, V. Costalat, N.A. Vora, A.J. Yoo, A.M. Malik, A.J. Furlan, M. Rubiera, A. Aghaebrahim, J.-M. Olivot, W.G. Tekle, R. Shields, T. Graves, R.J. Lewis, W.S. Smith, D.S. Liebeskind, J.L. Saver, and T.G. Jovin, for the DAWN Trial Investigators*

This article was published on November 11, 2017, at NEJM.org.

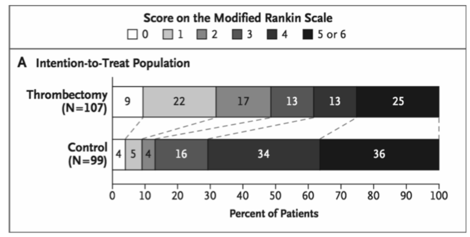
DOI: 10.1056/NEJMoa1706442

Berry Consultants

13

13

DAWN



14

14

RESULTS

A total of 206 patients were enrolled; 107 were assigned to the thrombectomy group and 99 to the control group. At 31 months, enrollment in the trial was stopped because of the results of a prespecified interim analysis. The mean score on the utility-weighted modified Rankin scale at 90 days was 5.5 in the thrombectomy group as compared with 3.4 in the control group (adjusted difference [Bayesian analysis], 2.0 points; 95% credible interval, 1.1 to 3.0; posterior probability of superiority, >0.999), and the rate of functional independence at 90 days was 49% in the thrombectomy group as compared with 13% in the control group (adjusted difference, 33 percentage points; 95% credible interval, 24 to 44; posterior probability of superiority, >0.999). The rate of symptomatic intracranial hemorrhage did not differ significantly between the two groups (6% in the thrombectomy group and 3% in the control group, P=0.50), nor did 90-day mortality (19% and 18%, respectively; P=1.00).

Table 2. **Efficacy Outcomes.***

Outcome	Thrombectomy Group (N=107)	Control Group (N=99)	Absolute Difference (95% CI)†	Adjusted Difference (95% Credible Interval)‡	Posterior Probability of Superiority
Primary end points					
Score on utility-weighted modified Rankin scale at 90 days§	5.5±3.8	3.4±3.1	2.1 (1.2–3.1)	2.0 (1.1–3.0)	>0.999
Functional independence at 90 days — no. (%)¶	52 (49)	13 (13)	36 (24–47)	33 (21–44)	>0.999

Berry Consultants

15

15

Summary

- Complex enrichment design – results was very strong and no enrichment occurred
  - Did the right thing!
- Could have run trial in only smaller group and left 'uncertainty' in where effect
- Sample size flexibility allowed success at 40% of maximum
- Designed and optimized by simulation

Berry Consultants

16

16